HMM-based Speech Synthesis Using an Acoustic Glottal Source Model

João Paulo Serrasqueiro Robalo Cabral

Doctor of Philosophy
The Centre for Speech Technology Research
Institute for Communicating and Collaborative Systems
School of Informatics
University of Edinburgh
2010

Abstract

Parametric speech synthesis has received increased attention in recent years following the development of statistical HMM-based speech synthesis. However, the speech produced using this method still does not sound as natural as human speech and there is limited parametric flexibility to replicate voice quality aspects, such as breathiness.

The hypothesis of this thesis is that speech naturalness and voice quality can be more accurately replicated by a HMM-based speech synthesiser using an acoustic glottal source model, the Liljencrants-Fant (LF) model, to represent the source component of speech instead of the traditional impulse train.

Two different analysis-synthesis methods were developed during this thesis, in order to integrate the LF-model into a baseline HMM-based speech synthesiser, which is based on the popular HTS system and uses the STRAIGHT vocoder. The first method, which is called Glottal Post-Filtering (GPF), consists of passing a chosen LF-model signal through a glottal post-filter to obtain the source signal and then generating speech, by passing this source signal through the spectral envelope filter. The system which uses the GPF method (HTS-GPF system) is similar to the baseline system, but it uses a different source signal instead of the impulse train used by STRAIGHT. The second method, called Glottal Spectral Separation (GSS), generates speech by passing the LF-model signal through the vocal tract filter. The major advantage of the synthesiser which incorporates the GSS method, named HTS-LF, is that the acoustic properties of the LF-model parameters are automatically learnt by the HMMs.

In this thesis, an initial perceptual experiment was conducted to compare the LF-model to the impulse train. The results showed that the LF-model was significantly better, both in terms of speech naturalness and replication of two basic voice qualities (breathy and tense). In a second perceptual evaluation, the HTS-LF system was better than the baseline system, although the difference between the two had been expected to be more significant. A third experiment was conducted to evaluate the HTS-GPF system and an improved HTS-LF system, in terms of speech naturalness, voice similarity and intelligibility. The results showed that the HTS-GPF system performed similarly to the baseline. However, the HTS-LF system was significantly outperformed by the baseline. Finally, acoustic measurements were performed on the synthetic speech to investigate the speech distortion in the HTS-LF system. The results indicated that a problem in replicating the rapid variations of the vocal tract filter parameters at transitions between voiced and unvoiced sounds is the most significant cause of speech distortion. This problem encourages future work to further improve the system.

I dedicate this thesis to my family, whom I love very much.

Acknowledgements

Firstly, I would like to thank my supervisors, Prof. Steve Renals, Dr. Korin Richmond and Dr. Junichi Yamagishi, for their invaluable advice, deep multi-disciplinary knowledge, and their generosity of time in discussing my work during this thesis. In particular, I would like to thank Dr. Junichi Yagamishi for his support on HMM-based speech synthesis. I am also grateful for the motivation and confidence they transmitted to me throughout the thesis. It was very exciting to work in CSTR and I would like to thank all the people from the group for creating a friendly atmosphere at the lab and for making it a stimulating place to conduct research.

I am also indebted to Vasilis Karaiskos from the School of Informatics, in the University of Edinburgh, for his help in adjusting the Blizzard computer interface for a perceptual evaluation I conducted during this thesis.

I am grateful to Prof. Simon King for helping me with my research visit to India and to the British Council for the financial support for this visit. I would also like to thank all the people in the speech processing labs of the IIT Guwahati and the IIIT Hyderabad for welcoming me so warmly during my time there. Particularly, I would like to thank to Prof. B. Yegnanarayana, to Prof. Mahadeva Prasanna, to Govind, and to Dhanu for the discussions about research topics and all the help they gave me during my stay.

I am also grateful for the financial support provided by the Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568), which has given me the opportunity to conduct this research.

Last but not least, I would like to thank my family for all the love and emotional support. While living in Edinburgh I had also the opportunity to meet friends besides my work colleagues. I am not going to list you all here but I am pleased to have met you. Especially, I am lucky I have met a wonderful person who is my best friend Davinia Anderson.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*João Paulo Serrasqueiro Robalo Cabral*)

Table of Contents

Chapter 1

Introduction

Speech is one the most important forms of communication between humans. The message to be spoken is formulated in a person's mind and expressed in the form of speech signals in a structured way, i.e. using the symbolic representation of the human language (phones, words, etc.), so that it can be interpreted and understood by the listener. The speech production system is commanded by the brain which controls a series of movements of articulators, such as vocal folds, tongue, and lips. The energy necessary for producing the airflow in the respiratory system is generated by a pressure drop in the lungs. For *voiced sounds*, the flow of air through the glottis causes the vocal folds to vibrate and the air stream is modulated into pulses. The rate of vibration of the vocal folds is called *fundamental frequency* (F_0) and its main perceptual effect is the *pitch*. Voiced sounds, such as vowels, are characterised by a periodicity pattern. The frequency structure of these sounds is also regular and it is characterised by a set of *harmonics*, i.e. frequency components multiples of the fundamental frequency. These harmonics are emphasised near the resonance frequencies of the vocal tract (pharyngeal and oral cavities), which are called *formants*. If there is passage of air through the nasal cavity, then the resonances of the nasal cavity are also excited. Variations in the vocal tract shape, such as lips opening, and tongue placement, change the formants and contribute to differentiation between different types of speech sounds (e.g. the phones /aa/ and /b/). *Unvoiced sounds* are excited either by creating a rapid flow of air through one or more constrictions, at some point between the trachea and the lips, or by making a closure at the point of constriction and abruptly releasing it. The first acts like a *turbulent noise* source while the second produces a *transient excitation* followed by turbulent flow of air, such as the excitation of the *stop consonant* /p/.

For a long time humans have developed systems to produce "human-like" speech.

Nowadays, automatic text-to-speech synthesisers can produce speech which sounds intelligible and natural. Although the quality of the synthetic speech has yet to fully match the quality of human speech, these systems have been successfully used in day-to-day applications, like screen readers to help people with visual impairments, text-to-speech systems to help people with speech impairments to communicate, and systems to convert written news to speech.

1.1 Speech Synthesis Methods

The earliest text-to-speech systems are based on a parametric speech production model, which represents speech by two components: the *glottal source* and the *vocal tract transfer function*. The traditional systems represent the vocal tract transfer function as a sequence of *formant resonators*, such as the Parametric Artificial Talker (PAT) synthesiser (Lawrence, 1953) and the MITalk system (Klatt, 1982). For this reason, they are often called *formant synthesisers*. These systems generate the speech signal using a set of acoustic rules derived by human experts, which describe how the parameters (fundamental frequency, formants, etc.) vary from one speech sound to another. *Articulatory speech synthesis* is another method which uses the knowledge about the speech production system for producing speech. However, this method uses the physical theory to describe the vocal tract shape and to model how the articulators of the speech production system change with time.

Techniques based on concatenating pre-recorded fragments of speech have been rising in popularity since the 1970s until today. These methods avoid the difficult task of deriving acoustic rules, because natural speech segments contain the phonetic information and the dynamic properties of speech sounds. However, for synthesis by concatenation it is necessary to record a relatively large amount of speech data. The traditional *concatenative synthesisers* use a speech model to represent the recorded speech fragments in terms of acoustic features. This technique allows the size of the speech database to be reduced and acoustic aspects of speech to be modified, such as pitch and formants. From the mid 1990s, the concatenation of units of natural speech started to become more popular than using a parametric model of speech. This was facilitated with the development of the storage and processing power of computers, which permitted to use more complex algorithms for searching the speech fragments and larger speech databases. State-of-the-art concatenative synthesisers, which are called *unit-selection synthesisers*, concatenate speech units of variable length without

applying signal processing (or very little processing), in order to obtain high speech naturalness (Campbell and Black, 1996).

Statistical speech synthesis is a relatively recent approach in which a statistical model, typically the *Hidden Markov Model* (HMM), is used to learn automatically the acoustic properties of the different speech sounds. This method uses a speech model as in formant synthesis, but does not require acoustic rules derived by humans. Hybrid systems which combine the concatenation method with the formant and statistical speech synthesises methods respectively, have also been successfully used, e.g. Högberg (1997); Plumpe et al. (1998).

1.1.1 Formant Synthesisers

Formant speech synthesisers generate the speech signal entirely from rules on the acoustic parameters, which are derived by human experts from speech data. Most of the parameters describe the pitch, formant/antiformant frequencies and bandwidths. In general, the synthetic speech sounds smooth since the variation of the formant frequencies is also driven by rules, which are determined using physical constraints. For example, the maximum allowable slopes of the formant in the transition between two sounds is determined by the speed of the articulators which produce those sounds (Huang et al., 2001).

Voiced sounds, such as vowels, are synthesised by passing a periodic source signal through a filter which represents the formant frequencies of the vocal tract. For unvoiced speech, the source signal is usually modelled as *white random noise* instead. The synthesis filter can be constructed by cascading second-order filters (each representing a resonance of the vocal tract). For example, the Parametric Artificial Talker (PAT) synthesiser (Lawrence, 1953) consists of a sequence of formant filters in parallel and the source (excitation of the filter) is either an impulse train or noise. Alternatively, a parallel structure of the format resonators can also be used, such as in the different versions of the Orator Verbis Electris (OVE) system (Fant, 1953; Liljencrants, 1968). The most sophisticated formant synthesisers use different structures to model the vocal tract of vowels, nasals and consonants. For example, the cascade structure is commonly used to model voiced sounds, whereas the parallel model is commonly used to synthesise unvoiced consonants. Formant synthesisers often use a sophisticated excitation model. For example, a *mixed excitation model* which is the combination of a periodic and a noise component of the source, is typically used to synthesise *voiced*

fricatives and to add *aspiration noise* in voiced sounds. Excitation models which include glottal parameters to control the shape of the *glottal pulse*, are also commonly used in these systems, e.g. Klatt (1987).

The large number of parameters (up to 60) and the difficulty in estimating formant frequencies and bandwidths makes the analysis stage of formant synthesisers complex and time-consuming. In general, speech generated using these systems is intelligible. They can also synthesise speech which sounds very close to the original speech by manual tuning the acoustic parameters of the systems, as shown by Holmes (1972) who synthesised a number of utterances using his system by manually adjusting the formant tracks. However, automatic formant synthesis does not sound natural, mainly due to incomplete phonetic knowledge and limitations of the acoustic model used in the systems to describe the variability and details of speech. The major advantage of this speech synthesis method is that is offers a high degree of *parametric flexibility* which allows voice characteristics to be controlled and expressive speech to be modelled by deriving specialised rules. For example, the Affect Editor program (Cahn, 1989) uses a formant synthesiser, the DECTalk synthesiser of Allen et al. (1987), in order to produce emotional speech by controlling several parameters related to pitch, timing, articulation and *voice quality* (e.g. breathiness). This synthesiser uses a *glottal source model* which allows different voice effects to be produced. Formant speech synthesisers are also suitable for memory-constrained applications because they require a small memory footprint.

Although most formant synthesisers are driven by rules, statistical modelling of the formant parameters using HMMs has also been explored (Acero, 1999). Even using a full data-driven approach to generate the parameters, it has proved difficult to further improve formant synthesisers.

1.1.2 Articulatory Synthesisers

Articulatory synthesisers describe speech in terms of articulatory features of the vocal generation system, as opposed to acoustic parameters in formant synthesisers. They use a physical theory to describe the vocal tract shape and to simulate how the articulators of the speech production system change with time, such as the Dynamic Analog of the Vocal Tract (DAVO) synthesiser of Rosen (1958) and the *VocalTractLab* synthesiser (Birkholz, 2010). The main issue in articulatory synthesisers is how to control the articulatory parameters in order to produce a certain speech sound, e.g. parameters of

the *vocal tract tube area function* and parameters which describe the tongue position. Typically these systems are driven by rule and use an acoustic source-filter model, in a similar way to formant synthesisers. However, the complexity of the *articulatory-acoustic mapping* is complex and makes it hard to determine what articulatory parameter should be used in order to produce a given acoustic signal. For example, the same speech sound can be produced with very different combinations of articulator positions, which makes the articulatory-to-acoustic mapping a difficult problem to solve (many-to-one possible mappings). State-of-the-art articulatory speech synthesisers can produce high-quality speech for isolated sounds, such as vowels. However, speech quality is significantly degraded when these systems are used to synthesise continuous speech, due to problems in modelling *co-articulation effects* and more complex sounds. Despite the progress of articulatory speech synthesis in recent years, this method is not yet feasible enough for text-to-speech applications.

1.1.3 Concatenative Synthesisers

In concatenative speech synthesis the problem is to select the fragments of recorded speech for a given phonetic sequence. In general, the segments to be concatenated have different *phonetic contexts*, since they are generally extracted from different words. As result, there is usually an acoustic and prosodic mismatch at the concatenation points which might produce distortion. In principle, the larger the speech database, the more likely it is that a good sequence of units may be found, and the better is the quality of the output speech. Typically, short speech units, such as *diphone* (starting at the middle of one phone and ending at the middle of the next phone) or phone units, are used so as to obtain a speech database of an affordable size.

Diphone concatenation synthesisers were widely used in the 1990s, as they could produce intelligible speech with a relatively small amount of speech data. The diphone join points are in the most stable part of the phone, which reduces the effect of audible discontinuities which occur at the join points. A careful corpus design is usually performed in order to obtain a relatively small (e.g. one hour long) and phonetically balanced inventory of diphone units. These systems typically use an *analysis-synthesis method*. For example, the *Linear Predictive Coding (LPC) model* (Markel and Gray, 1976) and the *harmonic model* (Dutoit, 1993), which are described in Section 2.1, are commonly used in diphone concatenation synthesisers to parameterise the speech signal and *resynthesise* speech using the speech parameters. The main advantages of using

a parametric model when compared to a speech waveform is the lower storage requirement and the parametric flexibility which enables the transformation of acoustic properties of speech. For example, speech parameters can be interpolated in order to obtain smoother transitions at the concatenation points and they can be transformed in order to reproduce prosodic and voice quality variations. Diphone concatenation systems often use *signal processing* techniques to manipulate acoustic characteristics of the units, such as the Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA) for pitch and duration transformations (Moulines and Charpentier, 1990). Although diphone synthesisers can produce more natural speech than formant synthesisers, the use of a parametric model and signal processing usually produce unnatural speech quality. For example, LPC diphone synthesisers are characterised by a "*buzzy*" speech quality.

The concatenation-based systems which produce the most natural sounding speech are the unit-selection synthesisers, e.g. the Festival Multisyn system (Clark et al., 2007b). In these systems, units of variable size are selected from a large speech database upon minimisation of the *target and the concatenation costs*. The target cost indicates how well each unit matches the ideal unit segment for each utterance, while the concatenation cost refers to how well each unit joins to the previous unit. In the unit-selection method, the speech units are usually not modified and a large speech corpus is used (usually not less than 6 hours of speech), in order to obtain high-quality speech. However, it is impossible for the speech database to cover all aspects of speech variability. Therefore, occasionally there are bad joins which result in audible speech artifacts. The tradeoff of using natural speech units to improve speech naturalness by unit-selection synthesisers is the lower control of voice characteristics due to reduced parametric flexibility. For example, another large speech corpus needs to be recorded in order to build a voice for a new speaker. Also, it is hard to synthesise speech with different speaker styles or voice qualities using these systems. One way to overcome this problem is to use signal processing to transform acoustic properties of the speech signal. However, the required degree of speech modifications often degrade speech quality, e.g. Murray and Edgington (2000). An alternative to signal processing is to use different speech inventories for each speaking style, e.g. Iida et al. (2000). However, recording additional speech corpus is demanding in terms of time and money. Also, the complexity of the speech corpus preparation, storage requirements and unit search techniques of these systems usually increase with the number of different speech inventories used.

1.1.4 Statistical Synthesisers

Statistical parametric speech synthesis is a relatively recent approach which has been summarised by Black et al. (2007) as "generating the average of some set of similarly sounding speech segments". The statistical model which has been used more often for speech synthesis is based on the Hidden Markov Model (HMM). HMMs have been applied successfully to speech recognition from the late 1970s. However, they have been used for speech synthesis for only about two decades. In comparison with formant synthesisers, HMM-based speech synthesisers are also fully parametric and require a small footprint, but they have the advantage that they are fully automatic. In other words, the difficult task of deriving the rules in formant synthesisers is overcome by the automatic training of the HMMs. These systems typically use *vocoding* techniques to extract the speech parameters from recorded speech and to generate the speech signal using a *source-filter model*, which is generally different from the formant model used by formant synthesisers.

HMM-based speech synthesisers can produce high-quality speech. In particular, they permit more natural sounding speech to be obtained than from conventional rule-based synthesisers or diphone concatenation synthesisers. However, the synthetic speech generated by current statistical speech synthesisers does not sound as natural as that generated by state-of-the-art unit-selection systems (Black et al., 2007; King and Karaiskos, 2009), mainly because the statistical speech synthesisers produce a "buzzy" and *muffled* speech quality. The "buzzy" or robotic quality is mainly associated with the vocoding technique used to generate speech from the parameters. In particular, the excitation of voiced sounds is typically modelled using a simple *impulse train*, which often produces the "buzzy" speech quality. On the other hand, the muffled quality is related to *over-smoothing* of the speech parameter trajectories measured on the recorded speech, which is caused by statistical modelling. Nevertheless, HMM-based speech synthesis is considered to be more robust than unit-selection (Black et al., 2007). This difference between the two methods is because unit-selection produces speech artefacts, when occasional bad joins occur, while HMM-based speech synthesisers produce speech which sounds smoother.

The major advantage of HMM-based speech synthesisers is their higher parametric flexibility compared to unit-selection systems. The HMM parameters of the synthesiser can be *interpolated* (Yoshimura et al., 1997) or *adapted* (Tamura et al., 1998, 2001) from one speaker to another using a small amount of the target speak-

ers speech data. HMM adaptation techniques have also been used to transform voice characteristics, e.g. specific voice qualities and basic emotions (Yamagishi et al., 2003, 2007a; Barra-Chicote et al., 2010), in HMM-based speech synthesis. However, HMM-based speech synthesisers typically do not model glottal source parameters, which are strongly correlated with voice quality. In contrast, formant synthesisers often use a glottal source model which enables the control of voice characteristics related to the glottal source.

HMM-based speech synthesisers can be classified into two general types. Traditional systems are *speaker dependent*, i.e. they are built using a large speech corpus from one speaker. The other type is called *speaker independent* HMM-based speech synthesis. In this case, statistical *average voice models* are created from several speakers' speech data and are adapted using a small amount of speech data from the target speaker (Yamagishi and Kobayashi, 2007).

1.1.5 Hybrid Systems

There have been several attempts to combine the advantages of rule-based or statistical approaches with the naturalness obtained using unit-selection. Several hybrid approaches using formant synthesis and data-driven methods have been proposed. For example, Högberg (1997); Öhlin and Carlson (2004) proposed data-driven formant synthesisers which use a *unit library of formant parameters* extracted from recorded speech in order to better model detailed gestures than the original rules of the formant synthesiser. These systems keep the parametric flexibility of the original rule-based model and the possibility to include both linguistic and extralinguistic knowledge sources. Another type of hybrid approach uses HMMs to calculate the costs for unit-selection systems (Rouibia and Rosec, 2005; Ling and Wang, 2006) or as a *probabilistic smoother* of the spectrum of the vocal tract across speech unit boundaries (Plumpe et al., 1998).

1.2 Contributions of the Thesis

Nowadays, automatic text-to-speech synthesisers can produce speech which sounds intelligible and natural. However, there is still a gap between synthetic and human speech which seems hard to bridge with the formant, articulatory, and concatenative synthesis methods. HMM-based speech synthesis is a more recent method which can

produce speech of comparable quality to the unit-selection method and it has a great potential of development.

Emerging applications, such as spoken dialogue systems, e-books, and computer games, demand expressive speech and high parametric flexibility from the speech synthesisers to control voice characteristics. Also, there has been an increasing interest from manufacturers to integrate the latest speech technology in portable electronic devices, such as PDAs and mobile phones. Unit-selection and rule-based synthesis methods have significant limitations for these applications. On one hand, formant and articulatory synthesisers traditionally offer parametric flexibility to control the type of voice, but they typically produce unnatural speech quality. On the other hand, the unit-selection systems, which provide the most natural quality, are very limited in terms of the control of voice characteristics and the synthesis of expressive speech. Also, these systems typically require a large inventory of speech units and high computational complexity which are inappropriate for the small memory footprint requirement of portable devices. Meanwhile, HMM-based statistical speech synthesisers are fully parametric and can produce high-quality speech. The main characteristics of these systems are summarised below:

- high-quality speech and robustness to variations in speech quality.
- fully parametric.
- fully automatic.
- small footprint.
- easy to transform voice characteristics.
- new languages can be built with little modification.
- speaking styles and emotions can be synthesised using a small amount of data.

These characteristics make this technique very attractive, especially for applications which expect variability in the type of voice and a small memory footprint.

In terms of speech quality, HMM-based speech synthesisers can produce more natural sounding speech than formant synthesisers. Also, they are typically more robust to variations in speech quality than unit-selection systems. Whereas concatenative synthesisers occasionally produce speech segments with very poor quality, statistical synthesisers produce speech which sounds smooth. However, speech synthesised using

HMMs does not sound as natural as speech obtained using unit-selection. This effect is related to the limitations of the parametric model of speech used by HMM-based speech synthesisers. In particular, these systems commonly use a simple impulse train to model the excitation of voiced speech, which produces a buzzy quality.

The major advantage of the statistical method when compared with unit-selection is that it offers the flexibility to synthesise speech with different speakers' voices and speaking styles, by using speech data spoken with the target voice characteristics. However, these systems generally allow a more limited control of voice characteristics than formant synthesisers. The main reason for this is that most statistical synthesisers use a speech model which does not separate the different components of speech (glottal source, vocal tract resonance, and radiation at the lips), unlike formant synthesisers. As a consequence, current HMM-based speech synthesisers do not allow glottal source parameters which are important for voice transformation to be controlled.

The objective of this thesis is to improve the *excitation model* in HMM-based speech synthesis. The method is to develop a synthesiser which uses an *acoustic glottal source model*, instead of the traditional impulse train. This work is based on the following hypothesis: *A glottal source model improves the quality of the synthetic speech when compared to the simple impulse train*.

The motivations to use glottal source modelling in HMM-based speech synthesis are:

- Reduce buzziness of synthetic speech.
- Better modelling of prosodic aspects which are related to the glottal source.
- Control over glottal source parameters to improve voice transformations.

The speech production system, which consists of exciting a vocal tract filter with a glottal source signal, has been extensively studied in the literature. However, speech models which use a simpler representation of the excitation, instead of the glottal source, are often preferred in speech technology applications. The main reason for this is that the methods to estimate the glottal source and the vocal tract filter are usually complex and not sufficiently robust. Therefore, the problem of improving the speech quality in HMM-based speech synthesis by using an acoustic glottal source model is not expected to be easy to solve. The following are important factors to be considered in this work:

- Degradation in speech quality due to errors in the glottal and vocal tract parameter estimation.
- Degradation in speech quality due to statistical modelling of the glottal and vocal tract parameters.
- Incorporation of the source-filter model into the HMM-based speech synthesiser.

The contributions of this thesis are:

Glottal Post-Filtering (GPF): transforms the *Liljencrants-Fant (LF) model* of the glottal source derivative into a spectrally flat signal. This method allows speech to be generated using the LF-model and a synthesis filter which represents the *spectral envelope*. The major advantage is that it allows voice transformations by controlling the LF-model parameters. This method is described in Section 6.3. The results of a HMM-based speech synthesiser which uses GPF for generating speech are presented in Section 8.4.

Glottal Spectral Separation (GSS): analysis-synthesis method to synthesise speech using a glottal source model (e.g. the LF-model) and the vocal tract transfer function. This method can be divided into three processes: 1) parameters of the glottal source model are estimated from the speech signal; 2) spectral effects of the glottal source model are removed from the speech signal; 3) vocal tract transfer function is estimated as the spectral envelope of the signal obtained in 2). The description and results of this method are presented in Sections 6.4 and 6.6 respectively.

Robust LF-model parameter extraction: method for estimating the LF-model parameters, which uses a *non-linear optimisation algorithm* to fit the LF-model to the glottal source derivative signal. The initial estimates of the iterative method are obtained using amplitude-based techniques which were developed during this work. They are used to estimate the parameters directly from the glottal source derivative. The LF-model parameter estimation method is described in Section 6.5.

HMM-based speech synthesiser using LF-model: system which models the excitation of voiced sounds as a mix of the LF-model signal and white noise. This synthesiser also uses the GSS method to estimate the vocal tract parameters from the

speech signal and the LF-model parameters. The LF-model, noise, and spectral parameters are modelled by HMMs and used by the system to generate speech. The first version of this system is described in Chapter 7. Improvements which were made to the system are described in Section 8.2. The evaluation of the first and second versions of the synthesiser are presented in Sections 7.4 and 8.4 respectively.

Chapter 2

Speech Modelling

The speech waveform can be used as a speech model, such as in unit-selection speech synthesisers (concatenate fragments of recorded speech). However, a more suitable and convenient speech model than the recorded speech waveform is often employed in speech applications, such as the extraction of acoustic or linguistic information from the speech signal, transformation of acoustic properties of speech, speech coding (compacted representation of speech), or speech synthesis (e.g. in formant and HMM-based speech synthesis systems). A speech analysis method is used to convert the speech signal into a different representation, i.e. to estimate the parameters of the speech model. This method usually decomposes the speech signal into the source and filter components, which are considered to be independent. For example, the acoustic model of speech production typically represents the source as the derivative of the signal produced at the glottis and the filter as the vocal tract system. The speech waveform can be reconstructed from the speech parameters using a synthesis method. In the case of the source/filter model, speech is generated by passing the source signal through the synthesis filter.

The next section gives an overview of the general types of speech models. Subsequently, Section 2.2 describes in more detail the acoustic model of speech production, focusing on the glottal source component. Specifically, this section reviews the general types of glottal source models (in Section 2.2.2), the most commonly used methods to estimate the glottal source and the vocal tract components from the speech signal (in Section 2.2.3), and the methods to parameterise the glottal source signal (in Section 2.2.4).

2.1 Parametric Models of Speech

Most parametric speech synthesisers use a source-filter model of speech. In this model, an excitation signal passes through a synthesis filter to generate the speech signal. The excitation is typically assumed to be aperiodic for voiceless speech and quasi-periodic for voiced speech. There are two general types of source-filter model. One is based on the speech production model, which represents the excitation of voiced sounds as the glottal signal produced at the vocal folds and the synthesis filter as the transfer function of the vocal tract system. For example, formant synthesisers typically use this speech model, e.g. Klatt and Klatt (1987). The other type of source-filter model consists of representing the source as a spectrally flat signal and the synthesis filter as the spectral envelope of the speech signal. For example, state-of-the-art HMM-based speech synthesisers typically use this type of source-filter model. Both types of source-filter model traditionally represent the excitation of unvoiced speech as white noise.

The next section gives a general overview of the speech production model. Then, three parametric models of speech which are commonly used in speech synthesis are described: the *harmonic/stochastic model*, the *linear prediction spectrum* and the *cepstrum*.

2.1.1 Speech Production Model

The speech production model assumes that speech is a linear and stable system, which consists of an excitation, a vocal tract filter and a radiation component.

The vocal tract transfer function can be represented by the z-transform (Quatieri, 2001):

$$V(z) = A\frac{\prod_{k=1}^{M_i}(1-a_k z^{-1})\prod_{k=1}^{M_o}(1-b_k z)}{\prod_{k=1}^{C_i}(1-c_k z^{-1})\prod_{k=1}^{C_i}(1-c_k^* z^{-1})}, \tag{2.1}$$

where $(1-c_k z^{-1})$ and $(1-c_k^* z^{-1})$ are *complex conjugate poles* inside the unit circle with $|c_k| < 1$. These complex conjugate poles model the resonant or formant structure of the vocal tract. The zeros $(1-a_k z^{-1})$ and $(1-b_k z)$ are due to the oral and nasal tract constrictions. The vocal tract shape determines the acoustic realisation of the different classes of sounds (phones /aa/,/b/,etc.).

The excitation of unvoiced sounds, $E(z)$, can be modelled as white noise. In the case of voiced speech, the excitation represents the glottal source signal, $g(n)$. This ex-

citation is modelled as an impulse train convolved with $g(n)$. That is, $E(z) = P(z)G(z)$, where $P(z)$ represents the spectrally flat impulse train. The glottal source signal is characterised by a decaying spectrum. It is often approximated by two time-reversed exponentially decaying sequences over one glottal cycle (Quatieri, 2001), that has z-transform

$$G(z) = \frac{1}{(1-\beta z)^2} \tag{2.2}$$

For $\beta < 1$, $G(z)$ represents two identical poles outside the unit circle. The duration of the glottal pulse is perceptually related to the pitch, while its shape is strongly correlated with voice quality.

The models in (2.1) and (2.2) assume infinite *glottal impedance*. All loss in the system is assumed to occur by radiation at the lips. The radiation has a high-pass filtering effect, which is typically modelled with a single zero, i.e.

$$R(z) = 1 - \alpha z^{-1} \tag{2.3}$$

Under the assumption of vocal tract linearity and time-invariance, speech production can be expressed as the convolution of the excitation and the vocal tract impulse response. Then, the z-transform of the speech output can be represented as

$$S(z) = E(z)V(z)R(z) \tag{2.4}$$

This model can be simplified by representing the excitation by a spectrally flat signal and the synthesis filter by the spectral envelope, $H(z)$, i.e.

$$S(z) = E(z)H(z) \tag{2.5}$$

For voiced speech, $H(z)$ includes the vocal tract transfer function, the radiation effect, and aspects of the glottal source. For example, the *spectral tilt* (decaying spectrum characteristic) of the glottal source is incorporated into $H(z)$, since the excitation is spectrally flat.

The simplified source-filter model of (2.5) is widely used in speech coding, synthesis and recognition. The main reasons for the popularity of this model are that the spectral envelope representation is typically sufficient for these applications and it can be estimated using efficient techniques, such as linear prediction and cepstral analysis. These two methods are described in Sections 2.1.3 and 2.1.4 respectively. In contrast, techniques which accurately estimate the vocal tract transfer function are

typically more complex and less robust than the spectral envelope estimation methods. The methods to analyse the glottal source and vocal tract are described later in Section 2.2.3.

2.1.2 Harmonic/Stochastic Model

The spectral representation of the speech signal is often used in speech synthesis and coding applications. For example, the *channel vocoder* developed by Dudley et al. (1939), which is the earliest speech vocoder, uses a bank of *analog bandpass filters* to represent the time-varying spectral magnitudes of the speech signal in different frequency bands. Each filter has a bandwidth between 100 Hz and 300 Hz. For covering the frequency band $0 - -4$ kHz, 16 to 20 filters are commonly used (Deller et al., 1993). During synthesis, the input of the bandpass filters is obtained using pulse or noise generators. The outputs of the bandpass filters are then summed to produce the speech signal.

The spectral periodicity characteristic of voiced sounds can be used to model speech more effectively than using the whole spectrum (as in the filterbank speech model of the channel vocoder). The *harmonic model* takes into account this periodicity information. It represents the speech signal $s(n)$ as a periodic signal, $\tilde{s}_p(n)$, which is a sum of L harmonic sinusoids:

$$\tilde{s}_p(n) = \sum_{l=0}^{L-1} A_l \cos(nlw_0 + \phi_l), \tag{2.6}$$

where A_l and ϕ_l are the amplitudes and phases of the harmonics, respectively. The frequency of each harmonic is an integer multiple of the fundamental frequency $w_0 = 2\pi F_0$.

During analysis, the problem of estimating the set of parameters $\{w_0, A_l, \phi_l\}$ can be solved by calculating the *least-squares minimisation* of the following squared error, e.g. Dutoit (1997):

$$E(w) = |S(w) - \tilde{S}_p(w)|^2, \tag{2.7}$$

where $S(w)$ and $\tilde{S}_p(w)$ are the short-time Fourier transforms of $s(n)$ and $\tilde{s}_p(n)$, respectively. The error $E(w)$ can be interpreted as a *stochastic component* of the signal, which can be modelled as white Gaussian noise. In this case, a voiced/unvoiced decision can be computed from the ratio between the energies of $S(w)$ and $E(w)$, that is, a

measure of the *signal-to-noise ratio* (SNR). For SNR values below a given threshold, the speech frame is classified as unvoiced.

Speech can also be represented as the sum of a harmonic and stochastic components, i.e. $s(n) = \tilde{s}_p(n) + \tilde{s}_r(n)$. The stochastic signal $\tilde{s}_r(n)$ is typically modelled using band limited noise signals, whose energy is computed from $E(w)$, e.g. Dutoit (1997). In general, hybrid *harmonic/stochastic (H/S) models* produce more natural speech than purely harmonic models.

In speech synthesis, H/S models have been mainly used to increase the degree of parametric flexibility of concatenative speech synthesisers. For instance, they have been used to allow large prosodic variations and to modify a speaker's voice. However, the effect of spectral variations between concatenation segments degrades speech quality. This effect can be reduced using a spectral smoothing algorithm, but the problem of phase discontinuities in these models is more difficult to solve.

2.1.3 Linear Predictive Coding

Linear predictive coding (LPC) or *linear auto-regressive* (AR) modelling represents the speech samples, $s(n)$, as a linear combination of past samples plus some error (Makhoul, 1975), that is,

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + e(n), \tag{2.8}$$

where a_k is the k-th order LPC coefficient, and $e(n)$ is the *LPC residual*. In the frequency domain, this model represents an *all-pole filter* applied to the residual $E(z)$, that is,

$$S(z) = E(z)\frac{1}{A(z)}, \tag{2.9}$$

where

$$A(z) = \frac{K}{1 + \sum_{k=1}^{p} a_k z^{-k}}, \tag{2.10}$$

and K is the gain of the filter. The coefficients of $A(z)$ can be calculated by minimising the error between the actual speech samples and predicted ones, i.e. by minimising the following prediction error:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \tag{2.11}$$

The most commonly used methods to calculate the prediction coefficients are the *auto-correlation* and the *covariance* methods (Deller et al., 1993). The first requires analysis windows at least 15 ms long (typically they have a duration comparable to that of several glottal cycles). In this interval, the losses vary as a function of the time-varying glottal impedance and the vocal tract might also change. Both factors may cause the source and vocal tract estimates obtained using the autocorrelation method to be less accurate. The advantage of the covariance method is that it can estimate the filter using a very short-time window, corresponding to the *closed phase* of a single pulse (phase during which the vocal folds are closed and there is no airflow through the glottis).

LP analysis assumes that speech can be represented as an all-pole model, i.e. the all-pole filter $1/A(z)$ represents the different speech components of speech production (glottal source, vocal tract and radiation). In this model, the LPC spectrum is an approximation of the spectral envelope of the short-time signal. On the other hand, the LP residual $E(z)$ is an approximately flat signal. In the frequency domain, the residual can be calculated from the speech signal $S(z)$ using the *inverse filtering* technique, which can be represented using (2.9) as follows:

$$E(z) = S(z)A(z) \tag{2.12}$$

A typical criterion to select p, the order of the LPC analysis, is to use 1 complex pole per each kHz of the total speech bandwidth (equal to half the sample rate) to model the resonances of the vocal tract, plus 2 to 4 poles to model the radiation and glottal effects (Huang et al., 2001). For example, 12 to 14 poles are typically used for the LPC analysis of speech sampled at 16 kHz (8 kHz frequency band). The higher the p, the lower the prediction error. However, for too high p values the LPC filter fits to the amplitude spectrum of the speech signal. As result, the glottal source and vocal tract components are poorly separated. For example, it is desirable to separate the periodicity of the speech signal from the LPC filter, because the periodicity is assumed to be modelled by the residual.

The conventional *LPC vocoder* models the residual of voiced sounds as an impulse train (Deller et al., 1993). Figure 2.1 shows the source-filter model of this vocoder. During analysis, the vocoder estimates F_0, and performs a voiced/unvoiced classification. During synthesis, F_0 is used to generate an impulse train, for voiced speech. Then, this signal is filtered by the all-pole filter to generate the speech waveform. Unvoiced speech is synthesised using white noise as the input into the all-pole filter. The all-pole filter is usually *minimum-phase* (contains only poles inside the unit circle) so

that it is stable.

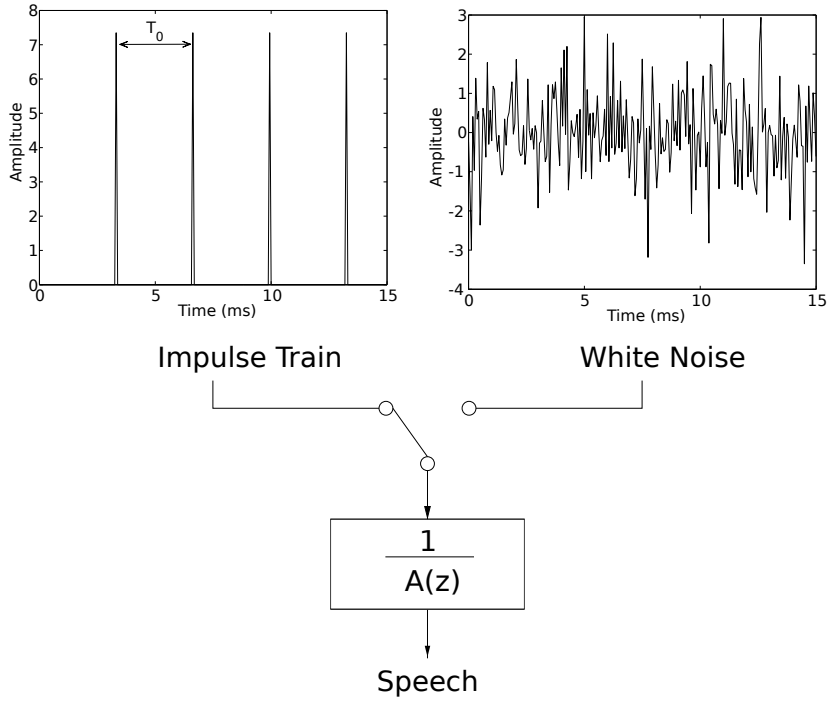

Figure 2.1: Speech synthesis using LPC.

One limitation of the all-pole model of speech is that it does not accurately model voiced sounds which contain zeros in the speech model, such as nasals or voiced fricatives. Another problem of the conventional LPC vocoder is that speech synthesised with the impulse train does not sound natural. This can be explained by the strong harmonic structure of the impulse train, which has the effect of producing a robotic or "buzzy" speech quality.

There are other LPC-based vocoders which use a better representation of the residual than the impulse train (Deller et al., 1993). For example, the *residual excited linear prediction* (RELP) vocoder transmits a low-pass filtered version of the residual in addition to the parameters of the basic LPC vocoder. A low-pass filtered version of the residual permits to transmit the speech parameters at a lower bit rate than using the original residual signal. The residual is regenerated using a *bandwidth regeneration algorithm* and then the resulting signal is passed through the all-pole filter to generate the speech waveform. An alternative to the RELP vocoder is the *code-excited linear prediction* (CELP) vocoder, which produces speech at a lower bit rate than RELP. In the CELP vocoder a relatively large number of residual signals are computed from recorded speech and stored in a *codebook of zero-mean Gaussian sequences*. Speech

is synthesised by passing a residual sequence which belongs to the codebook through a filter that is defined by the LPC coefficients. During analysis, the vocoder compares the original speech with speech synthesised with the residual sequences of the codebook, in order to find the sequence which minimises the residual error (difference between the two signals). The index of this sequence is transmitted by the vocoder together with the LPC coefficients and used to select the excitation from the codebook during the synthesis part. *Multipulse LPC* vocoders use a short sequence of pulses whose amplitudes and locations are optimised during speech analysis, in order to obtain higher speech quality compared to the conventional LPC vocoder which uses a sequence of simple identical pulses (impulse train). Another method for improving the speech quality of the basic LPC synthesis method is to mix noise with the impulse train for generating voiced speech. For example, the *mixed-excitation linear prediction* (MELP) vocoder classifies wide frequency bands of a speech segment as voiced or unvoiced. The voiced bands are modelled by the spectrum of the impulse train, whereas unvoiced bands are modelled by the noise spectrum.

In speech coding, the LPC coefficients are often converted to equivalent representations. For example, *line spectral frequency* (LSF) coefficients are often obtained from the LPC coefficients (Deller et al., 1993). LSFs have the property that their complex conjugate zeros lie on the unit circle. The advantages are that these parameters have better quantisation properties, result in low spectral distortion than conventional LPC coefficients, and the LPC filter obtained using LSFs is stable. LSFs have also been successfully used in HMM-based speech synthesis, e.g. Ling et al. (2006a), whereas LPC parameters appear to be less suited to statistical modelling.

2.1.4 Cepstrum

The cepstrum can be described as a *homomorphic transformation* (Deller et al., 1993), in which a convolution $z(n) = x(n) * y(n)$ is converted into a sum $\hat{z}(n) = \hat{x}(n) + \hat{y}(n)$. The speech signal $s(n)$, is assumed to be the convolution of two components. One component, the excitation signal $e(n)$, has its energy concentrated at the high frequencies of the spectrum. Conversely, the other component, which is the impulse response of the vocal tract system $h(n)$, has its energy concentrated at the low-frequency part of the spectrum. The speech cepstrum, $c_s(n) = c_e(n) + c_h(n)$, can be used to separate these excitation and vocal tract components. The cepstral analysis of speech is usually performed by calculating the short-term *real cepstrum* of the speech signal.

This can be computed using the short-term *discrete Fourier transform* (DFT) and the logarithm function, as shown in Figure 2.2. The analysis window, $w(n-m)$, which ends at time m, is typically implemented as a *Hamming window*, with duration of 20-40 ms. The function of the logarithm is to decompose the magnitude of the speech spectrum, $|S(w)|$, into a linear combination of the magnitudes of the excitation and impulse response parts, $|E(w)|$ and $|H(w)|$ respectively. That is,

$$\log|S(w)| = \log|E(w)| + \log|H(w)| \tag{2.13}$$

The real cepstrum discards phase information, which makes the analysis simpler (avoids the process of *phase unwrapping*). The phase information is usually neglected in speech processing applications because it is not considered to be important to the perceptual speech quality. For example, the phase is not necessary to calculate the minimum-phase impulse response of the vocal system, $h(n)$. Nevertheless, the phase information can be preserved using the *complex spectrum* of the speech signal. The complex spectrum is calculated similarly to the real spectrum, but the logarithm is applied to $S(w)$, instead of computing the logarithm of $|S(w)|$.

The two components of the cepstrum, $c_e(n)$ and $c_h(n)$, can be separated by *liftering* (analogous to the filtering in the frequency domain) the speech cepstrum, $c_s(n)$. The component $c_h(n)$ has its energy concentrated at smaller values on the time axis, whereas $c_e(n)$ has its energy concentrated at larger values on the time axis. Next, $c_h(n)$ and $c_e(n)$ can be obtained by using a "low-time" and "high-time" lifters respectively. For example, $c_h(n)$ can be estimated by multiplying the cepstrum $c_s(n)$ by a low-time lifter given by

$$l(n) = \begin{cases} 1, & 0 < n < L \\ 0, & \text{otherwise} \end{cases}, \tag{2.14}$$

where L is a value chosen, such that $\hat{h}(n) \approx 0$ for $n \geqslant L$ and $\hat{h}(n) \approx c_h(n)$ for $0 < n < L$. Typically, $l(n)$ is a time window of 2-3 ms.

Deller et al. (1993) describes the basic *cepstral vocoder*, which uses a simple model of the excitation. In this vocoder, it is assumed that voiced speech can be generated by exciting a slowly varying vocal system filter by a periodic signal, while unvoiced speech is generated by exciting the filter with white noise. The vocal system response is calculated as shown in Figure 2.2. F_0 and the voiced/unvoiced classification are also estimated during analysis. The synthesis system is shown in Figure 2.3. First, the cepstral component $c_h(n)$ is processed in order to calculate an estimate of the filter

impulse response, $\hat{h}(n)$. Next, the speech signal is obtained as the convolution of $\hat{h}(n)$ with the excitation, which is an impulse train or noise. Speech can also be synthesised by passing the excitation through the spectral envelope synthesis filter, $\hat{H}(w)$.

As in the case of LPC vocoders, the speech quality of the cepstral vocoders can be improved using a better model of the excitation. For example, Deller et al. (1993) describes another cepstral vocoder, which uses an *iterative analysis-by-synthesis method* to determine the optimal voiced excitation (Chung and Schafer, 1990). Speech is synthesised by exciting the vocal tract impulse response, using a different excitation for unvoiced, voiced and mixed speech.

In speech recognition, *mel-frequency cepstral coefficients* (MFCCs) are commonly used to represent of the vocal system impulse response (Mermelstein, 1976). The difference of the *mel-cepstrum* (defined by the MFCCs) to the real cepstrum is that a non-linear frequency scale is used. This *mel-scale* approximates the perceptual characteristics of the human auditory system. MFCCs are also known to perform well in HMM-based speech synthesis.

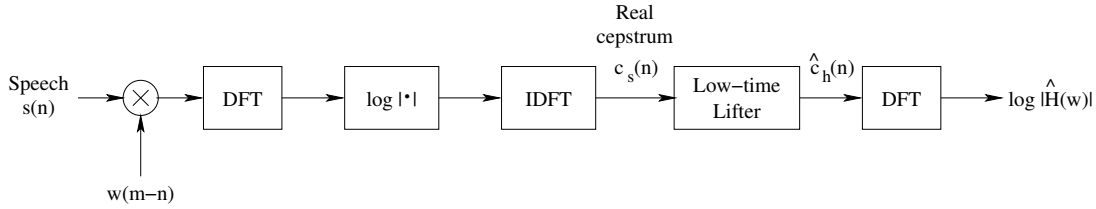

Figure 2.2: Block diagram of the method for estimating the impulse response of the vocal system by cepstral analysis, where $w(m-n)$ is an analysis window.

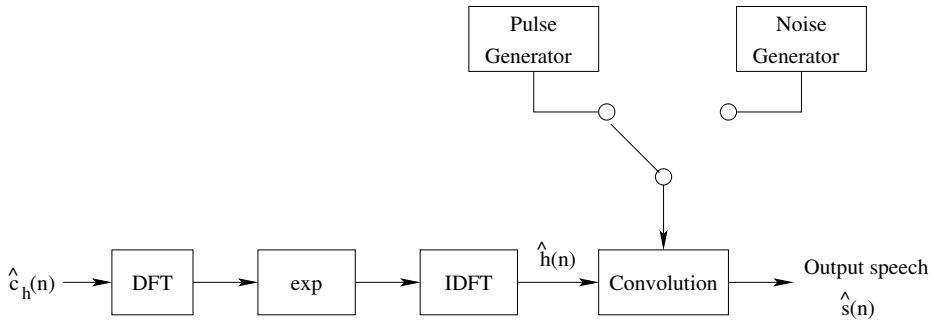

Figure 2.3: Block diagram of a typical cepstral vocoder synthesiser.

2.2 Glottal Source Modelling

This section initially presents a more detailed description of the speech production model (which was introduced in Section 2.1.1), focusing on the glottal source component of speech. In the subsequent sections, the main types of glottal source models and typical analysis methods which are used to estimate both the glottal source parameters and the vocal tract filter parameters will also be described.

2.2.1 Source-Filter Theory of Speech Production

The mechanical properties that influence the generation of sounds in the vocal tract are often described in terms of *elementary electrical theory* (e.g. impedance per unit area) and well known results of *waves on transmission lines* (Stevens, 1998; Flanagan, 1972). The next section reviews the quantitative description of the speech production system based on this type of analysis and how it relates to the linear acoustic source-filter model which is explored in this thesis (the model defined by the glottal source derivative and the vocal tract filter). Section 2.2.1.2 describes the glottal source component of speech in more detail using the electrical theory formalism. Although this voice source representation was not used in this work, it helps to show the complexity of the glottal flow and to explain its important acoustic characteristics, such as the *asymmetry of the glottal pulse*. Then, Section 2.2.1.3 discusses one of the limitations of the source-filter model which is the assumption that the source and filter components are independent. One of the effects of neglecting the source-filter interaction is the *ripple* component of the glottal source signal, which cannot usually be correctly modelled.

2.2.1.1 Speech Production Model

The acoustic analysis of speech production describes the propagation of the sound wave through the vocal cavities from the lungs to the radiating surface at the lips. To simplify the analysis, the vocal cavities are divided into contiguous parts (Stevens, 1998). This model depends on the assumption that the cross-sectional area perpendicular to the air stream is approximately constant and that the length l of the approximating sections are kept short compared to the minimum *wavelength* of the sound wave λ ($8l < \lambda$). Each section can be described in terms of the electric theory by the impedance Z:

$$Z = 1/wC_s + jwL_s + R_s \tag{2.15}$$

The *compliance* C_s represents the compressibility of the air, the *inertia* L_s is associated with the mass of the air which opposes acceleration, the *resistance* R_s represents the energy losses that can occur in the walls due to *viscous friction* and *heat conduction*, and w is the frequency (rad/s).

A model of the respiratory system can be divided into *subglottal*, *glottal* and *supraglottal* systems. Meanwhile, the supraglottal system can be divided into the following general major regions: larynx tube, the vocal tract (pharynx region and the oral cavity), the nasal tract and the radiating ports (formed by the lips and teeth, and by the nostrils). The simplified electric circuit of this system (Flanagan, 1972) is shown in Figure 2.4. In this model the sound pressure is analogous to the voltage and the volume velocity to the current in an electric line. The pressure drop in the bronchial and tracheal tubes due to the subglottal impedance Z_s is small, because they are relatively large. Consequently, the subglottic pressure source P_s is approximately equal to the lung pressure P_l, which is in general nearly constant to maintain a certain vocal effort throughout the utterance. The air flow through the glottis can make the vocal folds vibrate because of their mass and elastic characteristics. The quasi-periodic opening and closing of the cords varies the series impedance $Z_g = R_g + jwL_g$. This impedance is time-varying and non-linear. While the subglottal system can be considered to have an unconstricted configuration, there are changes in the configurations of the supraglottal cavities (they are equivalent to the impedance Z_t in Figure 2.4). For example, the narrow passage at the place where the tongue is humped, the variable constriction of the velum at the nasal tract entry, and the constriction at the lips or teeth.

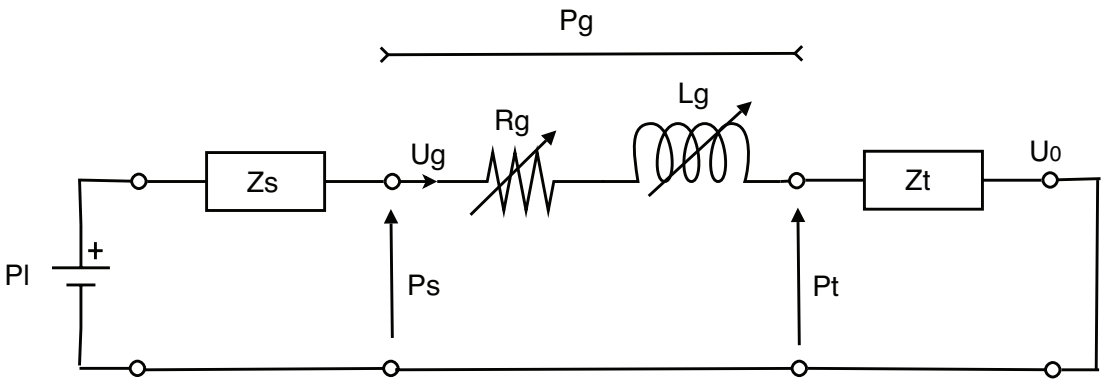

Figure 2.4: Equivalent circuit of the general parts of the respiratory system.

The source of excitation of the vocal tract can be approximated by a *volume velocity source*. According to Flanagan (1972), this approximation is valid under the assumption that the acoustic impedance of the glottis, Z_g, is usually large compared to the impedance of the supra- and sub-glottal cavities, Z_s and Z_t, respectively. He indicates that this assumption is true at least over most of the glottal cycle and over most of the frequency range of interest for speech.

If the output of the vocal tract is taken as the volume velocity u_0 at the lips, then the transfer function of the vocal tract with volume excitation u_g at the glottis is given by u_0/u_g. This is an all-pole transfer function for non-nasalised vowels (Stevens, 1998). Assuming time-invariant linearity of the vocal tract, the Fourier transform (FT) of the sound pressure $p_r(w)$ at distance r from the lips is given by

$$P_r(w) = G(w)V(w)R(w), \tag{2.16}$$

where *G(w)* is the FT of the source, *V(w)* is the transfer function of the vocal tract, and *R(w)* is the radiation characteristic.

The block diagram in Figure 2.5 shows the filtering of the source by the vocal tract. This source-filter model is equivalent to the speech production model of (2.4). It is often convenient to refer to the time derivative of the glottal flow, $u'_g(t)$, as the source. It has the same meaning of shifting the differentiation of the radiation function to the source. With $u'_g(t)$ as source, the filter function is constrained to the all-pole function of the supraglottal pathways (Fant, 1982).

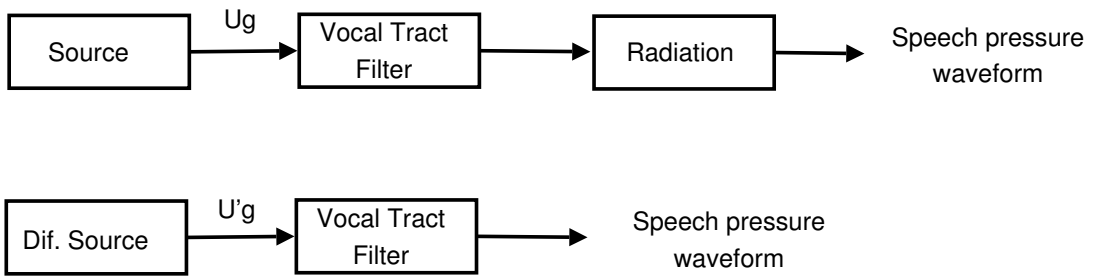

Figure 2.5: Source-filter model of speech production.

2.2.1.2 Voice Source

The glottal flow can be computed as a function of the *glottal area A(t)* and the *pressure source* P_s (Flanagan, 1972). According to the circuit of Figure 2.4 and assuming

the subglottic pressure P_s is equal to the transglottal pressure P_t (at most frequencies the driving point impedance Z_t of the vocal tract is small compared with the glottal impedance), then the volume velocity $u_g(t)$ satisfies

$$P_s = u_g(t)R_g(t) + \frac{d}{dt}\left[L_g(t)u_g(t)\right] \tag{2.17}$$

From Flanagan (1972), the inertia of the acoustic mass of the air in the glottis can be approximated by $L_g(t) = l\rho/A(t)$, where l is the length of the glottis, ρ is the *air density* and $A(t)$ is the cross-sectional area of the adjacent tubes to the glottis airways (larynx and pharynx). The main effect of the inertia is to cause a slower increase of the volume velocity when the area is increasing and a more rapid decrease in volume velocity during the closing phase of the glottis. That is, the inertia contributes to the skewness of the volume velocity waveform and causes a steeper slope during the glottal closing phase (Stevens, 1998). There are other factors which might introduce additional skewness in the waveform of the glottal air flow, such as the effect of a considerably narrow vocal tract constriction. Figure 5.1 shows an example of the glottal flow waveform calculated using an acoustic glottal source model, the Liljencrants-Fant (LF) model. The skewness characteristic of the glottal flow can be observed in this figure.

The glottal resistance $R_g(t)$ of (2.17) can be approximated by a linear combination of viscous and dynamic terms, $R_v(t)$ and $R_d(t)$, respectively. That is, $R_g(t)$ is given by (Flanagan, 1972):

$$R_g(t) = R_v(t) + kR_d(t) \simeq \frac{12d\mu}{A^3(t)} + 0.875\frac{\rho u_g(t)}{2A^2(t)}, \tag{2.18}$$

where μ is the *coefficient of viscosity*, d is the *thickness of the glottis*, and k is a real constant. The numerical approximation of $R_g(t)$ in (2.18) was obtained from steady flow measurements on models of the human larynx (Flanagan, 1972). The approximation holds within 10% for $0.1 \leqslant w \leqslant 0.2$ (mm), $P_s \leqslant 64$ cm H2O at small w and for $u_g \leqslant 2000$ cc\ sec. Over most of the open cycle of the vocal cords, the glottal resistance is determined by the kinetic part, $R_d(t)$. However, if the area and flow velocity are sufficiently small, the viscous term $R_v(t)$ predominates.

Equations (2.18) and (2.17) show that the calculation of the glottal source signal is complex. For example, (2.18) is a non-linear, first-order equation with non-constant coefficients. For an arbitrary glottal area *A(t)*, this equation is not easily integrated. Nevertheless, Flanagan (1972) calculated a rough estimate of the glottal volume velocity from the resistance expression (2.18) and by neglecting the effects of inertia $L_g(t)$

in (2.17). For these calculations, the glottal area $A(t)$ was measured from high speed motion pictures of the glottis and the subglottic pressure P_s was estimated from the sound intensity and direct tracheal pressure measurements.

The viscous term in R_g has the effect of sharpening the leading and trailing edges of the volume velocity wave. This is equivalent to increasing the amplitude of the high-frequency components in the glottal spectrum. Meanwhile, the asymmetry of the glottal volume flow produces an irregular spectrum. That is, the spectral minima are neither equally spaced nor as pronounced as for the case of a symmetrical glottal signal (Flanagan, 1972). The correlation between properties of the glottal source waveform and its spectrum will be further discussed in Section 5.3.1 for the case of the LF-model of the glottal source derivative.

2.2.1.3 Source-Filter Interaction

The volume velocity of airflow $u_g(t)$ is related to the subglottic pressure P_s through the non-linear equation (2.17). This equation assumes the transglottal pressure is constant and neglects the pressure drops at the sub- and supraglottal loads. However, the interaction between source and filter can cause significant changes in the volume velocity air flow, e.g. Ananthapadmanabha and Fant (1982). In general, only the effects of the first formant of the vocal tract and subglottal system on $u_g(t)$ are significant, because the inductance associated with higher formants is small.

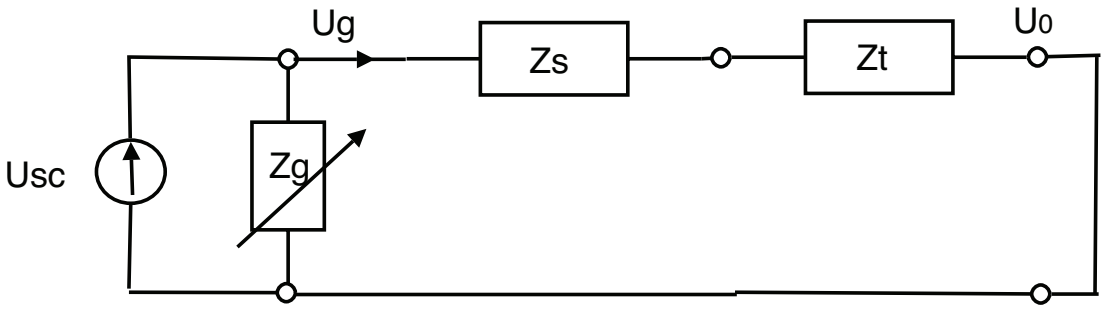

Figure 2.6: Norton's equivalent circuit of the respiratory system shown in Figure 2.4.

In the previous section, the volume velocity $u_g(t)$ represented the true glottal flow in the circuit of Figure 2.4. The voice source can also be modelled by using its *Norton's equivalent circuit* (Flanagan, 1972; Ananthapadmanabha and Fant, 1982), which is shown in Figure 2.6. In this case, the voice source is represented by the *short-circuit source* (Fant, 1981). The current source u_{sc} is calculated by short-circuiting the input

load. From (2.17), it is given by

$$u_{sc}(t) = \frac{P_l}{Z_g(t)} \tag{2.19}$$

The *true glottal flow* can be related to the short-circuit flow by the following equation (Ananthapadmanabha and Fant, 1982):

$$U_g(s) = U_{sc}(s)\left[\frac{s^2 + 2\alpha_0 s + \omega_0^2}{s^2 + 2\alpha_1 s + \omega_1^2}\right], \tag{2.20}$$

where α_0 and ω_0 are variables of a complex conjugate zero which depend on the parameters of the load, while α_1 and ω_1 are the variables of the complex conjugate pole which depend both on the load and the Norton's source impedance. The above equation was obtained by assuming the glottal impedance to be equal to the dynamic glottal resistance and also stationary for calculating the *Laplace transform*.

The complex conjugate pole pair in (2.20) is responsible for a *ripple* component of the source (Ananthapadmanabha and Fant, 1982). In the time domain, this can be interpreted as the source of the transient response of the vocal tract load, where the transients are excited at the points of discontinuity or *epochs* in u_{sc}. In the frequency domain, Ananthapadmanabha and Fant (1982) indicate that ripple is equivalent to a time varying bandwidth and resonant frequency modulation. That is, the spectrum of the ripple component is a bandpass type signal with a peak close to the first formant frequency F_1. For high F_1 vowels the maximum glottal bandwidth component could be large, causing the "truncation" of the F_1 response.

The use of the short-circuit source $u_{sc}(t)$ instead of the true glottal flow $u_g(t)$ has the advantage of being determined independently of the articulation, avoiding the linear source-filter dependency. Thus, $u_{sc}(t)$ can be easily modelled from a knowledge of glottal area function and lung pressure (Ananthapadmanabha and Fant, 1982), as it does not contain superimposed ripple components. However, the vocal tract filter function becomes very complex because it is time varying and non-linear as a consequence of the glottal impedance. Also, separate transfer functions have to be specified for open and closed phases of the glottal cycle. A practical problem is that this vocal tract filter is difficult to estimate from recordings of real speech using techniques such as inverse filtering. Alternatively, if the source function is defined by $u_g(t)$ the filter transfer function is simpler and constrained to the all-pole filter for non-nasal vowels. The use of the true glottal flow also has the advantages that it can be studied experimentally using inverse filtering and requires the specification of only the closed phase

transfer function. For example, the coefficients of the all-pole filter can be calculated efficiently using LPC analysis. However, the source $u_g(t)$ depends on the particular vocal tract configuration, which will introduce ripple components whenever the glottis is open. In this case, modelling the glottal flow signal is more difficult because the true glottal source signal is more complex, as shown in Section 2.2.1.2, and the superimposed ripple makes it more difficult to accurately estimate the glottal flow parameters.

2.2.2 Glottal Source Models

2.2.2.1 Physical Models

Most aerodynamic-mechanical vocal fold models are inspired by the *two-mass model* of Ishizaka and Flanagan (1972). This model approximates each vocal fold by a *self-oscillating* system characterised by a lower mass, an upper mass, a mechanical compliance for each mass and a coupling compliance. The cycle of vibration of the vocal folds is described by aerodynamic equations of motion of the mass-spring-damper system in terms of the *glottal rest area*, *sub-glottal pressure*, *cord-tension* parameters and the vocal tract shape. Moreover, the glottal excitation of this model is computed by incorporating source-tract interaction.

Typically, physical models can simulate very well a large variety of shapes of the glottal flow. They can also produce several natural effects related to the vocal tract interaction, such as oscillatory ripple. For these reasons, physical models are typically appropriate to study the mechanisms responsible for the behavior of the source and to be integrated into a full articulatory model. However, the price paid for the high flexibility and detailed description of the source is the high complexity of the models, such as the number of parameters involved. For example, nineteen parameters have to be estimated in the two mass model (Ishizaka and Flanagan, 1972). Physical models are often difficult to control and typically require manual tuning of the parameters. Also, the relationship between acoustic and physical parameters is not well known, which brings limitations to the use of a physical model to generate different voice qualities. Nevertheless, there have been studies which contributed to a better understanding of the variation of the acoustic parameters with those of a production model, e.g. Sciamarella and d'Alessandro (2002); Hirtum et al. (2003).

Improvements to the conventional two-mass model of Ishizaka and Flanagan (1972) have been proposed in literature, such as the *three-mass model* of Story (2003) and the adapted two-mass model of Pelorson et al. (1994). These models generally give a

more detailed description of the glottal system, but they result in increased complexity and number of parameters. On the other hand, simpler physical models than the conventional two-mass model, such as the *one-mass model* proposed by Avanzini et al. (2001), require fewer parameters and allow better controllability, but are typically less accurate.

2.2.2.2 Glottal Area Models

In the two-mass model, the vocal folds are observed from above the glottis. Glottal area models are characterised by an additional vertical cross-sectional description of the movement of the vocal folds, which permits a more realistic description of the folds contact.

The glottal area model of Titze (1984) is a good example of this type of model. It uses a kinetic description of the air and the vocal fold tissue. Titze (1984) derived a function for the tissue displacement from the glottal midline in terms of three configuration parameters which have physiological significance (*abduction quotient*, *shape quotient*, and *phase quotient*), the fundamental frequency of vibration, and three other parameters related to the geometry of the glottis and the vocal folds. The *displacement function* is used to determine the glottal area, the vocal fold contact area and the glottal volume velocity. For the glottal airflow estimation a *first-order non-linear interaction* between source and vocal tract is assumed and two additional parameters are used. They are the lung pressure and the *effective vocal tract area* that combines the subglottal and supraglottal areas.

An advantage of the typical glottal area models is that the model parametrisation of the glottal area and vocal fold contact area can be done from *electroglottography* (EGG) and *photoglottography* (PGG) measurements, respectively. When compared with physical models, the typical glottal area models have the disadvantages that they do not explore the self-oscillatory nature of vocal fold vibration and their description of the vocal fold movement is less detailed.

2.2.2.3 Acoustic Time-domain Models

A typical way of describing the source signal is in terms of a small set of parameters which are often coefficients of mathematical functions. Such models stylise the glottal pulse either in terms of the glottal flow signal or in terms of the glottal flow derivative.

Models of the glottal flow, $u_g(t)$, are often based on the analysis of the integrated

inverse filtered sound pressure signal or the *inverse filtered volume velocity waveform at the mouth*, e.g. Rothenberg et al. (1975). Rosenberg (1971) studied several pulse shape models with adjustable pulse amplitude, width, and skew. One of them, known as the "*Rosenberg model*", was composed of two trigonometric segments to model the glottal opening and closing phases, respectively, which had a slope discontinuity at *glottal closure*. Hedelin (1984) proposed a LPC vocoder which used a similar model. Fant (1979) also used a model described by cosine functions in order to control the pulse shape of the glottal source, by varying the amplitude of the cosine segment over the closing phase.

Models of the glottal flow derivative, u'_g, are used more often than models of the glottal flow, u_g. A great advantage of the first type of models is that they can be obtained directly from the inverse filtered speech signal. The glottal flow derivative also has the advantage of modelling the characteristics of the airflow around the significant instants of *glottal onset* and glottal closure more accurately.

The glottal flow derivative can be represented by a unique function, such as the exponential decreasing sine of the *Liljencrants model* (L-model), which is described by Fant et al. (1985). This model has an abrupt flow termination and does not represent the progressing closure after this flow discontinuity, which is an important aspect of the flow derivative shape. In order to overcome this limitation, u'_g is often described by a piecewise linear representation of u'_g. For example, the *A-model* proposed by Ananthapadmanabha (1984) uses two independent cosine functions, which model the rise and fall of u'_g by a smooth curve respectively. This model has the advantage that it allows for a progressive closure after the maximum closing discontinuity, by using an additional parabolic function. Fant et al. (1985) proposed the *LF-model* which is an extension of the L-model. The difference between the two is that the LF-model has an additional exponential function to model the final part after the flow discontinuity. Fant et al. (1985) argued that the LF-model provides a better overall fit to the flow waveforms obtained by inverse filtering compared with the A-model. The LF-model is very popular as it gives a good approximation of u'_g, can represent a wide variety of glottal flow shapes, and is simple (defined by six independent parameters). This model is described in detail in Chapter 5. Other acoustic models have also been developed from the point of view of being computationally more efficient or to overcome some of the limitations of the LF-model, e.g. Qi and Bi (1994); Veldhuis (1998); Schoentgen (1993).

Polynomial functions are also often used to model the glottal source. Fujisaki and

Ljungqvist (1986) proposed a source derivative model composed of a set of polynomial segments, in which the level of detail was controlled by varying the number of parameters from three to six. Other polynomial models can be found in literature with varying complexity (number of parameters typically vary from four to nine), e.g. Price (1989); Funaki and Mitome (1990); Lobo (2001).

Milenkovic (1993) proposed a glottal source representation which is more general than a polynomial model. It consists of representing a glottal pulse waveform as the *weighted sum of basis functions* $p_k(t)$, as follows:

$$g(t) = \sum_{k=1}^{m} w_k p_k(t), \quad 0 < t < T, \tag{2.21}$$

where w_k are the weighting coefficients of the basis functions, which control the pulse shape, and T is the pulse length. Milenkovic (1993) used four polynomial basic functions ($m = 4$) of order $n = 4$. The coefficients of the polynomials were calculated using a set of assumptions about the glottal pulse shape. Other papers have also proposed source models which use polynomials as basis functions, such as Thomson (1992); Kaburagi and Kawai (2003); Schnell (2006).

Another way of modelling the voice source is to use *wave shape functions*. A wave shape function transforms a sinusoid into any desired waveform. For example, Schoentgen (2003) represents the glottal signal as the combination of two wave shape polynomial functions. In this model, the source is represented by a sum of power series of sines and cosines.

2.2.2.4 Acoustic Frequency-domain Models

Voice source modelling in the frequency domain allows those spectral characteristics of the source with perceptual significance to be modelled, which simple time-domain models cannot represent. For example, the *spectral tilt*, amplitude of the first few harmonics and bandwidth of the first formant are important spectral parameters of the source, which can be modelled in the frequency domain. A general disadvantage of these models is that the details of the pulse shape cannot be described as well as in the time domain, especially around the glottal closure.

A typical method of modelling the voice source spectrum consists of representing it by an impulse response. This is the case of the model proposed by Doval et al. (2003), which represents the glottal flow signal as the impulse response of a *causal-anticausal linear filter*. The filter is an all-pole that has two anticausal poles to represent the

"*glottal formant*" and one causal pole for the spectral tilt filter. The "glottal formant" represents the maximum peak in the spectrum located at lower frequencies, while the spectral tilt is equivalent to a first order low-pass filter with a relatively high cut-off frequency and slope -6 dB/oct. Instead of an all-pole filter, Hong et al. (1994) models the voice source as the output of an *all-zero filter*. This filter is driven by an excitation signal that is the sum of an impulse train with noise.

2.2.3 Methods to Estimate the Source and the Vocal Tract

The parameterisation of the glottal source is usually performed using an estimate of the glottal source signal. Several methods have been proposed for the estimation of the glottal source and vocal tract filter from the speech signal. However, this problem is not easy to solve, because it is difficult to effectively separate the source from the vocal tract. The glottal parameters can also be measured from other signals obtained during the speech production process, like the EGG signal. In this thesis, the glottal source derivative is estimated from the speech signal in order to estimate the parameters of an acoustic glottal source model. The following sections give an overview of the main methods for separation and estimation of the glottal source signal and the vocal tract filter, from the speech signal.

2.2.3.1 Inverse Filtering Using Pre-emphasis

An estimate of the voice source signal can be obtained using inverse filtering. This technique consists of applying a filter to the speech signal, $S(z)$, with a transfer function which corresponds to the inverse of the vocal tract system, $V(z)$. This technique requires the calculation of the vocal tract. A simple method to estimate $V(z)$ is to perform LPC analysis on the speech signal, as described in Section 2.1.3. The estimated LPC parameters are used for inverse filtering the speech signal in order to obtain the residual, i.e. $E(z) = S(z)/V(z)$, where $V(z) = 1/A(z)$ is an all-pole model of speech. However, this method does not accurately separate the source from the vocal tract, because the all-pole filter models the spectral envelope of the speech signal instead of the true vocal tract. The spectral envelope incorporates the vocal tract component, the radiation effect and glottal source characteristics, such as the spectral tilt. As a result, the residual is approximately a spectrally flat signal, instead of having the characteristic decaying spectrum of the voice source.

Pre-emphasis of the speech signal prior to the LPC analysis is a technique often

used to obtain a better estimate of the vocal tract transfer function. This method consists of passing the speech signal through a pre-emphasis filter, which increases the relative energy of the speech spectrum at higher frequencies. Typically, the filter has the following form:

$$M(z) = 1 - \alpha z^{-1}, \tag{2.22}$$

where α is set close to one (the typical values range from 0.96 to 0.99) for voiced sounds and approximately equal to zero for unvoiced speech (not emphasised). The pre-emphasis filter is similar to the filter used to model the radiation effect (a zero near $z = 1$) of the speech production system which is described in Section 2.1.1. In the all-pole model of speech, the glottal source component is usually represented as a *minimum-phase glottal filter* with two real poles near $z = 1$ (Deller et al., 1993). Although this representation of the source is compatible with the all-pole model of speech, it does not model the *maximum-phase* component of the glottal source which has the effect of producing an asymmetric pulse shape. In this case, the zero of the lip radiation is assumed to cancel the spectral effect of one of the glottal poles. By using pre-emphasis, the effect of the second glottal pole is also cancelled. For this reason, LPC spectrum calculated using pre-emphasis approximates better the vocal tract. That is, the glottal source effects are better removed from the LPC spectrum.

The residual obtained from pre-emphasis LPC analysis can be represented by:

$$E(z) = S(z)A(z)/M(z) \tag{2.23}$$

This residual has a decaying spectrum due to the effect of $1/M(z)$. As result, $E(z)$ approximates better the glottal source signal than the conventional LPC residual, which is spectrally flat. However, the attenuation due to $1/M(z)$ is not a correct model of the spectral tilt. For example, the attenuation in the spectrum of $E(z)$ is always the same, whereas the tilt of the source varies.

Inverse filtering can also be performed using a different filter than the all-pole filter calculated through LPC analysis. For example, Alku and Vilkman (1994) uses the *discrete all-pole* (DAP) modelling (El-Jaroudi and Makhoul, 1991) to estimate the vocal tract transfer function by first eliminating the effect of the voice source to the speech spectrum with the help of a filter library. The DAP-technique gives a better estimate of the spectral envelope that is less biased towards harmonic frequencies than the conventional LPC-analysis.

2.2.3.2 Closed-phase Inverse Filtering

When the glottis is closed, the speech waveform is only a function of the vocal tract. Therefore, the vocal tract filter can be exactly estimated by performing the LPC analysis on the closed phased, e.g. by using the covariance method, and inverse filtering the speech signal. This approach, known as *closed-phase inverse filtering* (Wong et al., 1979), is often used to calculate the glottal flow waveform. The main difficulties with this technique are to estimate the closed phase (instants of glottal closure and opening) and the effects of the source-tract interaction when there is airflow through the glottis during the closed phase (the glottis does not close completely). This latter effect is more common in higher pitched voices (females, children) and non-modal voice qualities such as breathy and whispery voices. The parameterisation of the glottal source may also be difficult in *closed-phase analysis* of high fundamental frequency speech because the number of speech samples is small.

In traditional LP inverse filtering and closed-phase inverse filtering the vocal tract filter is assumed to be time-invariant and the source is considered to be the true glottal flow (derivative) signal, as explained in Section 2.2.1. In this case, the source contains the effects of the vocal tract interaction, specifically the ripple. This random component makes more difficult to fit a source model to the inverse filtered signal when estimating the model parameters (Milenkovic, 1986). Plumpe et al. (1999) proposed a model-based method for estimation of the glottal source which takes into account certain source characteristics such as ripple or non-typical glottal waveform shapes (which influence the inverse filtering results). They estimate the glottal flow derivative using closed-phase inverse filtering and use the LF-model to capture its coarse structure. The fine structure of the waveform is obtained by subtracting the LF-model signal from the inverse filtered signal.

2.2.3.3 Iterative Inverse Filtering

In the time domain, the influence of the source on the vocal tract estimation can be avoided by performing the analysis on the closed-phase, such as in the closed-phase inverse filtering technique. The source can also be separated from the vocal tract using iterative methods, in the frequency domain.

In the *iterative adaptive inverse filtering* (IAIF) method (Alku et al., 1991), the glottal source and the vocal tract are estimated iteratively using the inverse filtering technique. The glottal flow is first modelled as a low-order all-pole signal (2 poles).

This model is estimated by LPC analysis and its spectral effects are removed from the speech signal. Then, the resulting signal is used to obtain the initial estimate of the vocal tract using linear prediction. The glottal source waveform is also estimated by inverse filtering the speech signal using the estimated all-pole model. Next, a second estimate of the vocal tract and glottal source is performed similarly using a higher order parametric model of the glottal flow. The IAIF method is described in more detail in Section 4.5.2.2, given that it was adopted in this thesis. Unlike the closed-phase inverse filtering method, the IAIF method performs the analysis on the whole pitch period. Thus, due to source-filter interaction, the linear prediction will contain slight formant frequency and bandwidth errors which results in formant ripple in the estimated excitation.

Another iterative approach is to use a glottal source model to first eliminate the source effect on the input speech, then a *pitch-synchronous analysis* can be performed over the whole pitch period. In general, this method needs an adequate initialisation of source parameters and an iterative adaptive algorithm to optimise the parameters of the source model and the vocal tract filter simultaneously. For example, Fröhlich et al. (2001) uses the LF-model to represent the glottal source derivative and the DAP algorithm for inverse filtering. A similar method was used by Alku and Vilkman (1994), but using LP analysis and a low-order *finite impulse response* (FIR) filter to model the glottal source.

2.2.3.4 Glottal Inverse Filtering

The glottal source can be explicitly represented using the *autoregressive with exogenous input* (ARX) model of speech production as follows:

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + g(n) + e(n), \tag{2.24}$$

where $s(n)$ is the speech signal, $g(n)$ is the glottal source derivative (glottal source combined with the radiation effect), a_k are pth-order time-invariant coefficients of the all-pole filter, and $e(n)$ is the prediction error. Since $g(n)$ is not known (it is the *exogenous input*), it is usually described using a glottal source model. The glottal and vocal tract parameters can be calculated simultaneously using an optimisation algorithm to minimise an error measure. This error is often equal to the predicted *mean-square error* (MSE) of one pitch period, i.e. $\varepsilon = \sum_{m=1}^{p} e^2(m)$. Several methods using the ARX process combined with a glottal source model have been proposed to estimate the source

and the vocal tract, such as the *glottal AR* (GAR) of Fujisaki and Ljungqvist (1986), the *glottal LPC* of Hedelin (1984) and the AR-model proposed by Isaksson and Millnert (1989). The method proposed by Fröhlich et al. (2001) is similar to glottal inverse filtering using the ARX model. However, it is based on the DAP technique for inverse filtering, which was modified to include a model of the glottal flow as integral part.

The source and the vocal tract filter can also be estimated using a pole-zero representation of the speech signal, instead of the all-pole model used by conventional inverse filtering. The following *autoregressive moving average* (ARMA) process models speech with both poles and zeros:

$$s(n) = -\sum_{k=1}^{p} a_k s(n-k) + \sum_{j=1}^{q} b_j g(n-j) + g(n) + e(n), \tag{2.25}$$

where b_j are qth-order coefficients (MA coefficients). In the frequency domain, this model can be represented by

$$S(z) = \frac{B(z)}{A(z)} G(z) + \frac{E(z)}{A(z)}, \tag{2.26}$$

where $S(z)$, $G(z)$ and $E(z)$ are the z-transform of $s(n)$, $g(n)$ and $e(n)$, respectively. The vocal tract transfer function, $H(z) = B(z)/A(z)$ is equivalent to an *infinite impulse response* (IIR) filter. By setting $G(z) = 0$, (2.26) is equivalent to the LPC model of (2.9) and by setting $B(z) = 1$ it corresponds to the ARX model of (2.24). The ARMA model allows a better representation of speech than AR models, especially for nasals, fricatives and stop consonants. The main disadvantage is the increased computational complexity to estimate the parameters of the pole-zero model. Another difficulty is to determine which poles and zeros model the glottal source excitation. However, the ARMA model can be combined with a glottal source model to estimate the glottal source and vocal tract filter, e.g. Fujisaki and Ljungqvist (1987). Krishnamurthy (1992) also uses a pole-zero model for the vocal tract but he uses different transfer functions during the closed phase and the open phase, to avoid the ripple effect. For representing the glottal source derivative, he uses the LF-model.

Time-varying ARX and ARMA models have also been used to estimate the parameters of a glottal source model and the vocal tract jointly. In this case, the AR and MA time-varying coefficients are represented as $a_j(n)$ and $b_j(n)$, respectively. These extended models are able to better represent the time-varying characteristic of the vocal tract and the source-tract interaction. In this case, the resulting glottal source signal is not expected to have ripple effects. For example, Ding et al. (1995) used

the *Rosenberg-Klatt (RK) model* to represent the glottal source derivative and a time-varying AR model. The source and vocal tract parameters are estimated simultaneously using the *Kalman filtering algorithm*. Fu and Murphy (2006) also used a method based on the ARMA model and Kalman filtering for the estimation of the glottal and vocal tract parameters, but they use the LF-model to represent the glottal source derivative.

The *Glottal-ARMAX model* used by Funaki et al. (1999) is an extension of the time-varying Glottal-ARMA model, which also models white Gaussian inputs. This model can be represented by:

$$S(z) = \frac{B(z)}{A(z)}U(z) + \frac{B(z)}{A(z)}G(z) + E(z), \tag{2.27}$$

where $U(z)$ is an unknown white Gaussian input. Funaki et al. (1999) adopted the RK-model to represent the glottal source excitation $G(z)$ and used an extended Kalman filter to estimate the glottal source, the white noise and the vocal tract parameters jointly.

Inverse filtering using Glottal-AR and Glottal-ARMA models can give more accurate estimates of the vocal tract and glottal source, than inverse filtering using AR-based models. The main disadvantages of using a more complete model of speech production are the increased complexity and convergence problems of the iterative optimisation algorithms. Also, the performance of the methods usually depends on a good estimation of the number of poles and zeros, which is a difficult problem to solve.

2.2.3.5 Causal and Anticausal Component Separation

The glottal source signal has characteristics of *anticausality*, as explained by Doval et al. (2003). They indicated that if this signal is extended "to the right (towards positive times) as if it was causal, this will result in an indefinitely increasing (eventually oscillating) waveform". On the other hand, when the glottal flow signal is extended "to the left (towards negative times) as if it was anticausal, then this will result in a decreasing (eventually oscillating) waveform". According to them, the skewness of the glottal pulse towards the right part is also a characteristic of anticausality. Based on the assumption that the glottal source is a mixed phase signal, Doval et al. (2003) proposed a *causal-anticausal linear model* (CALM) of the voice source. In this model, the minimum- and maximum-phase components of the glottal flow pulse are described as an anticausal and causal linear filter, respectively. The spectral effect of the minimum-

phase characteristic is the spectral tilt at higher frequencies, while the maximum-phase effect is mainly related to a peak in amplitude spectrum at lower frequencies ("glottal formant"). In this case, the source-filter model of speech can be divided into the impulse train excitation, the causal-anticausal linear component of the glottal source and the minimum-phase transfer function of the vocal tract.

The source and filter components can also be described as the anticausal and causal components of speech respectively. In this case, the minimum-phase part of the glottal source is combined with the minimum-phase transfer function of the vocal tract. The advantage of this model is that there are analysis methods that can effectively separate the causal and anticausal components of the speech signal in the frequency domain. Bozkurt (2005) proposed to separate the causal and anticausal components of speech using the *zeros of the z-transform* (ZZT) signal representation. The ZZT is an all-zero representation of the z-transform of the speech signal $x(n)$, which is defined as the set of roots, Z_m, of the z-transform polynomial $X(z)$, as follows:

$$X(z) = \sum_{n=0}^{N-1} x(n)z^{-n} = x(0)z^{-N+1} \prod_{m=1}^{N-1} (z - Z_m), \tag{2.28}$$

where N is the length of the times series. The ZZT-decomposition method to separate the anticausal and causal components consists of splitting the roots of $X(z)$ into two subsets, Z_{AC} and Z_C. That is,

$$X(z) = x(0)z^{-N+1} \prod_{k=1}^{M_0} (z - Z_{AC,k}) \prod_{k=1}^{M_i} (z - Z_{C,k}) \tag{2.29}$$

The first group of roots, Z_{AC}, is determined as the roots which have modulus greater than one (lie outside the unit circle) and correspond to the anticausal component. Conversely, the second group, Z_C, corresponds to the roots which have modulus less than one (lie inside the unit circle) and correspond to the causal component. Bozkurt (2005) used a *Blackman window* with a size of two pitch periods and centered at the *glottal closing instant* (GCI) to obtain the short-time signals. The roots of a high order polynomial were then calculated and separated. The glottal source and the vocal tract transfer function (combined with the spectral tilt of the source) can be obtained from the Z_{AC} and Z_C by computing the DFT, respectively. According to Bozkurt (2005), the GCI detection is required to obtain separate patterns of the minimum and maximum-phase contributions. The decomposition algorithm also uses a voiced/unvoiced classification, as the analysis can only be performed for voiced frames. The main limitations of this source-tract estimation method is the computational complexity to compute the roots

of the high degree polynomial and the incomplete separation of the source component from the vocal tract, i.e. the minimum-phase contribution of the voice source (related to the spectral tilt) is not separated.

Drugman et al. (2009a) showed that the minimum- and maximum-phase components of the speech signal $x(n)$ can also be effectively separated using the complex cepstrum $\hat{x}(n)$. This method is based on the characteristic of $\hat{x}(n)$ that is either positive or negative, depending on whether $x(n)$ is causal or anticausal respectively. Then, the causal and anticausal components of $x(n)$ can be estimated as the positive and negative parts of $\hat{x}(n)$. The following relationship between $\hat{x}(n)$ and the ZZT of $x(n)$ (Steiglitz and Dickinson, 1977; Drugman et al., 2009a) shows that the source-tract decomposition is similar using the cepstrum and ZZT representations:

$$\hat{x}(n) = \begin{cases} \sum_{k=1}^{M_0} \frac{(Z_{AC,k})^n}{n}, & n < 0 \\ \sum_{k=1}^{M_i} \frac{(Z_{C,k})^n}{n}, & n > 0 \end{cases} \tag{2.30}$$

Drugman et al. (2009a) compared the cepstrum decomposition method with the ZZT decomposition method and the results showed that they produced similar estimates of the glottal source and the vocal tract transfer function. The cepstrum decomposition method has the advantage that it is computationally more efficient, but it requires a robust phase unwrapping algorithm.

2.2.4 Parameterisation of the Glottal Source

Glottal source parameters can be estimated directly from the glottal waveform, e.g. Gauffin and Sundberg (1989); Alku et al. (2002). Usually they are calculated from measurements of the glottal signal like zero crossings, minima, maxima, amplitudes, etc. These methods are typically simple but they have some disadvantages. One is that the integer values of the estimated sample points or amplitudes of the samples do not always coincide with the values of the time and amplitude parameters (may be non-integers), respectively. Consequently, the intrinsic errors can be large. The disturbance present in the estimated flow signals, e.g. aspiration noise and formant ripple, can also influence the position and amplitude of the parameters and contributes to the total error. For example, methods based on empirical derived amplitude thresholds or determination of zero crossings, e.g. Arroabarren and Carlosena (2003), usually are not robust to noise.

Another approach consists of fitting a voice source model to the glottal source signal. In general, fitting methods use a glottal source model, unlike direct estimation

methods which may or may not use a voice source model. A major advantage of using a source model is that the estimated source parameters can be used for speech synthesis. The fitting method is often performed in the time domain, e.g. Ananthapadmanabha (1984); Strik and Boves (1994). However, there are also methods which optimise the parameters in the frequency domain, e.g. Oliveira (1993); Alku and Vilkman (1996); Kane et al. (2010); Ó Cinnéide et al. (2010), or both in the time and frequency domain, e.g. Fant (1993); Ní Chasaide and Gobl (1993).

The fitting procedure tries to minimise the error between the samples of the fitted signal and the samples of the glottal source signal. A simple root-mean-square error can be used, or more sophisticated error functions may be needed to emphasise relevant aspects (e.g. the slope of the spectrum). For the fitting procedure a non-linear optimisation technique is usually employed. Also, an initial estimate of the parameters is necessary, which is often obtained using direct estimation methods.

Fitting a glottal source model to the data has many advantages compared with direct estimation techniques (Strik, 1998). For example, the use of a glottal source model permits to determine the optimal model fit for the whole period, which makes the method robust for disturbances present in the glottal signals (e.g. ripple). In contrast, direct methods try to locate events in the glottal source signal, such as maximal amplitude or zero crossing, and disturbances may lead to significant errors in the estimated parameters. Also, fitting methods make it possible to estimate parameters which are difficult to estimate from direct measurements, such as the spectral tilt. Another advantage of fitting methods is that they can estimate an exact parameter value because they fit a continuous curve of the source model to the glottal source signal, whereas a parameter value estimated by direct methods corresponds to a sample point of the glottal source signal (the time-resolution depends on the sampling frequency). However, the major problem of model matching methods is to define the trade-off between the accuracy of the temporal and the spectral match. Another important problem is that a glottal source model cannot describe all the observed glottal flow signals. This problem may be overcome by using a more detailed source model.

Chapter 3

HMM-based Speech Synthesis

3.1 Introduction

HMMs have been successfully used in automatic speech recognition (ASR) from the mid-1970s, e.g. Baker (1975), but recently they have been used for speech synthesis too. At first, HMMs were used to automatically estimate synthesis parameters for the selection of sub-word units in a concatenation speech synthesiser, e.g. Donovan and Woodland (1995). This type of hybrid synthesiser was often called *trainable speech synthesiser*, because speech data was used to train a set of *decision-tree* state-clustered HMMs. For example, Donovan and Woodland (1995) aligned the training data to the state-clustered HMMs and used the HMM state segmentation to define the speech units for unit selection. Moreover, Tokuda et al. (1995a) proposed a fully automatic and parametric speech synthesiser using HMMs. Both HMM-based speech synthesisers and hybrid systems (which combine HMMs with the concatenation of recorded units) have been increasing in popularity in the recent years.

Although the same underlying HMM technology has been used for speech synthesis and ASR, there are differences between the two applications (Zen et al., 2007a, 2009; Ostendorf and Bulyko, 2002; Dines et al., 2009). HMM-based speech recognition and synthesis systems share the type of parameters of the probabilistic models and use similar methods to learn the probability distribution. More specifically, they train the HMMs by optimising the HMM probability distribution given the sequence of speech features vectors and the sequence of sub-word units, e.g. phones. However, text-to-speech using the trained HMMs can be viewed as the inverse problem of speech recognition. ASR is related to the estimation of a word sequence from the input acoustic features using the HMMs. In contrast, speech synthesis relates to the estimation of

speech parameter sequences from input text using the HMMs.

Statistical models for ASR aim to normalise away speech parameter variations, to improve the recognition accuracy. For example, aspects of speech related to prosodic and noise variations are typically avoided, because they are not important to the word and sub-word units classification and they might degrade the performance of the speech recogniser. Conversely, statistical speech synthesis tries to preserve those aspects of speech variation which contribute to speech naturalness. For example, the F_0 parameter is used by HMM-based speech synthesisers to reproduce prosodic aspects of speech, whereas this parameter is not typically used in speech recognition. In general, the *contextual factors* used to model short-term dependencies between the phone units represented by the HMMs are also more detailed in speech synthesis than recognition. This is related to the fact that contextual dependencies have an important effect on synthetic speech quality.

The duration model of the conventional HMM, which is used for speech recognition, is also not adequate for synthesis because it does not capture the temporal structure of speech correctly. Therefore, improved duration models are typically used in HMM-based speech synthesis.

This chapter first introduces the general definitions of HMMs, which are characteristic of speech recognition. Then, the main extensions of the HMM commonly used in speech synthesis are described.

3.2 Overview of Basic HMMs

3.2.1 Definition

3.2.1.1 Structure

A hidden Markov model (HMM) is a finite state machine which changes from state i to state j each time step. At each time t that a state j is entered, a *continuous observation vector* $\boldsymbol{o}_t$ is generated from the *state output probability distribution* $b_j(\boldsymbol{o}_t)$. For a state sequence of length T, $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$, the sequence of observations is defined as $\boldsymbol{O} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T\}$. For example, mel-cepstral coefficients are often the elements of the continuous observation vector $\boldsymbol{o}$ in ASR. A HMM λ is defined by the *transition probabilities* from state i to state j, a_{ij}, the state probability distribution, $b_j(\boldsymbol{o})$, and the *initial state probabilities*, π_j. Figure 3.1 shows an example of a 3-state left-to-right HMM. In speech recognition and speech synthesis applications, left-to-right models

are typically used. Within this chapter, HMMs are assumed to be left-to-right. The transition probabilities of a left-to-right HMM satisfy $a_{ii} + a_{ij} = 1$, where a_{ii} is the probability of remaining in the same state i. The following parameter constraints are also assumed:

$$\sum_{j=1}^{N} \pi_j = 1 \tag{3.1}$$

$$\int_{-\infty}^{\infty} b_j(\boldsymbol{o}) d\boldsymbol{o} = 1 \tag{3.2}$$

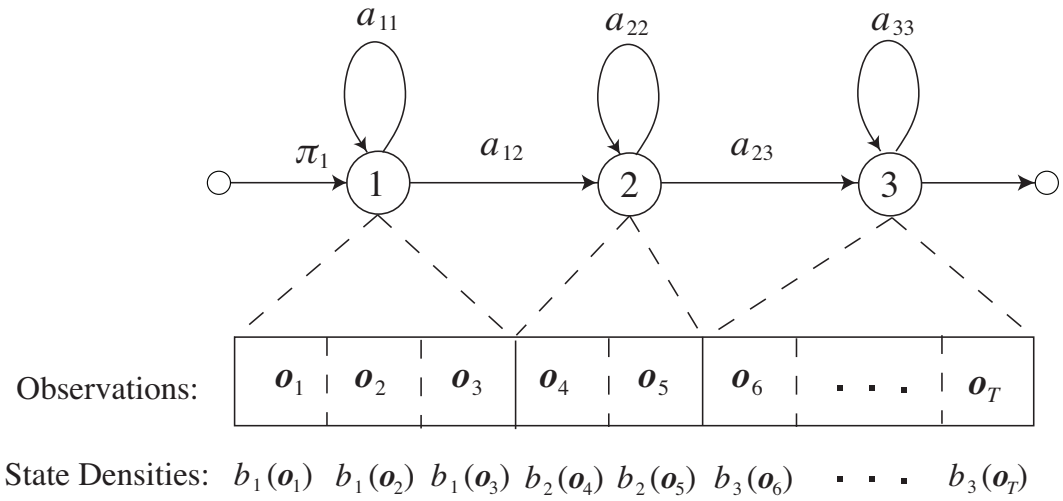

Figure 3.1: A 3-state left-to-right HMM with illustration of an observation sequence and the state output probability distributions associated with each state.

3.2.1.2 Output Probability Distribution

In continuous distribution HMM, the probability distribution $b_j(\boldsymbol{o})$ is usually modelled by K-mixtures of Gaussian distributions as follows:

$$b_j(\boldsymbol{o}) = \sum_{k=1}^{K} r_{jk} \mathcal{N}\left(\boldsymbol{o}, \boldsymbol{m}_{jk}, \boldsymbol{U}_{jk}\right) \tag{3.3}$$

$$\mathcal{N}\left(\boldsymbol{o}, \boldsymbol{m}_{jk}, \boldsymbol{U}_{jk}\right) = \frac{1}{\sqrt{(2\pi)^L |\boldsymbol{U}_{jk}|}} exp\left(-\frac{1}{2}(\boldsymbol{o} - \boldsymbol{m}_{jk})^\top \boldsymbol{U}_{jk}^{-1}(\boldsymbol{o} - \boldsymbol{m}_{jk})\right), \tag{3.4}$$

where r_{jk}, $\boldsymbol{m}_{jk}$, and $\boldsymbol{U}_{jk}$ are the mixture weight, S-dimensional mean vector (S is the dimension of $\boldsymbol{o}$), and $L \times L$ *full covariance matrix* of mixture component k of state j,

respectively. $|\boldsymbol{U}_{jk}|$ represents the determinant of $\boldsymbol{U}_{jk}$. When the elements of the continuous observation vector $\boldsymbol{o}$ are assumed to be independent and a single Gaussian is used ($K = 1$), the full covariance matrix can be restricted to its diagonal elements (*diagonal covarince matrix*). The mixture weights of a N-state HMM satisfy the following stochastic constraint:

$$\begin{cases} \sum_{k=1}^{K} r_{jk} = 1, & 1 \leq j \leq N \\ r_{jk} \geqslant 0, & 1 \leq j \leq N, \quad 1 \leq k \leq K, \end{cases} \tag{3.5}$$

so that $b_j(\boldsymbol{o})$ satisfies the constraint (3.2).

3.2.2 Assumptions

The operation of a HMM is based on the following conditional independence assumptions (Rabiner, 1989):

- a state, given the previous state, is statistically independent of all other states.
- an acoustic observation, given the state that generated it, is statistically independent of all other observations.

The first assumption can be used to calculate the probability of a state sequence, $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$, given the model λ, by multiplying the state transition probabilities:

$$P(\boldsymbol{q}|\lambda) = \prod_{t=1}^{T} a_{q_{t-1}q_t}, \tag{3.6}$$

where $a_{q_0q_1}$ is the initial state probability, which can also be represented by π_{q_1}.

Under the observation independence assumption, the probability of an observation sequence, $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$, given the HMM λ and the state sequence $\boldsymbol{q}$, can be calculated by multiplying the output probabilities of each state, as follows:

$$P(\boldsymbol{O}|\boldsymbol{q}, \lambda) = \prod_{t=1}^{T} b_{q_t}(\boldsymbol{o}_t) \tag{3.7}$$

3.2.3 Duration Model

The conventional HMM has no *explicit duration model*. However, the temporal structure of the continuous observations $\boldsymbol{o}$ can be *modelled implicitly* by the transition probabilities. The following exponential probability distribution of each state i arises from the model structure:

$$p_i(d_i) = a_{ii}^{d_i-1}(1-a_{ii}), \tag{3.8}$$

where d_i is the state duration (number of consecutive observations in state i), and a_{ii} is the state self-transition probability.

The implicit duration model of a HMM is associated with a basic segment. For example, if the basic segment is a phone, the duration model is a phone-based duration model. The prior probability of a state sequence $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$ can be calculated as:

$$P(\boldsymbol{q}|\lambda, T) = \prod_{i=1}^{N} p_i(d_i), \tag{3.9}$$

with the constraint:

$$\sum_{i=1}^{N} d_i = T, \tag{3.10}$$

where N is the total number of states and T is the total length of the sequence of states.

3.2.4 Observation Probability Calculation

3.2.4.1 Optimisation Problem

A common problem for HMMs is the computation of $P(\boldsymbol{O}|\lambda)$, i.e. the probability of the continuous observation sequence $\boldsymbol{O}$ given the model λ. For example, this problem is solved in the decoding part of a speech recogniser. In this case, the probability of the observation sequence of an unknown word is calculated for every word model (sequence of HMMs estimated in the training part) and the word model which maximises a given criterion is selected (Rabiner, 1989). Although the continuous observation sequence $\boldsymbol{O}$ is known, the underlying state sequence is hidden. Therefore, the probability of $\boldsymbol{O}$ given the model λ can be calculated by summing over all possible state sequences $\boldsymbol{q}$, that is,

$$P(\boldsymbol{O}|\lambda) = \sum_{\boldsymbol{q}} P(\boldsymbol{O}, \boldsymbol{q}|\lambda), \tag{3.11}$$

where $P(\boldsymbol{O}, \boldsymbol{q}|\lambda)$ is the probability that the observation sequence $\boldsymbol{O}$ is generated by the model λ moving through the state sequence $\boldsymbol{q}$. This joint probability can be obtained by using the *Bayes' theorem* and the statistical independence assumptions of (3.6) and (3.7), as follows:

$$P(\boldsymbol{O},\boldsymbol{q}|\lambda) = P(\boldsymbol{O}|\boldsymbol{q},\lambda)P(\boldsymbol{q}|\lambda) = \prod_{t=1}^{T} b_{q_t}(\boldsymbol{o}_t)a_{q_{t-1}q_t} \tag{3.12}$$

Then, from (3.11) and (3.12), $P(\boldsymbol{O}|\lambda)$ is given by:

$$P(\boldsymbol{O}|\lambda) = \sum_{\boldsymbol{q}} \prod_{t=1}^{T} b_{q_t}(\boldsymbol{o}_t)a_{q_{t-1}q_t} \tag{3.13}$$

However, this equation is not practical to solve, because it is too computationally demanding. A more effective way to compute $P(\boldsymbol{O}|\lambda)$ is to use a recursive algorithm such as the *forward-backward algorithm*, e.g. Rabiner (1989). The probability $P(\boldsymbol{O}|\lambda)$ can also be approximated by finding the optimum state sequence, $\boldsymbol{q}^*$, which maximises $P(\boldsymbol{q}|\boldsymbol{O},\lambda)$, e.g. Rabiner (1989). This problem is equivalent to maximising $P(\boldsymbol{q},\boldsymbol{O}|\lambda)$ and can be solved by using the *Viterbi algorithm* (Viterbi, 1967; Rabiner, 1989). The Viterbi algorithm is typically used to compute $\boldsymbol{q}^*$ and $P(\boldsymbol{O}|\lambda)$ in speech recognition.

3.2.4.2 Viterbi Algorithm

The Viterbi algorithm computes the *optimum state sequence*, $\boldsymbol{q}^*$, given an observation sequence $\boldsymbol{O}$ and the model λ, i.e. solves the problem

$$\boldsymbol{q}^* = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q},\boldsymbol{O}|\lambda) \tag{3.14}$$

The best state sequence is calculated by using the following recursion (Rabiner, 1989):

$$\delta_{t+1}(j) = b_j(\boldsymbol{O}_{t+1}) \max_{1\leqslant i\leqslant N}\{\delta_t(i)a_{ij}\}, \tag{3.15}$$

where

$$\delta_t(i) = \max_{\boldsymbol{Q}_t} P(q_t = i, \boldsymbol{O}_t|\lambda) \tag{3.16}$$

is the maximum probability of the partial observations sequence $\boldsymbol{O}_t = \{\boldsymbol{o}_1,\boldsymbol{o}_2,...,\boldsymbol{o}_t\}$, along a single path $\boldsymbol{Q}_t = \{q_1,q_2,...,q_t\}$ which ends in state i, at time t. It is initialised by setting $\delta_0(i) = 1$ for the initial entry state and zero for all other states, where $1 \leqslant i \leqslant N$ and N is the number of states of λ. The *maximisation argument* is recorded in each iteration, i.e.

$$\Psi_t(j) = \arg\max_{1\leqslant i\leqslant N} \delta_t(i-1)a_{ij} \tag{3.17}$$

At the end of the induction process, the probability of the most likely path is calculated as:

$$P^* = \max_{1 \leqslant i \leqslant N} \delta_t(i) \tag{3.18}$$

and the best state sequence is obtained by backtracking, as follows:

$$q_T^* = \arg \max_{1 \leqslant i \leqslant N} \delta_T(i) \tag{3.19}$$

$$q_t^* = \Psi_{t+1}(q_{t+1}^*), \quad t = T-1, T-2, \cdots, 1 \tag{3.20}$$

3.2.5 Model Parameter Estimation

3.2.5.1 Optimisation Problem

Another important problem for HMMs is to calculate the optimal model parameters, which best describe a given observation sequence. This problem can be described as:

$$\lambda^* = \arg\max_{\lambda} P(\boldsymbol{O}|\lambda) = \arg\max_{\lambda} \sum_{\boldsymbol{q}} P(\boldsymbol{O}, \boldsymbol{q}|\lambda) = \arg\max_{\lambda} \sum_{\boldsymbol{q}} P(\boldsymbol{O}|\boldsymbol{q}, \lambda) P(\boldsymbol{q}|\lambda) \tag{3.21}$$

The HMM λ which globally maximises $P(\boldsymbol{O}|\lambda)$ for a certain optimisation criterion, such as the *maximum likelihood*, is difficult to determine because both the λ parameters and $\boldsymbol{q}$ are unknown. However, the parameters of λ can be estimated by calculating the solution which maximises $P(\boldsymbol{O}|\lambda)$ locally. The *Baum-Welch algorithm* (Baum et al., 1970), also called the *expectation-maximisation (EM) algorithm*, is typically used to find this solution, e.g. Rabiner (1989); Young et al. (2006). This method is described in the next section.

The HMM *training part* of a speech recogniser can be regarded as an optimisation procedure with a known word sequence $\boldsymbol{Z} = \{z_1, z_2, ..., z_S\}$. In this case, (3.21) can be written as

$$\lambda^* = \arg\max_{\lambda} P(\boldsymbol{O}|\boldsymbol{Z}, \lambda) \tag{3.22}$$

In general, a text analysis procedure is used to assign contextual factors to the word sequence $\boldsymbol{Z}$ and to map it into a sequence of context-dependent sub-word units, such as sequence of phones. Each context-dependent unit is then modelled by a different context-dependent HMM, e.g. a triphone HMM. The context-dependent factors are

related to accent, lexical stress, part-of-speech, etc. In the training part, the phonetic context of the models needs to be initialised and the observation sequences segmented into states. The initial model could be one already created from another set of speakers, or it could be obtained from a uniform distribution of each word into states (Rabiner, 1989). The segmentation can be performed by using the Viterbi algorithm to find the best state sequence.

3.2.5.2 Baum-Welch Algorithm

The HMM parameter estimation method using the Baum-Welch algorithm consists of maximising the following auxiliary function of current model λ' and new λ:

$$A(\lambda',\lambda) = \sum_{\boldsymbol{q}} P(\boldsymbol{O},\boldsymbol{q}|\lambda') \log P(\boldsymbol{O},\boldsymbol{q}|\lambda) \tag{3.23}$$

$A(\lambda',\lambda)$ is maximised over λ to improve λ' in the sense of increasing the likelihood of the HMM λ, $P(\boldsymbol{O},\boldsymbol{q}|\lambda)$, e.g. Rabiner and Juang (1993). The parameters of λ are the initial i-th state probability π_i, the transition probabilities, a_{ij}, and the coefficients of the mixture density function of (3.4): r_{jk}, m_{jk}, and U_{jk}. From (3.12), the likelihood of a continuous HMM λ for the hidden state sequence $\boldsymbol{q}$, $P(\boldsymbol{O},\boldsymbol{q}|\lambda)$, can be given by

$$\log P(\boldsymbol{O},\boldsymbol{q}|\lambda) = \log \pi_{q_0} + \sum_{t=1}^{T} \log a_{q_{t-1}q_t} + \sum_{t=1}^{T} \log b_{q_t}(\boldsymbol{o}_t) \tag{3.24}$$

Equations (3.23) and (3.24) can be used to derive the formulae to calculate the HMM parameters, as described in (Rabiner, 1989). The re-estimation formulae of the initial and transition probabilities are given by:

$$\pi_i = \frac{\alpha_0(i)\beta_0(i)}{\sum_{j=1} N\alpha_T(j)} \tag{3.25}$$

$$a_{ij} = \frac{\sum_{t=1}^{T} \alpha_{t-1}(i) a_{ij} b_j(\boldsymbol{o}_t) \beta_t(j)}{\sum_{t=1}^{T} \alpha_{t-1}(i)\beta_{t-1}(i)}, \tag{3.26}$$

where $\alpha_t(i)$ is the probability of the partial observation sequence from one to t of state i, at time t. On the other hand, $\beta_t(i)$ is the probability of the partial observation from t to T, i.e.

$$\alpha_t(i) = P(\boldsymbol{o}_1,\boldsymbol{o}_2,...,\boldsymbol{o}_t,q_t = i|\lambda) \tag{3.27}$$

$$\beta_t(i) = P(\boldsymbol{o}_{t+1},\boldsymbol{o}_{t+2},...,\boldsymbol{o}_T|q_t = i,\lambda) \tag{3.28}$$

The forward-backward algorithm (Rabiner, 1989) can be used to calculate recursively $\alpha_t(i)$, the *forward probability*, and $\beta_t(i)$, the *backward probability*, as follows :

$$\alpha_1(i) = \pi_i b_i(\boldsymbol{o}_1), \qquad 1 \leqslant i \leqslant N \tag{3.29}$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(\boldsymbol{o}_{t+1}), \qquad 1 \leqslant t \leqslant T-1 \tag{3.30}$$

$$1 \leqslant j \leqslant N \tag{3.31}$$

$$\beta_T(i) = 1, \qquad 1 \leqslant i \leqslant N \tag{3.32}$$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\boldsymbol{o}_{t+1}) \beta_{t+1}(j), \qquad t = T-1, T-2, ..., 1 \tag{3.33}$$

$$1 \leqslant i \leqslant N \tag{3.34}$$

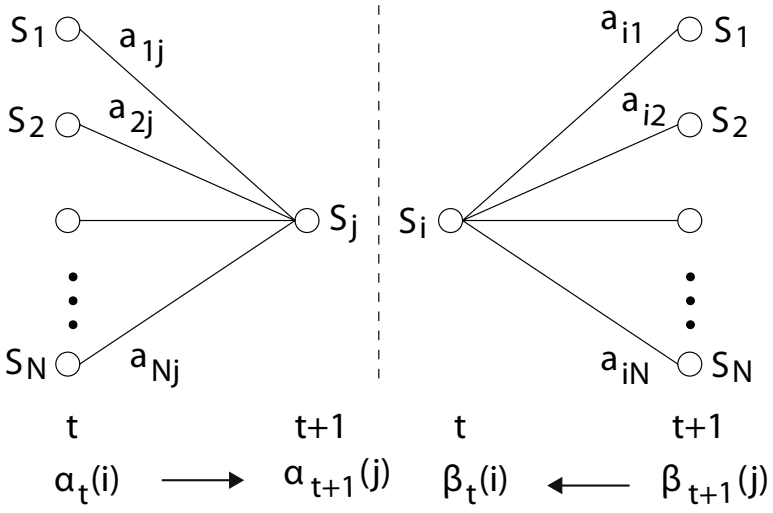

Figure 3.2: Illustration of the computation of the forward and backward probabilities, $\alpha_t(i)$ and $\beta_t(j)$, respectively.

Figure 3.2 illustrates the calculation of $\alpha_{t+1}(i)$ and $\beta_t(i)$. The probability of reaching state S_j at time $t+1$ via state S_i at time t is obtained by summing $\alpha_t(i)a_{ij}$ over all the N possible states S_i at time t, with $1 \leqslant i \leqslant N$. Then, the forward probability $\alpha_{t+1}(j)$

is obtained by multiplying this sum by the probability of the observation $\boldsymbol{o}_{t+1}$ at state j, $b_j(\boldsymbol{o}_{t+1})$. At each time t the forward probability $\alpha_{t+1}(j)$ is computed for all states j. On the other way, the backward probability $\beta_t(i)$ is calculated by summing over i the product of the transition probability a_{ij}, the probability of the observation $\boldsymbol{o}_{t+1}$ in state j, and the probability of the partial observation sequence $\boldsymbol{o}_{t+1}, \boldsymbol{o}_{t+2}, \ldots, \boldsymbol{o}_T$.

The parameters of the output probability distribution, which is given by (3.3) and (3.4), can be calculated using the forward and backward variables. The maximum likelihood re-estimation formulae for these parameters are given by

$$r_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_t(j,k)} \tag{3.35}$$

$$\boldsymbol{m}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \boldsymbol{o}_t}{\sum_{t=1}^{T} \gamma_t(j,k)} \tag{3.36}$$

$$\boldsymbol{U}_{jk} = \frac{\sum_{t=1}^{T} \gamma_t(j,k) \cdot (\boldsymbol{o}_t - \boldsymbol{m}_{jk})(\boldsymbol{o}_t - \boldsymbol{m}_{jk})^\top}{\sum_{t=1}^{T} \gamma_t(j,k)}, \tag{3.37}$$

where k indexes the mixture component of $P(\boldsymbol{o}_t)$ and $\gamma_t(j,k)$ is the probability of being in state j and component k at time t, which is given by

$$\gamma_t(j,k) = \left[\frac{\alpha_t(j)\beta_t(j)}{\sum_{j=1}^{N} \alpha_t(j)\beta_t(j)} \right] \left[\frac{r_{jk} \mathcal{N}(\boldsymbol{o}_t, \boldsymbol{m}_{jk}, \boldsymbol{U}_{jk})}{\sum_{k=1}^{K} \mathcal{N}(\boldsymbol{o}_t, \boldsymbol{m}_{jk}, \boldsymbol{U}_{jk})} \right] \tag{3.38}$$

3.3 Extension to Speech Synthesis

3.3.1 Speech Feature Generation Algorithm

In HMM-based speech synthesis, the generation of the *optimal speech feature sequence*, $\boldsymbol{O}^*$, given the model λ is more complex than the problem of finding the best state sequence in the decoding operation of ASR. This is because for speech synthesis both the observation and the state sequences are unknown. In this case, the optimisation problem is the following:

$$\boldsymbol{O}^* = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\lambda, T) \tag{3.39}$$

3.3.1.1 Optimisation Problem

For a given continuous HMM λ, the problem of generating the speech parameter vector sequence $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, ..., \boldsymbol{o}_T)$ from λ is to maximise the likelihood function $P(\boldsymbol{O}|\lambda, T)$

with respect to $\boldsymbol{O}$, as follows:

$$\boldsymbol{O}^* = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\lambda, T) = \arg\max_{\boldsymbol{O}} \sum_{\mathbf{q}} P(\boldsymbol{O}, \boldsymbol{q}|\lambda, T) \tag{3.40}$$

The problem of calculating the probability $P(\boldsymbol{O}, \boldsymbol{q}|\lambda, T)$ for a known observation $\boldsymbol{O}$ and hidden state sequence $\boldsymbol{q}$, which was described by (3.11) and (3.12), can be solved using the Viterbi algorithm. However, in (3.40) both $\boldsymbol{O}$ and $\boldsymbol{q}$ are unknown and there is no known method to analytically solve this problem. Nevertheless, the optimum $\boldsymbol{O}^*$ which locally maximises $P(\boldsymbol{O}|\lambda, T)$, can be calculated using an EM-based iterative optimisation algorithm (Tokuda et al., 2000). In this case, the state sequence (state and mixture sequence for a multi-mixture HMM) is unobservable.

Another method to estimate the optimal speech parameter sequence, consists of maximising $P(\boldsymbol{O}, \boldsymbol{q}|\lambda, T)$ with respect to $\boldsymbol{O}$ and $\boldsymbol{q}$, e.g. Tokuda et al. (2000). This method approximates the optimum sequence in a similar manner to the Viterbi algorithm, as follows:

$$\boldsymbol{O}^* \simeq \arg\max_{\boldsymbol{O}} \left(\max_{\boldsymbol{q}} P(\boldsymbol{O}, \boldsymbol{q}|\lambda, T) \right) = \arg\max_{\boldsymbol{O}} \left(\max_{\boldsymbol{q}} P(\boldsymbol{O}|\boldsymbol{q}, \lambda, T) P(\boldsymbol{q}|\lambda, T) \right) \tag{3.41}$$

This problem cannot be solved using the Viterbi algorithm described in Section 3.2.4.2, because $\boldsymbol{q}$ and $\boldsymbol{O}$ have to be determined simultaneously. However, it can be divided into the following two optimisation problems:

$$\boldsymbol{q}^* = \arg\max_{\boldsymbol{q}} P(\boldsymbol{q}|\lambda, T) \tag{3.42}$$

$$\boldsymbol{O}^* = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\boldsymbol{q}^*, \lambda, T) \tag{3.43}$$

It is computationally expensive to obtain the analytical solution of these problems, because of a too high combination of possible state sequences. In order to overcome this limitation, Tokuda et al. (1995b,a) proposed an effective method, which is typically faster than using the EM-based algorithm to solve the optimisation problem of (3.40). The Viterbi and the EM-based algorithms used to calculate the optimum $\boldsymbol{O}^*$ will be described in Sections 3.3.1.3 and 3.3.1.4 respectively. First, the importance of the *dynamic features* for parameter generation in HMM-based speech synthesis is explained.

In general, the speech parameter trajectories generated using static features only are not smooth. For example, this can be shown by considering the optimisation problem

of (3.41). Assuming that the state output probabilities are independent, the solution for the optimisation problem $\boldsymbol{O}^*$, given $\boldsymbol{q}^*$, can be obtained by using the following equation

$$P(\boldsymbol{O}|\mathbf{q}^*,\lambda,T) = \prod_{t=1}^{T} b_{\boldsymbol{q}^*}(\boldsymbol{o}_t) \tag{3.44}$$

The optimal speech parameter vector sequence $\boldsymbol{O}^*$ is the one that maximises $b_{\boldsymbol{q}}(\boldsymbol{o}_t)$ for $t = \{1,2,...,T\}$. The result is a sequence of mean vectors of the optimum state sequence $\boldsymbol{q}^*$ (Tokuda et al., 1995b).

Figure 3.3 shows an example of the sequence of mean output vectors obtained from the HMMs to synthesise a speech segment. This figure was obtained from Masuko (2002), with permission of the author. The variations between mean parameter vectors at transitions of states are often sufficiently high to produce discontinuities in the parameter trajectories, which cause degradation of the synthetic speech quality. This parameter discontinuity problem can be avoided using dynamic features, which are explained next.

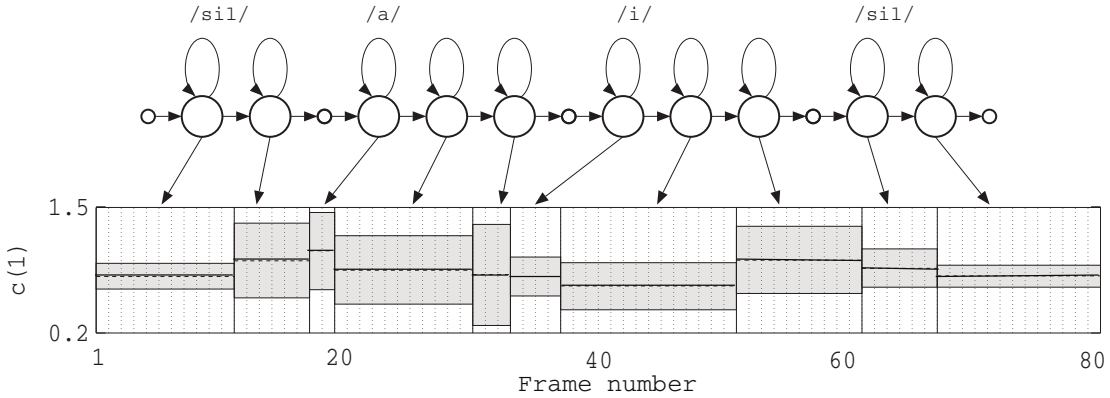

Figure 3.3: Example of the mean vector of the 1^{st}-order mel-cepstral coefficient generated by the HMMs to synthesise a segment of speech which consists of two phones and is delimited by segments of silence. This figure is a modified version of Figure 4.1 from Masuko (2002), which is used in this thesis with permission of the author. The original figure was modified by the author of this thesis in order to only show the trajectory of the mean vectors for the 1^{st}-order mel-cepstral coefficient.

3.3.1.2 Dynamic Features

The HMM has important limitations to model the time-varying characteristic of the speech observations. On one hand, the time-dependency of the observation vector sequence within a state cannot be represented, because the statistics of the observations of each state are stationary. This is the reason why the HMM generates a stepwise mean trajectory. On the other hand, the dependency between the output density function of a state and other states cannot be modelled under the observation independence assumption of (3.7). These problems can be overcome by using a different model from HMM which takes into account explicit dynamics of the speech signal, such as *segmental HMMs* (Russell, 1993) and *Hidden Dynamic Models* (Deng, 1998). However, the use of such models generally results in increased computational complexity. In ASR and HMM-based speech synthesis the typical method used to capture time dependencies is to augment the original static feature vector with dynamic features. The dynamic features are calculated as a linear combination of several adjacent static features. This augmented feature vector is able to capture short-term dependencies, because it depends on the adjacent frames. When the HMM is used as a generative model, the speech feature sequence is determined so as to maximise the likelihood of the output probability distribution using the constraints between static and dynamic features. By using the relationships between static and dynamic features, the HMM generates a smooth parameter trajectory instead of the piecewise stationary sequence of mean vectors.

The advantage of using augmented feature vectors is that the typical dynamic programming algorithms used to solve the HMM statistical problems can be used, e.g. the Viterbi and EM algorithms. However, the observation vectors are assumed to be statistically independent and the correlations between them are not taken into account in the training. As result, the constraints imposed on the generation of the speech features are from the output static features and do not represent the temporal constraints of the training data. This problem can be overcome by using the *trajectory-HMM* (Tokuda et al., 2004; Zen et al., 2007b). In this trajectory model, the probability density function is defined as a function of the static features and explicit relationships between the static and dynamic features are imposed through the normalisation of the original likelihood $P(\boldsymbol{O}|\boldsymbol{q},\lambda)$. The Viterbi and EM algorithms can also be used for trajectory-HMM but the computations are typically more complex than for the standard HMM.

The SD-dimensional parameter vector with static and dynamic features, $\boldsymbol{o}_t$, can be

represented by

$$\boldsymbol{o}_t = \left[\boldsymbol{c}_t^\top, \Delta^{(1)}\boldsymbol{c}_t^\top, ..., \Delta^{D-1}\boldsymbol{c}_t^\top\right]^\top, \tag{3.45}$$

where $\boldsymbol{c}_t$ and $\Delta^d\boldsymbol{c}_t$ are the S-dimensional static and the d-th dynamic feature vectors, respectively. In HMM-based speech synthesis these vectors are usually calculated as follows:

$$\boldsymbol{c}_t = [c_t(1), c_t(2), ..., c_t(M)]^\top \tag{3.46}$$

$$\Delta^{(d)}\boldsymbol{c}_t = \sum_{\tau=-L_-^d}^{L_+^d} w^{(d)}(\tau)\boldsymbol{c}_{t+\tau}, \tag{3.47}$$

where $w^{(d)}(\tau)$ is a window coefficient for calculating the d-th dynamic feature, $L_-^0 = L_+^0 = 0$ and $w^{(0)}(0) = 1$. The number of dynamic feature vectors is often two ($D = 3$). That is, the observation feature vector $\boldsymbol{o}_t$ is defined by the static coefficients, its delta and delta-delta coefficients. These delta and delta-delta features are typically obtained by using the following equations:

$$\Delta\boldsymbol{c}_t = \frac{\sum_{\tau=-l}^{l}(\boldsymbol{c}_{t+\tau} - \boldsymbol{c}_t)}{\sum_{\tau=-l}^{l}\tau^2} \tag{3.48}$$

$$\Delta^2\boldsymbol{c}_t = \frac{1}{2}\frac{\sum_{\tau=-l}^{l}\tau^2\boldsymbol{c}_{t+\tau} - \frac{1}{L}(\sum_{\tau=-l}^{l}\tau^2)(\sum_{\tau=-l}^{l}\boldsymbol{c}_{t+\tau})}{\sum_{\tau=-l}^{l}\tau^4 - 1}, \tag{3.49}$$

where $L = 2l + 1$ is the width of the window used to calculate the dynamic features at frame t. For example, a three-frame window is used in the HTS synthesiser (Tokuda et al., 2009), which is defined by the following formulas:

$$\Delta\boldsymbol{c}_t = 0.5\boldsymbol{c}_{t-1} - 0.5\boldsymbol{c}_{t+1} \tag{3.50}$$

$$\Delta^2\boldsymbol{c}_t = 0.25\boldsymbol{c}_{t-1} - 0.5\boldsymbol{c}_t + 0.25\boldsymbol{c}_{t+1} \tag{3.51}$$

In general, the dynamic features used in ASR are different from those given by (3.48) and (3.49). For example, the HTK toolkit (Young et al., 2006) uses the following formulas to calculate the delta and delta-delta features for ASR:

$$\Delta\boldsymbol{c}_t = \frac{\sum_{\tau=1}^{n}(\boldsymbol{c}_{t+\tau} - \boldsymbol{c}_t)}{2\sum_{\theta=1}^{n}\theta^2} \tag{3.52}$$

$$\Delta^2 \boldsymbol{c}_t = \frac{\sum_{\tau=1}^{n} (\Delta \boldsymbol{c}_{t+\tau} - \Delta \boldsymbol{c}_t)}{2 \sum_{\theta=1}^{n} \theta^2}, \tag{3.53}$$

where n is the half size of the window used to compute the dynamic feature at frame t. For example, choosing $n = 2$ yields the following formulas to calculate the dynamic features (Zhang, 2009):

$$\Delta \boldsymbol{c}_t = -0.2\boldsymbol{c}_{t-2} - 0.1\boldsymbol{c}_{t-1} + 0.1\boldsymbol{c}_{t+1} + 0.2\boldsymbol{c}_{t+2} \tag{3.54}$$

$$\begin{aligned} \Delta^2 \boldsymbol{c}_t &= 0.04\boldsymbol{c}_{t-4} + 0.04\boldsymbol{c}_{t-3} + 0.01\boldsymbol{c}_{t-2} - 0.04\boldsymbol{c}_{t-1} - 0.1\boldsymbol{c}_t \\ &\quad -0.04\boldsymbol{c}_{t+1} + 0.01\boldsymbol{c}_{t+2} + 0.04\boldsymbol{c}_{t+3} + 0.04\boldsymbol{c}_{t+4} \end{aligned} \tag{3.55}$$

The dynamic coefficients which are typically used for speech synthesis seem to produce smoother trajectories than the coefficients used for ASR (Zhang, 2009). In general, the smoother trajectories are preferred for the speech synthesis application. However, Zhang (2009) showed that the Δ and Δ^2 features used in ASR performed better in the recognition task than the Δ and Δ^2 used in speech synthesis.

Figure 3.4 shows an example of the first order mel-cepstral parameter and its dynamic parameters generated by a HMM-based speech synthesiser. This figure was obtained from Masuko (2002), with permission of the author. Dashed lines indicate means of output distributions, grey areas indicate the regions within standard deviations, and solid lines represent the parameter trajectories generated by the HMMs. In general, the generated trajectories are close to the mean of static features in the central states of the HMMs, since the variances of static and dynamic features are small. In contrast, at the first and last states of the HMMs, the trajectories are more dependent on the values of the previous and preceding frames. Nevertheless, the parameter variances at the transition states are sufficiently high to obtain smooth trajectories.

3.3.1.3 Method using the Viterbi Algorithm

Tokuda et al. (1995b) proposed a method to solve the optimisation problem given by (3.41). That is, the maximisation of $P(\boldsymbol{O}, \boldsymbol{q}|\lambda, T)$ with respect to the sequence of observation vectors $\boldsymbol{O}$ and the state sequence $\boldsymbol{q}$. This problem is solved in a similar manner to the Viterbi algorithm. It consists of searching for the optimum state sequence and solving a set of linear equations. In order to obtain smooth trajectories using dynamic trajectories, $P(\boldsymbol{O}, \boldsymbol{q}|\lambda, T)$ is optimised under the constraints of (3.47).

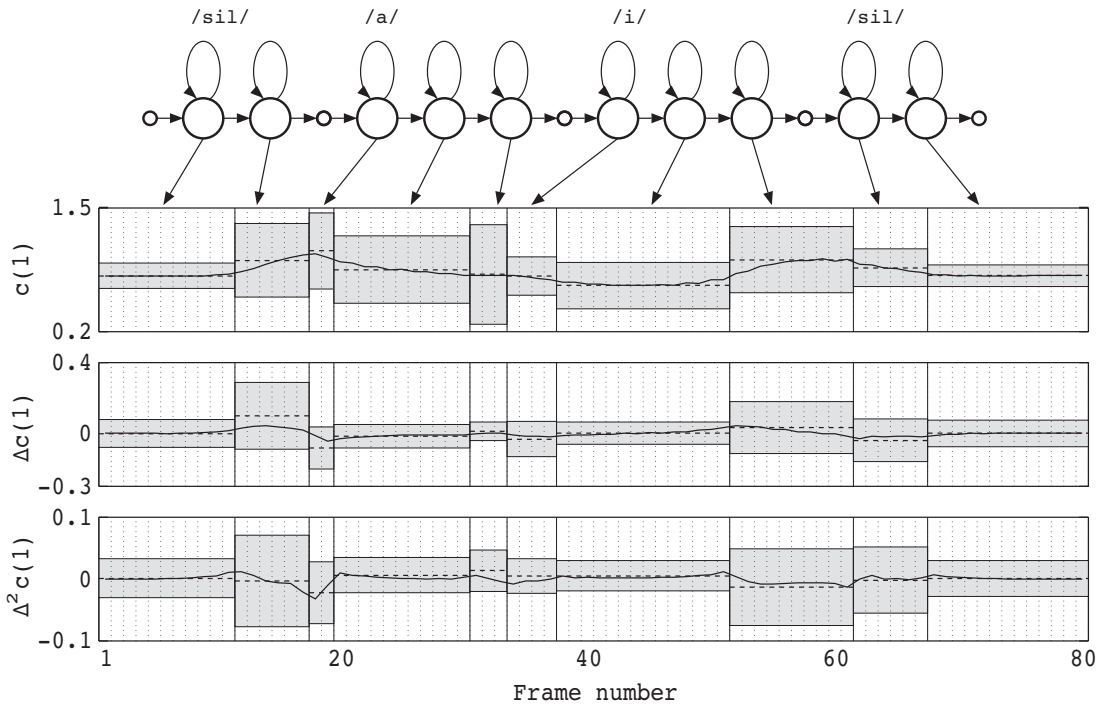

Figure 3.4: Example of the trajectory of the 1^{st}-order mel-cepstral coefficient generated by the HMMs using dynamic features, to synthesise a segment of speech. The last two plots represent the dynamic features generated by the HMMs. The variance is higher in the transition between states, which permits to obtain smooth trajectories using the parameter generation algorithm. This figure is part of Figure 4.1 from Masuko (2002) which is used in this thesis with permission of the author.

In the case of a continuous mixture HMM, λ, Tokuda et al. (1995b) considered the mixtures components of the output distribution $b_{\boldsymbol{q}}(\boldsymbol{o}_t)$ to be sub-states. Under this assumption, $P(\boldsymbol{O},\boldsymbol{Q}|\lambda,T)$ is maximised with respect to $\boldsymbol{O}$ and $\boldsymbol{Q}$, where

$$\boldsymbol{Q} = \{(q_1,k_1),(q_2,k_2),\cdots,(q_T,k_T)\} \tag{3.56}$$

is the state and mixture sequence, i.e. (q,k) is the k-th mixture of state q. The method described in the following paragraphs considers multi-mixture components but the same method can be used for a single mixture HMM (Tokuda et al., 1995a).

For finding the linear equations used to solve the optimisation problem, the super-vector made from all of the continuous parameter vectors, i.e. $\boldsymbol{O} = \left[\boldsymbol{o}_1^\top,\boldsymbol{o}_2^\top,...,\boldsymbol{o}_T^\top\right]^\top$, is arranged in the following matrix form, by using the conditions of (3.47):

$$\boldsymbol{O} = \boldsymbol{W}\boldsymbol{C}, \tag{3.57}$$

where

$$\mathbf{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_T]^\top, \tag{3.58}$$

$$\mathbf{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_T]^\top, \tag{3.59}$$

$$\boldsymbol{w}_t = \left[\boldsymbol{w}_t^{(0)}, \boldsymbol{w}_t^{(1)}, \boldsymbol{w}_t^{(2)}\right], \tag{3.60}$$

$$\begin{aligned}\boldsymbol{w}_t^{(n)} = \Bigg[& \underset{1st}{\mathbf{0}_{S\times S}}, \cdots, \mathbf{0}_{S\times S}, \\ & \underset{(t-L_-^{(n)})-th}{w^{(n)}(-L_-^{(n)})\boldsymbol{I}_{S\times S}}, \cdots, \underset{t-th}{w^{(n)}(0)\boldsymbol{I}_{S\times S}}, \cdots, \underset{(t+L_+^{(n)})-th}{w^{(n)}(-L_+^{(n)})\boldsymbol{I}_{S\times S}}, \\ & \mathbf{0}_{S\times S}, \cdots, \underset{T-th}{\mathbf{0}_{S\times S}}\Bigg]^\top, \quad n = 0,1,2\end{aligned} \tag{3.61}$$

and $\mathbf{0}_{S\times S}$ and $\boldsymbol{I}_{S\times S}$ are the $S\times S$ zero matrix and identity matrix, respectively. The dimensions of $\boldsymbol{O}$, $\boldsymbol{C}$, $\boldsymbol{w}$ and $\boldsymbol{W}$ are respectively $3MT$, MT, T, and $3MT\times MT$. By using (3.57), the optimisation of $\boldsymbol{O}$, being $\boldsymbol{Q}^*$ known, is given by

$$\boldsymbol{O}^* \simeq \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\boldsymbol{Q}^*, \lambda, T) = \arg\max_{\boldsymbol{C}} P(\boldsymbol{WC}|\boldsymbol{Q}^*, \lambda, T) \tag{3.62}$$

This problem can be solved by maximising $\log P(\boldsymbol{WC}|\boldsymbol{Q}^*, \lambda, T)$ with respect to $\boldsymbol{C}$, that is:

$$\frac{\partial \log P(\boldsymbol{WC}|\boldsymbol{Q}^*, \lambda, T)}{\partial \boldsymbol{C}} = \mathbf{0} \tag{3.63}$$

Typically, the probability density $P(\boldsymbol{WC}|\boldsymbol{Q}^*, \lambda, T)$ is assumed to be a Gaussian distribution, which can be represented by

$$P(\boldsymbol{WC}|\boldsymbol{Q}^*, \lambda, T) = \frac{1}{\sqrt{(2\pi)^{3ST}|\boldsymbol{U}|}} exp\left(-\frac{1}{2}(\boldsymbol{WC}-\boldsymbol{M})^\top \boldsymbol{U}^{-1}(\boldsymbol{WC}-\boldsymbol{M})\right), \tag{3.64}$$

with

$$\boldsymbol{M} = \left[\boldsymbol{m}_{q_1,k_1}^\top, \boldsymbol{m}_{q_2,k_2}^\top, ..., \boldsymbol{m}_{q_T,k_T}^\top\right]^\top \tag{3.65}$$

$$U^{-1} = \text{diag}\left[\boldsymbol{U}_{q_1,k_1}^{-1}, \boldsymbol{U}_{q_2,k_2}^{-1}, ..., \boldsymbol{U}_{q_T,k_T}^{-1}\right], \tag{3.66}$$

where $\boldsymbol{m}_{q_t,k_t}$ is the $3S\times 1$ mean vector and $\boldsymbol{U}_{q_t,k_t}$ is the $3S\times 3S$ covariance matrix, associated with the k_t-th mixture of $\boldsymbol{c}_t$ at the state q_t. The speech parameter vector

sequence, $\boldsymbol{C}$, which maximises $P(\boldsymbol{O},\boldsymbol{Q}|\lambda,T)$ can be calculated by solving the following set of linear equations, which are obtained from (3.63) and (3.64):

$$\boldsymbol{W}^\top \boldsymbol{U}^{-1}\boldsymbol{W}\boldsymbol{C} = \boldsymbol{W}^\top \boldsymbol{U}^{-1}\boldsymbol{M} \tag{3.67}$$

Then, the problem of maximising $P(\boldsymbol{O},\boldsymbol{Q}|\lambda,T) = P(\boldsymbol{O}|\boldsymbol{Q},\lambda,T)P(\boldsymbol{Q}|\lambda,T)$, with respect to $\boldsymbol{c}$ and $\boldsymbol{Q}$, is solved by evaluating $P(\boldsymbol{O}|\boldsymbol{Q},\lambda,T)$ for all $\boldsymbol{Q}$, using (3.67). However, these computations are very complex because there are too many combinations of sub-state sequences. Tokuda et al. (1995b,a) proposed a fast recursive algorithm to obtain an optimal or sub-optimal solution of $\boldsymbol{c}$ and $\boldsymbol{Q}$, by using special properties of (3.67).

The optimum state sequence $\boldsymbol{q}^*$ of (3.42) can be estimated independently of $\boldsymbol{O}$, by maximising $P(\boldsymbol{q}|\lambda,T)$ with respect to $\boldsymbol{q}$, as given by (3.42). Considering mixture components, $\boldsymbol{q}^*$ can also be estimated by $P(\boldsymbol{q}|\lambda,T)$. In this case, $P(\boldsymbol{O},\boldsymbol{Q}|\lambda,T) = P(\boldsymbol{O},k|\boldsymbol{q},\lambda,T)P(\boldsymbol{q}|\lambda,T)$ and $P(\boldsymbol{O},k|\boldsymbol{q},\lambda,T)$ is maximised with respect to $\boldsymbol{O}$ and $\boldsymbol{k}$. For solving the optimisation problem $\boldsymbol{q}^*$, the probability of a state sequence $\boldsymbol{q}$, given the HMM λ, can be calculated as:

$$P(\boldsymbol{q}|\lambda,T) = \prod_{n=1}^{N} p_{q_n}(d_{q_n}), \tag{3.68}$$

where $p_{q_n}(d_{q_n})$ is the state duration probability distribution associated with state q_n. The state sequence $\boldsymbol{q}^*$, which maximises $P(\boldsymbol{q}|\lambda,T)$ is calculated by solving a set of linear equations obtained from (3.68), e.g. by using the Viterbi algorithm.

A key difference between the parameter generation algorithm in HMM-based speech synthesis and the decoding process in ASR is that the optimal state $\boldsymbol{q}^*$ is calculated without reference to the observations for speech synthesis, unlike in ASR. This difference is clear by comparing the conditional probability of the state sequence $P(\boldsymbol{q}|\lambda,T)$ given by (3.68) with that used in ASR, which is given by (3.14). Since (3.68) depends only on the state duration probability $p_{q_n}(d_{q_n})$, an explicit duration model is typically used in HMM-based speech synthesis, e.g. a Gaussian density function. Duration modelling is discussed later in Section 3.3.4. On the other hand, accurate duration modelling is not as important to ASR as to statistical speech synthesis.

3.3.1.4 Method Using the Forward-Backward Algorithm

Tokuda et al. (2000) proposed another method to estimate $\boldsymbol{O}$, which consists of solving the optimisation problem of (3.40). That is, the problem of maximising the likelihood

function $P(\boldsymbol{O}|\lambda,T)$, with respect to $\boldsymbol{O}$. The critical point of this likelihood is estimated by maximising the following auxiliary function of the Baum-Welch algorithm:

$$A(\boldsymbol{O},\boldsymbol{O}') = \sum_{\mathbf{q}} P(\boldsymbol{O},\boldsymbol{q}|\lambda,T)\log P(\boldsymbol{O}',\boldsymbol{q}|\lambda,T), \tag{3.69}$$

where $\boldsymbol{O}$ and $\boldsymbol{O}^{'}$ are the current and new parameter vector sequences, respectively. Tokuda et al. (2000) used the same matrix form given by (3.57) for the calculation of the optimal sequence of static features vectors $\boldsymbol{C}^{'}$, that is, $\boldsymbol{O}^{'} = \boldsymbol{W}\boldsymbol{C}^{'}$. Under this condition, $\boldsymbol{C}^{'}$ which maximises $A(\boldsymbol{O},\boldsymbol{O}')$ is given by the following equations:

$$\boldsymbol{W}^{\top}\overline{\boldsymbol{U}^{-1}}\boldsymbol{W}\boldsymbol{C}' = \boldsymbol{W}^{\top}\overline{\boldsymbol{U}^{-1}\boldsymbol{M}}, \tag{3.70}$$

where

$$\overline{\boldsymbol{U}^{-1}} = \text{diag}\left[\overline{\boldsymbol{U}_1^{-1}},\overline{\boldsymbol{U}_2^{-1}},...,\overline{\boldsymbol{U}_T^{-1}}\right], \tag{3.71}$$

$$\overline{\boldsymbol{U}_t^{-1}} = \sum_{q,k}\gamma_t(q,k)\boldsymbol{U}_{q,k}^{-1}, \tag{3.72}$$

$$\overline{\boldsymbol{U}^{-1}\boldsymbol{M}} = \left[\overline{\boldsymbol{U}_1^{-1}\boldsymbol{m}_1}^{\top},\overline{\boldsymbol{U}_2^{-1}\boldsymbol{m}_2}^{\top},...,\overline{\boldsymbol{U}_T^{-1}\boldsymbol{m}_T}^{\top}\right]^{\top}, \tag{3.73}$$

$$\overline{\boldsymbol{U}_t^{-1}\boldsymbol{m}_t} = \sum_{q,k}\gamma_t(q,k)\boldsymbol{U}_{q,k}^{-1}\boldsymbol{m}_{q,k}, \tag{3.74}$$

and the occupancy probability $\gamma_t(q,k)$ is defined by

$$\gamma_t(q,k) = P(q_t = (q,k)|\boldsymbol{O},\lambda,T) \tag{3.75}$$

The set of equations given by (3.70) has the same form as (3.67). The optimum $\boldsymbol{O}^*$ is calculated by using (3.70) and an EM algorithm to maximise the likelihood function $P(\boldsymbol{O}|\lambda,T)$, with respect to $\boldsymbol{O}$. Tokuda et al. (2000) proposed the following recursive algorithm to calculate $\boldsymbol{O}^*$:

1. Choose an initial parameter vector sequence $\boldsymbol{C}$.
2. Calculate $\gamma_t(q,k)$ using the forward-backward algorithm.
3. Calculate $\overline{\boldsymbol{U}^{-1}}$ and $\overline{\boldsymbol{U}^{-1}\boldsymbol{M}}$ by (3.71) to (3.74), and solve (3.70).
4. Set $\boldsymbol{C} = \boldsymbol{C}'$. Go to 2 until a certain convergence condition is satisfied.

When compared with the Viterbi-based method of the previous section, this EM-based method has the advantage that $\boldsymbol{Q}^*$ can be considered unobservable, i.e. both the mixture sequence and the state sequence can be marginalised. In this method the optimum state sequence, $\boldsymbol{q}^*$, can also be calculated independently of $\boldsymbol{O}$ by maximising $P(\boldsymbol{q}|\lambda, T)$ with respect to $\boldsymbol{q}$. In this case $\boldsymbol{q}^*$ can be calculated by using the Viterbi algorithm as in Section 3.3.1.3 and the mixture sequence k is assumed to be unobservable.

3.3.2 Multi-space Distribution HMM

The observation vector used in HMM-based speech synthesis consists of a speech parameter vector, which describes the acoustic properties of a speech segment. For example, mel-cepstral coefficients and F_0 are parameters often used to describe the spectrum and to model the pitch of a speech segment, respectively. The spectral parameters of the observation vector typically represent the spectral envelope of the speech signal. They can be modelled by a continuous HMM because the spectral envelope is assumed to vary slowly across contiguous speech frames. However, F_0 patterns cannot be modelled by conventional discrete or continuous HMMs, because the values of F_0 are not defined in unvoiced regions of speech (unvoiced speech is considered to be non-periodic). Tokuda et al. (1999) proposed a solution to this problem which consists of using a hidden Markov model based on a *multi-space probability distribution* (MSD-HMM) to model F_0. This multi-space probability distribution (MSD) is more general than either a discrete or continuous mixture distribution and allows a probability distribution to be represented as a mix of discrete and continuous distributions. Figure 3.5 shows the structure of a MSD-HMM.

The MSD consists of G spaces, $\Omega = \{\Omega_1, \Omega_2, ..., \Omega_G\}$, and each space Ω_g has its probability w_g, where $\sum_{g=1}^{G} w_g = 1$. In general, the MSD used to model F_0 consists of two spaces: $\Omega = \{\Omega_1, \Omega_2\}$. Ω_1 is a zero-dimensional space associated with the unvoiced regions, while Ω_2 has one-dimensional normal distribution to model F_0 in the voiced regions. An F_0 observation is represented by a continuous random variable $\boldsymbol{y}$ and a set of space indices Y, as represented by

$$\boldsymbol{o} = (Y, \boldsymbol{y}), \tag{3.76}$$

where $Y = 1$ for the unvoiced region and $Y = 2$ for the voiced region. The output probability distribution of an N-state MSD-HMM is defined by

$$b_j(\boldsymbol{o}) = \sum_{g \in X(\boldsymbol{o})} w_{jg} \mathcal{N}_{jg}(V(\boldsymbol{o})), \tag{3.77}$$

where $V(\boldsymbol{o}) = \mathbf{y}$, $X(\boldsymbol{o}) = Y$, w_{jg} is the weight of $\mathcal{N}_{jg}$ and $\mathcal{N}_{jg}(V(\boldsymbol{o}))$ is the probability density function of the continuous observation vector $V(\boldsymbol{o})$ of state j and space g. Although, $\mathcal{N}_{jg}$ does not exist for Ω_1, $\mathcal{N}_{j1}$ is assumed to be equal to one, for simplicity of notation.

The parameters of MSD-HMMs can be estimated using the Baum-Welch algorithm, e.g. Yoshimura (2002). Each state i is assumed to have G probability density functions ($G = 2$ to model F_0 in voiced and unvoiced regions). The formulae to calculate the model parameters are derived from the auxiliary function given by (3.23) and the following likelihood:

$$\log P(\boldsymbol{O}, \boldsymbol{q}, \boldsymbol{l}|\lambda) = \sum_{t=1}^{T} \left(\log w_{q_t l_t} + \log a_{q_{t-1} q_t} + \log \mathcal{N}_{q_t l_t}(V(\boldsymbol{o}_t)) \right), \tag{3.78}$$

where $\boldsymbol{q} = \{q_1, q_2, ..., q_T\}$ is a possible state sequence and $\boldsymbol{l} = \{l_1, l_2, ..., l_T\}$ is a sequence of spaces indices which is possible for the observation sequence $\boldsymbol{O}$. The stochastic constraints of w_g are given by $\sum_{g=1}^{G} w_g = 1$.

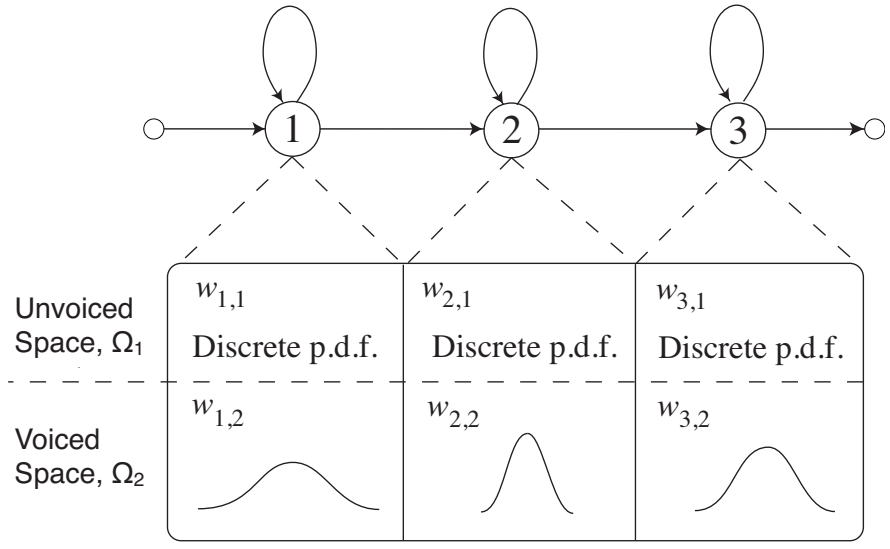

Figure 3.5: A 3-state left-to-right MSD-HMM, which uses a discrete probability density function (p.d.f.) for the unvoiced space and continuous p.d.f. for the voiced space.

3.3.3 Detailed Context Classes

In ASR and HMM-based speech synthesis, a HMM typically represents a phone unit and an utterance or word is associated with a sequence of HMMs. However, the use of a phone as a context-independent unit (called *monophone*) has the limitation of not modelling the contextual variation between phones which is characteristic of natural speech. For example, if a vowel is followed by 'n' or 'm' its pronunciation is influenced by the nasalisation effect. In general, context-dependent phone models are used to model short-term dependencies.

Typically, the method used to model the context-dependency of a phone is to use a unique phone model for every possible pair of left and right neighbours (called *triphone*). This is a practical method because it still uses a phone model. However, the number of triphones models is much higher than the number of monophones. That is, if the the number of phones is P, the number of triphones is P^3. This increase in the number of models usually causes data sparsity problems. In order to avoid this problem, model parameters are typically clustered using decision trees and the parameters are tied together in each cluster. Figure 3.6 shows an example of a decision tree, which is a modified version of Figure 3.4 from Yamagishi (2006). The author of the original figure gave me permission to modify and use the original figure in this thesis. In Figure 3.6, the notation $a-p+b$ denotes the triphone corresponding to the phone p, preceded by phone a and followed by phone b.

A decision tree is built using a top-down optimisation procedure. Starting from the root, each state is split into two by finding the question which partitions the states in the parent node so as to maximise a given criterion, e.g. increase in log likelihood (Young et al., 2006). Once the trees are built each node has a context related question, except the terminal nodes, which have state output distributions. For speech synthesis, *unseen models* can be obtained by going down the tree until the unseen context reaches a leaf node. The stopping decision rule used in decision tree construction is important because an overly large tree will be overspecialised to training data, whereas a small tree gives a poor modelling of the data. All states in each leaf node are then tied to form a set of clustered models.

Decision tree-based clustering smoothes the model parameters, since the parameters associated with the same tree leaf-node are averaged in order to re-estimate the model parameters. Smoothing is important to the robustness of the statistical modelling by HMMs. For example, it is effective in reducing speech variability due to

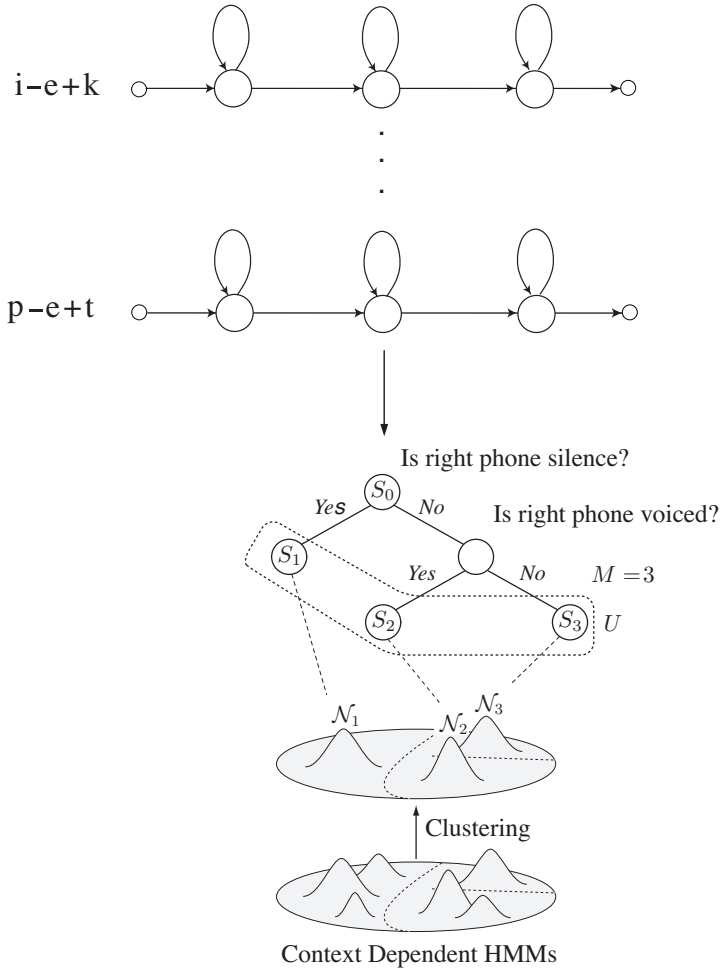

Figure 3.6: Illustration of the decision tree-based clustering typically used in HMM-based speech synthesis. This figure is a modified version of Figure 3.4 from Yamagishi (2006), which is used in this thesis with permission of the author of the original figure.

the speaker and environment conditions, in speech recognition. However, too much smoothing of the model parameters is one of the problems in HMM-based speech synthesis, e.g. Yan et al. (2009). Due to the over-smoothing problem, HMMs cannot accurately model the speech variability, which is important to the quality of the synthetic speech. This over-smoothing effect makes the synthetic speech sound blurred and muffled.

In ASR, each phone model usually has three states and the models are clustered at state-level with phonetic decision trees. That is, all states *i* of a phone are grouped at the root of the tree. Then, they are split according to the questions in each node, until all states have reached the leaf nodes. This type of tree avoids the confusion between phones, which is important to the word recognition accuracy. For example, the ques-

tions in each node are commonly related to the phonetic class of the left and right phone of the tri-phone model, e.g. if the right phone is a nasal or a consonant. On the other hand, HMM-based speech synthesis typically uses five-state models with a shared decision tree per state. In this case, the states in the leaf nodes are tied to states from other models and clustered together. This type of tree-based clustering avoids the data sparsity problem and the parameters are shared more efficiently across models. Also, it allows more effective modelling of supra-segmental effects, which is particularly important to model the F_0 parameter (Dines et al., 2009). HMM-based speech synthesis typically uses richer contextual information for building the decision tree than ASR. The use of a wide range of contextual information avoids the over-smoothing effect of the HMMs, because it allows the prosodic aspects and the speech variability to be more accurately modelled. Other techniques to reduce the over-smooting effect of the model parameters have been proposed. For example, Tokuda et al. (2000) proposed to increase the number of Gaussian mixtures in each leaf node. The size of the decision trees can also be increased to reduce the over-smoothing effect, but they might produce perceived discontinuities in the synthetic speech, if they get overspecialised to the training data (Tokuda et al., 2000). Yan et al. (2009) also proposed to use rich context models to model the training data. In this approach, the conventional parameter tree-based tying is used to estimate the optimal rich context model sequence, obtain the variance parameter of the models, and map unseen labels into seen models.

3.3.4 Duration Modelling

The exponential state duration distribution of the conventional HMM, given by (3.8), is usually inappropriate to model the duration of speech. The main problem with this model is that the probability of state occupancy decreases exponentially with time. For example, Vaseghi (1995) argues that "the likelihood of emerging from the current state increases with the increasing state residency and at a rate that depends on the distribution of state duration". Ferguson (1980) extended the HMM theory to include the explicit duration model HMM, in which duration is modelled by a *non-parameteric mass function* for each state. The explicit duration HMM is also often called *Hidden Semi-Markov Model* (HSMM). Different types of continuous distributions have been proposed to model the duration in HSMM, e.g. the *Poisson distribution* (Levinson, 1986) and the *Gamma distribution* (Russell and Moore, 1985).

Several applications of HSMMs in speech recognition can be found in the litera-

ture, such as Levinson (1986); Ratnayake et al. (1992). In general, the re-estimation algorithms for a HSMM are considerably more computationally complex than the conventional Viterbi and backward-forward algorithms used for estimation of HMM parameters (Yu, 2010). In order to avoid this complexity, speech recognisers typically use the Viterbi and backward-forward algorithms to estimate the HSMM parameters. However, the extension of the HMM parameter estimation methods to HSMM is still more complex than the conventional methods (Gales and Young, 2007; Yu, 2010). In general, the improvement in recognition accuracy due to explicit duration modelling is also not significant (Gales and Young, 2007). For these reasons the HSMM is rarely used in current speech recognisers.

In general, state-of-the-art HMM-based speech synthesisers use explicit duration distributions. One reason is that for the generation of speech parameters from the HMMs, the optimum state sequence is calculated using only the state duration probability (observation sequence is unknown), as explained in Section 3.3.1.3. On the contrary, in the decoding operation of a speech recogniser the probability distribution of the observation sequence is calculated and used to obtain the optimum state sequence using the Viterbi algorithm. Moreover, duration modelling is important in speech synthesis, because it significantly improves the quality of the synthetic speech, as shown by Zen et al. (2004). Also, HSMM increases the performance in the speaker adaptation of the average models of an independent HMM-based speech synthesiser to the target speaker, e.g. Yamagishi and Kobayashi (2005).

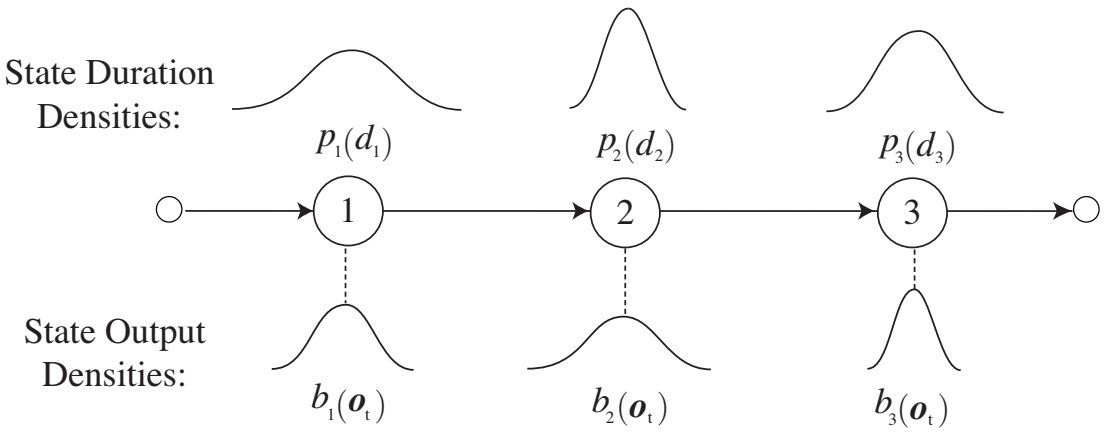

Figure 3.7: A 3-state left-to-right HSMM.

Figure 3.7 shows the structure of the HSMM. Zen et al. (2004) suggest to use the Gaussian distribution for the duration model, in order to be consistent with the proba-

bility distribution used for the acoustic model. In this case, the duration distribution of a state i with length d_i (equal to the number of frames in state i) can be represented by a Gaussian density function, as follows:

$$p_i(d_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} exp\left(-\frac{(d_i-\mu_i)^2}{2\sigma_i^2}\right), \tag{3.79}$$

where μ_i and σ_i are the mean and variance of the duration distribution of state i, respectively. The distribution $p_i(d_i)$ represents the probability of being d_i frames at state i.

In HMM-based speech synthesis, the parameters of an N-state HSMM are typically estimated using the backward-forward algorithm, e.g. Zen et al. (2004); Yamagishi and Kobayashi (2005). However, the backward and forward probabilities of (3.29) to (3.34) are modified to take into account the duration probability distribution, as follows:

$$\alpha_0^*(j) = \pi_j \tag{3.80}$$

$$\alpha_t^*(j) = \sum_{d=1}^{t} \sum_{i=1, i\neq j}^{N} \alpha_{t-d}^*(i) a_{ij} p_j(d) \prod_{s=t-d+1}^{t} b_j(\boldsymbol{o}_s), \qquad 1 \leqslant t \leqslant T \tag{3.81}$$

$$\beta_T^*(i) = 1 \tag{3.82}$$

$$\beta_t^*(i) = \sum_{d=1}^{T-t} \sum_{j=1, j\neq i}^{N} a_{ij} p_j(d) \prod_{s=t+1}^{t+d} b_j(\boldsymbol{o}_s) \beta_{t+d}^*(j), \qquad 1 \leqslant t \leqslant T \tag{3.83}$$

In the previous equations, the sum over all possible state durations increases the complexity of the forward and backward probabilities computation, when compared with the conventional algorithm. The formulae to re-estimate the mixture weights, mean, and covariance matrix of the state output probability distribution (Zen et al., 2004; Yamagishi and Kobayashi, 2005) suffer from the same increase in complexity, when compared to the formulae given by (3.36) to (3.37). According to Yamagishi and Kobayashi (2005), the duration distribution parameters can be calculated in the Baum-Welch re-estimation as:

$$\mu_j = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^*(j) d}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^*(j)} \tag{3.84}$$

$$\sigma_j^2 = \frac{\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^*(j)(d-\mu_j)^2}{\sum_{t=1}^{T}\sum_{d=1}^{t}\gamma_t^*(j)}, \tag{3.85}$$

where μ_j and σ_j^2 are the mean and variance of the duration Gaussian distribution at state j, respectively. $\gamma_t^*(j)$ is a probability of generating a serial observation sequence $\{\boldsymbol{o}_{t-d+1},...,\boldsymbol{o}_t\}$, that is,

$$\gamma_t^*(j) = \sum_{i=1,i\neq j}^{N}\alpha_{t-d}^*(i)a_{ij}p_j(d)\prod_{s=t-d+1}^{t}b_j(\boldsymbol{o}_s)\beta_t^*(j) \tag{3.86}$$

The re-estimation formulas for HSMM can also be extended to model F_0, by using MSD-HSMM (Zen et al., 2004).

For speech synthesis, the state duration d_i is determined by the parameter generation algorithm. Each state duration probability distribution of the N-state HSMM λ can be estimated by maximising the following log likelihood, with respect to the state sequence $\boldsymbol{q}$:

$$\log P(\boldsymbol{q}|\lambda,T) = \sum_{i=1}^{N}\log p_i(d_i), \tag{3.87}$$

under the constraint

$$T = \sum_{i=1}^{N} d_i \tag{3.88}$$

Assuming that the duration density $p_i(d_i)$ in state i is modelled by a single Gaussian distribution with mean μ_i and variance σ_i, the duration of each state of the optimal $\boldsymbol{q}$ can be calculated as (Yoshimura, 2002):

$$d_i = \mu_i + \rho\sigma_i^2 \tag{3.89}$$

$$\rho = \left(T - \sum_{i=1}^{N}\mu_i\right) / \sum_{i=1}^{N}\sigma_i^2 \tag{3.90}$$

The speaking rate of the synthetic speech can be controlled by ρ, because it is associated with T through (3.88). For example, Yoshimura et al. (2000) indicate that the speaking rate becomes faster or slower when ρ is set to a negative or positive value, respectively, and equal to the average speaking rate when $\rho = 0$.

3.4 HTS System

3.4.1 System Overview

The HTS system is a popular HMM-based speech synthesiser, which is available online (Tokuda et al., 2009). The basic structure of this system is shown in Figure 3.8. Most HMM-based speech synthesisers have a similar structure, which can be divided into the analysis, training and synthesis parts.

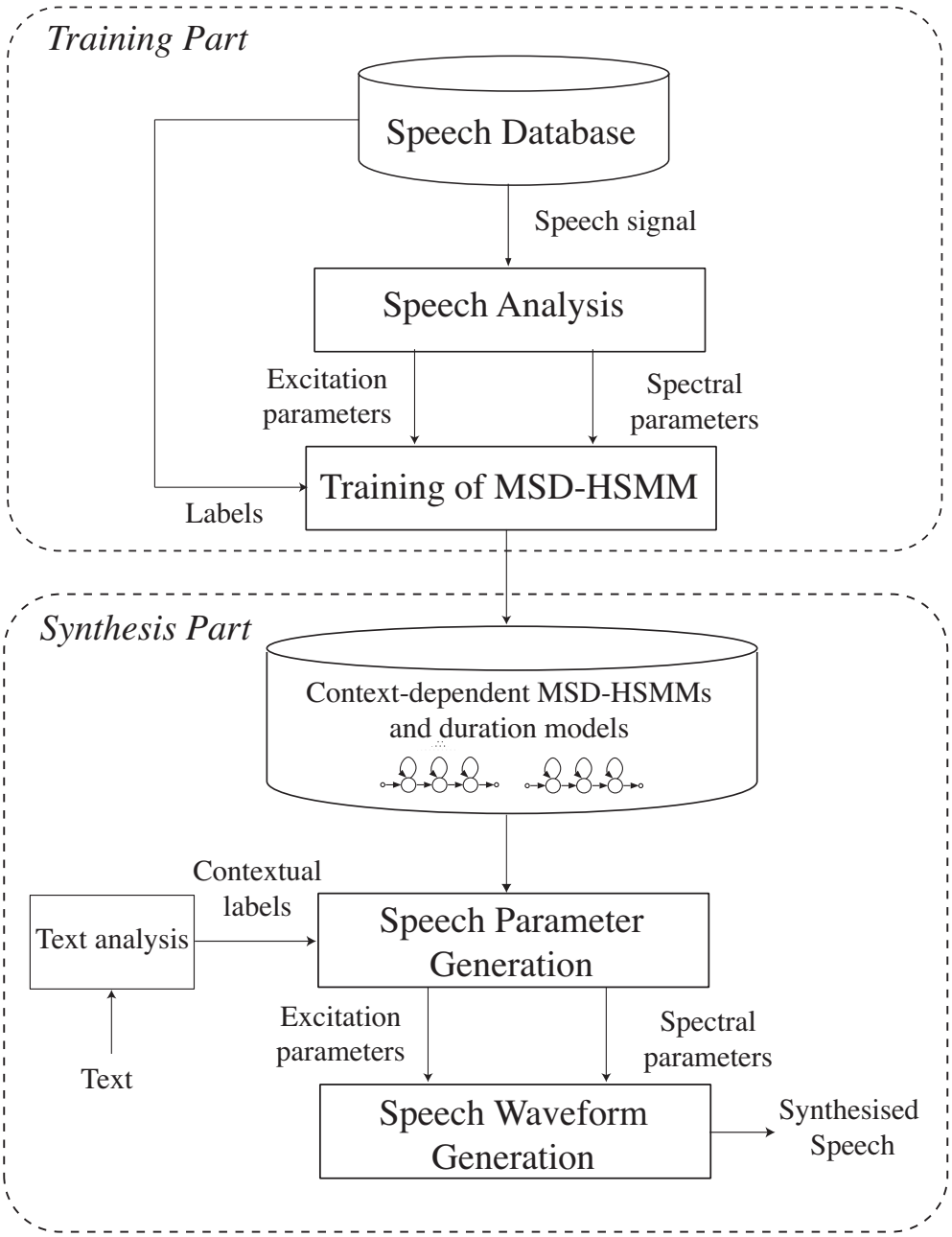

Figure 3.8: An overview of the basic HMM-based speech synthesis system.

During analysis, excitation and spectral parameters are extracted for each utterance of the speech corpus. For example, $\log F_0$ is generally used as an excitation parameter. The spectral parameters are often defined by mel-cepstral coefficients or line spectral frequencies, which are adequate features for statistical modelling. The phonetic labels can be obtained from the text, e.g. by using a *text analyser*. Typically, they also have context information, such as phone identity, phone boundaries, syllable, etc. The time label boundaries do not need to be estimated if the speech database is phonetically labelled or if a *flat-start training* of the HMMs is to be used. Otherwise, they can be calculated from the recorded utterances and their text transcriptions using a *time alignment* technique, such as the Viterbi algorithm, e.g. Young et al. (2006); Yoshimura (2002).

In the training part, the phonetic labels and the speech features are used to model context-dependent HMMs. In this process the statistical parameters of the HMMs are calculated. Then, decision trees which describe all the contextual factors are used to cluster the trained HMMs.

At the synthesis stage, the context-dependent labels are obtained from the input text and they are used by the speech parameter generation algorithm to generate the speech features. The excitation signal is calculated using the excitation features, which then passes through the synthesis filter to obtain the speech signal. The synthesis filter used in HTS is defined by the spectral features.

3.4.2 Analysis

Phonetic, linguistic, and prosodic parameters are estimated from the sentences of the recorded speech corpus using the text analysis tools of the FESTIVAL unit-selection speech synthesiser (Black et al., 2004). This information is represented by the HTS system in the form of labels which are used for training the context-dependent phone models (HMMs). Most contexts are related to counts, positions and distances of stressed and accented syllables, and stretches from phone to utterance level context. Examples of the contextual information used for English are given below:

- preceding, current, succeding phones.
- position of current phone in current syllable.
- number of phones in preceding, current, succeding syllable.
- accent of preceding, current, succeding syllable.

- number of preceding, succeding stressed syllables in current phrase.
- position of current word in current phrase.
- number of syllables in current utterance.

Spectral and excitation parameters are also estimated from the speech corpus. The spectrum estimated by HMM-based speech synthesisers typically represents the spectral envelope of the speech signal. The conventional method to estimate the envelope in HTS is mel-cepstral analysis. However, the spectral envelope can also be computed using other methods in HTS. For example, there is also a HTS demo which uses the STRAIGHT vocoder to compute the spectral envelope. The fundamental frequency is extracted using an F_0 estimation algorithm, e.g. the F_0 detector of the Entropic Speech Tools (ESPS) which uses the Robust Algorithm for Pitch Tracking (RAPT) of Talkin and Rowley (1990). The HTS demo using STRAIGHT also extracts aperiodicity measurements, which are used to generate the excitation signal during speech synthesis.

3.4.3 Statistical Modelling

Typically, the HMM topology used in HTS is a five-state left-to-right HMM. Each state output density function can be modelled by a single Gaussian or Gaussian mixture distributions. In general, the covariance matrix of each Gaussian mixture component takes the form of a diagonal covariance matrix. This covariance matrix is significantly more advantageous than the full covariance matrix, in terms of computational complexity. The spectrum is modelled by a continuous HMM while F_0 is modelled by a MSD-HMM.

The observation feature vector at time t, $\boldsymbol{o}_t$, has a *multi-stream* structure. F_0 and mel-cepstrum are modelled by different streams because they are assumed to be independent. The dynamic features, Δ and Δ^2, of the $\log F_0$ and spectral parameters are also included in the feature vector. The state duration densities are modelled by Gaussian distribution and the dimension of state duration density is equal to the number of states in the HMM.

In HTS, the re-estimation of the model parameters is performed using the Hidden Markov Model Toolkit (HTK) version 3.4 (Young et al., 2006). This training procedure uses the maximum likelihood estimation criterion. Finally, the spectral parameters, F_0 and state duration are clustered independently because they have their own influential contextual factors.

3.4.4 Speech Feature Generation Algorithm

The problem of generating the speech parameter vector sequence $\boldsymbol{O}$ from the HMM λ, for a given word transcription $\boldsymbol{W}$, is to maximise the output probability distribution with respect to $\boldsymbol{O}$, as follows:

$$\boldsymbol{O}^* = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\boldsymbol{W}, \lambda, T) \tag{3.91}$$

One way to solve this problem is to use the recursive method based on the expectation-maximisation (EM) algorithm, which was described in Section 3.3.1.4. The HMGenS tool of the HTS system allows speech parameters to be generated using this algorithm.

Another way to solve the optimisation problem is to use the Viterbi-based method described in Section 3.3.1.3. HTS includes a small run-time synthesis engine, called *hts_engine*, which generates speech parameters based on this method. The synthesis engine works without the HTK/HTS libraries and it is faster than HMGenS. The *hts_engine* program is indicated for application development purpose.

3.4.5 Synthesis

3.4.5.1 Source-filter Model

The speech waveform generation technique which is conventionally used in HTS is to pass the excitation signal through a synthesis filter, which is defined by the spectral parameters. For voiced speech, the excitation is the impulse train generated using the F_0 parameter. The synthesis filter is a variable *Mel Log Spectrum Approximation* (MLSA) filter (Imai, 1983). For unvoiced speech, the excitation is modelled as white noise.

A more sophisticated method to generate speech in HTS is to use the STRAIGHT vocoder (Kawahara et al., 1999b). In this case, the excitation of voiced speech is obtained from the F_0 and the aperiodicity parameters by mixing the impulse train with noise. STRAIGHT uses a minimum-phase filter which is different from the MLSA synthesis filter. The STRAIGHT vocoder is described in Sections 4.3.3 and 6.2.

HTS and most HMM-based speech synthesisers produce speech by shaping a spectrally flat excitation signal with the spectral envelope. This is the type of source-filter model used to synthesise speech by the MLSA filtering and STRAIGHT methods. Figure 3.9 shows an example of the transfer function of the HTS synthesis filter, the spectra of the impulse train, the noise, and the synthetic speech signal. In this ex-

ample, the voiced excitation is modelled as an impulse train, without adding a noise component to this signal (as in STRAIGHT).

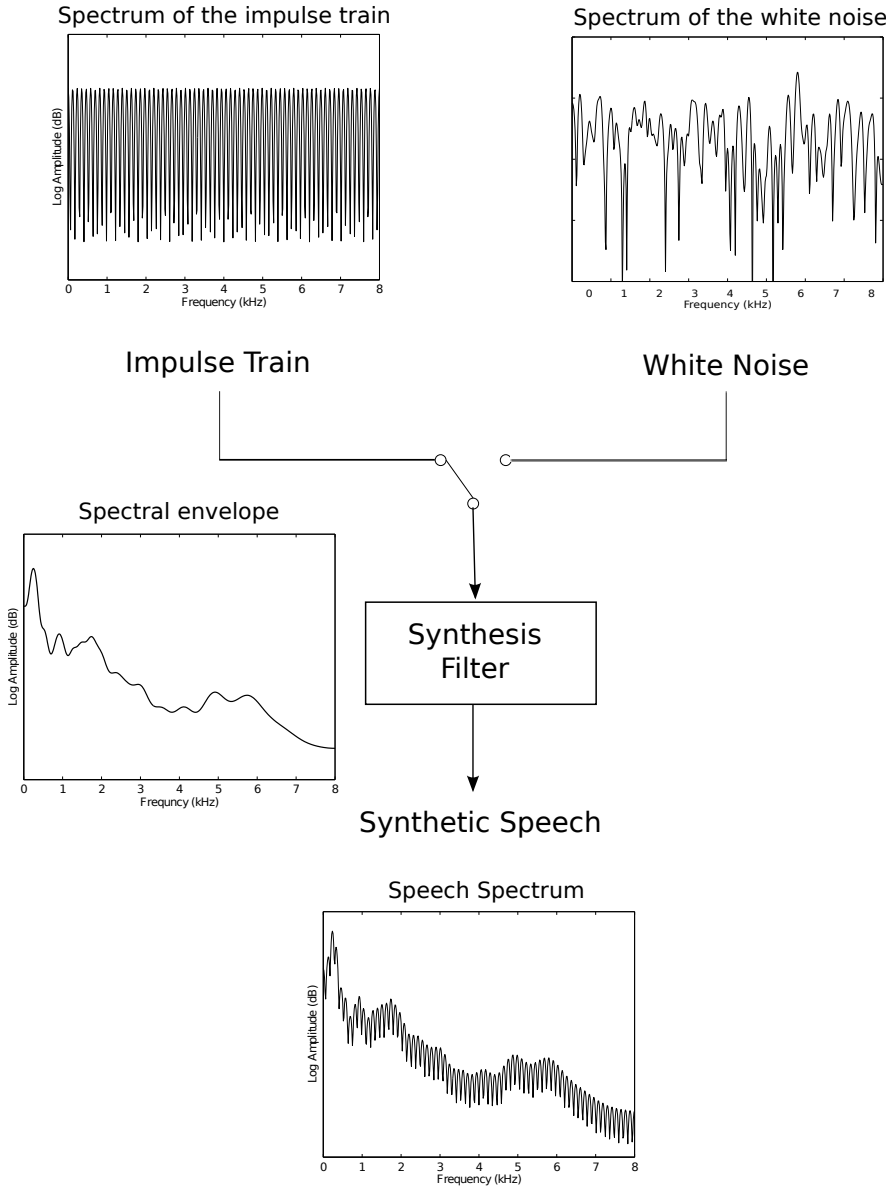

Figure 3.9: Example of the speech synthesis method used by HTS, which consists of shaping a spectrally flat excitation with the spectral envelope.

3.4.5.2 MLSA Filter

The MLSA filter used in HTS to synthesise speech is obtained from the mel-cepsturm $H(e^{jw})$, which is represented by the M-order mel-cepstral coefficients $c(m)$ as follows:

$$H(z) = \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m}, \tag{3.92}$$

where $\tilde{z}^{-1}$ is an all-pass function given by

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \quad |\alpha| < 1 \tag{3.93}$$

The phase response of this *all-pass function* has characteristics related to the perceptual model of the human auditory system. For example, it approximates the mel-scale (Fant, 1973), for the sampling rating of 16 kHz, when $\alpha = 0.42$. The minimum phase transfer function of the mel-cepstrum, $D(z)$, is estimated from the mel-cepstrum by using the unbiased estimation of the logarithmic spectrum (Imai and Furuichi, 1988), as follows:

$$H(z) = KD(z) = \exp \sum_{m=0}^{M} b(m)\Phi_m(z), \tag{3.94}$$

where $K = \exp b(0)$ is a gain factor and

$$D(z) = \exp \sum_{m=1}^{M} b(m)\Phi_m(z) \tag{3.95}$$

$$b(m) = \begin{cases} c(m) & m = M \\ c(m) - \alpha b(m+1) & 0 \leq m < M \end{cases} \tag{3.96}$$

$$\Phi_m(z) = \begin{cases} 1, & m = 0 \\ \frac{(1-\alpha^2)z^{-1}}{1-\alpha z^{-1}}\tilde{z}^{-(m-1)}, & m \geqslant 1 \end{cases} \tag{3.97}$$

However, $D(z)$ cannot be realised directly as a digital filter because it is not a rational function. HTS uses the algorithm proposed by Fukada et al. (1992) to perform the quantisation of the mel-generalised cepstrum $D(z)$. Basically, it consists of approximating the exponential transfer function $D(z)$ by a rational function $R_L(F(z))$, as follows:

$$D(z) = \exp(F(z)) \simeq R_L(F(z)) \tag{3.98}$$

$$R_L(F(z)) = \frac{1 + \sum_{l=1}^{L} A_{L,l} F(z)^l}{1 + \sum_{l=1}^{L} A_{L,l} (-F(z))^l}, \tag{3.99}$$

where $A_{L,l} (l = 1, 2, ..., L)$ are the coefficients of the function R_L and

$$F(z) = \sum_{m=1}^{M} b(m)\Phi_m(z) \tag{3.100}$$

The coefficients $A_{L,l}$ are optimised so as to minimise the maximum of the log approximation errors $|E_L(F(z))| = |\log D(z) - \log R_L(F(z))|$.

The MLSA filter proposed by Fukada et al. (1992) is implemented as a stable minimum-phase IIR filter with a two stage cascade structure, i.e.

$$D(z) \simeq R_L(F_1(z))R_L(F_2(z)), \tag{3.101}$$

where

$$F_1(z) = b(1)\Phi_1(z) \tag{3.102}$$

$$F_2(z) = \sum_{m=2}^{M} b(m)\Phi_m(z) \tag{3.103}$$

The cascade form is used to obtain a more accurate approximation of the rational function $D(z)$. Fukada et al. (1992) indicates that this cascade filter approximates the exponential transfer function $D(z)$ with sufficient accuracy ($|E_L(F(z))| \leq 0.24$ dB).

3.5 Conclusion

HMMs have been used for ASR since several decades ago. Meanwhile, HMM-based speech synthesis is a more recent application of the HMM in speech technology. The two types of technologies use the same generative model and similar algorithms for computing the HMM parameters and to evaluate the likelihood $P(\boldsymbol{O}|\lambda)$ in the decoding part of the speech recogniser and in the feature generation of the statistical synthesiser, respectively. However, the statistical models in speech recognition are used to decode an unknown word sequence from a sequence of observed speech feature vectors, whereas they are used to estimate the speech parameters from the input word sequence in statistical speech synthesis. For this reason, the speech recogniser is implemented in a way that maximises the discrimination between classes of sounds and to be robust to speech variability factors, such as speaker, environmental, and pronunciation variability. On the other hand, the statistical speech synthesiser aims to generate the most natural sounding speech as possible and to model the speech variability details which are characteristic of human speech, such as aspects related to speaker identity and

expressiveness. This contrast between the properties of speech synthesis and recognition using HMMs, yields to several differences between the implementation of the two methods. The main characteristics of HMM-based speech synthesis which differ from ASR are reviewed in the next paragraphs and summarised by the following list:

- speech feature generation algorithm takes into account derivative constraints not required in ASR.
- explicit duration modelling which is not required in ASR.
- context-dependent HMMs with richer contextual information than that used in ASR.
- multi-space distribution HMMs to model F_0 parameter, which is not typically modelled in ASR.
- higher order of the spectral feature vector.

The training parts of a speech synthesiser and recogniser are very similar. In both technologies, the HMM parameters are typically estimated using the forward-backward algorithm. The Viterbi algorithm is commonly used in the decoding part of the speech recogniser to maximise the likelihood $P(\boldsymbol{q},\boldsymbol{O}|\lambda)$, with respect to the state sequence $\boldsymbol{q}$, given a sequence of observation feature vectors $\boldsymbol{O}$ and the model λ. The Viterbi method can also be used for speech feature generation in statistical speech synthesis, but in this case it is used to maximise the likelihood $P(\boldsymbol{q}|\lambda)$ with respect to $\boldsymbol{q}$, because the observation sequence is unknown. The optimum state sequence, $\boldsymbol{q}$*, is then used to generate the sequence of speech features by maximising $P(\boldsymbol{O}|\boldsymbol{q}^*,\lambda)$. The forward-backward algorithm is also often used to generate the speech features in speech synthesis by maximising locally the likelihood $P(\boldsymbol{O}|\lambda)$ with respect to $\boldsymbol{O}$. The speech feature generation algorithm based on the Viterbi algorithm is simpler than the forwad-backward method, but the second typically gives better results.

The dynamic features, e.g. Δ and Δ^2, are typically used in ASR in order to improve the acoustic modelling. In speech synthesis they are also used by the speech feature generation algorithm to impose derivative constraints on the speech parameters so that the parameter trajectories are smooth. This function of the dynamic features is crucial for statistical synthesisers to produce high-quality speech.

The traditional left-to-right continuous HMM used for ASR is extended to the left-to-right HSMM for speech synthesis. HSMM models the duration explicitly, e.g. by a

Gaussian probability distribution. The main reasons for this difference are that speech duration has an important effect on the synthetic speech quality and it is not accurately modelled using the implicit transition probabilities of the HMM. Explicit duration modelling is very important in speech synthesis because the duration of the synthetic speech is determined by the state probability density functions when the Viterbi algorithm is used to maximise the likelihood $P(\boldsymbol{q}|\lambda)$. In contrast, the improvement to the implicit duration model of the basic HMM has small impact on the increase of the recognition accuracy. This can be explained by the fact that the observation sequence is taken into account in the decoding part, i.e. the Viterbi algorithm is used to maximise $P(\boldsymbol{q}|\boldsymbol{O},\lambda)$.

Both ASR and statistical speech synthesis generally use context-dependent HMMs, e.g. tri-phone models, to better model contextual factors. Also, the tree-based clustering is used by both applications to avoid data sparsity and overcome problems with unseen models. In speech synthesis, it is important to model many details of the context dependencies between speech units. The reason for this is that they capture aspects of speech variability which are important for the perceptual quality of the synthetic speech. In contrast, speech recognisers usually obtain better results when speech variability effects are smoothed, e.g. variability due to the voice characteristics related to the speaker's identity. For these reasons, the context-dependent information used by HMM-based speech synthesisers is typically more detailed than that used by speech recognisers.

Another difference between the HMM structure of the speech synthesiser and the recogniser systems is that the first uses a MSD-HMM for modelling F_0. In general, this parameter is not modelled in ASR but it is very important in speech synthesis, especially to capture the prosodic aspects of speech. MSD-HMM are used to model F_0 by a discrete distribution for unvoiced speech and by a continuous distribution for voiced.

The typical structure of a HMM-based speech synthesiser can be divided into the analysis, training, and synthesis parts. For analysis, excitation and spectral parameters are extracted from the speech signal, e.g. F_0 and mel-cepstral coefficients. The spectral parameters usually represent the spectral envelope of the short-term speech signal. In the training part, the HMM parameters are calculated by using the sequence of observation feature vectors. Each feature vector consists of the excitation and spectral parameters, including the dynamic features. Its structure is a multi-stream, e.g. F_0 and spectral parameters are assumed to be independent and they are modelled in separate

streams. During synthesis, the excitation signal is obtained by using the speech parameters generated by the HMMs. For example, F_0 is often used to produce an impulse train, which models the periodic characteristics of the voiced excitation. For unvoiced speech, the excitation is typically modelled as white noise. Speech is usually obtained by shaping a spectrally flat excitation with the spectral envelope, which is represented by the spectral parameters. For example, the conventional synthesis method used in the HTS system (Tokuda et al., 2009) consists of passing a spectrally flat excitation through the MLSA filter, which is defined by the mel-cepstral coefficients.

Chapter 4

Source Modelling Methods in Statistical Speech Synthesis

4.1 Introduction

Typically, HMM-based speech synthesisers generate speech by passing a *spectrally flat excitation* signal through a synthesis filter. This filter represents the *spectral envelope* of the speech signal, such as in the HTS system which was described in Section 3.4.

The excitation of voiced speech can be modelled as an *impulse train*, which only enables to control the pitch of the synthetic speech. However, the quality of the synthetic speech obtained with the impulse train is poor. One way to improve the quality is to use a mixed excitation signal, which is obtained by adding a noise component to the periodic pulse train. In HMM-based speech synthesis, the weighting of the noise and periodic signals is typically performed in the frequency domain by using a *mixed-multiband excitation model*. The impulse train signal might also be processed, e.g. by using a *phase manipulation* technique, in order to represent non-periodic characteristics of the spectrum of the excitation.

Another problem of the impulse train is that it does not represent the shape characteristics of the glottal source signal. For example, the residual signal obtained by inverse filtering gives a better approximation of the glottal source derivative as the energy of the residual signal is distributed along the fundamental period whereas the energy of the impulse train is only concentrated at one instant of the period. As an attempt to further improve the quality of the synthetic speech, other source modelling approaches have been proposed for HMM-based speech synthesis, which try to better approximate the voiced excitation to the residual. These are called *residual modelling*

methods in this chapter.

Another method that has been used to improve the source modelling of the statistical speech synthesisers consists of using a more accurate estimate of the glottal source signal than the residual signal. The conventional inverse filtering method does not correctly separate important characteristics of the source from the vocal tract, such as the spectral tilt. Section 2.2.3 described methods which can more accurately separate the glottal source from the vocal tract components of speech. However, such methods consider a different source-filter model of speech, which was explained in Section 2.1.1. That is, the source signal is no longer spectrally flat (it has a decaying spectrum) and the synthesis filter represents the vocal tract transfer function instead of the spectral envelope. This type of source-filter representation has also been employed in HMM-based speech synthesis. In this case, speech can be synthesised using a glottal source model or a *real glottal flow signal* which is transformed using glottal source parameters (trained by the synthesiser).

This chapter gives an overview of the main types of excitation models which have been used in HMM-based speech synthesis and the way these models have been incorporated into the statistical speech synthesisers. That is, the analysis, synthesis and statistical modelling parts of the synthesisers will be mainly reviewed in terms of the excitation.

4.2 Simple Pulse/Noise excitation

The simplest excitation model used in HMM-based speech synthesis consists of switching between a sequence of delta pulses (impulse train signal) and white Gaussian noise for segments of voiced and unvoiced speech respectively. The first versions of the HTS system described in Section 3.4 used this type of excitation.

4.2.1 Analysis

An advantage of the pulse/noise excitation is that it only requires a voiced/unvoiced speech detector and the estimation of the F_0 parameter, which is used to model the pitch of the voiced excitation.

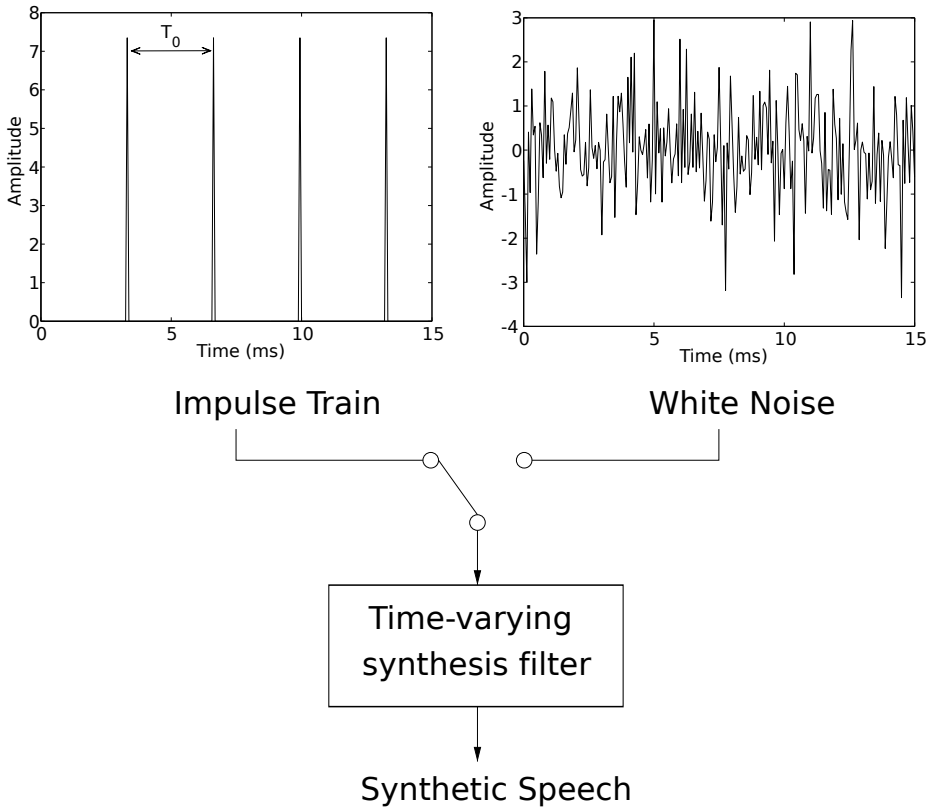

Figure 4.1: Speech synthesis using the simple excitation model.

4.2.2 Synthesis

Figure 4.1 shows the block diagram of the speech waveform generation technique of HMM-based speech synthesisers which use the simple pulse/noise excitation model, e.g. Yoshimura et al. (2000); Tokuda et al. (2002). This figure also shows an example of the impulse train and the noise excitation signals. The impulse train signal consists of single pulses, which are spaced by the pitch period $T_0 = 1/F_0$. Speech is generated by passing the excitation signal through the synthesis filter which is obtained from the spectral envelope parameters. For example, the MLSA filter which was described in Section 3.4.5.2 is typically employed in the HTS system.

The spectra of the noise excitation and the impulse train are approximately flat. These signals also have the same power as they are both shaped by the spectral envelope of the speech signal. Figure 4.2 shows an example of the white noise and impulse train spectra. Typically, the noise signal is obtained by generating random values from a normal distribution with zero mean and unit variance. On average, the power of this noise signal is one. The power of an impulse train excitation $x(n)$, with length N, can

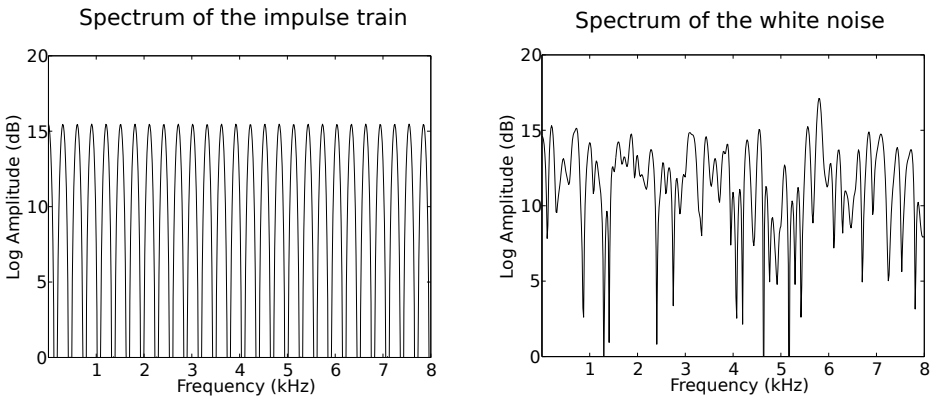

Figure 4.2: Spectra of the impulse train (left) and white noise (right) components of the simple excitation model, respectively.

be calculated as

$$P_x = \frac{1}{N}\sum_{n=1}^{N}(x(n))^2 \tag{4.1}$$

In order to ensure consistency between the energy of the noise and the impulse train signals, the amplitude of the pulses generated by the synthesiser is equal to $\sqrt{N_0}$, in which N_0 is the number of samples of the pitch period. From (4.1), the resulting impulse train has power equal to one, which matches the power of the noise signal. However, the pulse amplitude which is determined from an energy constraint of the excitation signal does not correctly model the amplitude variations which are characteristic of the voice source signal, such as amplitude variations at the instants of maximum excitation (maximum of the real glottal flow derivative). The amplitude variations of the delta pulse could have a negative effect on the speech quality if they were not modelled correctly. For example, Fant (1997) indicated that variations of the amplitude of maximum excitation are related to variations in voice effort. Also, he found that there are dynamic changes of this amplitude parameter within an utterance, which are related to intonation patterns, and that it has a characteristic phrase contour (initial rise, declination, and fall at the end).

The main problem of the simple impulse train excitation is that it produces a "*buzzy*" speech quality due to the strong harmonic structure of this signal. The strong periodicity of the impulse train is clear in the example of Figure 4.2. Also, the pulse/noise model is unable to correctly represent the excitation of speech sounds which are characterised by the mix of a periodic with a noise component, such as *voiced fricatives*.

4.2.3 Statistical Modelling

The acoustic modelling topology of HMM-based speech synthesisers which use the pulse/noise excitation is usually similar to that of the HTS system, described in Section 3.4.3. Table 4.1 summarises the structure of the HMM model in this type of synthesisers. The streams for spectral parameters (representing the spectral envelope) are modelled by Gaussian distributions. F_0, its Δ and Δ^2 are modelled by a Multi-Space probability Distribution (MSD) HMM. Each of these parameters is modelled using a Gaussian distribution in the voiced space and a discrete distribution in the unvoiced space. The distributions for spectral and F_0 parameters are typically clustered independently using different decision trees for each type of speech parameter.

Streams	Probability Distributions
Spectrum	Gaussian
F0	Multi-space

Table 4.1: HMM structure which is characteristic of HMM-based speech synthesisers using a simple pulse/noise excitation.

4.3 Multi-band Mixed Excitation

4.3.1 Introduction

Different types of Multi-Band mixed Excitation (MBE) models have been used in HMM-based speech synthesis in order to reduce the buzziness of the impulse train. In general, the MBE signal is modelled in the frequency domain using a technique that mixes the spectrum of a harmonic signal with the spectrum of a noise signal. This section gives an overview of the most relevant MBE models which have been used in HMM-based speech synthesis.

4.3.2 Mixed Multi-band Linear Prediction (MELP) Vocoder

The first statistical HMM-based speech synthesiser to use a MBE model was proposed by Yoshimura et al. (2001). This system was developed by incorporating a MELP

vocoder into the standard HMM-based speech synthesiser with simple excitation of Yoshimura et al. (2000). MELP was first used for low-bit rate speech coding (2.4 and 4.8 kHz sampling frequencies) by McCree and Barnwell III (1995). Recently, the MELP vocoder has also been integrated into the statistical speech synthesiser proposed by Gonzalvo et al. (2007). Abdel-Hamid et al. (2006) have also used a MBE model similar to the one of MELP in order to improve the speech naturalness of an Arabic HMM-based speech synthesiser.

4.3.2.1 Analysis

The excitation parameters used by the MELP vocoder are F_0, *voicing strengths* of the speech signal in different frequency bands and the *Fourier magnitudes* of the harmonics of the residual signal. The residual is calculated by inverse filtering the speech signal using the LPC coefficients, while the Fourier magnitudes are obtained by computing the *fast Fourier transform* (FFT) of the residual signal.

For the two HMM-based speech synthesisers which use MELP (Yoshimura et al., 2001; Gonzalvo et al., 2007), the voicing strengths are calculated from the speech signal (sampled at 16 kHz) using the analysis method of the wide-band MELP vocoder proposed by Lin et al. (2000). First, the speech signal was bandpass-filtered into five frequency bands: 0-1, 1-2, 2-4, 4-6, and 6-8 kHz. Next, the voicing strength in each frequency band is estimated by the maximum *autocorrelation* of the signal which is bandpass filtered in that frequency band. The autocorrelation is often used to measure the periodicity of speech signals, as it is high for voiced and low for unvoiced speech. The autocorrelation analysis is performed around the *pitch lag*, by calculating the correlation coefficient at delay t (Lin et al., 2000), as follows:

$$c_t = \frac{\sum_{n=0}^{N-1} s_n s_{n+t}}{\sqrt{\sum_{n=0}^{N-1} s_n s_n \sum_{n=0}^{N-1} s_{n+t} s_{n+t}}}, \tag{4.2}$$

where s_n represents the bandpass filtered signal at sample n and N is the the size of the pitch analysis window.

Both HMM-based speech synthesisers estimate the Fourier series magnitudes as the largest DFT magnitudes of the residual signal within the frequency bands corresponding to each pitch harmonic, as in the MELP vocoder of McCree and Barnwell III (1995). The synthesiser of Yoshimura et al. (2000) uses the first ten pitch harmonic magnitudes while the system of Gonzalvo et al. (2007) uses the first thirty magnitudes.

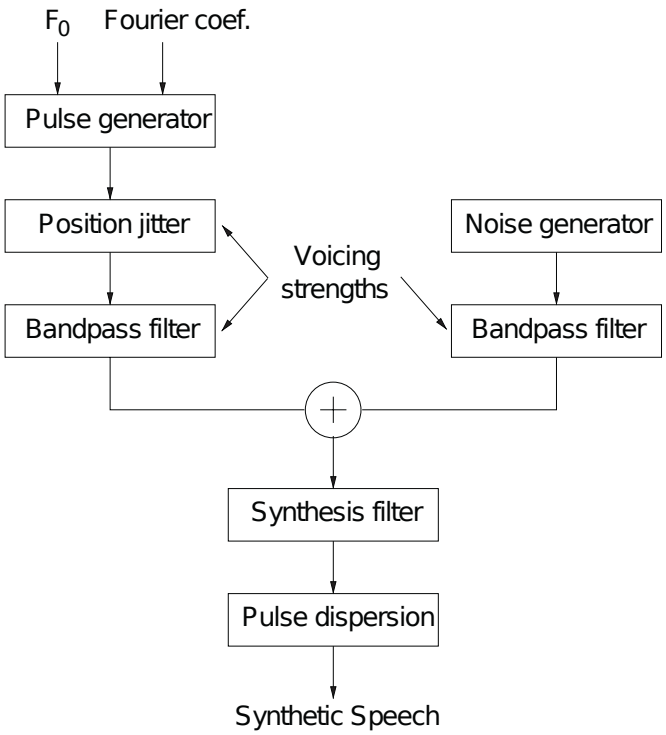

Figure 4.3: Typical speech waveform generation technique of HMM-based speech synthesisers which use the mixed excitation model of the MELP vocoder.

4.3.2.2 Synthesis

Figure 4.3 shows the general block diagram of the speech waveform generation method of HMM-based speech synthesisers which use the excitation model of MELP.

The MELP vocoder produces the spectrum of the periodic pulse signal from the input Fourier coefficients, by setting the magnitudes of the first harmonics (placed at frequencies multiples of F_0) equal to the normalised Fourier magnitudes and by synthesising the remaining harmonics with a fixed magnitude value of one. Each phase of the harmonics is set equal to zero, in order to align the harmonics into a single pulse per pitch period. This ensures phase coherence between the synthetic speech frames when they are concatenated using the Pitch-Synchronous Overlap-and-Add (PSOLA) technique (Moulines and Charpentier, 1990), as the single pulse is located always at the same position within the frame using this technique. Note that if all the magnitude values are equal to one, the resulting spectrum is equivalent to that of the impulse train (an example of this spectrum is shown in Figure 4.2). Finally, the pulse waveform is calculated from the Fourier magnitudes and F_0 by inverse DFT of one pitch period in length.

MELP uses an aperiodicity flag to decide if the pulse train of voiced speech is periodic or aperiodic. If speech is classified as aperiodic than each pitch period length of the pulse is varied with a pulse *position jitter*. The HMM-based speech synthesiser proposed by Gonzalvo et al. (2007) does not use position jitter, while the system of Yoshimura et al. (2001) performs this aperiodicity transformation by using the same method of the wideband MELP vocoder proposed by Lin et al. (2000). This vocoder estimates the aperiodicity flag according to the voicing strengths and synthesises the jittery speech by varying 25% of the pitch length. The aperiodic pulses and the noise component of the mixed excitation have different functions in the vocoder. The main goal of the mixed excitation is to reduce the buzzy quality while the jitter destroys the periodicity of the synthetic speech in order to reduce the *tonal noises*. Another function of jitter, indicated by McCree and Barnwell III (1995), is to reproduce the *erratic glottal pulses* in speech frames located at *voicing transitions* (transitions between voiced and aperiodic speech) or the vocal fry effect of speech.

The frequency bands of the filters used for mixing the pulse train (*voiced filter*) and white noise (*unvoiced filter*) are calculated from the bandpass voicing strengths. A frequency band is assigned to the voiced filter if the measure of voicing strength in that band is above a certain threshold, and to the unvoiced filter if the voicing strength is lower than the threshold. Figure 4.4 illustrates the mixing of an impulse train (without using position jitter) with a noise signal using the method of the MELP vocoder. The frequency bands assigned to each filter are represented in grey.

The two HMM-based speech synthesisers proposed by Yoshimura et al. (2000) and Gonzalvo et al. (2007), respectively, use the MLSA filter to synthesise speech instead of the conventional LPC synthesis filter of the MELP vocoder. The MLSA filter is often used in HMM-based speech synthesis because it can be obtained directly from the mel-cepstral coefficients and it is computationally efficient.

Finally, synthetic speech is filtered by a *pulse dispersion filter* in order to introduce time-domain spread of the energy over the pitch period and enhance speech quality. The dispersion filter reduces the *peak-to-valley ratio* (spectral parameter) of bandpass filtered signals in frequencies away from the formants. According to McCree and Barnwell III (1995), the smaller peakiness of the bandpass filtered natural speech compared with the synthetic speech could be "due to a secondary excitation peak from the opening of the glottis, aspiration noise resulting from incomplete glottal closure, or a small amount of background noise which is visible in between the excitation peaks". The pulse dispersion filter which is used by both HMM-based speech synthesisers

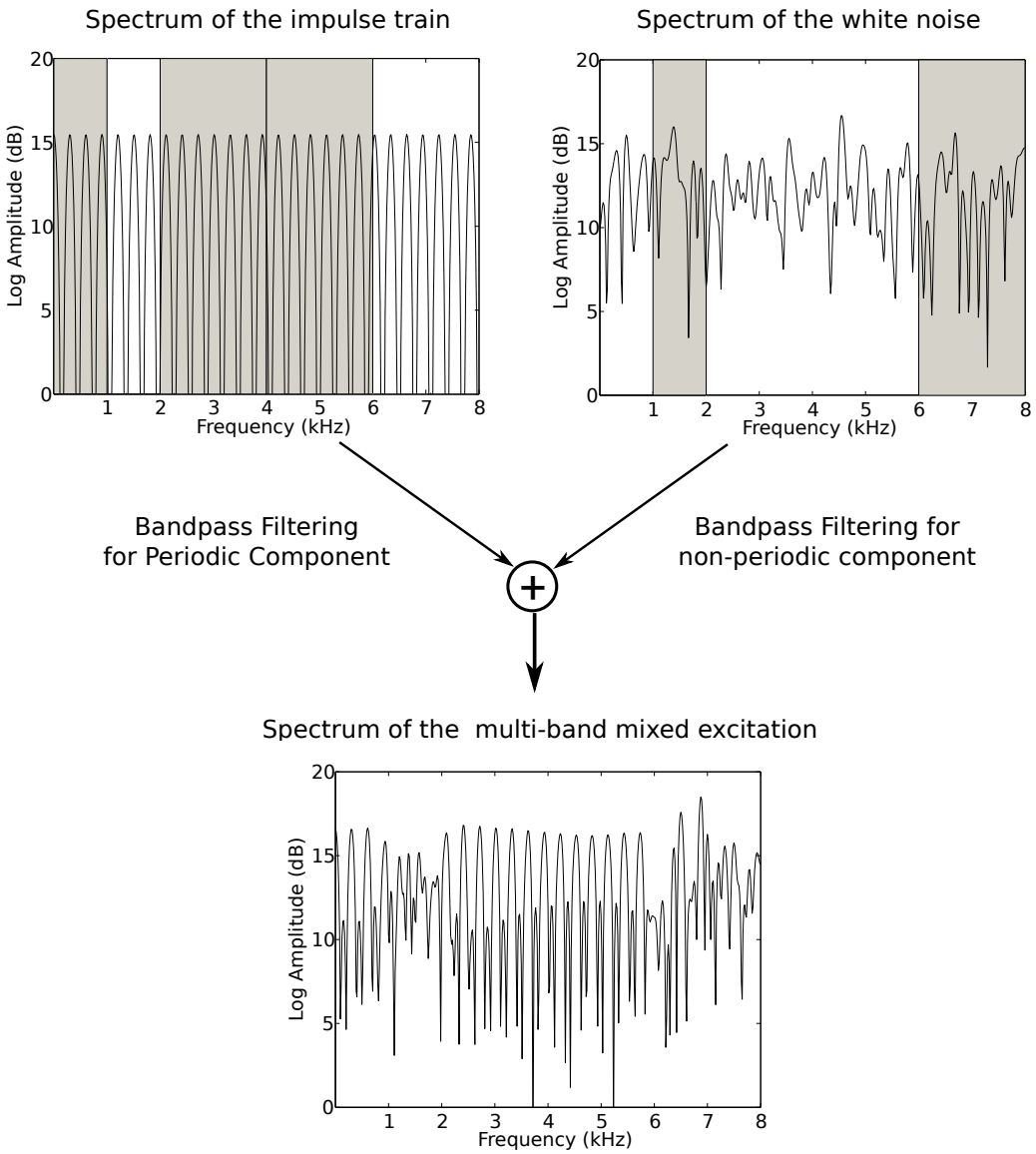

Figure 4.4: Example of the frequency bands assigned to the spectra of the pulse train and noise signals, respectively. The shaded regions in the plots represent the frequency bands of the bandpass filters which are used to obtain the periodic and non-periodic components of the excitation, respectively.

(Yoshimura et al., 2000; Gonzalvo et al., 2007) is a 130^{th} order FIR filter derived from a spectrally flattened triangle pulse.

4.3.2.3 Statistical Modelling

Both synthesisers of Yoshimura et al. (2000) and Gonzalvo et al. (2007) respectively, which use the MELP excitation model, have a similar HMM structure. These systems

model the spectral envelope of speech, FFT magnitudes and voicing strengths using Gaussian distributions. Each HMM state has four data streams for mel-cepstral coefficients, F_0, bandpass voicing strengths and Fourier magnitudes. Each stream contains the static, the first order and second order derivatives. The mel-cesptral coefficients, bandpass voicing strengths and Fourier magnitudes are modelled by diagonal Gaussian distributions, respectively. Meanwhile, the F_0 parameters are modelled by three multi-space distributions (MSD), for the static vector and its first and second order derivatives, respectively. The HMM structure of the speech synthesisers which use the MELP vocoder is summarised in Table 4.2. The main characteristic of this statistical model is the higher number of data streams, when compared with the synthesisers with simple excitation.

The context-dependent HMMs are clustered using decision trees, which were described in Section 3.3.3. HMM-based speech synthesisers which use the simple excitation model (pulse/noise model) typically use separate decision trees to model the mel-cepstrum and F_0, as they have different contextual factors. For the same reason, the distributions for the bandpass voicing strength and the Fourier magnitude are also clustered independently from F_0 and the spectrum, in the synthesisers which use the MELP vocoder. However, the state occupation statistics used for clustering the voicing strength and Fourier magnitude parameters are calculated from the mel-cepstrum and F_0 streams only.

Streams	Probability Distributions
Mel-cepstrum	Gaussian
F0	Multi-space
Voicing Strengths	Gaussian
Fourier Magnitudes	Gaussian

Table 4.2: Information about the statistical model used by the HMM-based speech synthesisers with MELP vocoder.

4.3.3 STRAIGHT Vocoder

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum) is a high-quality system for speech modification (Kawahara et al., 1999b). This system incorporates a mixed excitation model described by Kawahara et al. (2001), which consists of weighting the periodic and noise components using *aperiodicity measurements* of the speech signal.

4.3.3.1 Analysis

The Nitech-HTS 2005 system of Zen et al. (2007a) uses an implementation of the STRAIGHT vocoder to extract the spectral envelope and aperiodicity measurements from the speech signal. STRAIGHT represents both the spectrum and aperiodicity of the speech signal by FFT coefficients, which are not suitable for statistical modelling due to their high-dimensionality. Nitech-HTS 2005 overcomes this problem by converting the amplitude spectrum to mel-cepstral coefficients and by averaging the aperiodicity measurements in five frequency bands: 0-1, 1-2, 2-4, 4-6, and 6-8 kHz.

The aperiodicity measure used by STRAIGHT consists of the ratio between the lower and upper smoothed spectral envelopes of the short-time speech signal (Kawahara et al., 2001). The *upper envelope*, $|S_U|^2$, is calculated by connecting spectral peaks (typically located at the harmonic frequencies) and the *lower envelope*, $|S_L|^2$, is calculated by connecting spectral valleys (located around the middle point of two harmonic frequencies). Next, the aperiodicity is calibrated by a table-look-up, averaged and weighted by the speech power spectrum $|S(w)|^2$ to obtain the final aperiodicity measurement $P_{AP}(w)$:

$$P_{AP}(w) = \frac{\int w_{ERB}(\lambda; w) \left|S(\lambda)\right|^2 \Gamma\left(\frac{|S_U|^2}{|S_L|^2}\right) d\lambda}{\int w_{ERB}(\lambda; w) \left|S(\lambda)\right|^2 d\lambda}, \tag{4.3}$$

where $w_{ERB}(\lambda; w)$ represents a simplified *auditory filter shape* for smoothing the power spectrum at the center frequency w and $\Gamma()$ represents a table-look-up operation to calibrate the spectral ratio obtained from simulation results using known aperiodic signals.

The method used by the STRAIGHT vocoder to calculate the spectral envelope and the aperiodicity component are explained in more detail in Section 4.3.3.1.

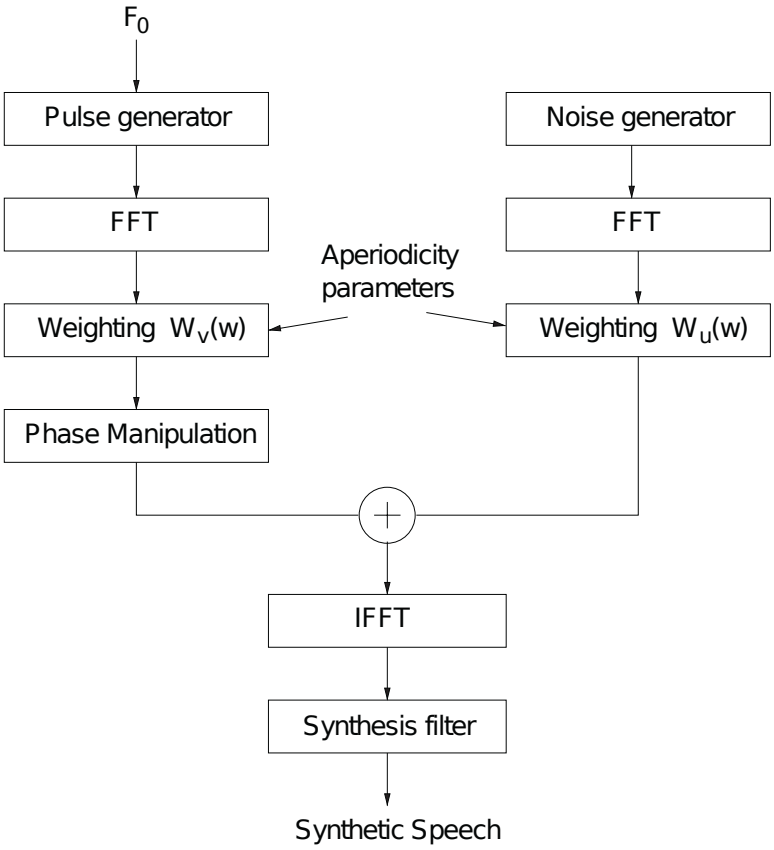

Figure 4.5: Speech waveform generation method of the Nitech-HTS 2005 system, which uses the mixed excitation model of the STRAIGHT vocoder.

4.3.3.2 Synthesis

The block diagram of the synthesis part of the Nitech-HTS 2005 system is shown in Figure 4.5. This system synthesises speech pitch-synchronously by using frames with length equal to twice the length of the pitch period (a fixed length for unvoiced speech). For unvoiced speech frames, white Gaussian noise is uniformly distributed along the unvoiced excitation frame. For voiced frames, the system generates a multi-band mixed excitation signal similarly to STRAIGHT, in order to reduce the buzzy quality caused by the impulse train signal. The weighting of the noise and the periodic components of the excitation is performed by multiplying the amplitude spectrum of each signal by a stepwise function, respectively. The two stepwise functions are different and they are defined by a constant weight value in each frequency band. The speech synthesiser obtains the stepwise functions from the aperiodicity parameters defined for the five frequency bands. Figure 4.6 shows an example of the voiced and unvoiced weighting functions used to synthesise a speech frame by the system. The

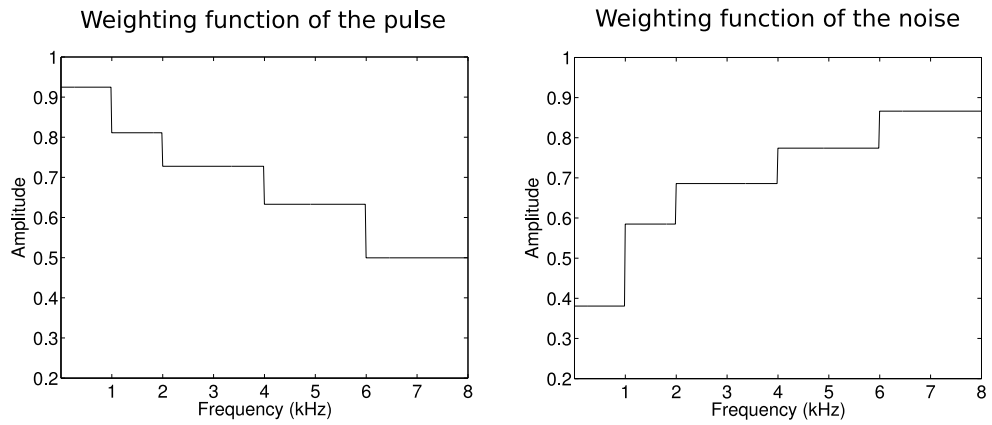

Figure 4.6: Example of the weighting functions of the periodic and noise components generated by the Nitech-HTS 2005 system, which uses a mixed excitation model similar to that of STRAIGHT.

amplitude spectrum of the signal obtained by mixing an impulse train with the noise, using these weighting functions, is shown in Figure 4.7. The weighting functions used by STRAIGHT are smoother than those of the HMM-based speech synthesiser, because the length of the aperiodicity parameters vector is the same as the number of the frequencies components of the Fourier transform of the speech (which is obtained with 1024 point FFT).

The Nitech-HTS 2005 system also employes the STRAIGHT method for manipulation of the phase of the delta pulse, in order to reduce the buzzy timbre. This method consists of using an all-pass filter function $\Phi(w)$ of the excitation pulse (delta pulse), which is based on the *group delay* design using random numbers. The desired spread σ_g of the target group delay function τ_g is calculated by the following equations (Kawahara et al., 2001):

$$\tau_g(w) = \rho(w) \frac{\sigma_g x(w)}{\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} |x(w)|^2 dw}} \tag{4.4}$$

$$x(w) = F^{-1}\left(W_s(\tau)N(\tau)\right), \tag{4.5}$$

where F^{-1} denotes the inverse fast Fourier transform (IFFT) and $N(\tau)$ is the initial random group delay function obtained by weighting Gaussian white noise, $n(t)$, with the function $W_s(\tau)$, in the spatial frequency domain. In this equation, $\rho(w)$ represents a frequency-weighting function used to control the *temporal energy spread* in each frequency region of the pulse excitation. The phase characteristic of the excitation

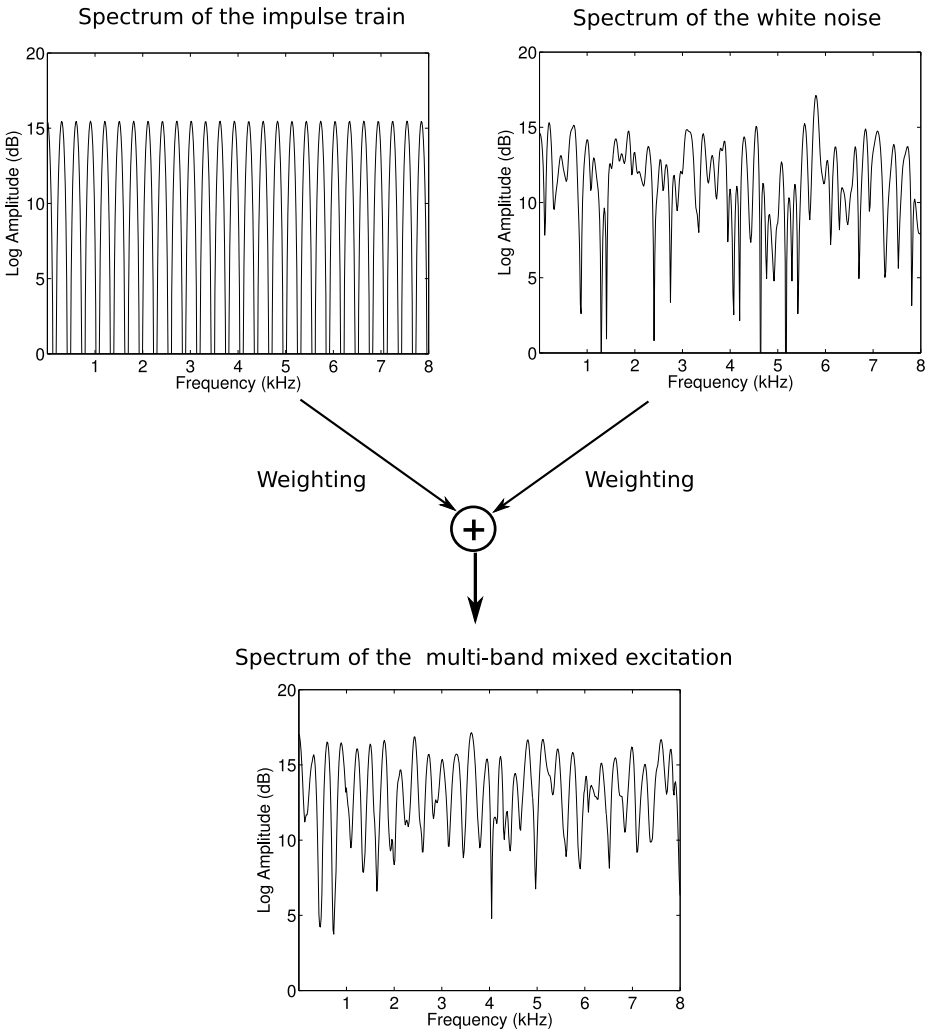

Figure 4.7: Example of the spectrum of the mixed excitation obtained with STRAIGHT (without phase manipulation).

$\Phi(w)$ is calculated by integrating $\tau_g(w)$. Figure 4.8 shows the effect of the group delay manipulation on the pulse signal. The two signals were obtained using the MATLAB version of STRAIGHT. Note that the pulse signal shown in Figure 4.8, without phase manipulation, is slightly different from the traditional delta pulse which is used by Nitech-HTS 2005. The segment of the pulse train $e(n)$ used by STRAIGHT, which is shown in Figure 4.8, is calculated as

$$e(n) = -\frac{h(n)}{\sum_{k=1}^{2N_0} h(n)} + 1, \tag{4.6}$$

where N_0 is the number of samples of the pitch period T_0 and $h(n)$ is a *Hanning window* with length equal to twice the pitch period length. The signal $e(n)$ is multiplied by $\sqrt{N_0}$

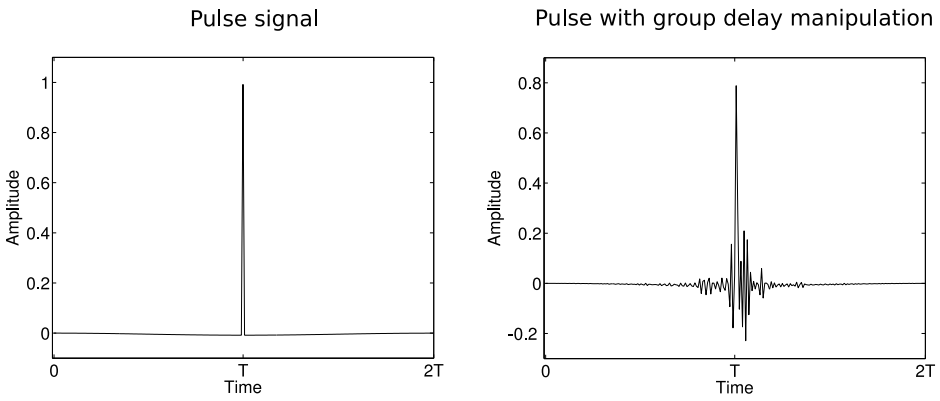

Figure 4.8: Effect of the group delay manipulation performed by STRAIGHT on the simple pulse signal. In this example, the pitch period of the pulse excitation is equal to 7.6 ms and the standard deviation of random group delay was set equal to the standard value used by STRAIGHT of 0.5 ms.

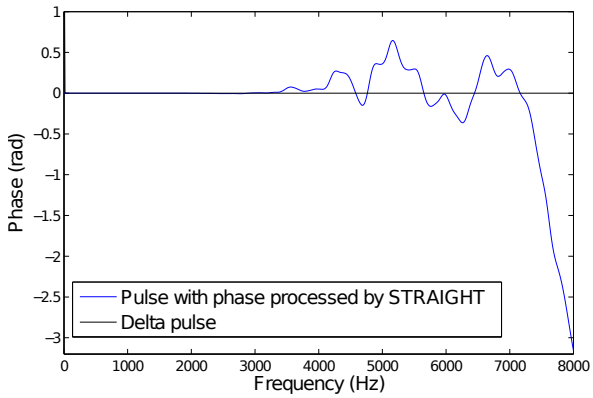

Figure 4.9: Phase spectra of a delta pulse and a pulse generated by STRAIGHT using group delay manipulation.

to have the same power as the noise signal. Figure 4.9 shows the phase spectra of a delta pulse and a phase-processed pulse used by STRAIGHT.

STRAIGHT reconstructs a speech signal by convolving the excitation signal with a minimum-phase impulse response which is obtained by calculating the complex cepstrum of the speech spectrum (Kawahara et al., 2001). This method is described in more detail in Section 6.2.3. The Nitech-HTS 2005 system uses a MLSA filter instead of the STRAIGHT minimum-phase filter for generating the speech waveform. Finally, the system concatenates the synthetic speech frames using the PSOLA technique (Moulines and Charpentier, 1990).

4.3.3.3 Statistical Modelling

The general characteristics of the statistical model of the Nitech-HTS 2005 system are presented in Table 4.3. The statistical model has three streams for mel-cepstrum, F_0 and aperiodicity parameters. The spectral and aperiodicity parameters are modelled by single diagonal Gaussian distribution while F_0 and its first and second derivatives are modelled by an MSD each.

Streams	Probability Distributions
Mel-cepstrum	Gaussian
F0	Multi-space
Aperiodicity	Gaussian

Table 4.3: Information about the statistical model used by the Nitech-HTS 2005 system.

A decision tree is separately constructed for each state position of spectrum, F_0, aperiodicity measurements, and state duration. Zen et al. (2007a) give information about the number of leaf nodes of constructed decision trees for the different types of features, for different voices built with this system. The number of nodes for the aperiodicity measurements (minimum of 676 and maximum of 924) is of the same order of the number of nodes for the spectral parameters (minimum of 859 and maximum of 1021) but it is significantly lower than the number of nodes for F_0 (minimum of 1691 and maximum of 2090) on average.

4.3.4 Harmonic-plus-Noise Model

The HMM-based speech synthesisers of Kim et al. (2006), Kim and Hahn (2007), and Drugman et al. (2009b) respectively employ the hybrid harmonic/stochastic or Harmonic-plus-Noise Model (HNM) of speech (Stylianou, 2001, 1996), in order to combine the periodic and noise components of the excitation. The HNM has also been used to represent the speech signal in HMM-based speech synthesis by Banos et al. (2008) and Hemptinne (2006), but the following sections only describe the methods which use the HNM for excitation modelling.

The HNM divides the spectrum of the speech signal into two bands separated by the *maximum voiced frequency*, F_m. The low-frequency band is composed of a harmonic

structure, while the high-frequency band contains a *modulated noise* component.

4.3.4.1 Analysis

The harmonic part of the speech signal in HNM is described by a sum of harmonics:

$$s_h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{jkw_0(t)t}, \tag{4.7}$$

where $L(t)$ represents the number of harmonics included in the harmonic part, $w_0(t)$ is the fundamental frequency and $A_k(t)$ is a complex number which represents the amplitude and phase of the harmonic k. The number of harmonics depends on $w_0(t)$ and F_m.

The method proposed by Stylianou (2001) to calculate F_m is based on a *peak picking algorithm*. Spectral peaks are searched for along the spectrum of the speech signal and they are classified as voiced or unvoiced depending on a threshold based test, called the "*harmonic test*". After applying a *smoothing filter* to the resulting values of the harmonic test, F_m is estimated as the highest voiced frequency.

Stylianou (2001) describes the noise component of the HNM, $s_n(t)$, using a time-varying autoregressive (AR) model, $h(\tau,t)$, and a time-domain modulation imposed by a *parametric envelope*, $e(t)$, as follows:

$$s_n(t) = e(t)\left[h(\tau,t) * b(t)\right], \tag{4.8}$$

where $*$ denotes convolution and $b(t)$ is white Gaussian noise. The HNM uses modulated noise in order to better represent the time-domain characteristic of the noise, as the noise in natural speech usually is not spread uniformly over the whole pitch period (it appears as noise bursts, instead). The noise parameters used by Stylianou (2001) are the coefficients of the AR filter (10^{th} order) and ten values of speech variance, which are estimated per speech frame using ten sub-windows.

The HMM-based speech synthesiser proposed by Kim et al. (2006) uses a two-band excitation model which is a simplification of the conventional HNM. The excitation parameters used by this synthesiser are F_0 and F_M only. The spectral parameters are the LSF coefficients calculated from the speech signal. The synthesiser also uses a different method to estimate the F_M parameter than the original HNM method (Stylianou, 2001), in order to improve the robustness of the analysis. This technique estimates F_M from the *normalised correlation* of the high-pass filtered speech $R^f_{n,HB}$, which is calculated using the following equations:

$$R^f_{n,HB}(\tau) = \frac{\sum_{n=0}^{N-1} s^f_{HB}(n) s^f_{HB}(n+\tau)}{\sqrt{\sum_{n=0}^{N-1} \left\{ s^f_{HB}(n) \right\}^2 \sum_{n=0}^{N-1} \left\{ s^f_{HB}(n+\tau) \right\}^2}} \tag{4.9}$$

$$s^f_{HB}(n) = h^f_{HPF} * s(n), \tag{4.10}$$

where τ is the number of samples of the pitch period, N is the pitch analysis window size, h^f_{HPF} is the high-pass filter with cut-off frequency f and $s^f_{HB}(n)$ is the filtered high-band speech. First, each speech frame is classified as voiced or unvoiced and F_0 is calculated. Next, if the input frame is voiced it is filtered sequentially with high-pass filters of increasing cut-off frequencies and $R^f_{n,HB}$ is calculated for each signal. F_M is estimated as the lowest cut-off frequency which satisfies $R^f_{n,HB} < 0.5$. This method is based on the assumption that speech is characterised by a more irregular harmonic structure at higher frequencies. The autocorrelation of a signal is excepted to increase with its degree of periodicity (autocorrelation is close to one for a periodic signal and close to zero for an aperiodic signal). Then, if the signal is aperiodic, a lower cut-off frequency f would result in higher $R^f_{n,HB}$.

Recently, a more accurate method to estimate the maximum voiced frequency has been proposed by Han et al. (2009), in order to improve the quality of HMM-based speech synthesis using HNM. This technique consists of employing an iterative analysis-by-synthesis scheme to minimise spectral distortion and estimate the optimal F_M. The initial estimate of F_M for the iterative algorithm is calculated from the normalised correlation of high-pass filtered speech.

The HMM-based speech synthesiser proposed by Drugman et al. (2009b) also uses the idea of HNM to model the excitation. However, the model of the harmonic component is different from the model described by (4.7). The parameters of the periodic excitation are the *principal components* (Jolliffe, 2002) of the residual calculated by inverse filtering, instead of the harmonic amplitudes and phases of the speech signal. Also, the maximum voiced frequency is set equal to a constant value $F_M = 4$ kHz.

In general, the statistical speech synthesisers which use a two-band excitation model (Kim et al., 2006; Kim and Hahn, 2007; Drugman et al., 2009b) do not estimate the all-pole coefficients and the variance parameters of the HNM (Stylianou, 2001). The AR parameters of the noise are not modelled by the synthesisers because the spectrum of the excitation model is assumed to be approximately flat. In these systems, the amplitude spectrum of the unvoiced speech signal is shaped by the synthesis filter, which represents the spectral envelope.

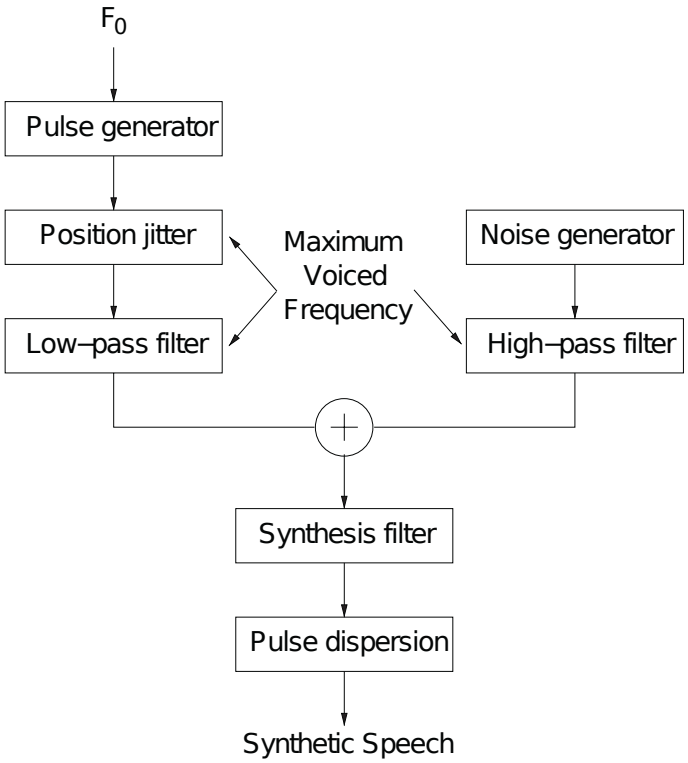

Figure 4.10: Synthesis part of a HMM-based speech synthesiser which uses the maximum voiced frequency parameter of the HNM to mix the harmonic and noise parts of the mixed excitation.

4.3.4.2 Synthesis

The harmonic part of the HNM is synthesised by a sum of sinusoids using (4.7), which are calculated from the estimated F_0, amplitudes of the harmonics and their phases. The noise component of the speech signal is obtained by filtering a unit-variance white Gaussian noise through the all-pole filter and modelling the envelope of the resulting signal using the variance parameters, as described by the noise model in (4.8). For synthesising a voiced speech frame, the noise component is also high-pass filtered with cut-off frequency F_M and then it is multipled by a time-domain envelope (parametric triangular function) synchronized with the pitch period. The noise and harmonic parts are shifted to be centered on the *center of gravity* of the harmonic part (Stylianou, 2001) . Then, the periodic and noise signals are added together pitch-synchronously.

Figure 4.10 shows the synthesis part of the HMM-based speech synthesiser proposed by Kim et al. (2006), which uses F_M to model the two-band mixed excitation. It is similar to the synthesis method used by the HMM-based speech synthesiser using

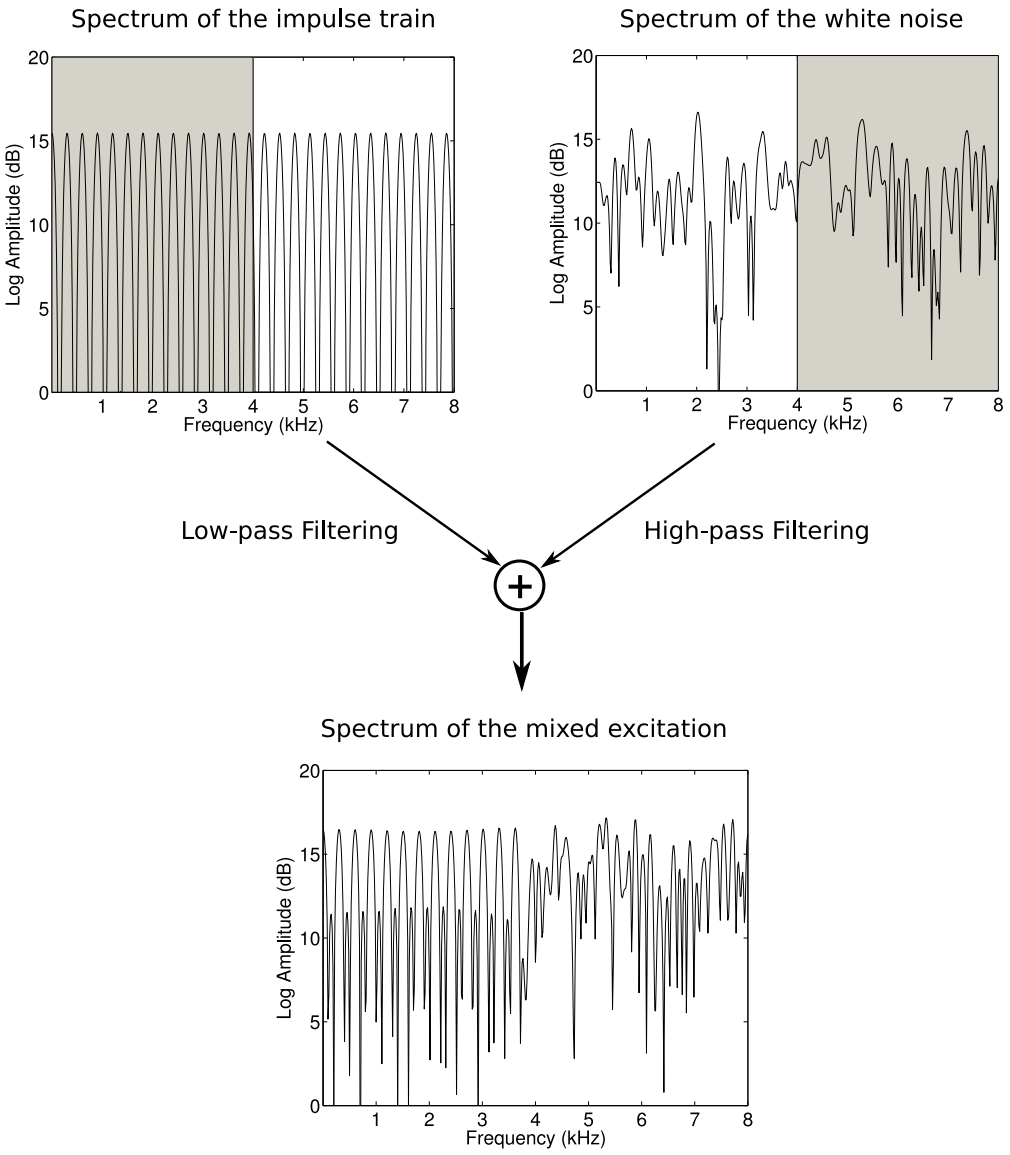

Figure 4.11: Mixing of a low-pass filtered impulse train with high-pass filtered white noise. In this example, the cut-off frequency of the filters is 4 kHz. The shaded regions in the upper left and right plots represent the frequency bands of the low- and high-pass filters, respectively.

MELP, which is illustrated in Figure 4.3, except the bandpass filters and the voicing strength decisions are different. Also, the synthesiser with HNM does not use the harmonic amplitudes of the periodic pulse (which are used by MELP) to generate the periodic pulse train. The position jitter and the pulse dispersion techniques are the same as used by MELP. The HMM-based speech synthesiser with HNM uses fifteen pairs of the 6th-order Butterworth low- and high-pass filters, which are designed with

0.5 kHz step increment within the 8 kHz frequency band. The maximum voiced frequency parameter is used to select the low- and high-pass filters during synthesis. The filters divide the full bandwidth into the lower and higher frequency band. The low- and high-pass filters are applied to the pulse and the white noise signals, respectively. Then, the mixed excitation is obtained by adding the filtered signals together. Figure 4.11 illustrates the mixing of an impulse train signal with a noise signal using the low- and high-pass filters of the excitation model, in the frequency domain. In this example, $F_M = 4$ kHz. This system does not perform the amplitude modulation of the noise, which is used in the conventional HNM (Stylianou, 2001).

The statistical speech synthesiser of Drugman et al. (2009b) uses the parameters of the residual signal to model the harmonic part of the HNM. The synthesis of this periodic excitation is described in the next section. The noise component is synthesised using (4.8). However, the autoregressive-model $h(\tau,t)$ is always the same and acts as a high-pass filter, with cut-off frequency $F_M = 4$ kHz and slightly attenuated in the very high frequencies (near 8 kHz). The variance parameter of the noise in the HNM is not modelled by this system. The noise is modulated by a *pitch-dependent triangular window* only.

4.3.4.3 Statistical Modelling

Table 4.4 shows the general characteristics of the statistical model used by the HMM-based speech synthesisers with HNM. Each data stream contains the static, delta and delta-delta features. The maximum voiced frequency parameters are modelled with a multi-space probability distribution because they are not estimated for unvoiced speech.

Streams	Probability Distributions
LSP	Gaussian
F0	Multi-space
Maximum Voiced Frequency	Multi-space

Table 4.4: Information about the statistical model used by the HMM-based speech synthesisers with HNM.

The acoustic modelling part of the system proposed by Drugman et al. (2009b), which uses the residual parameters to model the periodic excitation of the HNM, is described later in Section 4.4.3.3.

4.3.5 Speech Quality

By using a multi-band mixed excitation in HMM-based speech synthesis, the quality of the synthetic speech can be significantly improved compared with the simple pulse/noise excitation. However, the speech quality achieved by the state-of-the art synthesisers which use this type of excitation is still far from the quality of human speech.

The results of the experiment conducted by Yoshimura et al. (2001) in order to evaluate their HMM-based speech synthesiser which uses the MELP vocoder indicated that modelling the Fourier magnitudes, the jitter processing and the pulse dispersion had a small effect on the synthetic speech quality. According to these results, the main contribution to the improvement in speech quality by using MELP is the mix of the noise and periodic components of the excitation (by using the bandpass filters, which are controlled by the voicing strength parameters).

4.4 Residual Modelling

4.4.1 Introduction

The residual obtained by inverse filtering the speech signal contains more characteristics of the voice source than the pulse train signal. For example, the residual calculated for voiced speech better approximates the energy contour of the voice source, compared to the impulse train. The residual also contains more detail of the source, compared to both the simple pulse and the multi-band mixed excitation models which were described in Section 4.3. For example, the residual contains *phase information* and *non-linear effects* which are not represented by those excitation models.

HMM-based speech synthesisers which use the multi-band excitation model of the MELP and STRAIGHT vocoders perform signal processing on the pulse train in order to better mimic the non-harmonic characteristics of the source However, the parameters which are use to control the degree of voicing in these synthesisers are usually calculated heuristically from the speech signal, e.g. the voicing strength parameters of the MELP vocoder and the maximum voiced frequency of the HNM.

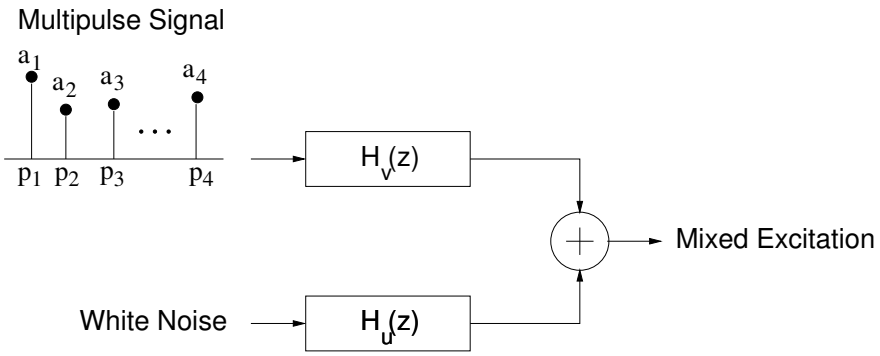

Figure 4.12: Mixed excitation model based on a multipulse signal and adaptive filtering.

This section presents an overview of the statistical speech synthesisers which use the waveform of the residual signal to model the relevant characteristics of the source.

4.4.2 Multipulse-based Mixed Excitation

4.4.2.1 Excitation Model

The speech synthesiser described by Maia et al. (2007a) is based on the Nitech-HTS 2005 system (Zen et al., 2007a) but it uses a different excitation model to the STRAIGHT multi-band mixed excitation. They proposed a model which is based on *state-dependent filters* and pulse trains. This model resembles *multipulse* excitation linear prediction coding algorithms, such as the one used by the Code Excited Linear Prediction (CELP) vocoder of Guerchi and Mermelstein (2000). Figure 4.12 shows the block diagram of this excitation model. The periodic component of the excitation is represented by a multipulse signal (defined by the positions p_j and the amplitudes a_j of the pulses) and the coefficients of a *voiced filter*, $H_v(z)$. The input of the pulse train to the voiced filter yields a signal which is intended to be as similar as possible to the residual. The noise component is modelled by the coefficients of an *unvoiced filter*, $H_u(z)$, which weights the white noise in terms of the spectral shape and power.

4.4.2.2 Analysis

Figure 4.13 shows the system used by Maia et al. (2007a) to estimate the filters and optimise the positions and amplitudes of the multipulse $t(n)$, by minimising the error $w(n)$ between the input residual signal $e(n)$ and the periodic component $v(n)$. The goal of pulse optimisation is to approximate the voiced excitation $v(n)$ to $e(n)$ as much as possible, in a way to remove the short and long-term correlation of the unvoiced

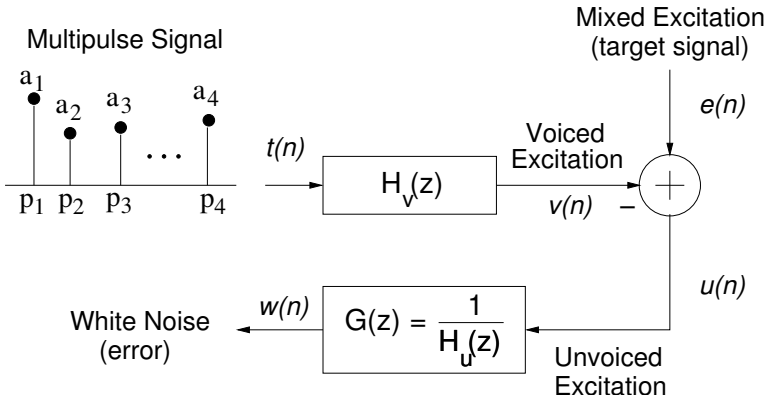

Figure 4.13: System used to maximise the likelihood of the residual given the excitation model.

excitation, $u(n)$, during the filter calculation. The function of $H_u(z)$ is to remove the remaining long term correlation from the signal $u(n)$.

The transfer functions of the voiced and unvoiced filters, respectively, are given by

$$H_v(z) = \sum_{l=-\frac{M}{2}}^{\frac{M}{2}} h(l) z^{-l} \tag{4.11}$$

$$H_u(z) = \frac{1}{G(z)} = \frac{K}{1 - \sum_{l=1}^{L} g(l) z^{-l}}, \tag{4.12}$$

where M and L are the respective orders of the filters, and K is the gain of the unvoiced filter.

The residual vector $\mathbf{e} = [e(0) \dots e(N-1)]^T$ is the sequence of all the residual samples, with length N, which are computed from the speech database. As shown in the analysis system in Figure 4.13, the unvoiced excitation vector $\mathbf{u} = [u(0) \dots u(N-1)]^T$ is given by

$$\mathbf{u}(n) = \mathbf{e}(n) - \mathbf{v}(n), \tag{4.13}$$

where $[\cdot]^T$ means transposition and $\mathbf{v} = [v(0) \dots v(N-1)]^T$ is the voiced excitation vector. The error vector $\mathbf{w}$ can be represented by

$$\mathbf{w} = \mathbf{G}\mathbf{u}, \tag{4.14}$$

where $\mathbf{G}$ is an $N \times (N+L)$ matrix containing the overall impulse response of the *inverse unvoiced filter* $G(z)$. Maia et al. (2007a) compute a voiced and unvoiced filter for all the HMM states, $\{1, \dots, S\}$, along the entire database. The residual segments which

are used to calculate the impulse responses for each state s are obtained using Viterbi alignment of the speech database. $\mathbf{G}$ contains the impulse responses $\tilde{G}_{j,s}$ for all the j-speech segments, which are assigned to a state s. Since the filters are state-dependent, the overall voiced excitation $\mathbf{v}$ is given by

$$\mathbf{v} = \mathbf{A}_1\mathbf{h}_1 + \ldots + \mathbf{A}_S\mathbf{h}_S, \tag{4.15}$$

where $\mathbf{h}_s = [h_s(-M/2)\ldots h_s(M/2)]^T$ is the impulse response vector of the voiced filter for state s and $\mathbf{A}_s$ is the overall pulse train matrix where only the pulse train positions belonging to state s are non-zero.

By using (4.13) to (4.15), the likelihood of $\mathbf{e}$ given the excitation model is

$$P\left[\mathbf{e}|H_v(z),H_u(z),t(n)\right] = \frac{1}{\sqrt{(2\pi)^N(\left|\mathbf{G}^T\mathbf{G}\right|)^{-1}}} e^{-\frac{1}{2}\left[\mathbf{e}-\sum_{s=1}^{S}\mathbf{A}_s\mathbf{h}_s\right]^T\mathbf{G}^T\mathbf{G}\left[\mathbf{e}-\sum_{s=1}^{S}\mathbf{A}_s\mathbf{h}_s\right]} \tag{4.16}$$

The state-dependent filter $H_v(z)$ is calculated by maximising the log likelihood. Thus, the vector of coefficients of the voiced filter $\mathbf{h}_s$ for each state s, can be obtained from

$$\frac{\partial \log P\left[\mathbf{e}|H_v(z),H_u(z),t(n)\right]}{\partial \mathbf{h}_s} = 0, \tag{4.17}$$

which results in

$$\mathbf{h}_s = \left[\mathbf{A}_s^T\mathbf{G}^T\mathbf{G}\mathbf{A}_s\right]^{-1}\mathbf{A}_s^T\mathbf{G}^T\mathbf{G}\left[\mathbf{e} - \sum_{k=1,k\neq i}^{S}\mathbf{A}_l\mathbf{h}_l\right] \tag{4.18}$$

Maia et al. (2007a) solve this linear system by considering the *least-squares formulation* for the design of a filter (Jackson, 1996).

The state-dependent filter $H_u(z)$ is obtained from another expression of the log likelihood. The coefficients of $G(z)$ are calculated as

$$\frac{\partial \log P\left[\mathbf{e}|H_u(z)\right]}{\partial K} = 0 \tag{4.19}$$

This equation can be solved by performing autoregressive spectral analysis on $u(n)$ over speech segments belonging to the state s. Maia et al. (2007a) first estimate the mean autocorrelation function for each state and then calculate the filter coefficients using the Levinson-Durbin algorithm (Markel and Gray, 1976). The all-pole structure based on LP coefficients, which is given by (4.12), was chosen because of its simplicity and to ensure the stability of $H_u(z)$.

Maia et al. (2007a) proposed a method to optimise the pulse positions and amplitudes of $t(n)$ similar to the technique used by the MELP coders of McCree and Barnwell III (1995); McCree et al. (1996). The algorithm consists of minimising the *mean squared error* $\varepsilon = \frac{1}{N}\mathbf{w}^T\mathbf{w}$ of the system shown in Figure 4.13. The expression for this error is

$$\varepsilon = \frac{1}{N}\left[\mathbf{e} - \sum_{s=1}^{S}\mathbf{A}_s\mathbf{h}_s\right]^T \mathbf{G}^T\mathbf{G}\left[\mathbf{e} - \sum_{s=1}^{S}\mathbf{A}_s\mathbf{h}_s\right] \tag{4.20}$$

Figure 4.14 shows the block diagram of the recursive algorithm proposed by Maia et al. (2007a) to estimate iteratively the filters $H_v(z)$ and $H_u(z)$, and optimise the positions and amplitudes of $t(n)$. In the HMM-based speech synthesiser of Maia et al. (2007a), the residual signal is extracted by inverse filtering the speech signal with the Mel Log Spectrum Approximation (MLSA) model. Figure 4.15 shows an example of the residual waveform calculated by inverse filtering the speech signal using the mel-generalised cepstral coefficients. The pulse positions are first obtained from the *pitch-marks* and each voiced filter is initialised by $h_s(n) = \delta(n)$, which means that the initial pitch pulses are given by the pitch-marks in $e(n)$. In the recursive algorithm, the pulses are optimised by calculating the pulse positions and amplitudes from (4.20) and keeping the filters $H_v(z)$ and $H_u(z)$ constant for each state. Next, the coefficients of the voiced filter $\left\{h_s(-\frac{M}{2}),\ldots,h_s(\frac{M}{2})\right\}$ and the coefficients of the inverse unvoiced filter $\{g_s(1),\ldots,g_s(l)\}$ and its gain K_s, are calculated for each state s using (4.18) and (4.19), respectively. The stop criterion is obtained from the *voiced filter variation tolerance* and the maximum number of iterations.

4.4.2.3 Synthesis

Speech is synthesised according to the excitation model represented in Figure 4.12. The input multipulse and white noise sequences are filtered through the voiced and unvoiced filters, respectively. The resulting noise component is high-pass filtered with cut-off frequency of 2 kHz (Maia et al., 2007b), in order to avoid the synthesis of rough speech. Next, the harmonic and noise components are added together to produce the mixed excitation. In the unvoiced regions, no pulses are assigned to the periodic component of the excitation. Finally, the excitation signal is passed through a MLSA filter defined by the mel-cepstral coefficients.

Although the estimation of pulse positions and amplitudes of the multipulse signal $t(n)$ is performed at the training phase in the HMM-based speech synthesiser of Maia

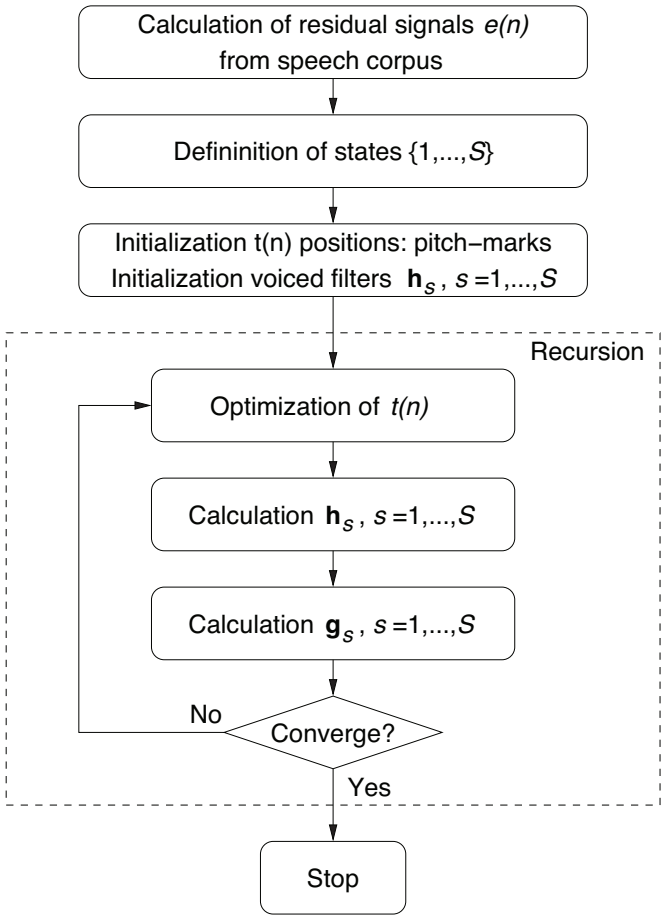

Figure 4.14: Closed-loop algorithm for joint filter calculation and multipulse optimisation.

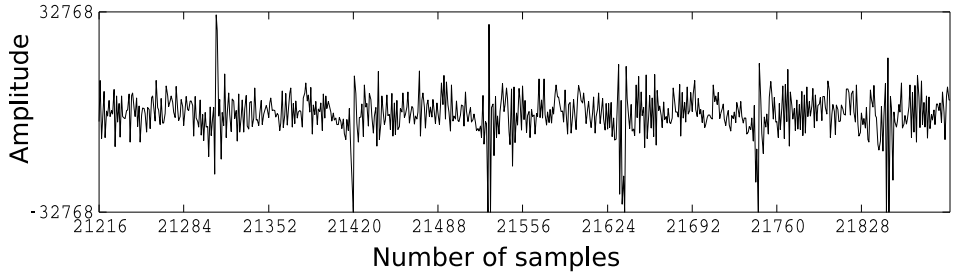

Figure 4.15: Segment of a residual signal calculated by inverse filtering the speech signal using the mel-generalised cepstral coefficients. The mel-generalised spectral analysis of speech was performed with $\alpha = 0.42$, $\gamma = -1/3$, order 39, and windows with duration 25 ms.

et al. (2007a), this system utilises the traditional impulse train generated from F_0 at run-time. They plan to introduce some *multipulse models* to be utilised at run-time synthesis.

4.4.2.4 Statistical Modelling

The HMM-based speech synthesiser of Maia et al. (2007a) models F_0 and the mel-cepstral coefficients using a method similar to that used by the HTS system, which is described in Section 3.4.3. The trained HMMs are clustered by building different decision trees for F_0 and spectrum parameters, in both systems.

The states used to train the voiced and unvoiced filter parameters are not the same as the states which are used to train the F_0 and the spectrum parameters. The states $\{1,\ldots,S\}$ of the multipulse-based excitation are obtained after the training of the HMM-based speech synthesiser and are regarded as leaves of some decision-trees generated for the spectrum stream. For filter calculation, each excitation state must constain a certain number of speech segments. These segments are obtained by mapping their corresponding *full-context labels* onto the clustered states of the referred decision-trees. The boundaries of the speech segment are generated by Viterbi alignment of the database after the training of F_0 and the mel-cepstral coefficients.

Maia et al. (2007a) use phonetic and phonemic questions only, and they adjust parameters which control the size of the trees to obtain small trees. The number of state clusters used by the system is $S = 131$. Using smaller decision-trees to represent the states of the multipulse excitation reduces the computational complexity of the system by using a smaller number of states than the number used to model F_0 and the mel-cepstral coefficients. The method of deriving the filter states from the spectrum stream relies on the assumption that the residual sequences are highly correlated with the spectral parameters from which they were obtained.

The number of additional parameters used to model the excitation compared with the simple excitation is equal to the sum of the number of voiced and unvoiced filters coefficients times the number of states S, that is $S(M+L+2)$. The filter orders in the synthesiser of Maia et al. (2007a) are $M = 512$ and $L = 256$, respectively. The general structure of the HMMs used by this system is given in Table 4.5.

4.4.3 Pitch-synchronous Residual Frames

The HMM-based speech synthesisers proposed by Drugman et al. (2009c,b) are a modified version of the HTS version 2.1 (Tokuda et al., 2009). The main alteration was the integration of a source model based on *pitch-synchronous (PS) residual* signals calculated from the recorded speech, which replaced the impulse train.

The two systems, which use the PS-residuals, parameterise the residual frames us-

Streams	Probability Distributions
Mel-cepstrum	Gaussian
F0	Multi-space
Voiced and Unvoiced filters	Gaussian

Table 4.5: Statistical model used by the HMM-based speech synthesisers which uses a multipulse and adaptive filters to model the excitation.

ing the same method. However, they differ in the way the residual signal is generated from the excitation parameters for synthesising speech. The system proposed by Drugman et al. (2009c) uses a *codebook* of typical residual frames to obtain real segments of the residual, from the excitation parameters. On the other hand, the system from Drugman et al. (2009b) uses a *deterministic stochastic model* of the residual.

4.4.3.1 Analysis

The residual is calculated by performing Mel-Generalised Cepstral (MGC) analysis on the speech signal and by inverse filtering the short-time signal using the MGC coefficients. Figure 4.15 shows a segment of a residual signal calculated using the MGC coefficients. This residual signal was calculated by choosing $\alpha = 0.42$ and $\gamma = -1/3$, which are the same values used for MGC analysis by the systems of Drugman et al. (2009c,b).

The analysis of the residual is perfomed pitch-synchronously by segmenting the signal into frames with duration equal to twice the fundamental period and centered at the *Glottal Closure Instants* (GCI). Both speech synthesisers (Drugman et al., 2009c,b) use the GCI detector proposed by Drugman and Dutoit (2009). This method first calculates the time intervals where the GCI are expected to occur, from the mean-based signal $y(n)$ of the speech waveform $s(n)$. The mean-based signal is given by

$$y(n) = \frac{1}{2N+1} \sum_{m=-N}^{N} w(m)s(n+m), \tag{4.21}$$

where $w(m)$ is a window of length $2N+1$. Drugman and Dutoit (2009) proposed to use a *Blackman window* whose duration is between 1.5 and 2 times the average pitch period $T_{0,mean}$ (they used a duration of $1.75T_{0,mean}$). The final step of the GCI estimation is

to estimate the glottal closure as the *strongest peak* of the linear prediction residual within each interval.

The source parameters used by the speech synthesisers of Drugman et al. (2009c,b) are F_0 and the coefficients calculated by *Principal Component Analysis* (PCA), e.g. Jolliffe (2002), of the PS-residual frames. PCA decomposes the short-time signal on an orthogonal basis defined by a set of *eigenvectors* and its coefficients. The residual parameters trained by the synthesisers are the coefficients of the eigenvectors.

The short-time residual signals are normalised in both length and energy before applying PCA, in order to ensure the coherence of the data set. In general, the number of PCA coefficients which are selected is lower than the length of the residual signal, in order to achieve dimensionality reduction. Drugman et al. (2009b) suggest that 15 eigenvectors calculated by PCA is sufficient to obtain high-quality coding results.

The effect of shortening the residual frames by *resampling* (decimation) is to expand the spectrum of the residual signal. Thus, the resulting normalised frames represent a *low-frequency signature* of the original residual frames. Drugman et al. (2009c,b) assume that the time-scale transformation of the residual preserves the shape parameters of the source signal, such as the open quotient (measures the normalised duration of the open phase) and the speed quotient (measures the asymmetry of the glottal pulse). However, the shape of the residual obtained by inverse filtering is not a correct representation of the shape of the glottal source, for the reasons explained in Section 2.2.3.1. Thus, it is not clear that the resampling of the residual frame preserves the shape characteristics of the source signal.

The HMM-based speech synthesiser of Drugman et al. (2009c) uses a codebook-based method to map the normalised residual frames to a set of residual frames which were extracted from the speech database. In this approach, the *Resampled and energy Normalized* (RN) frames have a length of 20 samples. Figure 4.16 shows the method used to build the codebooks. The RN frames are clustered using the *K-means algorithm*, resulting in approximately 100 *centroids*. The RN frame associated with each centroid is obtained by selecting the ten closest RN frames to each centroid and retaining the longest frame candidate. The longest frame is chosen in order to avoid the appearance of "energy holes" in the spectrum of the synthetic speech. That is, by choosing the longest residual frame the spectral compression effect of time-scaling the residual frame to have a normalised length is reduced. A codebook of real residual frames contains the residual signals assigned to each centroid of the RN codebook.

The speech synthesiser of Drugman et al. (2009b) uses the RN frames as part of

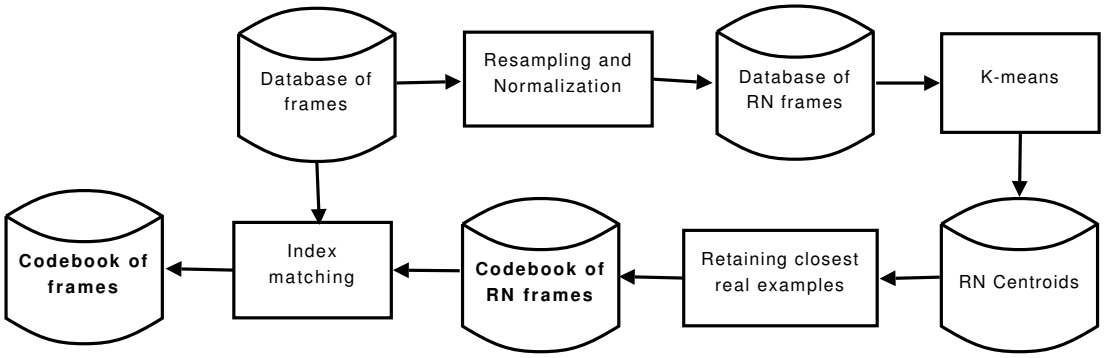

Figure 4.16: Method used to build the codebooks of residual frames.

a harmonic plus stochastic model (or HNM) of the excitation, which was described in Section 4.3.4. Basically, the HNM divides the speech spectrum into two parts and models the lower and higher frequency regions by a harmonic and a noise signal, respectively. The maximum voiced frequency F_M delimits the frequency bands of the two components. (Drugman et al., 2009b) uses the RN frames to represent the harmonic part of the excitation for voiced speech. In this system the maximum voiced frequency has a constant value, $F_M = 4$ kHz. This approach assumes that the low-frequency signature of the RN frames is a good approximation of the low-pass filtered (with cut-off frequency 4 kHz) version of the real residual signal. The use of this HNM for the excitation avoids the mapping of the RN frames to the real residual frames, e.g. using the codebook technique. The HMM-based speech synthesiser using HNM was developed in order to overcome problems associated with the codebook-based technique, in particular to improve the quality of female synthetic speech. This synthesiser chooses the pitch value F_0^* of the RN frames using the following condition:

$$F_0^* \leq \frac{F_N}{F_m} F_{0,min}, \tag{4.22}$$

where F_N and $F_{0,min}$ denote respectively the Nyquist frequency and minimum pitch value measured from the speech database, which is associated with a given speaker. The normalised pitch is restricted by the condition (4.22) in order to avoid the appearance of "energy holes" in the spectrum of the synthetic speech.

4.4.3.2 Synthesis

Figure 4.17 shows the method to synthesise voiced speech by the system which uses a codebook of pitch-synchronous residual frames (Drugman et al., 2009c). The excita-

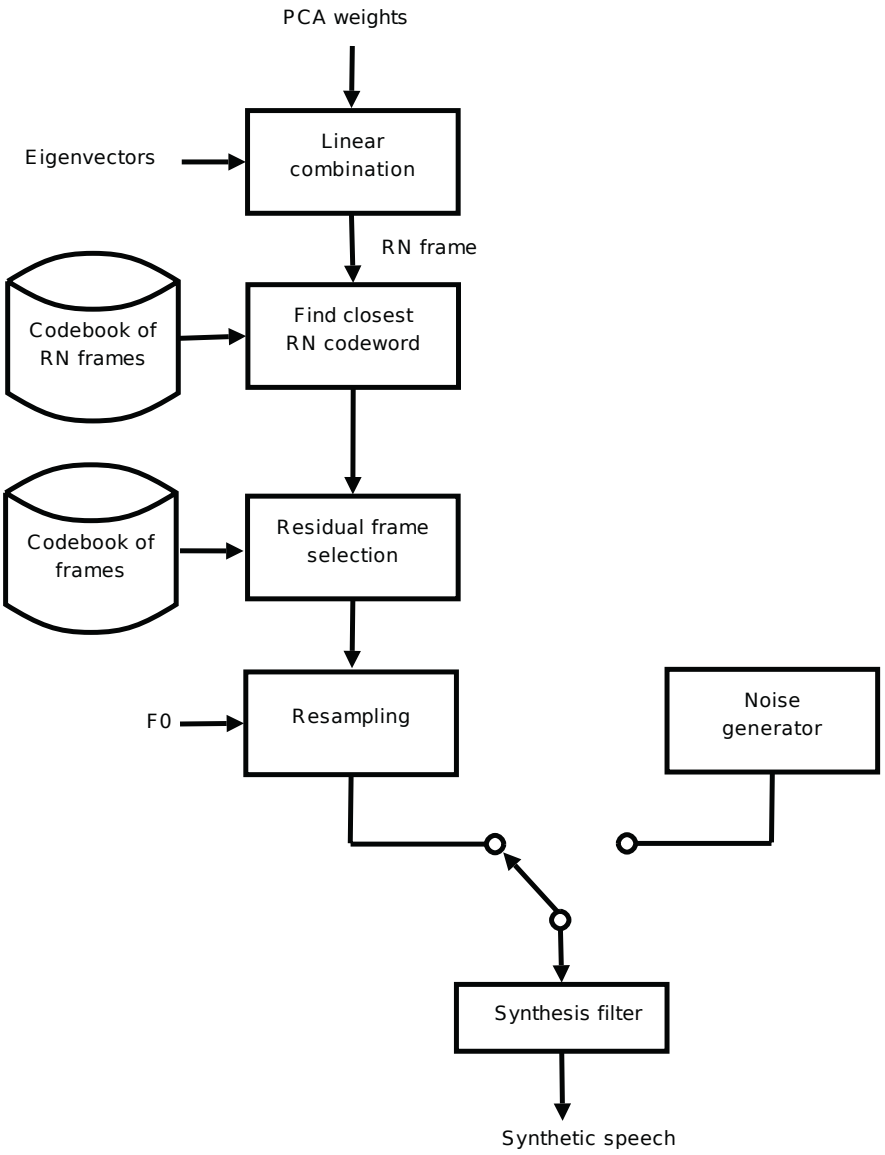

Figure 4.17: Speech synthesis part of the HMM-based speech synthesiser which uses a codebook of pitch-synchronous frames to model the excitation.

tion parameters are F_0 and the PCA coefficients. First, the RN residual frame (residual signal normalised in pitch and energy) is obtained by linear combination of the eigenvectors, using the PCA parameters. Then, the closest residual frame (with the original length and energy) to the RN frame is selected from the codebook by using the *mean square error* criterion. Next, the selected residual signal is resampled to the target pitch, which is given by F_0. Resampling of the residual changes its spectrum. Nevertheless, Drugman et al. (2009c,b) assume that the time-scaling transformation preserves the important parameters of the voice source, which are related to voice quality

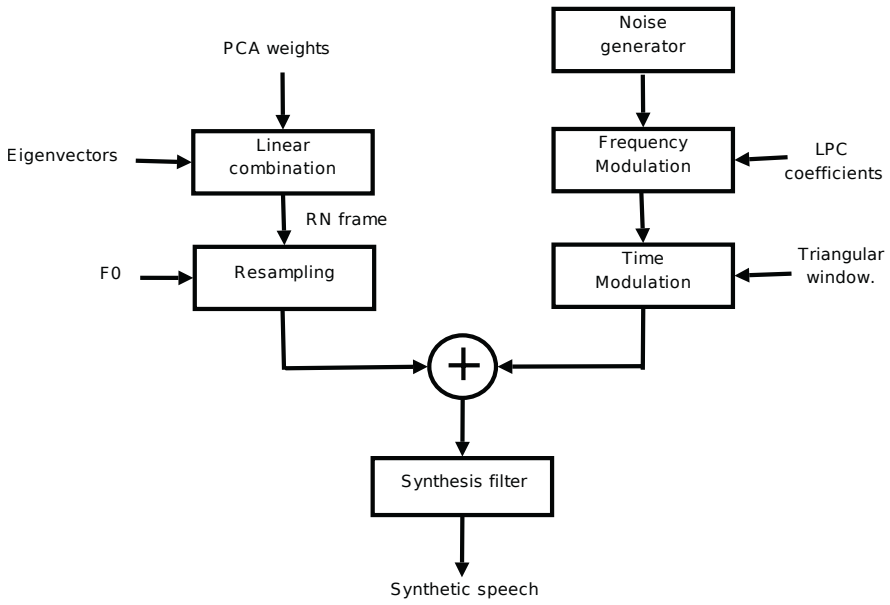

Figure 4.18: Synthesis of voiced speech frames by the statistical speech synthesiser which uses the HNM and pitch-synchronous residual frames to model the excitation.

(such as the open quotient and the speed quotient).

The block diagram of the method to synthesise voiced speech by the synthesiser which uses the HNM (Drugman et al., 2009b) is shown in Figure 4.18. The RN residual frame is generated using the PCA parameters, as in the codebook-based method. However, the RN residual signal is used to represent the harmonic component (low-frequency part) of the HNM instead of using a codebook. This signal is resampled to obtain the desired pitch period and then it is added to the noise component (high-frequency part) to obtain the excitation of voiced speech. For unvoiced speech, the excitation is modelled as white Gaussian noise only. The technique to synthesise the stochastic part of this mixed excitation model has been described in Section 4.3.4. Basically, it consists of high-pass filtering the white noise (beyond $F_m = 4$ kHz) using an autoregressive model and modulating the energy envelope of the signal with a pitch-synchronous triangular window.

4.4.3.3 Statistical Modelling

The acoustic modelling part of the two statistical speech synthesisers which use PS residual frames to model the excitation is similar since they use the same type of acoustic features. The main adjustment made to the training part of the HTS system was to integrate a new data stream for the PCA parameters of the excitation model. This

stream has the same structure as the F_0 stream. The PCA parameters and its first and second derivatives are modelled by a multi-space probability distribution, respectively. The general properties of this type of statistical model are presented in Table 4.6.

Streams	Probability Distributions
Mel-generalised cepstrum	Gaussian
F0	Multi-space
PCA coefficients	Multi-space

Table 4.6: Statistical model used by the HMM-based speech synthesisers which use PS-residual frames to model the excitation.

4.4.4 Speech Quality

The results of the evaluation of HMM-based speech synthesisers using residual modelling (Maia et al., 2007a; Drugman et al., 2009c,b) show that these systems performed considerably better than the standard HMM-based speech synthesiser which uses the simple pulse/noise model. These results give support to the hypothesis that the residual signal better approximates the glottal flow first derivative waveform and better models the source characteristics of voiced speech, when compared with the impulse train.

The results reported by Drugman et al. (2009b), for the system which uses PS residual frames, were obtained by using the first eigenvector of the normalised residual frames only. In this case, the PCA parameters of the excitation do not need to be trained by the statistical speech synthesiser. The higher order eigenvectors were not used because experiments showed that they did not produce audible differences in the synthetic speech.

The speech synthesiser using a multipulse-based excitation model (Maia et al., 2007a) obtained similar results to a conventional HMM-based speech synthesiser which uses the multi-band mixed excitation model. The disadvantage of the multipulse model compared with the mixed multi-band excitation is that it requires many more parameters to model the excitation, e.g. 512 coefficients for the voiced filter and 256 filter coefficients for the unvoiced filter (Maia et al., 2007a). Drugman et al. (2009c,b) evaluated the HMM-based speech synthesisers using PS residual frames against a baseline

system which uses simple excitation. However, they also plan to evaluate their systems against a standard HMM-based speech synthesiser using multi-band mixed excitation.

Residual-based models of the excitation do not represent all aspects of the glottal source signal, since inverse filtering does not accurately separate the source characteristics from the speech signal. For example, there are characteristics of the source, such as the spectral tilt (decaying spectrum at higher frequencies), which are not correctly modelled by the residual signal. In particular, the spectral tilt of the glottal source is incorporated into the spectral envelope of speech, as inverse filtering does not remove the spectral tilt from the speech spectrum.

The filter coefficients of the multipulse model and the PCA coefficients of the residual are not adequate for controlling voice quality in the HMM-based speech synthesisers, because their correlation with voice quality is not known. For example, they do not allow acoustic characteristics of the glottal source which are correlated with voice quality to be easily modified, such as the open quotient (duration of the glottal pulse) and the speed quotient (asymmetry of the glottal pulse).

4.5 Glottal Source Modelling

4.5.1 Introduction

The conventional inverse filtering technique, which was described in Sections 2.1.3 and 2.2.3.1, does not produce an accurate estimate of the glottal source signal. The HMM-based speech synthesiser utilising *Glottal Inverse Filtering* (GIF) proposed by Raitio et al. (2008) uses an analysis method that more accurately estimates the glottal source signal and the vocal tract transfer function than inverse filtering. This system uses a better approximation of the glottal source signal than the delta pulse, in order to produce higher-quality speech than the HMM-based speech synthesiser which uses the simple pulse/noise excitation. Another advantage of using GIF is that it separates the glottal source from the vocal tract components of speech. This enables the statistical speech synthesiser to model the source and the vocal tract independently, which is consistent with the theory of speech production.

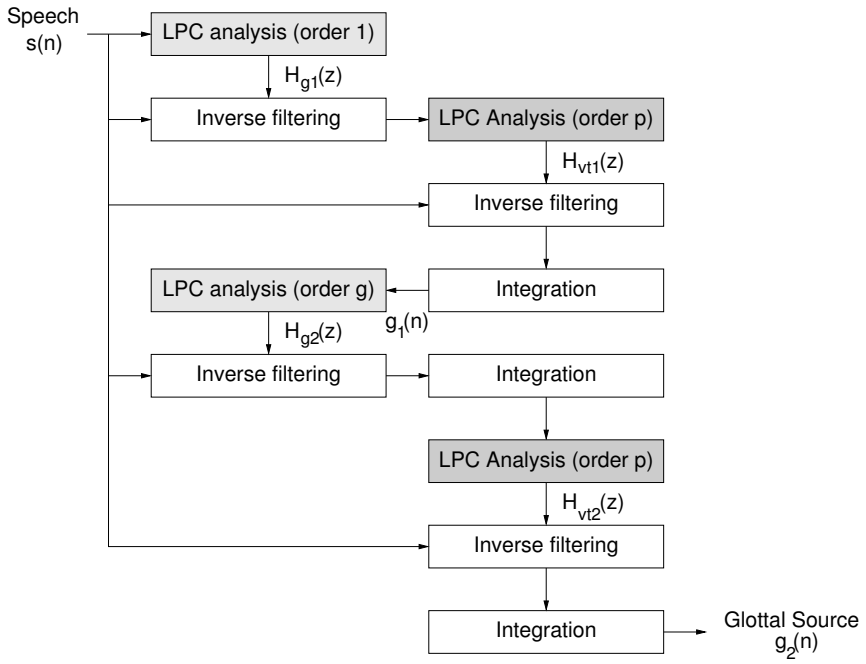

Figure 4.19: Flowchart of the IAIF method. The glottal source signal g(n) and the vocal tract transfer function $H_{vt2}(z)$ are calculated through an iterative algorithm using adaptive all-pole modelling.

4.5.2 Glottal Inverse Filtered Signal

4.5.2.1 Excitation Model

The HMM-based speech synthesiser of Raitio et al. (2008) represents the source by a signal calculated using a GIF method. The parameters calculated by GIF are used to model both the spectra of the vocal tract and the glottal source. For synthesising voiced speech, the system generates the excitation signal by using the source parameters to modify the *real pulse* calculated by GIF. The noise component of this excitation is modelled by the spectral energy of the noise in five frequency bands. Thus, this excitation model is comparable to a multi-band mixed excitation in which the traditional impulse train is replaced by a transformed real glottal pulse.

4.5.2.2 Analysis

The GIF method used by the speech synthesiser (Raitio et al., 2008) is the Iterative Adaptive Inverse Filtering (IAIF) method (Alku et al., 1991), which was introduced in Section 2.2.3.3. This is an automatic method for estimation of the glottal flow waveform and the vocal tract spectrum of voiced speech. Figure 4.19 shows the block

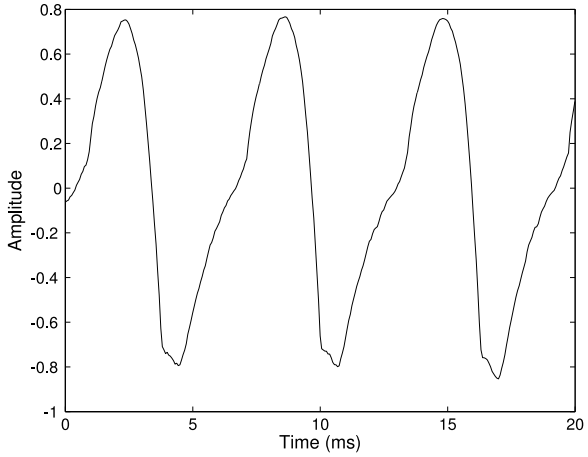

Figure 4.20: Waveform of the glottal source signal calculated using the IAIF method, from a speech frame with duration 20 ms. The analysis was performed pitch-synchronous using an analysis window centered at the estimated glottal epoch.

diagram of this technique. First, the speech signal is inverse filtered using the coefficients of the first order LPC analysis, in order to remove the effect of the spectral tilt associated with the glottal source and the lip radiation. Next, the initial estimate of the vocal tract, $H_{vt1}(z)$, is calculated by performing LPC analysis of order p on the output signal (typically, p is between 10 and 12 for 8 kHz sampling frequency). The glottal source signal $g_1(n)$ is calculated by inverse filtering the speech signal using $H_{vt1}(z)$ and by canceling the lip radiation through integration. The all-pole model of the resulting glottal source signal, $H_{g1}(z)$, is calculated by LPC analysis of order g (typically between 8 and 10 for 8 kHz speech). Then, a second estimation of the vocal tract and the glottal source is conducted. The spectral effect of the glottal source, which is represented by $H_{g2}(z)$, and the lip radiation are canceled from the speech signal through inverse filtering and integration respectively. The final model of the vocal tract, $H_{vt2}(z)$, is obtained by applying LPC analysis of order p to the filter output. Finally, the second estimate of the glottal flow signal, $g_2(n)$, is obtained by canceling the spectral effect of the vocal tract, given by $H_{vt2}(z)$, and the lip radiation from the speech signal. Figure 4.20 shows an example of a glottal source signal (sampled at 16 kHz) calculated by IAIF using LPC orders $p = 20$ and $g = 10$. These orders are equal to those used by Raitio et al. (2008) to analyse speech sampled at 16 kHz. In this example, the three peaks with maximal amplitude that can be observed in Figure 4.20 correspond to the instants of maximal amplitude of three glottal pulses, respectively.

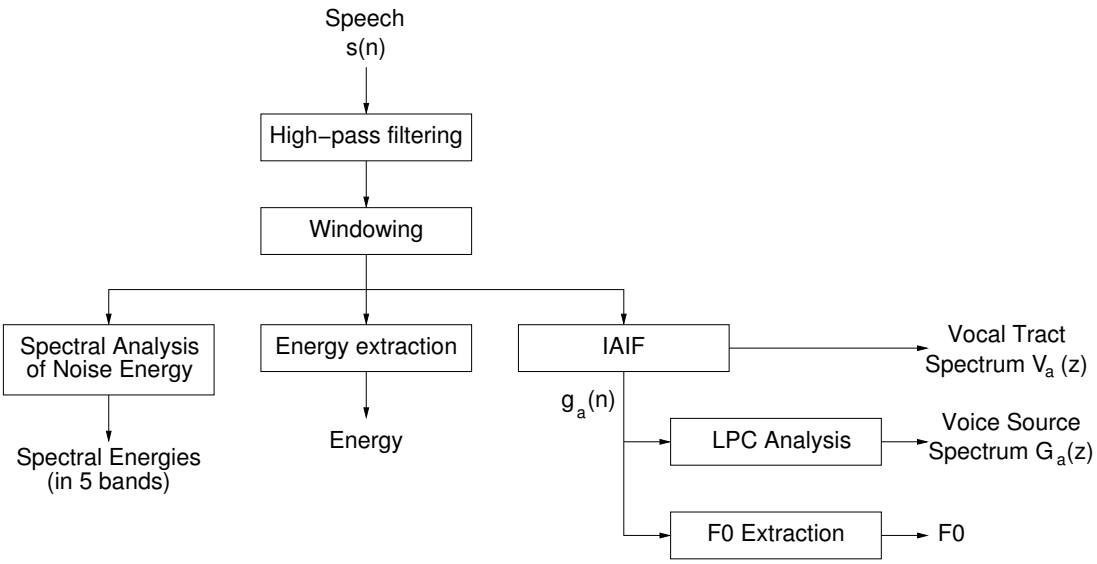

Figure 4.21: Block diagram of the analysis method used by the HMM-based speech synthesiser with glottal source modelling.

Figure 4.21 shows the block diagram of speech analysis which is performed by the HMM-based speech synthesiser using glottal source modelling (Raitio et al., 2008). First, speech is high-pass filtered in order to remove any low-frequency distortions and is segmented using rectangular windows. The energy parameter is calculated directly from the speech waveform, while the spectral energy of the noise is calculated for five frequency bands (0-1, 1-2, 2-4 4-6 and 6-8 kHz) from the amplitude spectrum of the speech signal obtained by FFT. The IAIF method is used to extract the vocal tract spectrum $V_a(z)$ and the glottal source signal $g_a(n)$ from the short-time speech signal (sampled at 16 kHz). Next, the spectral envelope of the glottal flow pulses $G_a(z)$ is parameterised using LPC analysis. The LPC orders of $V_a(z)$ and $G_a(z)$ are 20 and 10 respectively. Finally, the F_0 parameter is also estimated from the glottal source signal using the *autocorrelation function*. The LPC parameters of the voiced and unvoiced spectrum are converted to LSF parameters, as the LSF representation is more adequate for statistical modelling. During speech analysis, one glottal flow pulse is selected and stored as the *library pulse*, in order to be used for speech synthesis.

The IAIF method is not used for the analysis of unvoiced speech, as the source component of this type of speech does not represent the glottal source. For unvoiced speech, Raitio et al. (2008) uses conventional LPC analysis of order 20 for estimating the spectral envelope. In this case, the spectrum of the excitation spectrum is obtained by inverse filtering.

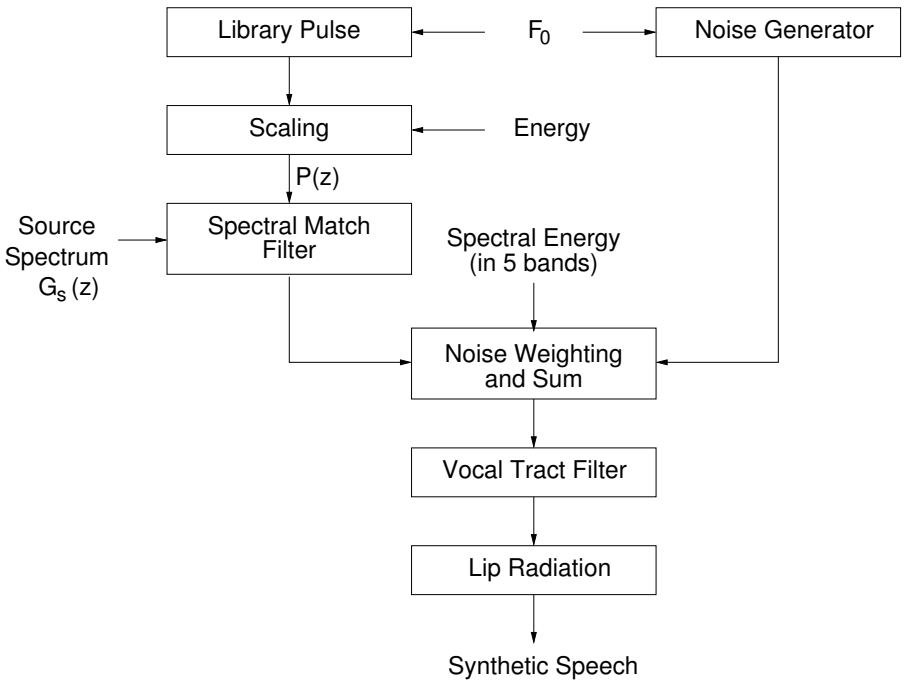

Figure 4.22: Overview of the method to synthesise voiced speech by the HMM-based speech synthesiser using glottal source modelling.

4.5.2.3 Synthesis

The HMM-based speech synthesiser using GIF (Raitio et al., 2008) generates the speech waveform using the method shown in Figure 4.22. A simple method to synthesise the periodic excitation from the spectral parameters of the glottal source is to shape an impulse train with the source spectrum. However, Raitio et al. (2008) proposed a different method. First, the glottal flow signal (from the pulse library) is interpolated and scaled in magnitude to obtain a glottal pulse with the desired period length and energy. Next, the glottal pulse train is filtered by an *adaptive IIR filter* which flattens the spectrum of the glottal pulse train and applies the source spectrum represented by the LPC coefficients. The transfer function of the filter is given by

$$H_{IIR}(z) = \frac{G_s(z)}{P(z)}, \tag{4.23}$$

where $G_s(z)$ is the target all-pole spectrum and $P(z)$ denotes the amplitude spectrum of the library pulse. The goal of using a real glottal source pulse is to capture the aperiodicity characteristics which exist in the real glottal signal. The noise is weighted using the spectral energy parameters in five frequency bands and added to the periodic excitation to obtain the multi-band mixed excitation. The two components of the ex-

citation are added pitch-synchronously by using F_0 to control the duration of the noise signal.

For unvoiced speech, the excitation is synthesised by scaling the energy of the white noise signal. In this case, the spectral energy parameters and the voice source spectrum are not used. Also, the synthesis filter is defined by the parameters of the unvoiced spectrum (represents the spectral envelope) instead of the vocal tract spectrum.

A *formant enhancement* method (Ling et al., 2006a) is applied to the LSFs generated by the speech synthesiser in order to compensate for the averaging effect of statistical modelling. The resulting voiced and unvoiced LSFs are converted to the LPC parameters of the synthesis filter.

4.5.2.4 Statistical Modelling

The training method of the statistical speech synthesiser using glottal source modelling (Raitio et al., 2008) is similar to that of the HTS system, which was described in Section 3.4.3. The main differences between the two systems are in the statistical model structure and the HMM parameter values, such as *stream weights*, and *global variance factors*. Another difference is that the system with glottal source modelling uses LSF parameters to represent the vocal tract and source spectra instead of the mel-cepstral coefficients used by the HTS system to represent the spectral envelope.

The HMM topology of the system which uses the glottal pulse is a 5-state left-to-right model. The feature vectors for mel-cepstrum, F_0, source spectrum, and spectral energies are each assigned to individual streams. Each feature and its derivatives (delta and delta-delta features) are modelled as a continuous probability distribution (Gaussian) streams, except F_0 and its derivatives. The F_0 parameters are modelled by the conventional MSD (because they are not defined in unvoiced regions). Note that the stream used for the vocal tract spectrum contains both the spectrum estimated by the IAIF method for voiced speech and the spectrum estimated by conventional inverse filtering for unvoiced speech. The stream of the voice source spectrum also contains two types of spectrum: the glottal source spectrum and the unvoiced speech spectrum. The structure of the HMM is summarised in Table 4.7.

In the synthesiser proposed by Raitio et al. (2008), the decision tree state-tying is performed for each stream. The contextual features used for the decision tree clustering, such as phone level and higher-level phonological features (e.g. word prominence, and clause type) were extracted using a front-end for Finnish, since the system was built to synthesise Finnish speech.

Streams	Probability Distributions
Vocal tract spectrum (LSP)	Gaussian
Voice source spectrum (LSP)	Gaussian
F_0	Multi-space
Spectral energies	Gaussian
Energy	Gaussian

Table 4.7: General characteristics of the statistical model used by the HMM-based speech synthesiser with glottal source modelling.

4.5.3 Speech Quality

The speech synthesiser which employes GIF uses a library pulse (with a single pulse) to synthesise speech for a given speaker. The excitation model used by this system can be used to modify voice characteristics of the synthetic speech, e.g. by building different library pulses for different speaking styles.

The system uses a real glottal pulse to generate the excitation signal, in order to reproduce the fine characteristics of the glottal source signal. However, the interpolation of the real glottal pulse for controlling the pitch may affect the speech quality, because it produces an "energy hole" in the spectrum of the synthetic speech. Also, this time-scaling transformation does not take into account the variation of the source characteristics with F_0. Several papers show that the glottal parameters are correlated with F_0, such as Strik and Boves (1992); Tooher and McKenna (2003); Fant (1997). This correlation is discussed in Section 5.3.3.

Raitio et al. (2008) conducted an evaluation to compare their statistical speech synthesiser utilising GIF to the HTS system which uses a simple pulse/noise excitation model (described in Section 3.4). The same Finnish voice was built using the two systems for the evaluation. The system of Raitio et al. (2008) was clearly better than the system which used the simple excitation model.

	Spectrum	Periodic excitation	Mixed excitation	Process. of periodic excit.
Pulse/Noise	spec. env.	impulse	-	-
MELP	spec. env.	harmonics of residual	voiced & unvoiced bandpass filters	jitter
STRAIGHT	spec. env.	impulse	spectral weighting	phase processing
HNM	spec. env.	harmonics	voiced LP filter & unvoiced HP filter	jitter
Residual filters	spec env.	filtered multipulse	sum with filtered noise	-
Residual frames	spec. env.	pitch-sync. residual [1]	voiced LP filter & unvoiced HP filter	time-scaling
Glottal source	vocal tract	real glottal pulse	spectral weighting	time-scaling

Table 4.8: General characteristics of the main excitation models used in HMM-based speech synthesis.

4.6 Conclusion

Recently, several methods have been proposed to improve source modelling in HMM-based speech synthesis. Table 4.8 summarises the general characteristics of the excitation models which were reviewed in this chapter. In general, the multi-band mixed excitation, the residual-based, and the glottal source models outperform the simple pulse/noise model. However, the speech quality achieved by the HMM-based speech

[1] Pitch-synchronous residual frames have been modelled using a codebook or HNM. In the first case, the voiced and unvoiced filters of the mixed excitation were not used.

synthesisers using improved excitation models to the pulse/noise is still far from the naturalness of human speech. Further improvements are necessary to produce more natural speech by statistical speech synthesisers and glottal source modelling is one of the aspects which has room for more developments.

The main limitations found in the current excitation models used in HMM-based speech synthesis are :

- correlation between F_0 and source parameters is not modelled.
- signal processing of the excitation signal may deteriorate speech quality.
- reduced control over voice quality.

The excitation models described in this chapter do not seem to be appropriate to model the correlation between the characteristics of the glottal pulse shape and the fundamental frequency, F_0. This explains the fact that all the parameters of the excitation models are trained separately from F_0 by the HMM-based speech synthesisers. For example, the LPC parameters of the glottal source spectrum, used by the synthesiser of Raitio et al. (2008), and the PCA coefficients of the residual signal, used by the synthesiser of Drugman et al. (2009c,b), are both modelled using an individual stream of the HMM. Since the correlation between the F_0 and the source parameters is not modelled by the HMM-based synthesisers, assumptions about the characteristics of the source signal are usually made by the systems during synthesis. For example, the systems proposed by Drugman et al. (2009c) and Raitio et al. (2008) generate the periodic excitation using the source parameters and then resample the resulting signal to reproduce the target pitch. However, this time-scale transformation relies on the assumption that the correlation between the important time parameters of the source pulse and its duration (the period T_0) is linear and has slope of one. In other words, this assumption means that when the pitch period changes, the important shape parameters of the glottal pulse (e.g. the relative duration of the pulse duration with the period and the asymmetry of the pulse) remain the same. Past studies have showed that the behaviour of the glottal parameters with F_0 might not be a direct proportion, e.g. Strik and Boves (1992) and Tooher and McKenna (2003). The correlation between the glottal source parameters and T_0 is discussed in Section 5.3.3.

The time-scale transformations of the excitation to obtain the desirable pitch, which are used by Drugman et al. (2009c,b) and Raitio et al. (2008), could deteriorate the quality of the synthetic speech because they cause compression and expansion of the

spectrum. The phase manipulation techniques used by some HMM-based speech synthesisers with multi-band mixed excitation could also produce speech artefacts if the amount of randomness added to the phase is not appropriate.

Another limitation of the excitation models which have been used in HMM-based speech synthesis, is that they do not offer parametric flexibility to easily control the voice quality of the synthetic speech. The multi-band mixed excitation models used by the statistical synthesisers allow the amount of noise of voiced speech to be controlled and, in some cases, the position jitter and the phase of the pulse train too. However, they do not represent many aspects of the excitation which are important for voice quality, such as the glottal pulse shape. The voice quality control offered by speech synthesisers which use the residual modelling methods is also limited. Both the coefficients of the adaptive filters used by Maia et al. (2007a), and the PCA parameters used byDrugman et al. (2009c,b), do not have acoustic meaning. Thus, the control of the acoustic properties which are related to voice quality (e.g. the waveform and the spectral characteristics of the glottal flow) using this type of parameter is difficult.

The HMM-based speech synthesiser using glottal source modelling of Raitio et al. (2008) uses the LPC parameters of the voice source to model a glottal pulse from a library pulse. One way to transform voice quality using this system could be to use a larger library of glottal pulses for different voice qualities. However, this technique would still have problems for modelling glottal source dynamics related to voice quality, e.g. the variation of voice quality along an utterance. Another way to modify voice quality using this system could be to transform the LPC coefficients. This option also has some difficulties because these parameters are not directly related to the time and spectral-characteristics of the glottal source signal.

In this thesis, an acoustic glottal source model, the *Liljencrants-Fant (LF) model* (Fant et al., 1985), is used to model the periodic component of the excitation in HMM-based speech synthesis. This glottal source model is described in the next chapter. One advantage of using the LF-model when compared with the source model represented by a real pulse is that it permits the correlation between F_0 and the other glottal parameters to be modelled, as the F_0 parameter is described by this model. Another advantage is that the LF-model parameters are strongly correlated with voice quality and they can be controlled for achieving voice transformation.

Chapter 5

Acoustic Glottal Source Model

5.1 Introduction

A wide variety of models have been proposed in the literature to represent the glottal source signal. For example, the most commonly used types of glottal source model were described in Section 2.2.2, such as physical, acoustic, and pole-zero models. In general, acoustic glottal source models use mathematical functions to represent the curves of the glottal source waveform. Typically, the parameters of these models describe acoustic properties of the source, e.g. the instant and amplitude of the glottal pulse peak. Acoustic models of the glottal source derivative are usually preferred over models of the glottal flow signal because they better describe relevant voice source characteristics, such as how rapid the vocal folds close.

This chapter describes the Liljencrants-Fant (LF) model (Fant et al., 1985), which is a popular acoustic model of the *glottal source derivative* signal. The LF-model is defined by a small set of parameters, including the fundamental period T_0. One advantage of using this model in HMM-speech synthesis is the possibility to model the correlation between the glottal parameters of the model and T_0. Another important aspect of this model is the correlation between its parameters and voice quality.

5.2 LF-model

5.2.1 Waveform

Figure 5.1 shows a segment of the glottal flow derivative, $e_{LF}(t)$, and the corresponding glottal flow waveform, $u_{LF}(t)$, which were obtained using the LF-model. The signal

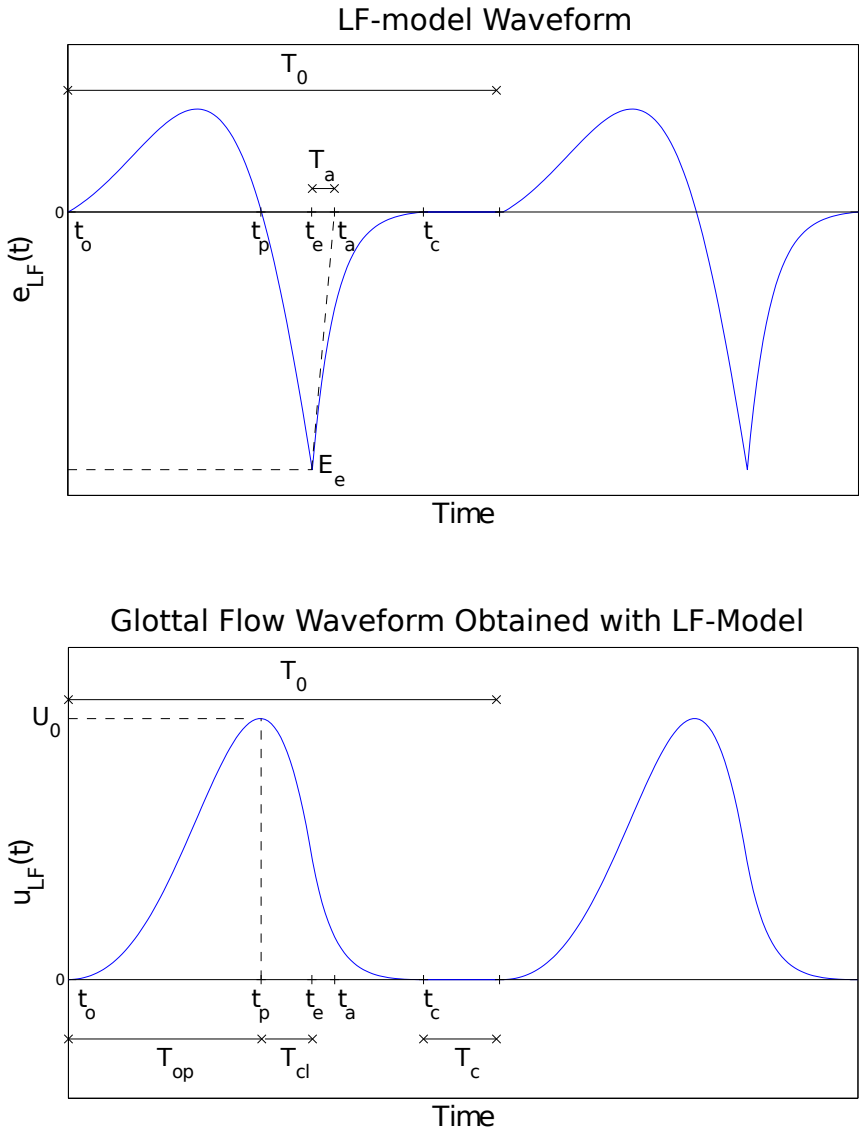

Figure 5.1: Top: segment of the LF-model waveform and representation of the glottal parameters during one fundamental period of the model. Bottom: segment of the glottal flow calculated by integration of the LF-model.

$u_{LF}(t)$ was calculated by integrating $e_{LF}(t)$. Analytically, the LF-model is defined by an exponentially increasing sine wave, followed by a decaying exponential function, and completed with a zero amplitude section, as described by the following equations:

$$e_{LF}(t) = \begin{cases} e_1(t) = E_0 e^{\alpha t} \sin(w_g t), & t_o \leq t \leq t_e \\ e_2(t) = -\frac{E_e}{\varepsilon T_a}[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}], & t_e < t \leq t_c \\ e_3(t) = 0, & t_c < t \leq T_0 \end{cases} \tag{5.1}$$

$$\int_0^{T_0} e_{LF}(t)dt = 0 \tag{5.2}$$

$$e_1(t_e) = e_2(t_e) = -E_e, \tag{5.3}$$

where $w_g = \pi/t_p$. Equations 5.2 and 5.3 represent the zero energy balance and amplitude continuity constraints, respectively. The value of the parameter t_o is arbitrary, as it represents the start of the LF-model. In this work, t_o is assumed to be zero and it is omitted in the formulas that describe the LF-model. In general, the parameters α, E_0 and ε are derived from (5.2) and (5.3). Therefore, the LF-model given by (5.1) can be defined by the six parameters: t_p, t_e, T_a, t_c, T_0, and E_e. Figure 5.1 represents these parameters for a cycle of the source model.

The LF-model parameters represent the following characteristics of the flow derivative waveform:

- t_o: instant of glottal opening, when the vocal folds start to open.
- t_p: instant of maximum flow, which corresponds to a zero of the flow derivative.
- t_e: instant of maximum excitation, when the vocal folds close abruptly.
- T_a: duration between t_e and t_a (t_a is the point where the tangent to the decaying exponential at $t = t_e$ hits the time axis).
- t_c: instant of complete closure of the vocal folds.
- T_0: duration of the glottal flow cycle (fundamental period).
- E_0: amplitude scaling of the sine wave.
- E_e: amplitude of maximum excitation.
- w_g: angular frequency of the sine wave, which is related to the rise time of the glottal flow.
- α: growth factor, which represents the ratio of E_e to the peak height of the exponentially increasing sine wave, E_i.
- ε: exponential time constant.

The region between the start of the glottal pulse and the instant of maximum airflow, is called the *opening phase* and has duration $T_{op} = t_p - t_o$. At t_p, the vocal folds start to close and the flow amplitude decreases until the abrupt closure of the glottis (discontinuity in the derivative of the LF-model) at the instant of *maximum excitation*, t_e. The time interval $T_{cl} = t_e - t_p$ is the duration of this *closing phase*. The time interval which corresponds to the duration when the vocal folds are opened and there is airflow through the glottis (duration equal to $T_{op} + T_{cl}$) is called the *open phase*. The next part represents the transition between the open phase and the closed phase, which is called *return phase* (the return phase is often assumed to be a part of the open phase). The duration of the return phase is given by $T_a = t_a - t_e$ and it measures the abruptness of the closure. Finally, the *closed phase* is the region of the glottal cycle when the vocal folds are completely closed and it has duration $T_c = T_0 - t_c$.

The decaying exponential function, given by $e_2(t)$ in (5.1), represents the return phase. Often, this expression is used to represent both the return phase and the closed phase. This simplification avoids the calculation of the parameter t_c, by making $t_c = T_0$, as suggested by Fant (1997). In general, this is a good approximation because $e_2(t)$ is close to zero for $t > t_c$. By using this approximation, (5.1) can be reduced to the terms $e_1(t)$ and $e_2(t)$ as follows:

$$e_{LF}(t) = \begin{cases} e_1(t) = E_0 e^{\alpha t} \sin(w_g t), & t_o \leq t \leq t_e \\ e_2(t) = -\frac{E_e}{\varepsilon T_a}[e^{-\varepsilon(t-t_e)} - e^{-\varepsilon(t_c-t_e)}], & t_e < t \leq T_0 \end{cases} \tag{5.4}$$

The LF-model parameters must satisfy physical constraints because they have an acoustic meaning (e.g. the time parameters must be positive) and there are parameter settings which produce a distorted flow derivative waveform. The following parameter limits are based on the LF-model parameter ranges reported by Doval and d'Alessandro (1997):

- $E_e > 0$
- $T_0 > 0$
- $0 < T_a \leq T_0 - t_e$
- $0 < t_e \leq 3t_p/2$ and $t_e \leq T_0$
- $0 < t_p \leq t_e$

The constraint $t_e \leq 3/2t_p$ ensures that the negative maximum of the flow derivative is E_e and the condition $T_a \leq T_0 - t_e$ ensures that the return phase is a decreasing exponential.

5.2.2 Parameter Calculation

The LF-model in (5.4) is defined by four time-parameters: t_p, t_e, T_a, and T_0. In addition, one of the two amplitude parameters needs to be given: E_e or E_0. Typically, E_e is chosen as the waveform parameter and (5.4) to (5.3) are solved for E_0, e.g. Fant et al. (1985). The angular frequency can be calculated directly from $w_g = \pi t_p$. The remaining parameters (ε and α, and E_0) are obtained using the energy and continuity constraints of (5.2) and (5.3).

5.2.2.1 Calculation of ε

The parameter ε is calculated by solving the equation below.

$$\varepsilon t_a = 1 - e^{-\varepsilon}(t_c - t_e), \tag{5.5}$$

which results from imposing the continuity constraint $e_2(t_e) = -E_e$ on (5.4).

5.2.2.2 Calculation of α and E_0

The values of α and E_0 can be calculated by solving the following equations:

$$\int_0^{t_e} e_1(t)dt + \int_{t_e}^{T_0} e_2(t)dt = 0 \tag{5.6}$$

$$E_e = -E_0 e^{\alpha t_e} \sin(w_g t_e) \tag{5.7}$$

Equation (5.6) is equivalent to (5.2), which represents the assumption that the energy balance of the glottal flow derivative is zero over the fundamental period. On the other hand, (5.7) is obtained from $e_1(t_e) = -E_e$.

The first integral in (5.6) is obtained using the indefinite integral of $e_1(t)$, which is given by

$$U_1(t) = \frac{E_0 e^{\alpha t}\left(\alpha \sin(w_g t) - w_g \cos(w_g t)\right) + w_g}{\alpha^2 + w_g^2} \tag{5.8}$$

Fant et al. (1985) proposed the following approximation to calculate the second integral in (5.6):

$$U_2(t) = \int_{t_e}^{T_0} e_2(t)dt \cong \frac{E_e t_a}{2} k_a, \tag{5.9}$$

where

$$K_a = \begin{cases} 2, & R_a < 0.1 \\ K_a = 2 - 2.34R_a^2 + 1.34R_a^4, & 0.1 \le R_a < 0.5 \\ K_a = 2.16 - 1.32R_a + 0.64(R_a - 0.5)^2 & 0.5 \le R_a \end{cases}, \tag{5.10}$$

with $R_a = t_a/(T_0 - t_e)$.

Then, the parameter α is the root of the following non-linear equation, which is obtained from (5.6) to (5.9).

$$\frac{e^{\alpha t_e}(\alpha \sin(w_g t_e) - w_g \cos(w_g t_e)) + w_g}{\alpha^2 + w_g^2} \cong \frac{-e^{\alpha t_e} \sin(w_g t_e t_a)}{2} K_a \tag{5.11}$$

After α is calculated, the scale factor E_0 can be determined from (5.7).

5.2.3 Dimensionless Parameters

The parameters of the LF-model can also be expressed as dimensionless quotients, which are often used to describe the shape of the glottal source signal. The following dimensionless parameters are based on the ratios of the glottal time intervals described in Section 5.2.1:

- *Open quotient*, which measures the relative duration of the open phase:

$$OQ = \frac{T_o + T_a}{T_0} = \frac{t_e + T_a}{T_0} \tag{5.12}$$

- *Speed quotient*, which measures the ratio between the opening and closing times:

$$SQ = \frac{T_{op}}{T_{cl}} = \frac{t_p}{t_e - t_p} \tag{5.13}$$

- *Return quotient*, which measures the relative duration of the return phase:

$$RQ = \frac{T_a}{T_0} = \frac{t_a - t_e}{T_0} \tag{5.14}$$

Amplitude ratios have also been used to describe the glottal pulse waveform. For example, the *amplitude quotient* is the ratio between the amplitude of the glottal flow peak, U_0, and the amplitude of maximum excitation, E_e:

$$AQ = \frac{U_0}{E_e} \tag{5.15}$$

Variations of the dimensionless parameters and additional parameters can also be found in the literature. They represent different quotients, which describe specific properties of the source signal. For example, the open quotient can also be defined by the reduced form $OQ_e = T_e/T_0$, in which the return phase is not included in the open phase. Often, the *closing quotient*, $CQ = \frac{t_c - t_p}{T_0}$, and the *opening quotient*, $OQ_{op} = \frac{t_p}{T_0}$, are used to describe the relative duration of the closed phase and the opening phase T_{op}, respectively.

Fant (1995) has also derived a set of dimensionless parameters which are more correlated with the relevant waveshape characteristics of the LF-model (such as glottal pulse asymmetry), than the time instants t_a, t_p, t_e and t_c. These dimensionless parameters are called the *R-parameters*:

$$R_g = \frac{T_0}{2T_{op}} = \frac{T_0}{2t_p} \tag{5.16}$$

$$R_k = \frac{t_e - t_p}{t_p} \tag{5.17}$$

$$R_a = \frac{T_a}{T_0} \tag{5.18}$$

The R-parameters are comparable to the parameters given by (5.12) to (5.14). In particular, RQ and R_a are equivalent. SQ is the inverse of R_k but they both represent the skewness of the glottal waveform. Fant (1995) also related the OQ with the R-parameters by the formula:

$$OQ = \frac{1 + R_k}{2R_g} + R_a \tag{5.19}$$

In addition, Fant (1995) also proposed the R_d parameter, which is closely related to the amplitude quotient:

$$R_d = \frac{U_0}{E_e}\frac{F_0}{110} \tag{5.20}$$

The R_d parameter can be used to control R_g, R_k, and R_a by using the following approximation:

$$R_d = (1/0.11)(0.5 + 1.2R_k)(R_k/4R_g + R_a) \tag{5.21}$$

Fant (1995) estimated this equation from the geometrical constraints of the LF-model. He reported that this approximation holds with an accuracy of 0.5 dB for $R_d < 1.4$ and with a maximum error of 1.7 dB at $R_d = 2.7$. An interesting property of R_d is that increasing values of this parameter result in increasing values of the OQ parameter.

5.2.4 Spectral Representation

The spectrum of the LF-model is characterised by a spectral peak at the lower frequencies, often called the "*glottal formant*", and the *spectral tilt* (attenuation at higher frequencies). Figure 5.2 shows the stylised spectrum of the LF-model proposed by Doval and d'Alessandro (1997). In this figure the spectral peak is centered at the frequency F_g and the spectral tilt is characterised by the attenuation above the frequency F_c.

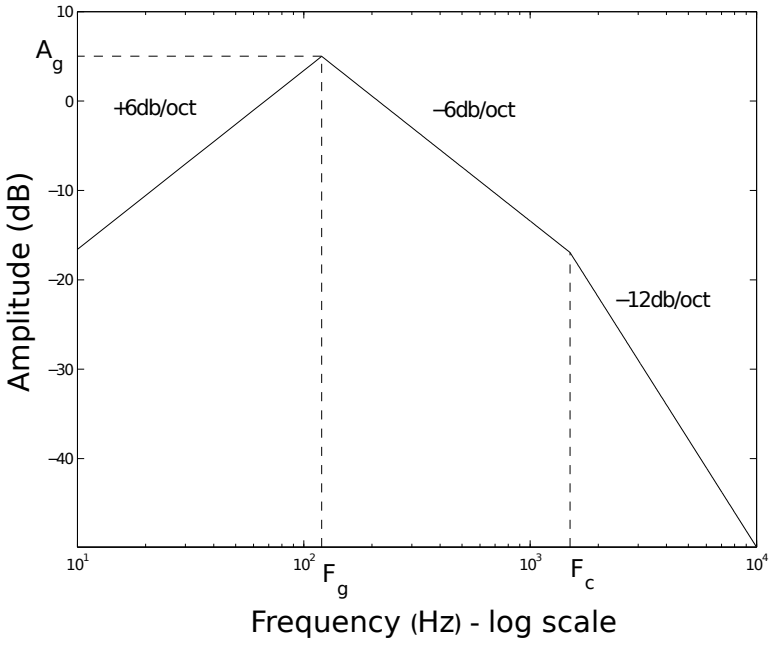

Figure 5.2: Linear stylization of the LF-model spectrum.

5.2.4.1 Glottal Formant

Typically, the spectrum of the glottal flow signal $u(t)$ is represented by two identical poles in the real axis (Doval and d'Alessandro, 1999). This model represents the effect of the glottal formant. It and can be described in the frequency domain by

$$U_g(s) = \frac{U_0}{(1+s/s_r)^2}, \tag{5.22}$$

where $U_g(s)$ represents the Laplace transform of $u(t)$, s_r is a real pole and G_0 is a gain factor. This spectral representation corresponds to a first order low-pass system with cut-off frequency $F_g = s_r/2\pi$ and gain U_0. The transfer function of the filter is defined by two asymptotic lines, with slopes of 0 dB/oct and -12 dB/oct for frequencies below and above the cut-off frequency, respectively.

The spectrum of the glottal flow derivative can be obtained by adding a zero to $U_g(s)$ at $f = 0$:

$$E_g(s) = U_g'(s) = \frac{sU_0}{(1+s/s_r)^2} + u(0) \tag{5.23}$$

Assuming that the glottal source waveform starts at the instant of glottal opening $t_o = 0$, then $u(0) = 0$ in (5.23), as in the LF-model ($u_{LF}(0) = 0$). The effect of adding the zero to $U_g(s)$ is to produce two asymptotic lines with slopes +6 dB/oct and -6 dB/oct, which are represented by the stylised spectrum of the LF-model shown in Figure 5.2. The crossing point of these lines is a spectral peak which is located at F_g. This frequency is equal to the cut-off frequency of $U_g(s)$. However, the asymptotic behavior of the spectral peak is equivalent to a second order linear filter, instead of the first-order low-pass filter of $U_g(s)$. For this reason, the spectral peak is often called the "glottal formant".

Doval and d'Alessandro (1999) showed that the spectrum of the time-domain glottal flow derivative models are generally characterised by the "*glottal formant*", although they are modelled by different equations. For example, the KLGLOTT88 model used by the Klatt speech synthesiser (Klatt and Klatt, 1987) has the same spectral characteristics as $E_g(s)$ but it can be represented by a 3-order low-pass filter, with a double real pole and a simple pole (Doval and d'Alessandro, 1997).

The spectrum $U_g(s)$ in (5.22) is characteristic of the glottal flow with an abrupt glottal closure that corresponds to the truncation of the waveform at the instant of maximum excitation. In the case of the LF-model, this is equivalent to setting the

duration of the return phase to zero ($T_a = 0$). When T_a is positive, the contribution of this parameter to the spectral tilt must be considered.

5.2.4.2 Spectral Tilt

Several glottal source models have a return phase component to simulate a smooth closure of the vocal folds, e.g. the LF-model and the KLGLOTT88 model of Klatt and Klatt (1987). Typically, this is an additional low-pass filter with order one or two. The return phase of the LF-model acts as a low-pass filter of order one. A first order filter with cut-off frequency F_c can be represented by the following transfer function:

$$H(s) = \frac{1}{1 + \frac{s}{2\pi F_c}} \tag{5.24}$$

This first order low-pass filter contributes to the spectral tilt with an additional -6 db/oct for frequencies above F_c.

The spectral representation of the LF-model is obtained by combining the glottal formant with the spectral tilt effects: $E_{LF}(s) = E_g(s)H(s)$. This spectrum is stylized into three lines with +6 db/oct, -6 db/oct and -12 db/oct slopes, respectively, as shown in Figure 5.2.

5.2.4.3 Spectral Parameters

Doval and d'Alessandro (1997) defined the general spectrum of the glottal flow derivative using five parameters:

- A_g: maximum amplitude of the *glottal spectral peak*.
- F_0: fundamental frequency.
- F_g: glottal spectral peak.
- Q_g: *quality factor* of the glottal spectral peak.
- F_c: spectral tilt cut-off frequency.

The parameters A_g, F_g, and F_c are represented in Figure 5.2. The quality factor Q_g is a characteristic of the second order low-pass filter associated with the glottal formant. Basically, this parameter measures the difference in dB between the maximum of the spectrum and the amplitude A_g. Doval and d'Alessandro (1997) also indicate that the variation of the quality factor mainly affects the amplitude of the first harmonics,

with the glottal formant frequency F_g and the asymptotes remaining approximately unchanged.

5.2.5 Phase Spectrum

5.2.5.1 Filter Transfer Function

Doval et al. (2003) proposed to describe the LF-model as the impulse response of an *anticausal filter* and a *causal filter*. They also showed that this representation is compatible with the time and spectral characteristics of the glottal source derivative, which were described in the previous sections.

In general, the glottal pulse is skewed to the right, which can be observed in the example of Figure 5.1. This time domain behavior is the evidence of anticausality.

For the LF-model, the open phase (defined as the duration until the instant of maximum excitation t_e) has the characteristics of a second-order anticausal filter and the return phase can be described as the impulse response of a first-order causal filter (Bozkurt, 2005; Doval et al., 2003). Under this assumption, Doval et al. (2003) defined the LF-model as the impulse response of a linear all-pole filter that has two anticausal poles to represent the glottal formant, one causal pole for the spectral tilt and a zero to get the glottal flow derivative. For this filter to be stable, the anticausal poles must be outside the unit circle and the causal pole inside the unit circle on the z-plane. The z-transform of this transfer function can be represented by:

$$H_{LF}(z) = \frac{Gz}{(1+a_1 z+a_2 z^2)(1-a_{TL} z^{-1})}, \tag{5.25}$$

where a_1 and a_2 represent the anticausal poles, a_{TL} is the causal pole, and G is the filter gain. Doval et al. (2003) also derived formulas to calculate the coefficients of this filter from the parameters of the LF-model: OQ, SQ, and E_e. Finally, Doval et al. (2003) suggested to model the truncation of the open phase with the return phase by the convolution with a *sine cardinal function* in the frequency domain. The spectral effect of this operation is to enlarge the glottal formant and create ripples.

5.2.5.2 Mixed-phase Model

In general, the source-filter models used in HMM-based speech synthesis are *minimum-phase*. Basically, the minimum-phase speech model is all-pole, in which the poles are causal and stable (inside the unit-circle in the z-plane). For the case of the simple

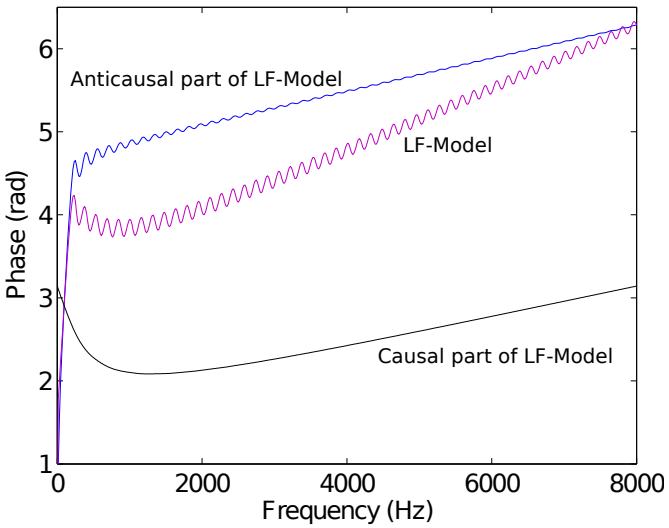

Figure 5.3: Phase spectra of the anticausal component of the LF-model signal, causal component, and LF-model signal.

excitation of voiced speech (impulse train), the speech signal is simply the impulse response of a minimum-phase filter. The transfer function of this filter is linear-phase and represents only the magnitude spectrum of the speech signal.

The importance of the phase information for speech quality and to model the voice characteristics of the speaker has been demonstrated by several papers in the past, e.g. Quatieri (1979); Murthy et al. (2004). Recent work by Gardner (1994) and Bozkurt (2005) suggested that a *mixed-phase* model of voiced speech (when a minimum-phase system is excited by a *maximum-phase* signal) is more appropriate than the minimum-phase model due to the maximum-phase characteristic (anti-causality) of the source signal.

The LF-model is a mixed-phase signal (has both causal and anticausal properties) and it can be represented by a stable all-pole linear filter, as explained in the previous section. Therefore, the convolution of the LF-model with the minimum-phase filter of the vocal tract produces a mixed-phase speech signal. This source-filter model of speech is expected to give a better representation of the phase spectrum, when compared with the traditional impulse response of the minimum-phase filter (represents the spectral envelope).

Figure 5.3 shows the phase spectra associated with the anticausal component of the LF-model (exponentially increasing sine wave, which represents the open phase), the causal component (decaying exponential, which represents the return phase), and the combination of the two components (phase spectrum of the LF-model signal).

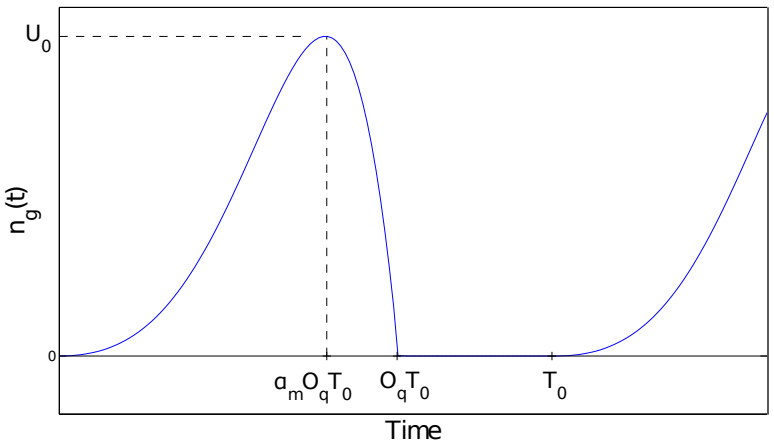

Figure 5.4: General form of the glottal flow pulse obtained with the shape and scale parameters of Doval and d'Alessandro (1999) for the LF-model. The return phase component of the LF-model is not considered in this representation (T_a is set equal to zero).

5.3 LF-model Correlates

5.3.1 Spectrum

5.3.1.1 Scale and Shape Parameters of the Glottal Waveform

Doval and d'Alessandro (1999) showed that the five spectral parameters of the glottal flow derivative spectrum (A_g, F_0, F_g, Q_g and F_c) were correlated with a set of five time-domain parameters of the glottal flow waveform:

- U_0: maximum amplitude of the glottal flow.
- T_0: fundamental period.
- OQ_e: open quotient, calculated without considering the return phase as part of the open phase ($OQ_e = t_e/T_0$).
- α_m: asymmetry coefficient, which is the ratio between the glottal opening duration, T_{op}, and the effective duration of the open phase, O_qT_0.
- T_a: return phase time constant.

Figure 5.4 shows the representation of these parameters for a cycle of the LF-model waveform. The time-domain parameters used by Doval and d'Alessandro (1999) are related to the LF-model parameters by the formulas: $\alpha_m = t_p/t_e = SQ/(1+SQ)$ and $OQ_e = t_e/T_0$.

The parameters U_0, T_0, and OQ_e are scale parameters. These parameters have the same effect on the spectrum of the glottal flow model, independently of the mathematical expressions used by each time-domain model. The spectral effects described by Doval and d'Alessandro (1999) for these parameters are:

- U_0: amplitude scaling of the glottal flow, which changes the spectral gain by the same proportion.
- T_0: scales the spectrum in the opposite direction. For example, depending on whether T_0 increases or decreases, the spectrum of the glottal flow signal is contracted or expanded by the same amount respectively.
- OQ_e: scales the spectral envelope in the opposite direction.

The parameter α_m is related to the specific shape characteristics of each glottal source model and affects mainly the characteristics of the glottal formant.

Doval and d'Alessandro (1999) characterised different glottal flow models by their normalised glottal flow waveform $n_g(t)$, which is obtained by setting $A_v = 1$, $T_0 = 1$, and $OQ_e = 1$. This waveform depends on the shape parameter α_m only. For example, Doval and d'Alessandro (1999) calculated the following expression for the normalised glottal flow of the LF-model:

$$n_g(t) = \frac{1 + e^{at}\left(a\frac{\alpha_m}{\pi}\sin(\pi t/\alpha_m) - \cos(\pi t/\alpha_m)\right)}{1 + e^{a}\alpha_m}, \tag{5.26}$$

where a is a parameter equivalent to the parameter α of the LF-model, which can be obtained from the implicit equation $n_g(1) = 0$ (energy balance condition).

Finally, Doval and d'Alessandro (1999) derived the following formulas that correlate the frequency parameters with the scale and shape parameters of the glottal flow waveform:

$$F_0 = 1/T_0 \tag{5.27}$$

$$A_g = E_0\sqrt{e_n(\alpha_m)i_n(\alpha_m)} \tag{5.28}$$

$$F_g = \frac{1}{2\pi OQ_e T_0}\sqrt{\frac{e_n(\alpha_m)}{i_n(\alpha_m)}} \tag{5.29}$$

$$Q_g = q_g(\alpha_m), \tag{5.30}$$

where $i_n(\alpha_m)$, $e_n(\alpha_m)$ and $q_g(\alpha_m)$ are functions of the asymmetry coefficient α_m. The parameter e_n represents the maximum excitation of the normalised glottal flow $n_g(t)$, while i_n is the integral of $n_g(t)$.

5.3.1.2 Glottal Formant

The glottal spectral peak or glottal formant depends mainly on the open quotient and the asymmetry coefficient. From (5.29), the open quotient is inversely proportional to the glottal formant. On the other hand, the asymmetry of the glottal waveform, which is quantified by α_m or SQ, is assumed to be directly proportional to the bandwidth of the glottal formant (d'Alessandro et al., 2006).

In general, the glottal peak affects the source spectrum in the low to mid-frequency range. For example, d'Alessandro et al. (2006) indicated that "a typical value of the asymmetry coefficient (2/3) and for normal values of the open quotient (between 0.5 and 1), the glottal formant is located slightly below or close to the first harmonic". They also suggested that for both lower values of the open quotient OQ and higher asymmetry coefficients, the glottal formant can reach higher order harmonics such as the fourth.

5.3.1.3 Spectral Tilt

The main spectral effect of the return phase is to change the cut-off frequency F_c of the low-pass filter associated with the spectral tilt. This frequency depends on the expression used to represent the return phase by the glottal flow model. A typical impulse response of this filter is a decreasing exponential with time constant T_a. The cut-off frequency of this filter is:

$$F_c = \frac{1}{2\pi T_a} \tag{5.31}$$

Doval and d'Alessandro (1997) calculated analytically the following expression for the F_c of the LF-model:

$$F_c = \frac{1}{2\pi T_a} + \frac{a}{2\pi} + \frac{1}{t_p}\cot\left(\pi\left(1+\frac{t_e - t_p}{t_p}\right)\right), \tag{5.32}$$

where a is the same as in (5.26). However, the F_c parameter of the LF-model mostly depends on the return phase parameter T_a and it is often approximated by the simpler expression given by (5.31), e.g. Fant et al. (1985).

5.3.1.4 Dimensionless Parameters

The dimensionless parameters of the LF-model (*OQ*, *SQ*, and *RQ*) are directly related to the glottal formant and the cut-off frequency of the spectral tilt filter. They can also be characterised by their effect on the overall spectrum of the glottal source. According to Fant (1995) and Doval et al. (2003), the spectral correlates of the LF-model parameters can be described by:.

- Open quotient (*OQ*): the main spectral effect is to shift the energy between the lower frequency and the higher frequency harmonics. An increase of the *OQ* has the spectral effect of expanding the frequency scale (equivalent to a frequency-scale operation) and shifting the energy from the lower to the higher frequencies. In the other way, a decrease of the *OQ* compresses the frequency scale and moves the energy from the higher to the lower frequencies. This parameter also affects the amplitude of the first harmonics in the voice source spectrum. An increased value of the *OQ* is correlated with an increase in the amplitude of the lower harmonics.

- Speech quotient (*SQ*): mainly affects the amplitude of the first harmonics. In general, increased *SQ* (asymmetry of the glottal pulse) results in increased amplitude of the lower frequency harmonics and a deepening of the spectral dips.

- Return quotient (*RQ*): the major effect is to change the spectral amplitudes at higher frequencies. The smaller the *RQ*, the more the energy in the higher frequency part of the spectrum.

5.3.1.5 Spectrum Measurements

In addition to the time- and frequency-domain parameters of the LF-model, different types of spectral parameters have also been used to represent the glottal source spectrum. In general, they are measurements of the spectral tilt and measurements of the relationship between the intensity of the fundamental frequency and its harmonics, e.g. Childers and Lee (1991); Hanson and Chuang (1999); Gobl (1989). This section describes the relevant source spectrum aspects found in the literature which are correlated with the LF-model.

One of the most perceptually important spectral measures which appears to be correlated with the LF-model parameters is the ratio between the amplitude of the first harmonic, H_1^*, and the second harmonic, H_2^*, of the source spectrum. The notation

H_1^* is used to distinguish this parameter from H_1, which is often used to represent the amplitude of the first harmonic in the speech spectrum. Typically, the amplitudes H_1^* and H_2^* are obtained by removing the spectral contribution of the vocal tract transfer function. For example, Fant (1995) used formulas to estimate and remove the spectral influence in the low frequency range of the first and second formant amplitudes. Fant (1995) derived the following numerical expression of the logarithmic ratios (in dB) as the result of a regression analysis of the R_d parameter:

$$H_1^* - H_2^* = -7.6 + 11.1R_d \tag{5.33}$$

From the experimental results of Fant (1995), this equation is a good approximation for $0.3 < R_d < 2.7$ and typical values of the LF-model parameters. Both R_d and OQ are related through (5.19) and (5.21). As expected, Fant (1995) also obtained a correlation between the open quotient and the amplitude ratio, which is given by:

$$H_1^* - H_2^* = -6 + 0.27\exp(0.055OQ_e), \tag{5.34}$$

where OQ_e is defined without the return phase, i.e. $OQ_e = t_e/T_0$. This equation is valid within 0.5 dB in the range $0.3 < OQ_e < 0.7$.

The glottal formant depends mainly on the OQ parameter and the asymmetry coefficient, $\alpha_m = SQ/(1+SQ)$, e.g. d'Alessandro et al. (2006). If the glottal formant is near F_0 it will change mainly the relative amplitudes of the first harmonics. Doval and d'Alessandro (1997) proposed another expression for the amplitude ratio of the first two harmonics, which depends on the OQ and the $SQ = 1/R_k$. It is given by the following expression for 1 dB approximation and common parameters ranges ($0.3 < R_k < 0.6$ and $1 < R_g < 1.3$):

$$H_1^* - H_2^* = 12\left(\frac{O_q}{0.7}^2\right)\left(1 - \left(1 - \frac{R_k}{0.7}^2\right)\right) - 6 \tag{5.35}$$

Fant (1995) derived the following equation for the ratio between the amplitude of the first harmonic and the harmonics of order n, H_n^*, of the glottal source derivative spectrum (for n well above one):

$$\frac{H_1^*}{H_n^*} = \frac{U_0 F_0}{E_e} k\pi^2 \sqrt{1 + \frac{f^2}{F_a^2}}, \tag{5.36}$$

where the parameters U_0 and E_e are the amplitude of the glottal pulse and the amplitude of maximum excitation, respectively. The constant k is close to one for normal

phonation (OQ around 0.5) but can reach values as small as 0.5 for $OQ = 0.35$ and R_d=0.3. Equation (5.36) was obtained using formulas of H_1^* and H_n^* respectively. The formula of H_1^* was derived from measurements of the radiation effect of a recorded voiced sound and from knowledge of the glottal flow and F_0 data (Fant and Lin, 1988). The second was obtained using the spectral representation of the LF-model (-6 dB/oct slope above F_g and additional -6 dB/oct above F_a). Equation (5.36) also indicates that U_0 is proportional to the level of the voice fundamental, H_1^*, and that E_e is proportional to the amplitude of the harmonics at higher frequencies than the fundamental.

The spectral tilt is often measured by the ratio of the amplitudes of the first harmonic and a formant of higher order than one. For example, the amplitude ratio between the first harmonic and the third formant has been used to measure the spectral tilt by Hanson and Chuang (1999). The ratio H_1^*/H_n^* can also be used to measure the spectral tilt, e.g. by choosing H_n^* close to the third formant. The effect of the return phase parameter $R_a = 1/T_a$ on the spectral tilt is also represented in (5.36).

Fant (1995, 1997); Stevens (1998) suggested that the parameters E_e and T_a are also correlated with the bandwidth of the formants. For example, Fant (1995) obtained the following empirical formulas from measurements on the estimated glottal source signal:

$$\Delta B_1 = 250 \left(\frac{F_1}{500} \right)^2 \frac{R_a}{12} \tag{5.37}$$

$$\Delta B_2 = \frac{\Delta B_1}{2} R_a \frac{F_1}{F_2}, \tag{5.38}$$

where ΔB_1 and ΔB_2 are the bandwidth variations of the first and second formants, ΔB_1 and ΔB_2 respectively. F_1 and F_2 represent the first and second formant frequencies respectively.

5.3.2 Voice Quality

The shape parameters of the LF-model, which were described in Section 5.2.3, have been widely used to study the voice quality of speech signals because they are strongly correlated with the type of phonation, e.g. Fant (1995); Keller (2005). The phonetic properties of these parameters are summarised below:

- Open quotient (OQ): the relative duration of the glottal pulse to T_0 is mainly related to the level of vocal folds *abduction/adduction* and the *pressed-lax* dimension of the glottis. Increased degrees of the OQ are associated with wider

opening of the glottis (when the vocal folds are more abducted) and lower tension in the glottis.

- Speed quotient (*SQ*): the asymmetry of the glottal pulse is affected by both the pressed-lax and *vocal effort* dimensions. In general, the skewness of the glottal pulse (higher *SQ*) increases with the tension of the vocal folds and with the vocal effort (voice loudness).

- Return quotient (*RQ*): the abruptness of the glottal closure is mainly related to the vocal effort dimension. A louder voice is typically associated with a longer return phase (larger return duration T_a) and higher spectral tilt. When the loudness is lower, the glottal closure tends to be more abrupt, resulting in a lower attenuation of the higher frequency region of the source spectrum.

This section reviews mainly the voice quality correlates of the *OQ*, *SQ*, and the *RQ* parameters because they are considered to be the most important LF-model parameters related to phonation type and they are used to synthesise speech with different voice qualities in this work. However, other acoustic correlates of voice quality can be found in the literature. These include amplitude based parameters, such as the amplitude of maximum excitation of the LF-model, E_e, the maximal amplitude of the glottal pulse, U_0, or the *peak-to-peak ratio*, U_0/E_e. Often, frequency-domain parameters are also used to study the type of voice. In general, they measure the variations in the spectral amplitude at the frequencies of the first harmonics, the overall spectral slope of the source spectrum, and the *harmonic-to-noise ratio*, e.g. Childers and Lee (1991); Hanson and Chuang (1999).

The research on acoustic correlates of voice quality is typically limited to a small group of "major" voice types. For example, Gobl (1989) studies modal, breathy, whispery and creaky voices. The following relations between the acoustic parameters of the LF-model and four major types of voice quality are based on the papers by Childers and Ahn (1995); Gobl (1989); Fant (1995); Keller (2005); Alku et al. (1997).

- Breathy: high symmetry of the glottal pulse that corresponds to a small *SQ*. There is a general lack of tension of the vocal folds and highly abducted phonation, which results in a high *OQ*. Typically, the vocal folds do not close completely, which is associated with a slow glottal closure (high *RQ*). The incomplete glottis closure also creates the effect of *glottal leakage* and the production

of *aspiration noise*. Also, the air flows through the glottis at a high rate when the vocal folds are widely opened which causes additional *turbulent noise*.

- Whispery: small OQ and RQ as a consequence of low adductive tension. This voice type mainly differs from the breathy voice by its lower OQ and higher skewness of the pulse (high SQ), due to a very small glottal opening. Audible *frication noise* is also a characteristic of whispery speech.

- Tense: very adducted phonation (short glottal open interval), with a small OQ and low RQ (short return phase). The asymmetry of the glottal pulse is large (as well as the SQ) as an effect of the increased vocal folds tension when compared with the modal voice (neutral voice quality). The lax voice quality has the opposite effect on these voice quality parameters of the LF-model.

- Creaky: similar characteristics to the tense voice, with high adduction of the vocal folds and high asymmetry of the glottal pulse. Therefore, the voice quality parameters show a similar behavior as those of the tense voice: small OQ, small RQ, and high SQ. This voice type is also characterised by the *diplophony* effect (two pulses appear during one fundamental period), in which two different pulses appear to occur within one glottal pulse cycle.

	RQ and R_a	SQ and $1/R_k$	OQ_e	$1/R_g$
Breathy	High	Low	Very High	Low
Whispery	Very High	High	High	High
Tense	Low	High	Low	Low
Creaky	Low	High	Low	Low

Table 5.1: Summary of the relations between the dimensionless parameters of the LF-model and four key voice qualities, obtained from the literature.

Table 5.1 summarises the voice quality correlates of the LF-model parameters, taking as reference the modal voice (normal phonation). In this table, the open quotient is defined without the return phase ($OQ_e = T_e/T_0$), because OQ_e appeared to be more commonly used for studying voice quality correlates than OQ, from the papers we

found. Nevertheless, the OQ parameter (defined with the return phase) has a similar behavior to the OQ_e parameter for these voice qualities, according to the results obtained by Childers and Lee (1991); Karlsson and Liljencrants (1996); Alku et al. (1997).

The OQ_e, SQ, and RQ parameters are closely related to the R-parameters with: $RQ = R_a$, $SQ = 1/R_k$, and $OQ_e = (1+R_k)/(2R_g)$. Therefore, the R-parameters have similar correlates to the different voice qualities, as shown in Table 5.1. The R_d parameter increases with OQ and decreases with SQ, according to (5.21). Fant (1997) suggested that this property was perceptually important to describe a range of voice qualities, from a tense male voice with low R_d (low OQ and high SQ) to a breathy voice with high R_d (high OQ and small SQ).

There are voice qualities which appear to be acoustically similar. For example, the patterns of the acoustic parameters for the tense and creaky voices are the same in Table 5.1. This could be a limitation of the LF-model parameters to model certain acoustic properties which are important to differentiate the two voice qualities. For example, the LF-model does not model aspiration noise and *diplophony*, which are distinguishable characteristics of a creaky voice when compared with a tense voice. Nevertheless, effects such as the aspiration noise have been successfully modelled by adding pitch-synchronously amplitude modulated noise to the LF-model signal, e.g. Gobl (2006).

Table 5.1 was derived from voice quality correlates reported in the literature. In general, these studies calculated averages of LF-model parameter estimates over different vowels, for each voice quality. In this type of analysis, the phonetic context and the dynamics of the parameters is not taken into account. For example, Ní Chasaide and Gobl (1993); Tooher and McKenna (2003) observed that voice quality varied along a vowel and is affected by the preceding phone. This type of voice quality variation is associated with aspects of prosody of which an overview is given in the next section.

5.3.3 Prosody

In general, there is a correlation between the LF-model parameters and $F_0 = 1/T_0$. In particular, if the main voice quality parameters of the LF-model (OQ, SQ, and RQ) are assumed to be constant along a speech segment, the parameters t_a, t_p, t_e, and t_c should vary by the same proportion as T_0. However, this is not the case for most of the time. For example, Strik and Boves (1992); Tooher and McKenna (2003) observed that the

time-parameters of the LF-model typically increase with T_0, by measuring the parameters for a small set of short speech segments. Most importantly, both studies found that the time parameters are characterised by different constants of proportionality (contour slopes) for the same acoustic sound, under a limited T_0 range. Nevertheless, the correlation between the LF-model parameters and F_0 is not well known and contradictory results have also been reported. For example, Strik and Boves (1992) found a high correlation between T_a and T_0, in contrast to the results obtained by Tooher and McKenna (2003).

The measurements of the voice quality parameters (R_g, R_k and R_a) by Strik and Boves (1992) and Tooher and McKenna (2003) also showed significant correlation of these parameters with F_0. Moreover, Tooher and McKenna (2003) found that the correlation between the time parameters of the LF-model and F_0 appeared to have been influenced by contextual factors, e.g. the preceding phone. This result is compatible with previous studies about the contextual effect on the voice source. For example, Ní Chasaide and Gobl (1993) indicated that when the vowel is preceded by a voiceless stop, it becomes increasingly breathy-voiced.

The maximum amplitude of the excitation, E_e, also shows a strong correlation with F_0. For example, Fant (1997) suggested that E_e increases proportionally to F_0^p (p in the range of 1.5 to 2) up to a maximum value, which is speaker dependent (e.g. depends wether the speaker is a male or female).

Finally, the voice source appears to be important for different aspects of prosody. For example, results have been published which show the correlation of the LF-model parameters with *stress*, *pitch accent* and the *phrase contour*, e.g. Carlson et al. (1989); Fant (1997); Fant and Kruckenberg (1996); Iseli et al. (2006); Ní Chasaide and Gobl (2004).

5.4 LF-model Compared with Other Source Models

The LF-model has been extensively used to study the voice source and it is often considered as the reference for comparison with other glottal source models. This model has been used in different areas of speech research, such as speech synthesis, analysis of voice qualities and pathological voices, models of speech production, etc. Therefore, its potential has been largely explored and its limitations have been reported in the literature.

5.4.1 Limitations

The limitations of the LF-model found in the literature are summarised as follows:

- complexity of the model parameter calculations.
- *parametric oscillator*, which requires external timing control.
- model parameters are not independent.
- limited parameters to control the shape of the glottal pulse.
- signal phase is not a parameter.

5.4.1.1 Complexity of the Model Parameter Calculations

The numerical complexity of the LF-model signal calculation is mostly related with solving the non-linear equations (5.5) and (5.7) to obtain the parameters ε and α, respectively.

The non-linear nature of the functions in (5.1) may also make the estimation of the LF-model parameters difficult. For example, when the parameters of the LF-model are calculated by fitting the model to observed glottal source signals, a *non-linear optimisation algorithm* is required. In general, the performance of this iterative method depends on good estimates of the *initial conditions* and might be affected by convergence problems, e.g. becoming stuck in local minima.

There are other types of source model which are simpler to calculate and to fit to data than the LF-model. For example, the coefficients of a polynomial based model can be easily calculated by fitting the observed glottal source signals linearly to the model, e.g. Fujisaki and Ljungqvist (1986); Thomson (1992). Simplified approximations of the LF-model have also been proposed in order to reduce the computational complexity, e.g. Qi and Bi (1994); Veldhuis (1998).

5.4.1.2 Parametric Oscillator

The LF-model was described by Schoentgen (1993) as a parametric oscillator "that is driven by periodically changing the values of one or more of its parameters" . This is a general characteristic of the acoustic models which represent the amplitude and shape of a glottal pulse over a fundamental period. The main limitation of parametric oscillators is that the cycle duration of the model is controlled externally. For example,

the calculation of the LF-model signal requires the estimation of T_0 and the instants of maximum glottal excitation (epochs) beforehand. The other parameters can be obtained pitch-synchronously from the observed glottal source signal, e.g. by fitting each of the two curves of the model pitch-synchronously to the derivative of the glottal pulse. However, pitch-synchronous analysis is typically affected by errors in the epoch estimates.

Another disadvantage of parametric oscillators is that the frequency and characteristics of the pulse shape cannot be controlled instantaneously (they are constant throughout the pitch cycle). Schoentgen (2002) reported that this limitation does not allow fine control over prosodic and phonatory timbre features.

In contrast to parametric oscillators, self-sustained oscillators generate their own timing. In general, this is the case of the physical models of the glottal source, e.g. Ishizaka and Flanagan (1972). There are also other types of glottal source model with the flow-induced oscillation property. For example, the *polynomial shaping model* (Schoentgen, 2002) and an adapted LF-model (Schoentgen, 1993).

5.4.1.3 Dependency Between Model Parameters

The five parameters of the LF-model (time and amplitude parameters) are not independent due to the constraint that the glottal flow derivative has energy balance zero over the pitch period. Thus, if any parameter changes, the LF-model waveform has to be calculated again. Furthermore, Schoentgen (2002) argued that the modification of one parameter requires the prediction of the remaining parameters because the relationship between the control parameters of the LF-model cannot be expressed analytically.

In general, physical models do not have this problem because the parameters have a physical meaning and they can be controlled independently. For example the two-mass model proposed by Ishizaka and Flanagan (1972) has nineteen independent parameters, such as the relative length of the vocal folds and the sub-glottal pressure.

There are also acoustic models in which the parameters control different acoustic aspects of the source signal and they can be modified without the need to readjust the other parameters of the model. For example, the model proposed by Schoentgen (1993) is represented by two linear mathematical expressions in which the coefficients are independent.

5.4.1.4 Limitation to Control the Glottal Pulse Shape

The LF-model cannot reproduce all the observed characteristics of the glottal pulse shapes. For example, there are glottal effects such as diplophony and aspiration noise, which cannot be represented by the LF-model.

In general, physical models can reproduce a more diverse range of pulse shapes than the LF-model because they are able to represent more complex shapes observed in the glottal source signal. Typically, the polynomial models can also produce a wider variety of shapes than the LF-model because they can fit to a wider range of curves. However, the LF-model parameters have acoustic meaning, in contrast to most polynomial models, which allows a more intuitive control of the glottal pulse shape.

5.4.1.5 Signal Phase

The LF-model does not allow the control of phase through its parameters. However, the control over the phase of the source signal is a relevant aspect to transform and synthesise more complex shapes of the glottal flow waveform and transform voice quality. For example, Hanquinet et al. (2005) synthesised disordered speech by manipulating several parameters of the source, including the phase. They used a glottal source model based on a sinusoidal shaping function that transformed a periodic input signal into the desired waveshape. This model allowed them to control the *vocal jitter* and the *vocal frequency tremor* characteristics of the excitation by manipulating the phase of the sinusoidal driving function.

5.4.2 Advantages

Despite the limitations described in Section 5.4.1, the LF-model also has attractive properties. The following list indicates the main characteristics which motivated the use of the LF-model in this work.

- good approximation of the glottal flow derivative.
- small number of parameters.
- good control over the source signal shape.
- can be represented using spectral parameters.
- correlation with voice quality and prosody.

- mixed-phase signal.
- good performance in speech synthesis applications.
- can be used to synthesise speech pitch-synchronously.
- popular and extensively studied in the literature.

5.4.2.1 Waveform

In general, the LF-model gives a good representation of the glottal source derivative. In this work, the LF-model is expected to accurately model the excitation signals which are calculated for different speech corpora (each corpus contains speech spoken by a speaker). The voice corpora used in this thesis were built for speech synthesis applications by asking a speaker to read text sentences. Typically, this type of corpus has limited speech expressiveness. The problem of fitting the LF-model to irregular source pulse shapes is assumed not to be important in this thesis because the voice quality variety of the speech corpus used to build the speech synthesisers is assumed to be relatively low.

The complexity due to LF-model parameter estimation is also not important in this work because the parameters are extracted from the speech corpus once during the speech analysis part of the HMM-based speech synthesiser (before the training of the statistical models).

5.4.2.2 Number of Parameters

Another great advantage of the LF-model when compared with other glottal source models, especially the physical models, is the small number of parameters. This is an important factor to take into account in HMM-based speech synthesis because the memory requirements and complexity of the system typically increases with the number of speech parameters used to train the statistical models. Also, the amount of data required to obtain good statistical modelling typically grows with the number of parameters modelled by the HMMs.

5.4.2.3 Voice Quality and Prosodic Correlation

The control over the pulse shape provided by the LF-model is considered to be large enough for this work. One of the objectives of this thesis is to use a glottal source model

for speech synthesis which gives a good parametric flexibility to transform basic voice qualities. Past work have already showed that the LF-model parameters can be used to model a set of "basic" voice qualities, e.g. Gobl (1989); Fant (1995).

In this thesis, the incorporation of the LF-model into a HMM-based speech synthesiser also enabled us to model the prosody and voice quality correlates of the LF-model parameters by the HMMs, in order to improve the quality of the synthetic speech.

5.4.2.4 Spectral Representation

The LF-model also gives the possibility of modelling the voice source using spectral parameters. In this work, the spectral representation of the LF-model is used to design a glottal post-filter which flattens the LF-model spectrum. This method is proposed for the integration of the LF-model into the HMM-based speech synthesiser in Section 6.3.

5.4.2.5 Mixed-phase Signal

The mixed-phase characteristic of the LF-model (related to the causal and anticausal characteristics of the glottal flow) is assumed to be a good model of phase for voiced speech.

The LF-model does not give the parametric flexibility to control the phase. Phase manipulation could be a useful feature to transform speech or introduce randomness to the phase of the harmonic part of the excitation, but it goes behind the scope of this work. Nevertheless, the HMM-based speech synthesisers developed during this thesis use both the LF-model and the STRAIGHT vocoder. It is possible to manipulate the phase of the speech signal in these systems by using STRAIGHT.

The LF-model limitation of not allowing fine control over the instantaneous frequency is not considered to be important in this work. Speech synthesised pitch-synchronously using glottal pulses with the duration of the pitch period for the excitation generally provides good time-resolution. For example, the HMM-based speech synthesiser of Zen et al. (2007a) generates speech pitch-synchronously by passing an excitation signal (such as the impulse train) with duration equal to two times the fundamental period through a synthesis filter and by using overlap-and-add to concatenate the short-time speech signals.

5.4.2.6 Pitch-synchronous Synthesis

Speech can be easily synthesised pitch-synchronously using the LF-model as the excitation. For example, the PSOLA technique (Moulines and Charpentier, 1990) can be effectively performed using the LF-model (by centering the overlap windows at the instants of maximum excitation t_e).

5.4.2.7 Reference Source Model

Another advantage of using the LF-model is that it is a reference glottal source model used in different speech research fields, such as speech synthesis, speech analysis, and voice quality transformation. For example, a model of the flow derivative has been successfully used in the popular synthesiser proposed by Klatt and Klatt (1987). This model allows the synthesiser to control several aspects related to voice quality, such as spectral tilt, the open quotient, and breathiness.

5.5 Conclusion

The LF-model is a popular acoustic model of the glottal source derivative. It gives a very good approximation to the glottal source waveform using a small number of parameters (five or six).

This model can also be represented in the frequency domain using a small set of parameters. Furthermore, the relationship between the time- and frequency-domain parameters of the LF-model can be described by equations, which is very useful in order to represent the glottal source signal either in terms of its shape or spectral properties.

The LF-model parameters are strongly correlated with voice quality and prosody. Formulae of these LF-model correlates have also been proposed in the literature. These correlations are important in this work because one of the goals is to improve voice quality modelling and control by using the LF-model.

The main problems of the LF-model are the complexity of the waveform generation and limitations in terms of representing some details of the glottal source signal. However, these factors were not considered to be relevant because the applications of the LF-model in this work did not require synthesis of speech in real-time and the LF-model signal appeared to generally fit well to the glottal source derivative signal in these applications. How well the LF-model signal fitted to the glottal source deriva-

tive signal in the application of an HMM-based speech synthesiser using glottal source modelling is further discussed in Section 8.4.5.5.

Chapter 6

Analysis/Synthesis Methods

6.1 Introduction

Three different methods for *speech analysis-and-synthesis* have been used in this work. One is the *STRAIGHT vocoder* (version V40_006b) and the other two have been developed in this thesis in order to synthesise speech using the LF-model parameters.

The source-filter model used by STRAIGHT (V40_006b) describes speech as the convolution of a spectrally flat excitation by the spectral envelope of the speech signal. For speech analysis, it extracts the spectral envelope, the F_0 and aperiodicity parameters from the speech signal. For synthesis of voiced speech, a mixed multi-band excitation is the input to the synthesis filter defined by the spectral parameters. In the case of unvoiced speech, the excitation is modelled as white noise.

The second method is called *Glottal Post-Filtering* (GPF) and uses the same source-filter model as STRAIGHT. It also uses STRAIGHT analysis to calculate both the spectral envelope and the aperiodicity parameters. However, this method generates the periodic component of the mixed excitation by passing a chosen LF-model signal through a *glottal post-filter*, instead of using an impulse train (as in STRAIGHT).

The third method, called *Glottal Spectral Separation* (GSS), uses a different source-filter model to represent voiced speech. In this model, the excitation is represented by the glottal source signal and the synthesis filter by the vocal tract transfer function. First, this method estimates the glottal parameters from recorded speech. Then, the vocal tract transfer function is estimated by separating the glottal source characteristics from the speech signal and calculating the spectral envelope of the resulting signal. In this work, the GSS method is implemented using the LF-model to represent the glottal source and STRAIGHT to compute the spectral envelope. The GSS method generates

the excitation signal by mixing the LF-model signal with a noise component and then performs the convolution of this excitation signal with the vocal tract transfer function to obtain the speech signal.

One advantage of combining the GPF method and the GSS method with the analysis method used by STRAIGHT is that the spectral envelope extraction technique of this vocoder is very robust and it also estimates aperiodicity measurements, which can be used to mix a noise signal with the LF-model signal, in order to improve the naturalness of the synthetic speech. Another advantage is that the LF-model can be consistently compared against the impulse train in terms of speech quality by comparing speech synthesised with the GSS and STRAIGHT methods, respectively.

6.2 STRAIGHT

In this work the STRAIGHT version V40_006b was used, because this was the only STRAIGHT version which was publicly accessible (through the following webpage: `http://www.wakayama-u.ac.jp/~kawahara/index-e.html`).

This section describes the methods used by STRAIGHT V40_006b. The latest version of the STRAIGHT vocoder is called TANDEM-STRAIGHT (Kawahara et al., 2008). This version uses a unified approach to estimate the F_0, aperiodicity and spectrogram, which is simpler than the methods used in STRAIGHT V40_006b.

6.2.1 Speech Model

A quasi-periodic speech signal $s(t)$ can be represented by a *sinusoidal model*, which is given by the sum of amplitude and phase modulated harmonics:

$$s(t) = \sum_{k \in N} \alpha_k(t) \sin(2\pi f_k(t) + \theta_k(t)), \tag{6.1}$$

where $\alpha_k(t)$, $f_k(t)$ and $\theta_k(t)$ are the amplitudes, frequencies, and phases of the harmonics, respectively.

The speech model used by STRAIGHT (Kawahara et al., 1999b) represents speech in terms of the *instantaneous angular frequency* of the harmonic component k, i.e. $\omega_k(t) = d\phi_k/dt$, where $\phi_k(t) = 2\pi f_k(t) + \theta_k(t)$ is the instantaneous phase. This model is similar to the sinusoidal model and it is described by Kawahara (1997) as

$$s(t) = \sum_{k \in N} \alpha_k(t) \sin\left(\int_{t_0}^{t} (k\omega(\tau) + \omega_k(\tau)) d\tau + \phi_k(t_o) \right), \tag{6.2}$$

where $\omega(\tau)$ is the instantaneous frequency and $\omega_k(\tau)$ is a slowly varying component of the k-th harmonic (frequency modulation).

6.2.2 Analysis

6.2.2.1 F_0 Estimation

Different methods to estimate F_0 based on instantaneous frequency have been employed in STRAIGHT (Kawahara, 1997; Kawahara et al., 1999a, 2005). This section describes the method called the "*Time-domain Excitation extraction based on a Minimum Perturbation Operator*" (TEMPO), which is proposed by Kawahara (1997) and is used by STRAIGHT (V40_006b). TEMPO estimates F_0 as the instantaneous frequency of the *fundamental component* of the signal. This corresponds to the instantaneous frequency of the harmonic $k = 1$, in (6.2).

The instantaneous frequency is calculated using a method based on the *continuous wavelet transform* (CWT) of the speech signal $s(t)$. This CWT is represented by:

$$D(t, \tau_c) = |\tau_c|^{-\frac{1}{2}} \int_{-\infty}^{\infty} s(t) \psi^* \left(\frac{t-u}{\tau_c} \right) du, \tag{6.3}$$

where $\psi(t)$ is the *wavelet function*, τ_c represents the *scale factor of the wavelet*, and * represents the operation of *complex conjugate*. Kawahara (1997) uses a *Gabor function* for the wavelet, $g(t)$, which is defined by the multiplication of a Gaussian by a sinusoidal function:

$$\psi(t) = g(t - 1/4) - g(t + 1/4) \tag{6.4}$$

$$g(t) = e^{-\pi\left(\frac{t}{\eta}\right)^2} e^{-j2\pi t}, \tag{6.5}$$

where $\eta > 1$ is a parameter that represents the frequency resolution of the wavelet transfer function.

The CWT represented in (6.3) is equivalent to filtering the speech signal with multiple bandpass filters , which have the shape of the wavelet function and cover different parts of the spectrum, respectively. The scale factor of the wavelet, τ_c, defines the frequency f_c at which the output of each filter channel is maximum. The output of

each filter, $D(t,\tau_c)$, represents the amplitude envelope and instantaneous phase of the spectral components of the signal in the frequency band centered at f_c.

Kawahara (1997) estimates the fundamental frequency by assuming that the signal-to-noise ratio of the output of the filters is higher for the filters which have a frequency f_c closest to F_0. He defines a parameter M_{τ_c}, called "*fundamentalness*", which measures this effect. M_{τ_c} is calculated from $D(t,\tau_c)$ and it is used to obtain the filter which maximises the "fundamentalness". Finally, F_0 is calculated as the average of the instantaneous frequency using the outputs of the obtained filter and its neighbours. The instantaneous frequency $f_0(t)$ of a filter output signal $D(t,\tau_c)$ is defined by

$$f_0(t) = \frac{1}{2\pi}\frac{d\arg D(t,\tau_c)}{dt} \tag{6.6}$$

The results reported by Kawahara (1997) indicate that this method is very accurate and its performance is comparable to other popular F_0 detection methods, such as the *Average Magnitude Difference Function* (AMDF) method of de Cheveigné (1996) and the RAPT algorithm (Talkin, 1995).

6.2.2.2 Spectral Envelope

The power spectrum of the speech signal is calculated by using a *pitch-adaptive Short-term Fourier Transform* (SFT) analysis. Kawahara et al. (1999b) propose to use two *compensatory time windows* to calculate the spectrogram.

First, a convolution of the speech signal with a pitch-adaptive window is performed. The time window is given by the convolution of a Gaussian function $w_g(t)$ with a second order *cardinal B-spline function* $h(t)$:

$$w_p(t) = w_g(t) \odot h(t/t_0) \tag{6.7}$$

$$w_g(t) = e^{-\pi\left(\frac{t}{\eta t_0}\right)^2} \tag{6.8}$$

$$h(t) = \begin{cases} 1-|t|, & |t|<1 \\ 0, & otherwise \end{cases}, \tag{6.9}$$

where $\odot$ represents convolution and t_0 is the instantaneous fundamental period (is a function of time). The resulting window $w_p(t)$ is also a second order spline function.

Figure 6.1 shows the shape of the Gaussian time window and the second order spline function.

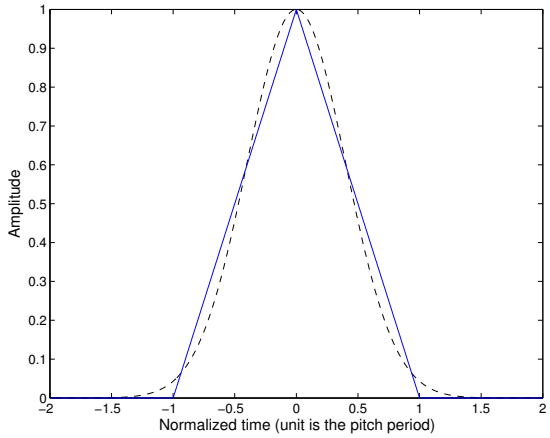

Figure 6.1: Gaussian time window (dashed line) and the basis function of the second-order cardinal B-spline window (solid line).

The main objective of the convolution of the speech signal with $w_p(t)$ is to smooth the spectrogram in the frequency domain. Kawahara et al. (1999b) argue that this type of smoothing is robust to variations and estimation errors of the fundamental period, T_0.

The periodicity of the speech signal along the time domain also produces *phase interference* in the spectrogram. This effect is reduced in STRAIGHT by setting the length of the window $w_p(t)$ equal to twice the fundamental period. For example, if a short window which provides good spectral resolution (length comparable to T_0) has a different length from a multiple of T_0, then the spectrogram shows periodicity along the time domain.

Another special property of the window $w_p(t)$ is the equivalent relative resolution in both time and frequency domain (Kawahara et al., 1999b). The following formula of the FT of $w_g(w)$ shows that the analysis window size also adaptively changes in the frequency domain, in terms of the fundamental frequency F_0.

$$W_g(w) = \frac{t_0^2}{\sqrt{2\pi}} e^{-\pi\left(\frac{w}{\eta w_0}\right)^2}, \tag{6.10}$$

where $w_0 = 2\pi f_0$. This characteristic also reduces phase interference caused by periodic variations in the frequency domain.

However, the smoothing operation is not enough to remove the periodic interference. According to Kawahara et al. (1999b), there is still periodic interference in the

spectral valley areas. In STRAIGHT this problem is overcome by using "a compensatory window that produces maxima where the original spectrogram has holes". The *compensatory window* of $w_p(t)$ is given by

$$w_c(t) = w_p(t) \sin\left(\pi \frac{t}{T_0}\right) \tag{6.11}$$

This window represents a *sinusoidal modulation* which converts the frequency of the harmonics and shifts their phases towards the opposite directions by the desired amounts. Figure 6.2 shows the general shape of the compensatory window.

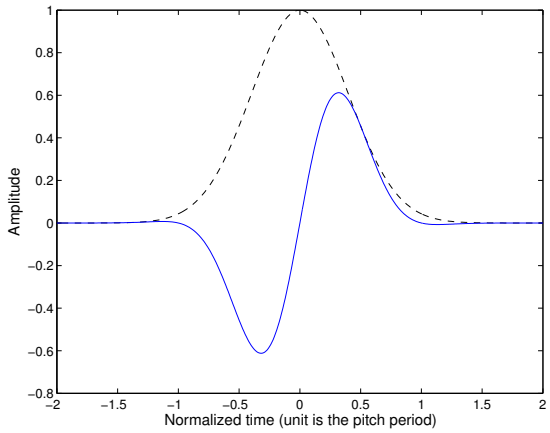

Figure 6.2: Gaussian time window (dashed line) and the respective compensatory window (solid line).

Two power spectra, $P_o(w,t)$ and $P_c(w,t)$, are calculated using the original window $w_p(t)$ and the compensatory window $w_c(t)$, respectively. Then, the power spectrum of the speech signal is represented as a weighted squared sum of the power spectra:

$$P_\tau(w,t) = \sqrt{P_o^2(w,t) + \xi P_c^2(w,t)}, \tag{6.12}$$

where ξ is a *blending factor*, which is selected so that it minimises the temporal variation of the resulting spectrogram. Figure 6.3 a) shows an example of the speech spectrum calculated by STRAIGHT using the compensatory windows to remove the periodicity.

The power spectrum $P_\tau(w,t)$ has minimal interferences from the speech spectrum periodicity but the resulting spectral envelope is typically over-smoothed. Kawahara et al. (1999b) indicate that the main reason for this over-smoothing effect is the isometric Gaussian time window $w_g(t)$, which also contributes to the smoothing of the

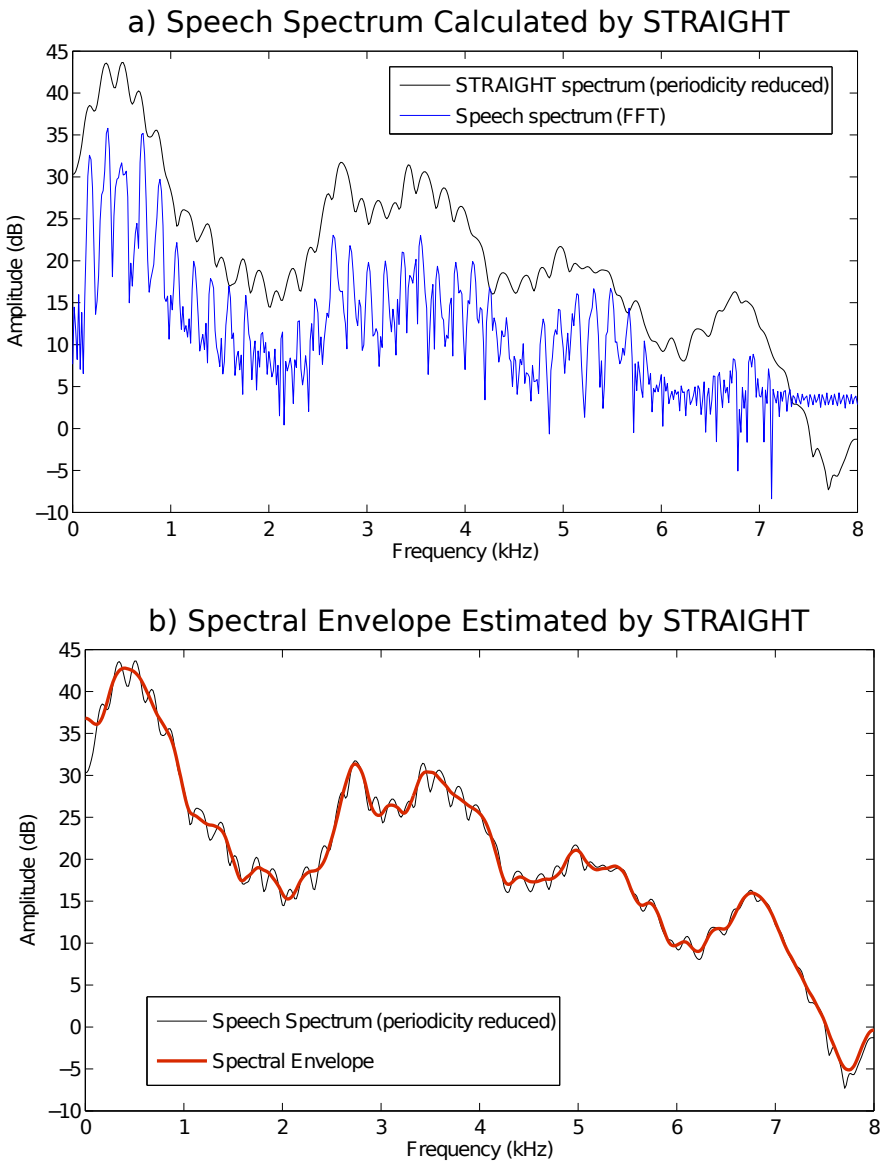

Figure 6.3: Top: Comparison of the amplitude spectrum calculated by STRAIGHT and the amplitude spectrum calculated by conventional SFT analysis using a Hamming window, for a 40 ms speech frame. Bottom: spectral envelope calculated by STRAIGHT from the speech spectrum.

spectrum calculated by SFT. This effect is associated with the limited frequency resolution caused by the *time-frequency trade-off problem* (a high frequency resolution implies low time resolution and vice-versa). The combined contribution of both $w_g(t)$ and $h(t)$ makes spectral smoothing excessively high. Kawahara et al. (1999b) propose a *quasi-optimal smoothing function* $h(t)$ which reduces the smoothing effect of $w_g(t)$. This function consists of three second-order cardinal B-spline functions. Figure 6.3 shows the spectral envelope calculated by STRAIGHT from the speech spectrum with

reduced periodicity, which was obtained using the quasi-optimal smoothing function.

Figure 6.4 also shows an example of the spectral envelopes calculated for a voiced speech frame, using STRAIGHT and LPC analysis respectively. STRAIGHT can more accurately estimate the spectral envelope than the LPC vocoder (Makhoul, 1975), in general. On of the reasons for this is that STRAIGHT better removes the periodicity effects of the speech signal than the conventional autocorrelation method for LPC analysis (Makhoul, 1975). Also, STRAIGHT analysis takes into account the fine variations in F_0 along the time, whereas the autocorrelation method has a poorer F_0 resolution.

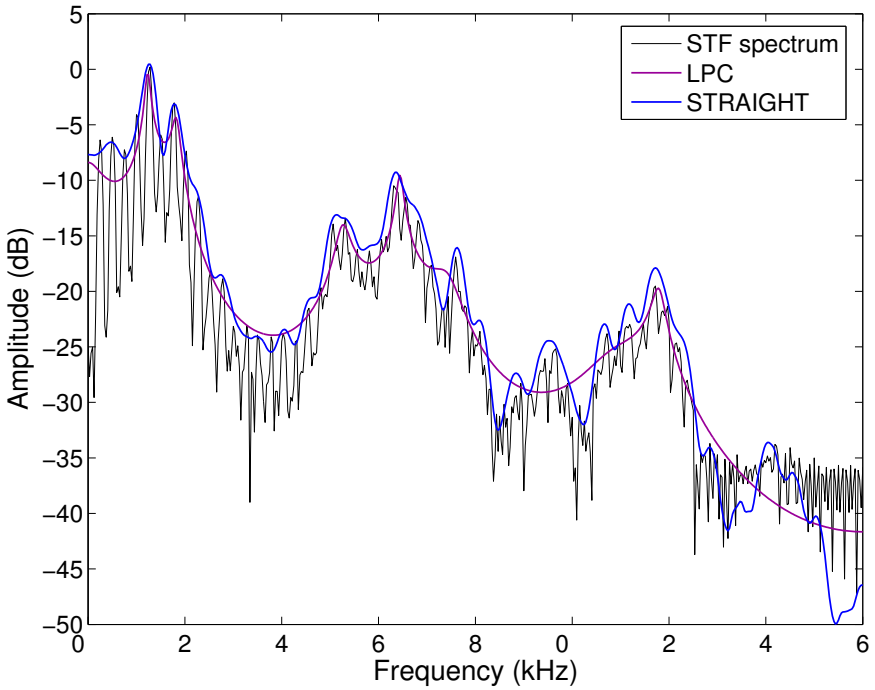

Figure 6.4: Spectral envelopes calculated with STRAIGHT and with the Levison-Durbin method of LPC analysis, for a speech frame.

6.2.2.3 Aperiodicity Measurements

STRAIGHT measures the aperiodicity of a speech signal using the phase of the fundamental component and the power spectrum calculated with appropriate time windows (Kawahara et al., 2001).

The TEMPO method described in Section 6.2.2.1 can be used by STRAIGHT to calculate the phase of the fundamental. Kawahara et al. (2001) propose another method to calculate the phase of the fundamental, which is based on the concept of *fixed-point analysis* of a mapping from the center frequencies of the analysing wavelet to their

output instantaneous frequencies.

Before the calculation of the power spectrum, the effects of F_0 variation along the time domain are removed from the speech signal by performing *time warping* using the inverse function of the phase of the fundamental. The resulting signal has approximately constant F_0 and a regular harmonic structure. Kawahara et al. (2001) assume that the aperiodic components are the frequency components between the harmonics in the amplitude spectrum of this signal.

The smooth power spectrum is calculated along the new time axis by using a method similar to the spectral envelope estimation method described in the previous section. The analysis time-window is also the convolution of a Gaussian function (slightly stretched) with a second-order cardinal B-spline function. In this case, the B-spline function is tuned to F_0 on the new time axis and it is designed to have zeros between harmonic components. Kawahara et al. (2001) indicate that "a power spectrum calculated with this window provides the energy sum of periodic and aperiodic components at each harmonic frequency and provides the energy of the aperiodic components at each in-between frequency". Based on this assumption, the aperiodicity is measured as the ratio between the *lower and upper smoothed spectral envelopes* of the short-time signal.

The upper envelope, $|S_U(w)|^2$, is calculated from the speech spectrum by connecting *spectral peaks* and the lower envelope, $|S_L(w)|^2$, is calculated by connecting *spectral valleys*. Figure 6.5 a) shows an example of the spectral peaks obtained by STRAIGHT for a speech frame. Next, the *aperiodicity measurement* $P_{AP}(w)$ is calculated from the upper and lower envelopes using (4.3). Figure 6.5 b) shows an example of the aperiodicity spectrum calculated for a voiced speech frame. Typically, the overall slope of the aperiodicity curve is positive because the SNR is lower at the high frequency region than at the lower part of the speech spectrum (for voiced speech).

6.2.3 Synthesis

6.2.3.1 Source-Filter Model

Speech can be synthesised from the STRAIGHT parameters using the sinusoidal model represented by (6.1). However, Kawahara (1997) proposes the method SPIKES (Synthetic Phase Impulse for Keeping Equivalent Sound), as it is easier to implement and allows more control over speech characteristics. Basically, this technique represents the synthesis filter by a minimum-phase impulse response, $H(w,t)$, and uses an all-

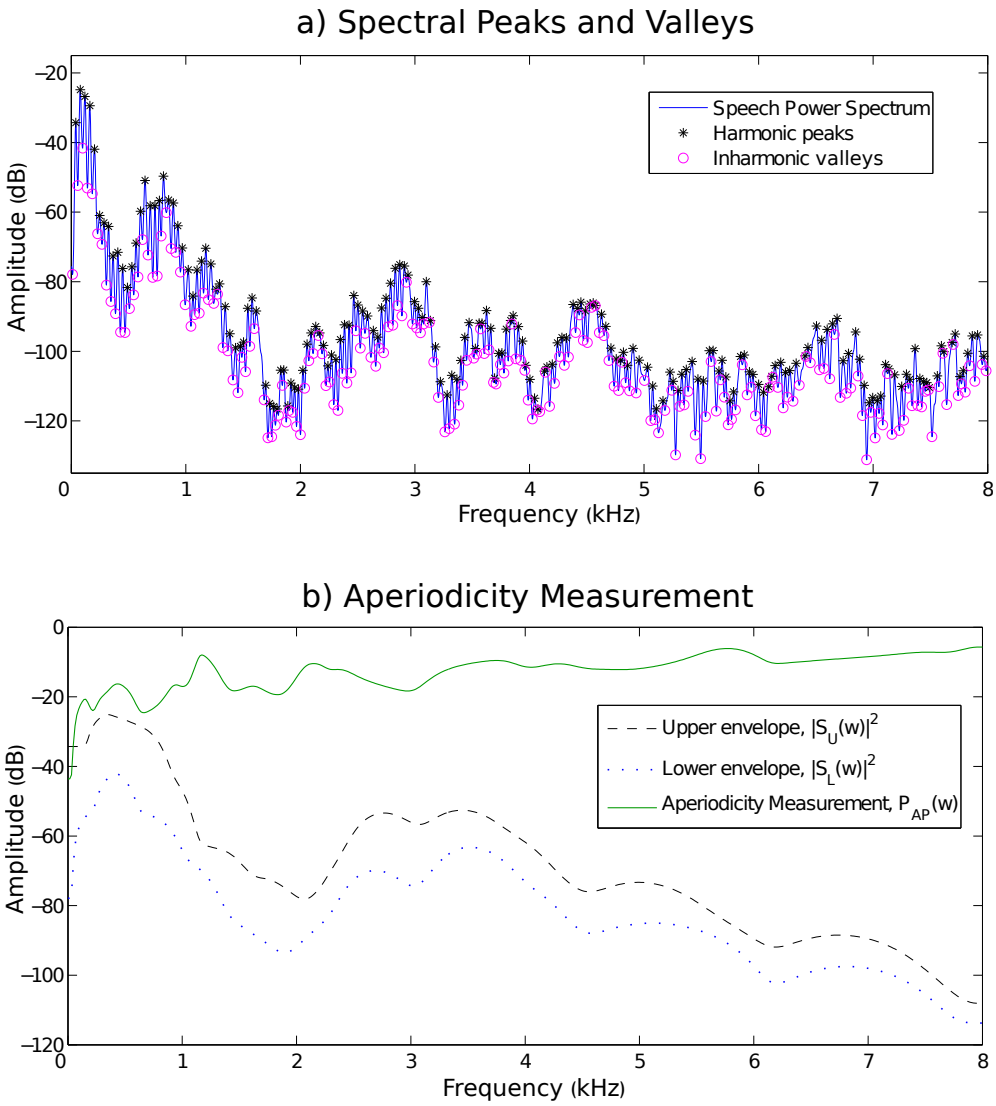

Figure 6.5: Example of the aperiodicity spectrum calculated for a voiced speech frame. Top: amplitudes of the spectral peaks and valleys obtained from the amplitude spectrum of the speech signal, by STRAIGHT. Bottom: lower and upper spectral envelopes calculated by STRAIGHT and the resulting aperiodicity spectrum.

pass filter, $\Phi(w)$, to transform the phase characteristics of the impulse train excitation. $H(w,t)$ is obtained by calculating the complex cepstrum of the speech spectrum. This type of impulse response is physically stable because the zeros of the z-transform are all inside the unit circle. Each short-time speech signal y_{t_i} is synthesised from one excitation pulse located at the position i by using the following equation (represents the inverse Fourier transform) from Kawahara et al. (2001):

$$y_{t_i}(t) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} H(w,t_i)\Phi(w)e^{jw(t)}dw \tag{6.13}$$

The all-pass filter function $\Phi(w)$ is used because it has a completely flat spectrum, allows a fine control of F_0 and reduces the buzzy timbre by manipulating the phase of the pulse excitation. $\Phi(w)$ is based on group delay design, as described in Section 4.3.3.2. For synthesis of unvoiced speech, the excitation is modelled as white noise only.

STRAIGHT also adds a noise component to the impulse train in order to reduce the "buzziness" effect caused by this signal. The weighting of the periodic and noise components of the excitation is controlled by the aperiodicity parameters (Kawahara et al., 2001).

Synthetic speech can be represented in terms of the minimum-phase impulse response, $H(w)$, and the FT of the mixed excitation signal, $X(w)$, by:

$$Y(w) = X(w)H(w), \tag{6.14}$$

where $Y(w)$ is the FT of the synthetic speech. The impulse response is obtained by calculating the complex cepstrum of the smooth spectral envelope. In other words, speech is synthesised by passing the mixed excitation through the minimum-phase filter, which represents the spectral envelope of the speech signal.

The mixed excitation is the sum of the periodic and noise components, which is given by:

$$X(w) = \sqrt{1/F_0}D(w)\Phi(w)W_p(w) + N(w)W_a(w), \tag{6.15}$$

where $D(w)$ is the FT of the delta pulse, $N(w)$ is the FT of white noise, and $\Phi(w)$ represents the all-pass filter function. Finally, $W_p(w)$ and $W_a(w)$ are the weighting functions of the periodic and noise components, respectively. The noise is modelled by a random sequence with zero mean and unit variance. For the impulse train to have the same energy as the noise signal, the pulse is multiplied by $\sqrt{1/F_0}$.

6.2.3.2 Phase Manipulation

The all-pass filter design is based on the group delay function. The method to derive the all-pass filter $\Phi(w)$ from the group delay was described in Section 4.3.3.2.

STRAIGHT uses all-pass filters in order to reduce the degradation in speech quality associated with the strong periodicity of the pulse train, $P(w)$. It introduces randomness in the phase of this signal by manipulating the group delay at higher frequencies.

6.2.3.3 Pulse/Noise Weighting

The weighting functions, $W_p(w)$ and $W_a(w)$, are obtained from the aperiodicity parameters. Figure 6.6 shows an example of how the spectra of the impulse and noise components of the excitation are mixed using the weighting functions. The impulse signal, the all pass filter function, and the noise are spectrally flat. The weighting operation determines the spectral energy balance between the pulse train and the noise. The resulting mixed excitation signal also approximates a spectrally flat signal.

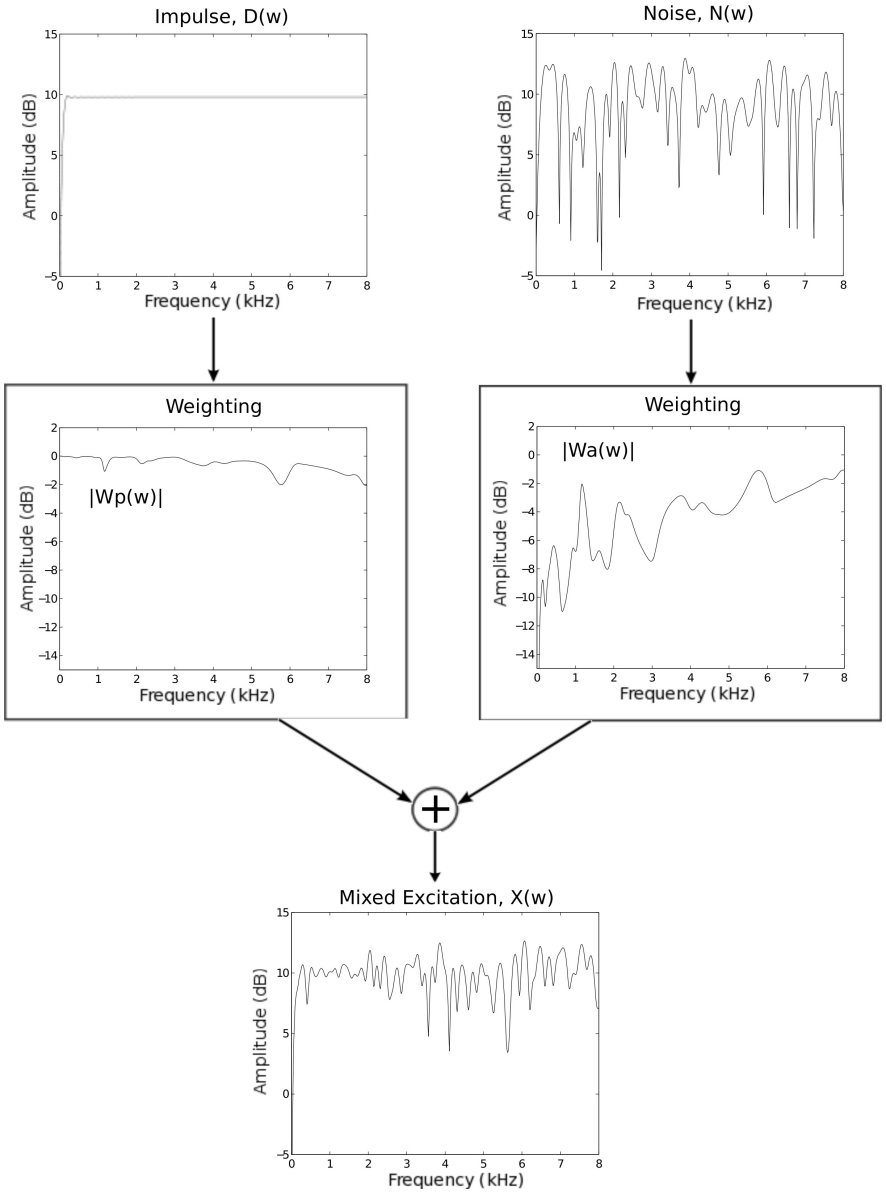

Figure 6.6: Mixing of the impulse signal (a phase manipulated delta pulse) with the noise to obtain the excitation signal.

6.3 Glottal Post-Filtering (GPF)

The GPF method was developed during this thesis to combine the LF-model with the spectral envelope of STRAIGHT. Basically, it consists of transforming the LF-model signal into a spectrally flat signal. The resulting signal can be used to synthesise speech instead of the impulse train. Although the excitation obtained using GPF does not represent the glottal source signal, this excitation is expected to produce more natural speech than the impulse train. This improvement is explained by the fact that the voiced excitation of the GPF method contains the phase information of the LF-model, whereas the phase of the impulse train is constant and equal to zero. Also, the GPF method can be used to transform voice characteristics of the synthetic speech by modifying glottal source parameters of the LF-model.

The GPF method was not directly compared against a baseline analysis/synthesis method in terms of speech naturalness and voice transformation, in this work. However, the perceptual experiment presented in Section 8.4 evaluates the speech quality of an HMM-based speech synthesiser using the GPF method and an HMM-based speech synthesiser using the STRAIGHT vocoder. Since these systems only differ in the analysis/synthesis method, the performance of the GPF method is evaluated in the application to HMM-based speech synthesis.

6.3.1 Speech Model

The speech model used by GPF is similar to the model used by STRAIGHT, which was described in Section 6.2.1. The main difference to STRAIGHT is that GPF represents the periodic component of the excitation by a transformed LF-model signal, instead of the impulse train. A *glottal post-filter* is used to perform whitening of the LF-model spectrum. This filter is computed during analysis and it is used to generate the excitation signal during synthesis of speech (it remains *unchanged* for synthesis).

The excitation signal is represented by a mixed multi-band model, in which the spectra of the periodic and noise components are weighted using the STRAIGHT aperiodicity parameters.

6.3.2 Analysis

In this work, STRAIGHT is used to extract the spectral envelope and aperiodicity parameters. Besides these parameters, the LF-model parameters are also estimated in

order to derive the transfer function of the glottal post-filter which is used to synthesise speech.

6.3.2.1 LF-model

The LF-model is used by the GPF method to derive the glottal post-filter. Also, the LF-model waveform is used to generate the excitation signal in order to synthesise speech. In both cases, the same set of LF-model parameters is used, unless speech is synthesised using voice quality transformation. In this case, the LF-model parameters used to generate the excitation are different, but the glottal post-filter remains the same. Voice transformation using GPF is described later in Section 6.3.4.

There is not a rule for the selection of the LF-model parameter values and different sets of parameter values could be used. However, these values must satisfy the constraints given in Section 5.2.1, in order to ensure that the LF-model waveform is not distorted. One period of the LF-model waveform is calculated from the parameter values of: t_p, t_e, T_a, T_0, and E_e. This signal is called the *reference LF-model* signal. Figure 5.1 shows an example of the LF-model waveform and its parameters.

The reference LF-model signal is used to calculate the glottal post-filter and it is also used for synthesising the speech signal. It has to be chosen carefully, because it might affect the quality of the synthetic speech. For example, the duration of the LF-model pulse (equal to the duration of the open phase) should not be much longer than the minimum fundamental period (T_0), which characterises the speaker's voice. This is to avoid problems with synthesis of speech with low T_0 values, which are explained in Section 6.3.3.2.

In this work, the LF-parameter values are obtained by measuring the average LF-parameters and the minimum T_0 values for the speaker's voice.

6.3.2.2 Parameters of the Glottal Post-Filter

In the spectral domain, the LF-model can be approximated by the stylised spectrum proposed by Doval and d'Alessandro (1997). This spectral representation was explained in Section 5.2.4. Basically, it represents the glottal source derivative using three asymptotic lines with +6 dB/oct, -6 dB/oct and -12 dB/oct slopes, respectively. Figure 6.7 a) illustrates this representation. The crossing point of the first two lines corresponds to a peak (called glottal spectral peak or glottal formant) at the frequency F_g. The second line is due to the spectral tilt which leads to an additional -6 dB/oct

above the frequency F_c. The spectrum of the LF-model is characterised by these two frequency parameters and a gain factor.

It is possible to design a filter which transforms the LF-model signal into an approximately spectrally flat signal if the frequencies F_g and F_c of this model are known. The stylised spectrum of the proposed filter is described by three linear segments, whose slopes are symmetric to the slopes of the LF-model spectrum: -6 dB/oct, +6 dB/oct and +12 dB/oct. The stylised transfer function of this filter is illustrated in Figure 6.7 b).

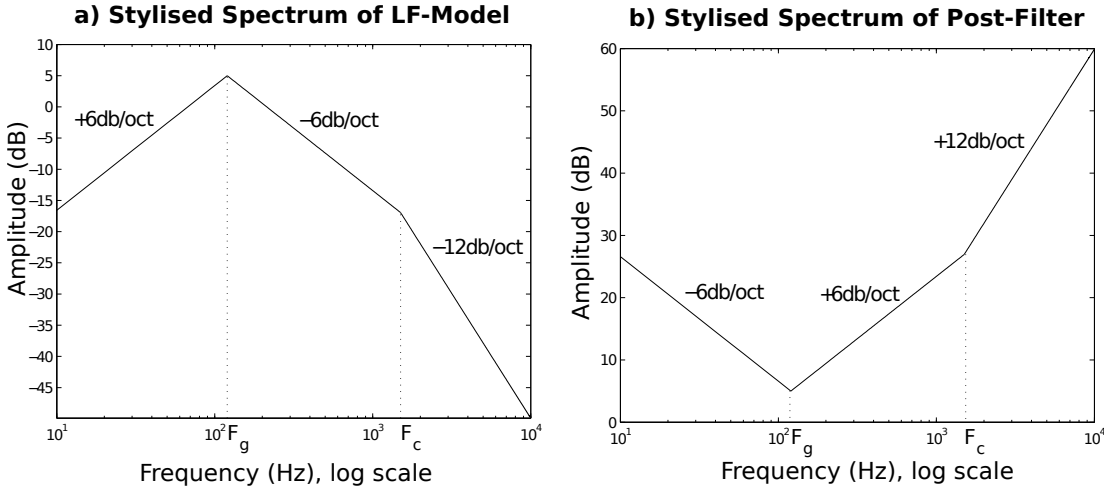

Figure 6.7: Stylised spectrum of the LF-model (a) and its corresponding post-filter spectrum (b).

The formulas which describe the spectral correlates of the LF-model (explained in Section 5.3.1) are used to calculate the frequency parameters of the glottal post-filter.

From Doval and d'Alessandro (1997), the frequency F_g is related to the LF-model parameters by the following formula:

$$F_g = \frac{1}{2\pi OQ_e T_0}\sqrt{\frac{e_n(\alpha_m)}{i_n(\alpha_m)}}, \tag{6.16}$$

where OQ_e is the open quotient ($OQ_e = t_e/T_0$), e_n represents the maximum excitation of the normalised glottal flow $n_g(t)$, and i_n is the integral of $n_g(t)$. Equation (6.16) can be used to calculate the variation of F_g in terms of the variation of the LF-model parameters relative to a reference $n_g(t)$.

In this work, F_g is calculated by using the following formula from Doval and d'Alessandro (1997), which is equivalent to (6.16):

$$F_g = \frac{1}{2\pi}\sqrt{\frac{E}{I}}, \tag{6.17}$$

where E is the amplitude of maximum excitation and I is the integral of the glottal flow pulse. First, the LF-model signal is calculated by using (5.1) to (5.3). This LF-model signal is defined by an abrupt closure ($T_a = 0$) because Doval and d'Alessandro (1999) assume that F_g does not depend on T_a in (6.16) and (6.17). The resulting cycle of the LF-model waveform is integrated to obtain the glottal flow pulse, $u_{LF}(t)$. Next, the parameter I is calculated as the integral of the resulting pulse. In discrete time, the integral of $u_{LF}(n)$ is equal to:

$$I_n = \frac{1}{F_s}\sum_{n=1}^{N_0} u_{LF}(n), \tag{6.18}$$

where F_s is the sampling frequency and N_0 is the length of the pulse. Finally, the frequency F_g is calculated as

$$F_g = \frac{1}{2\pi}\sqrt{\frac{E_e * F_s}{I}} \tag{6.19}$$

In this equation, the parameter E_e is multiplied by F_s, in discrete time, as it represents the slope of $u_{LF}(n)$ at the instant of maximum excitation, t_e. The other parameter used to design the glottal post-filter is the frequency F_c, which represents the cut-off frequency of a low-pass filter associated with the spectral tilt of the source. It is calculated as:

$$F_c = \frac{1}{2\pi T_a} \tag{6.20}$$

The frequencies F_g and F_c of the glottal post-filter are computed using the parameters of the reference LF-model signal (described in Section 6.3.2.1) and equations (6.19) and (6.20) respectively. In this work, the stylised spectrum of the filter shown in Figure 6.7 b) is implemented as a *linear phase FIR filter*.

6.3.3 Synthesis

6.3.3.1 Source-Filter Model

The source-filter model used by the GPF method to synthesise speech mainly differs from the STRAIGHT model in the excitation part. The block diagram of the speech synthesis method using the glottal post-filter is shown in Figure 6.8.

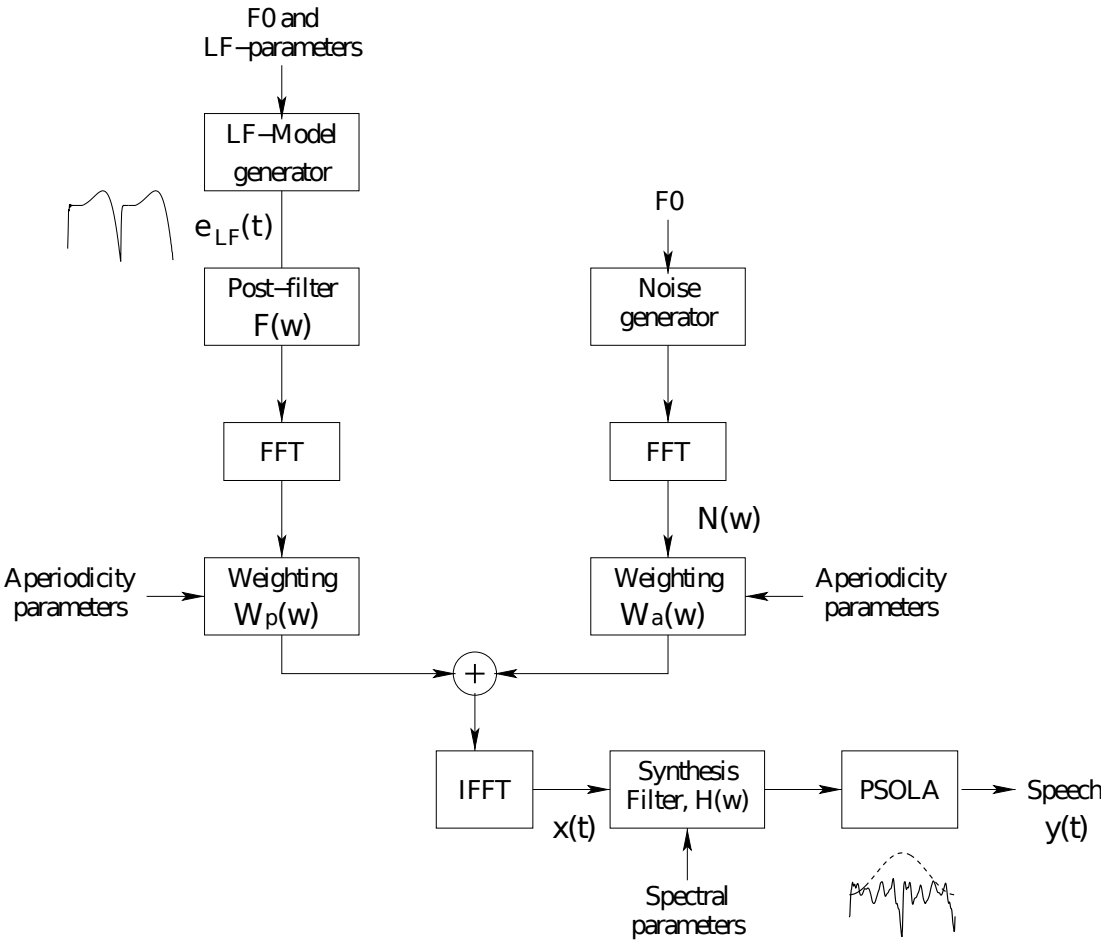

Figure 6.8: Block diagram of the speech synthesis method using GPF.

In the GPF method the synthetic speech, $Y(w)$, is obtained by:

$$Y(w) = X(w)H(w), \tag{6.21}$$

where $X(w)$ is the FT of a mixed multi-band excitation and $H(w)$ represents the transfer function of the synthesis filter. $H(w)$ models the spectral envelope (as in STRAIGHT) and it is calculated from the spectral parameters. The GPF synthesis method uses a technique based on PSOLA (Moulines and Charpentier, 1990) to concatenate the synthesised speech frames, unlike the STRAIGHT synthesis method.

The excitation model of the GPF method is represented by

$$X(w) = K_e E_{LF}(w)F(w)W_p(w) + N(w)W_a(w), \tag{6.22}$$

where $E_{LF}(w)$ represents the FT of a periodic LF-model signal, $F(w)$ represents the transfer function of the glottal post-filter, $N(w)$ is the FT of the white noise signal, and

K_e is a gain factor. The weighting functions $W_p(w)$ and $W_a(w)$, are calculated using the STRAIGHT aperiodicity measurements, as described in Section 6.2.3. The all-pass filter function, which transforms the phase of the impulse in STRAIGHT, is not used in this excitation model. The reason for this is to preserve the phase characteristics of the LF-model, which were explained in Section 5.2.5. The scale factor K_e adjusts the energy of the LF-model signal so that this signal has the same energy as the noise signal.

6.3.3.2 LF-model

For synthesising speech without voice transformation, the LF-parameter values used to generate the glottal source derivative waveform are the same as those used to derive the glottal post-filter during analysis.

The GPF method does not model the correlation between the glottal pulse shape and F_0, because the spectral characteristics of the LF-model signal are lost when this signal is transformed into a spectrally flat signal. However, when the duration of the reference LF-model signal is adjusted, it is important to preserve its shape in order to obtain a spectrally flat excitation. For controlling the pitch of the synthetic speech, the reference LF-model waveform is either padded with zeros, or its closed phase is truncated, in order to obtain a signal with duration equal to the fundamental period, T_0. This operation allows the pitch period to be controlled without affecting the spectrum of the LF-model signal, unless the truncation region is longer than the closed phase of the LF-model signal. If the length of the closed phase is not long enough to perform the truncation, then the open phase of the glottal signal has to be truncated or decimated, which alters the shape of the LF-model signal and its spectrum.

Interpolation or decimation of the LF-model are not used to control the pitch period because they change the spectrum of the reference LF-model signal. Equations (6.16) and (6.20) show that the spectrum of the LF-model changes if the duration of the glottal pulse, t_e, or the duration of the return phase, T_a, vary. Since the glottal post-filter is tuned to the reference LF-model spectrum, changes in the shape of the LF-model signal used for synthesis deteriorate the whitening effect of the post-filter.

The problem of truncating the LF-model behind the closed phase can be avoided by choosing the reference LF-model signal so that it has a short pulse duration (small duration of the open phase). For example, the reference LF-model signal could be selected so that it has the duration of the open phase (with duration equal to $t_e + T_a$) close to the minimum fundamental period $T_0 = 1/F_0$ (characteristic of the speaker).

The periodic component of the excitation is the concatenation of two LF-model signals, which start at the instant of maximum excitation t_e. These signals are obtained by adjusting the length of the reference LF-model signal (by truncating/padding with zeros) to the target T_0. That is, for synthesising the speech frame i, the first LF-model signal has the duration T_0^{i-1} (equal to the period of the previous frame) and the second has the duration T_0^i. The resulting LF-model waveform is approximately centered at the instant of maximum excitation, t_e. The synthetic speech frames are concatenated using the overlap-and-add technique with windows approximately centered at the instants of maximum excitation. The overlap windows are asymmetric, to obtain perfect overlap-and-add (they sum to one), as in the *Pitch-Synchronous Time-Scaling* (PSTS) method (Cabral and Oliveira, 2005). Each overlap window is obtained by concatenating the first half of a *Hanning window* with the second half of a Hanning window, which may have different durations. The first part has duration T_0^{i-1}, whereas the second has duration T_0^i.

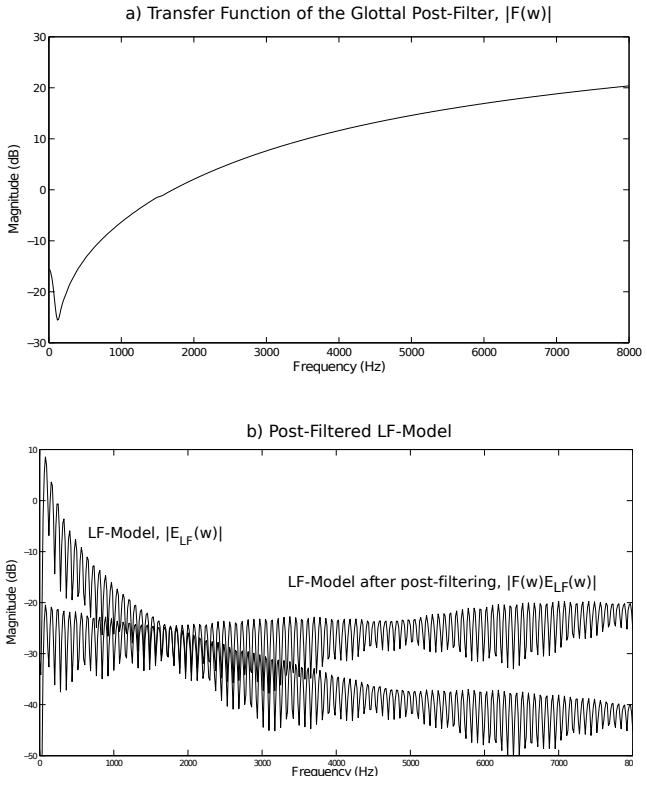

Figure 6.9: Transfer function of the glottal post-filter, on the top. On the bottom, the amplitude spectra of a segment of the LF-model signal (with duration 25 ms) and this signal after glottal post-filtering. The spectrum of the post-filtered LF-model signal is approximately flat.

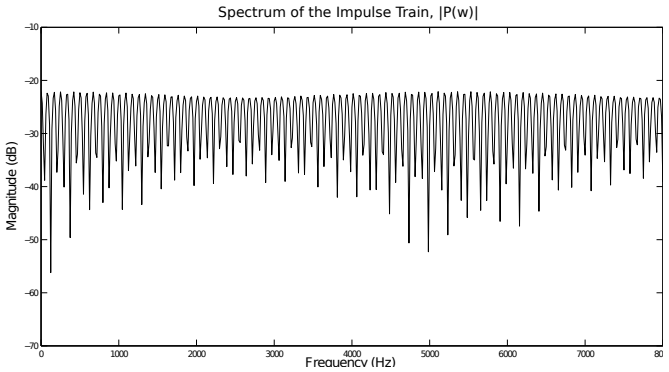

Figure 6.10: Spectrum of a segment of the pulse train (with duration 25 ms).

6.3.3.3 Glottal Post-Filtering

Figure 6.9 shows an example of the transfer function of the glottal post-filter and the spectral effect of this filter on the input LF-model signal. The spectrum of the impulse train is shown in Figure 6.10, for comparison with the LF-model spectrum. These figures show that the amplitude spectrum of the post-fitered LF-model is approximately flat, similar to the amplitude spectrum of the impulse train.

In this work the glottal-post filter is implemented as a FIR filter so that it produces a linear transformation of the phase of the LF-model signal, which does not affect the perceptual quality of the speech signal (corresponds to a time shift of the speech waveform).

Figure 6.11 shows an example of the signal obtained by passing the *reference* LF-model signal through the glottal post-filter. The resulting signal has an amplitude peak at the same point as the instant of maximum excitation of the LF-model signal, since the phase information of the LF-model signal is preserved in the filtering operation. The energy of the signal obtained using post-filtering is not concentrated into a single point as in the delta pulse shown in Figure 4.8. The phase of the signal obtained by post-filtering is also different from both the phase spectra of the delta pulse and the pulse obtained using STRAIGHT, which are shown in Figure 4.9. This variation in phase explains the difference between the waveforms of the STRAIGHT and GPF pulses (Figures 4.8 and 6.11 respectively). The GPF pulse has the advantage that it does not require phase processing and contains the mixed-phase characteristic of glottal source signals, which was explained in Section 5.2.5.2. This phase information of the excitation used in the GPF method is expected to reduce the "buzzy" quality, which is often perceived when listening to speech synthesised with the impulse train.

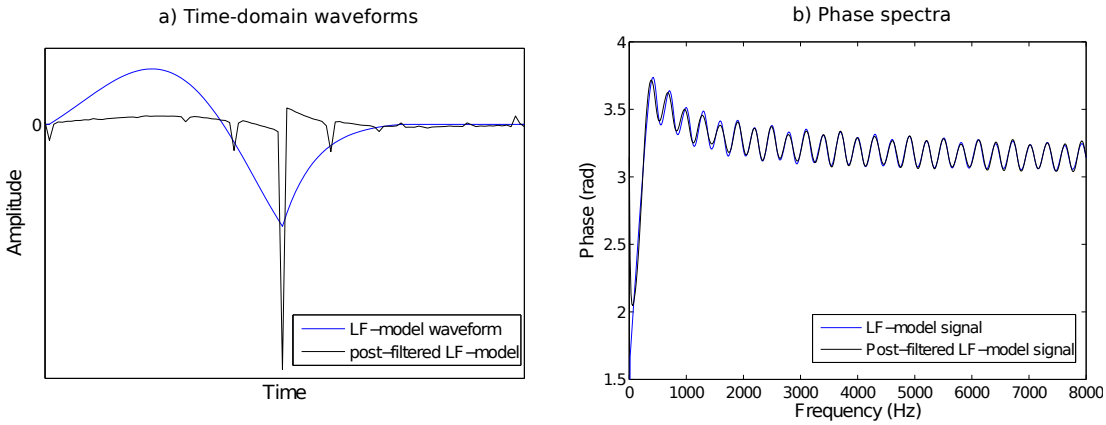

Figure 6.11: Example of the signal obtained by passing the *reference LF-model* signal through the glottal post-filter. This signal preserves the phase information of the LF-model signal.

6.3.4 Voice Quality Transformation

The characteristics of the glottal source signal (used to represent the excitation) can be modified using a different set of LF-parameter values to that which defines the reference LF-model signal. For example, if the return phase parameter T_a is decreased, the spectral tilt of the LF-model signal decreases (lower attenuation at higher frequencies). The variations in the LF-model spectrum produce similar changes in the spectrum of the synthetic speech, as the glottal post-filter remains the same. Therefore, the GPF method allows the voice characteristics of the synthetic speech to be modified. For example, voice quality can be modified by controlling parameters of the LF-model which are correlated with voice quality, such as the open quotient (OQ), speed quotient (SQ), and return quotient (RQ).

The GPF method gives a limited control over the glottal source signal. One limitation is that it does not allow the values of the glottal parameters to be directly set. Nevertheless, it can be used to produce variations of the glottal characteristics, relative to the speech signal which is synthesised using the reference LF-model signal. For example, if we take a reference LF-model signal with OQ=0.6, then by using a LF-model with lower OQ for synthesising speech, e.g. OQ=0.3, the resulting synthetic speech has the spectral effects of decreasing the OQ.

Another problem with the voice quality transformation using GPF is that the degree of glottal parameter transformations depends on the reference LF-model signal. For example, if the OQ of the reference LF-model signal is low, decreasing the OQ of this

signal has a small effect on the voice quality of the synthetic speech. Furthermore, a short reference model signal does not allow very low scale factors of the LF-parameters to be used, because the length of the LF-model signal is constrained by a minimum number of samples.

Nevertheless, this voice transformation method can be used to produce the same glottal parameter transformation effects on the synthetic speech along the utterance. For example, if the analysed speech is spoken with modal voice, then the voice quality parameters of the LF-model (e.g. OQ, SQ, and RQ) could be transformed by scale factors to modify the modal voice quality of the synthetic speech.

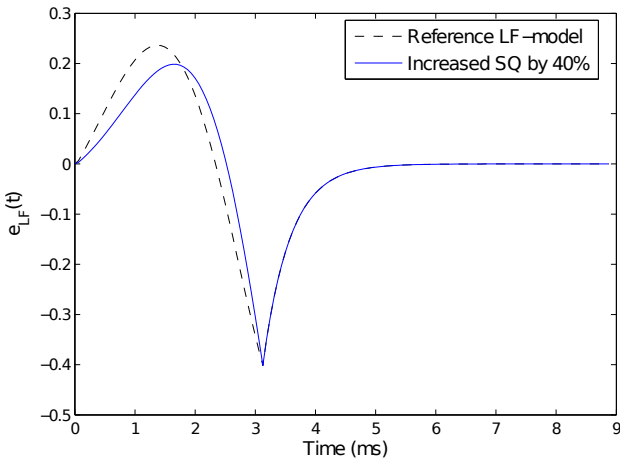

Figure 6.12: Reference LF-model waveform and LF-model signal obtained by increasing the SQ of the reference LF-model signal by 40%.

Figure 6.12 shows an example of the reference LF-model waveform and the signal obtained by increasing the SQ of the reference LF-model by 40%. Figure 6.13 a) shows the difference between the spectrum of these two signals. The effect of increasing the SQ of the reference LF-model signal is to decrease the spectral tilt (increase of energy at higher frequencies) and to change the frequency and amplitude of the glottal formant. The excitation is affected by the same variation, because the glottal post-filter does not change. Figure 6.13 b) shows the spectrum of the two filtered signals. When the input of the filter is the reference LF-model signal, the excitation is spectrally flat. Meanwhile, when the SQ of the reference LF-model is increased, the spectrum of the excitation is no longer flat. This variation in the spectrum of the excitation has the same effect on the spectrum of the synthetic speech. As result, by changing the SQ of the reference LF-model signal, the synthetic speech will exhibit a different voice quality.

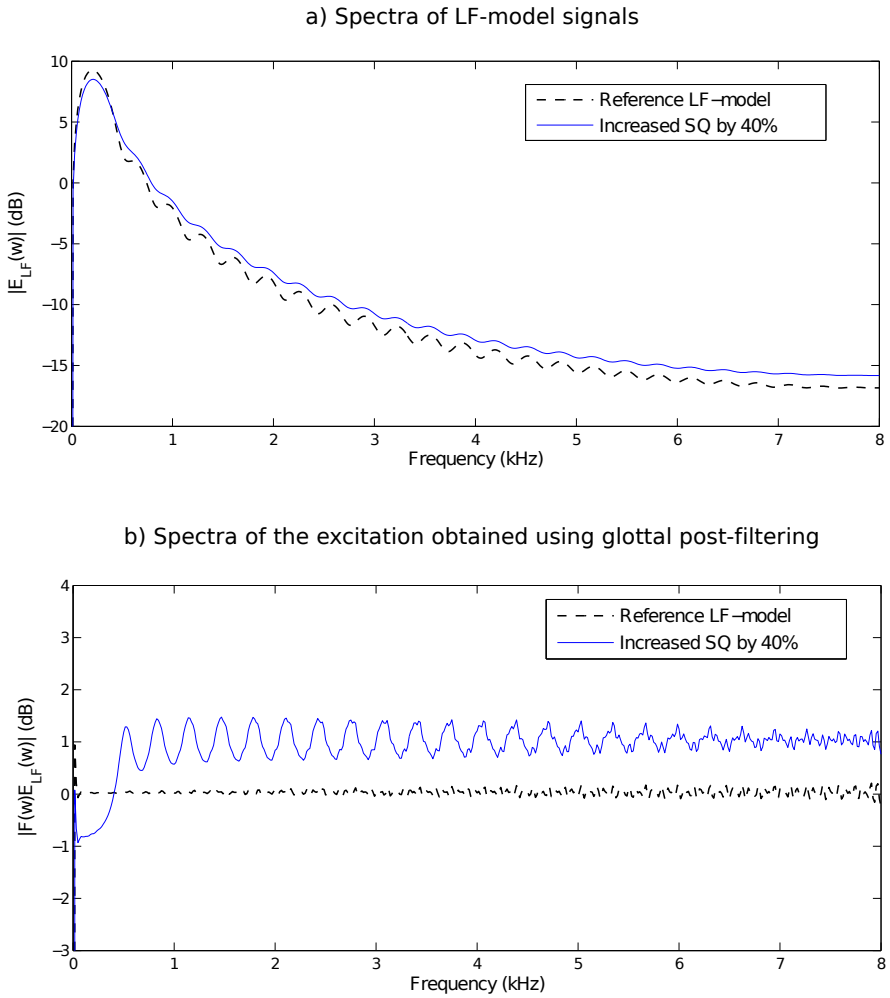

Figure 6.13: a) Spectra of the reference LF-model signal and its modified version with higher SQ; b) Spectra of the two glottal post-filtered LF-model signals.

6.4 Glottal Spectral Separation (GSS)

The GSS method developed in this thesis synthesises speech using an acoustic glottal source model and the vocal tract transfer function. In particular, this method was implemented using the LF-model to represent the glottal source.

6.4.1 Speech Model

The Glottal Spectral Separation (GSS) method assumes that voiced speech is the convolution of a glottal source signal with the vocal tract filter. In the frequency domain, this speech model can be represented by

$$S(w) = P(w)U(w)V(w)R(w), \tag{6.23}$$

where $P(w)$ is the FT of an impulse train, $U(w)$ is the FT of a glottal pulse, $V(w)$ is the vocal tract transfer function and $R(w)$ is the radiation characteristic, which can be modelled by a differentiating filter. In this work, $G(w) = U(w)R(w)$ is represented by a multi-band mixed excitation model, which is a model of the glottal source derivative. The LF-model was used in this work to represent the glottal source derivative in the GSS method, as described Section 6.5.

The speech production model of (6.23) is different from the model used by the LPC vocoder (Proakis and Manolakis, 1996) or STRAIGHT. These vocoders are based on the following model:

$$S(w) = P(w)H(w) \tag{6.24}$$

In this representation, the input excitation is represented by the impulse train and $H(w)$ represents the spectral envelope of $S(w)$. The vocal tract, the lip radiation and the glottal source effects are all incorporated into $H(w)$.

6.4.2 Analysis

The block diagram of the GSS analysis method is illustrated in Figure 6.14. The glottal source signal $v(t)$ is estimated from the speech signal $s(t)$ and the glottal parameters are extracted from $v(t)$. A *smoothing operation* on the glottal parameters is employed in order to reduce possible estimation errors. The smoothed parameters are then used to generate the spectrum of one glottal flow pulse, $E_p(w)$. This signal is equivalent to the spectral envelope of a periodic glottal source signal, $E(w)$, since it does not have harmonic components. Then, the spectral parameters are calculated by removing the spectral characteristics of the source from the speech spectrum and by estimating the spectral envelope of the resulting signal. In this work, the aperiodicity parameters and the spectral envelope are calculated using the STRAIGHT vocoder.

For separating the spectral properties of the glottal source from the speech, the speech spectrum is divided by the amplitude spectrum of one period of the glottal source derivative, $E_p(w)$. The FT of the resulting signal can be represented by $S(w)/E_p(w)$. From (6.23), this signal can be described by

$$\frac{S(w)}{E_p(w)} = P(w)V(w)\frac{U(w)R(w)}{E_p(w)} \tag{6.25}$$

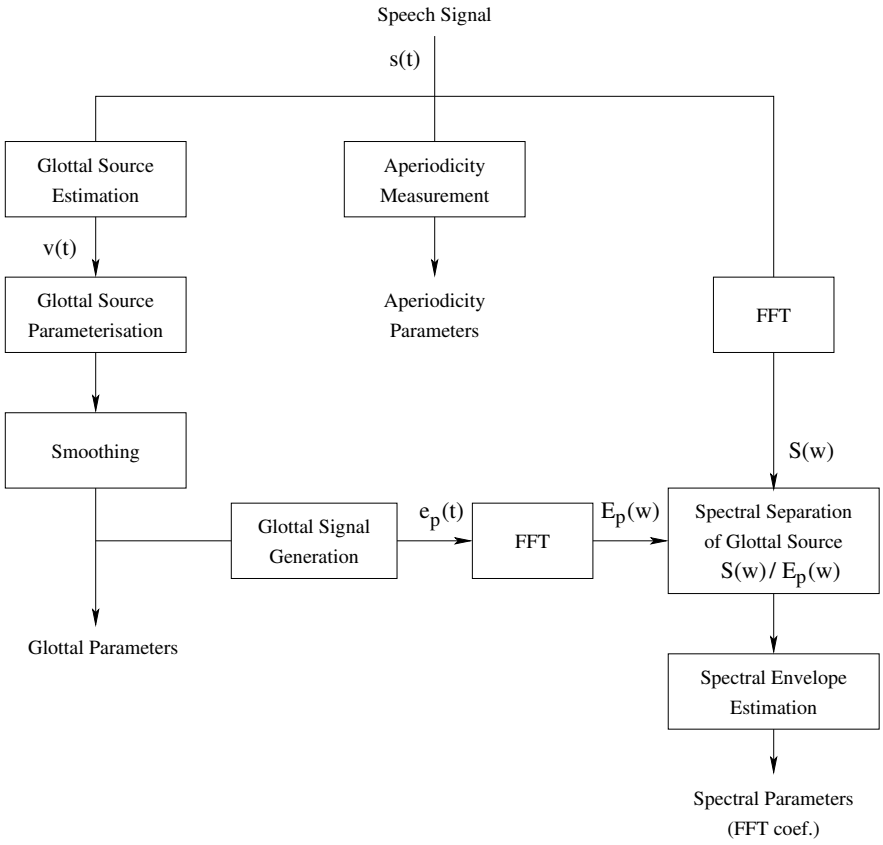

Figure 6.14: Block diagram of the analysis part of the GSS method.

Assuming that $R(w)$ is modelled by the derivative function and that the estimated $E_p(w)$ is a good approximation of the glottal source derivative, then $E_p(w) \simeq U(w)R(w) = G(w)$. Under this approximation, (6.25) can be rewritten as

$$\frac{S(w)}{E_p(w)} \simeq P(w)V(w) \tag{6.26}$$

This equation shows that the vocal tract filter $V(w)$ can be estimated as the spectral envelope of $S(w)/E_p(w)$, by comparison with the speech model of (6.24). This is how the GSS method estimates the vocal tract transfer function.

The GSS analysis could also be performed using a model of the glottal flow instead of its derivative. In this case, the glottal flow pulse generated from this model does not include the radiation effect, unlike $E_p(w)$. Then, the spectrum obtained using GSS is the combination of the vocal tract and the radiation effect, i.e. $V(w)R(w)$.

When the quotient between the speech and the source spectra is calculated, it is important that the duration of the glottal source signal is equal to the fundamental period. For example, if the glottal source signal is longer than the fundamental period, then its spectrum contains periodicity. A periodic source spectrum is not suitable for separat-

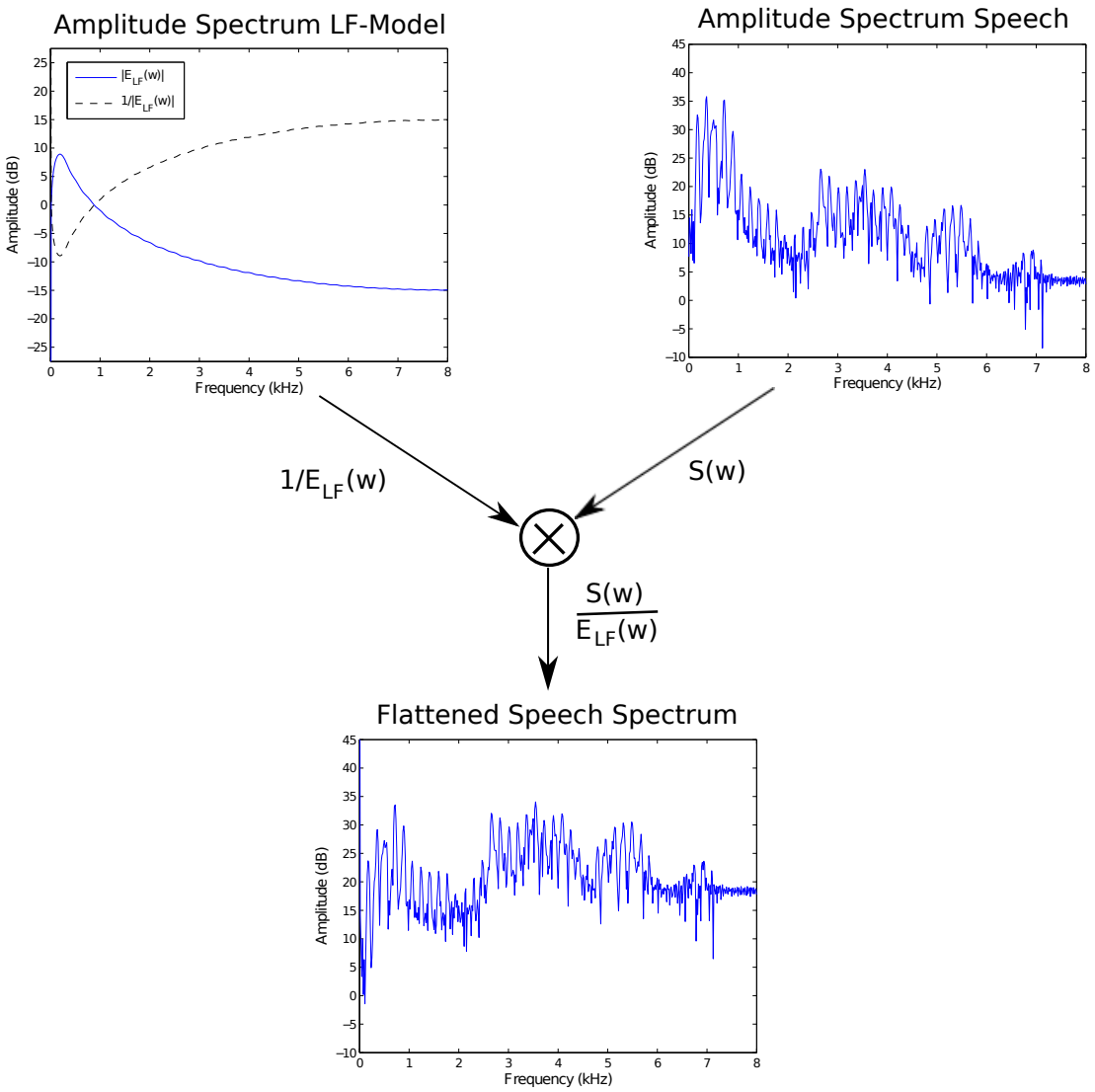

Figure 6.15: Separation of the LF-model amplitude spectrum from the speech signal. In this example, the spectrum of the speech signal is calculated by performing SFT on a 40 ms voiced speech segment and using a Hamming window with the same duration.

ing the glottal source effects from the speech signal, because the relative position of the source harmonics to the speech harmonics produces variations in the amplitude of the resulting spectrum.

Figure 6.15 shows an example of the separation of the LF-model spectral effects from the spectrum of a speech signal, $S(w)$. The overall slope of the resulting spectrum, $S(w)/E_{LF}(w)$, is close to zero (overall spectrum is approximately flat), because the spectral tilt of the LF-model has been removed from the speech spectrum. Figure 6.16 shows the spectral envelope of the signal $S(w)/E_{LF}(w)$, which was calculated using STRAIGHT. The estimated vocal tract is also flatter than the spectral envelope of the

original speech signal $S(w)$, due to the removal of the tilt characteristic of the LF-model. The frequency of the first maximum peak is also different between the two spectra because of the removal of the glottal peak characteristic of the LF-model by GSS. In general, the signal $S(w)/E_{LF}(w)$ has a high *DC component* due to the very low amplitude of $E_{LF}(w)$ near the zero frequency. This effect is because acoustic glottal source models typically have a DC value approximately equal to zero. The high DC component could affect the estimation of the spectral envelope. However, this problem is not relevant when using STRAIGHT to compute the spectral envelope, because it removes the DC component from the speech spectrum before computing the spectral envelope.

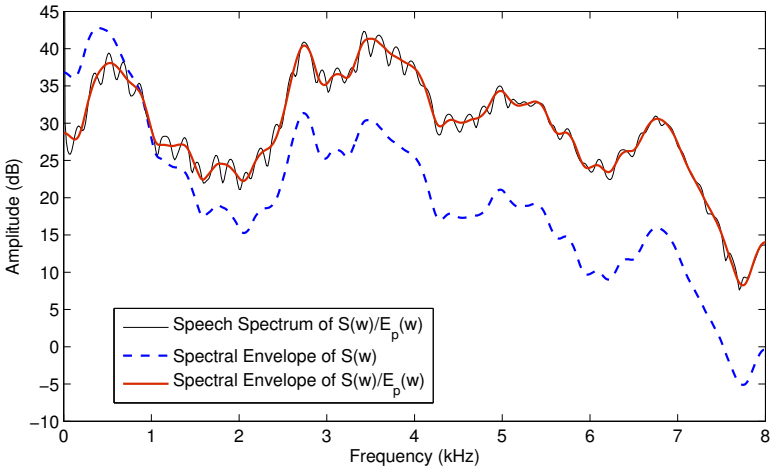

Figure 6.16: Spectral envelope of a 40 ms short-time speech signal calculated by the GSS method, using the LF-model and STRAIGHT. The LF-model spectral effects are first removed from the speech signal. This is the input signal to STRAIGHT, which calculates a speech spectrum with reduced periodicity and estimates the spectral envelope. The spectral envelope of the speech signal calculated only using STRAIGHT is also represented, for comparison.

6.4.3 Synthesis

6.4.3.1 Source-Filter Model

The GSS method synthesises voiced speech by using the following speech production model:

$$Y(w) = P(w)G(w)V(w), \tag{6.27}$$

where $P(w)$ represents the FT of a delta pulse train, $G(w)$ represents the FT of the glottal source derivative, $V(w)$ is the transfer function of the vocal tract filter and $Y(w)$ is the FT of the synthetic speech.

The vocal tract filter is defined by the spectral parameters estimated using the GSS method. For generating the source derivative, the following multi-band mixed model is used:

$$G(w) = E(w)W_p(w) + K_n N(w)|E_p(w)|W_a(w), \tag{6.28}$$

where $E(w)$ and $N(w)$ represent the FT of the periodic component of the glottal source derivative and white noise, respectively. $E_p(w)$ represents the spectral envelope of the glottal signal $E(w)$ and K_n is a scale factor to normalise the energy of the noise relative to the source signal. Finally, $W_p(w)$ and $W_a(w)$ are the weighting functions of the periodic and aperiodic components of the excitation, respectively. Figure 6.17 shows the flowchart of the speech synthesis method using this model.

Both $E(w)$ and $E_p(w)$ are calculated using the glottal parameters and F_0. The GSS method can be used with different types of glottal source models. In this work, the LF-model is used to represent the glottal source derivative signal. However, the glottal source model used for synthesis is expected to be the same as the model used in the GSS analysis. If the source signal represents the glottal flow signal instead of its derivative, the source-filter model described by (6.27) and (6.28) is still valid, because the radiation effect is included in the vocal tract filter.

Similarly to the GPF method, the synthetic speech frames are concatenated using the overlap-and-add technique with asymmetric windows approximately centered at the instants of maximum excitation of the LF-model signal, as described in Section 6.3.3.2.

6.4.3.2 Glottal Source/Noise Weighting

The weighting functions, $W_p(w)$ and $W_a(w)$, are calculated from the aperiodicity parameters. In this work, the aperiodicity measurements are estimated using STRAIGHT analysis. This vocoder applies $W_p(w)$ and $W_a(w)$ to the spectra of a delta pulse signal and white noise, respectively. Next, it adds them together to yield the mixed excitation, which is approximately flat. The delta pulse spectrum, $D(w)$, and the noise spectrum, $N(w)$, are approximately flat and have the same energy. The noise has power one (zero mean and unit variance noise), whereas the delta pulse has amplitude $\sqrt{N_0}$ so that it

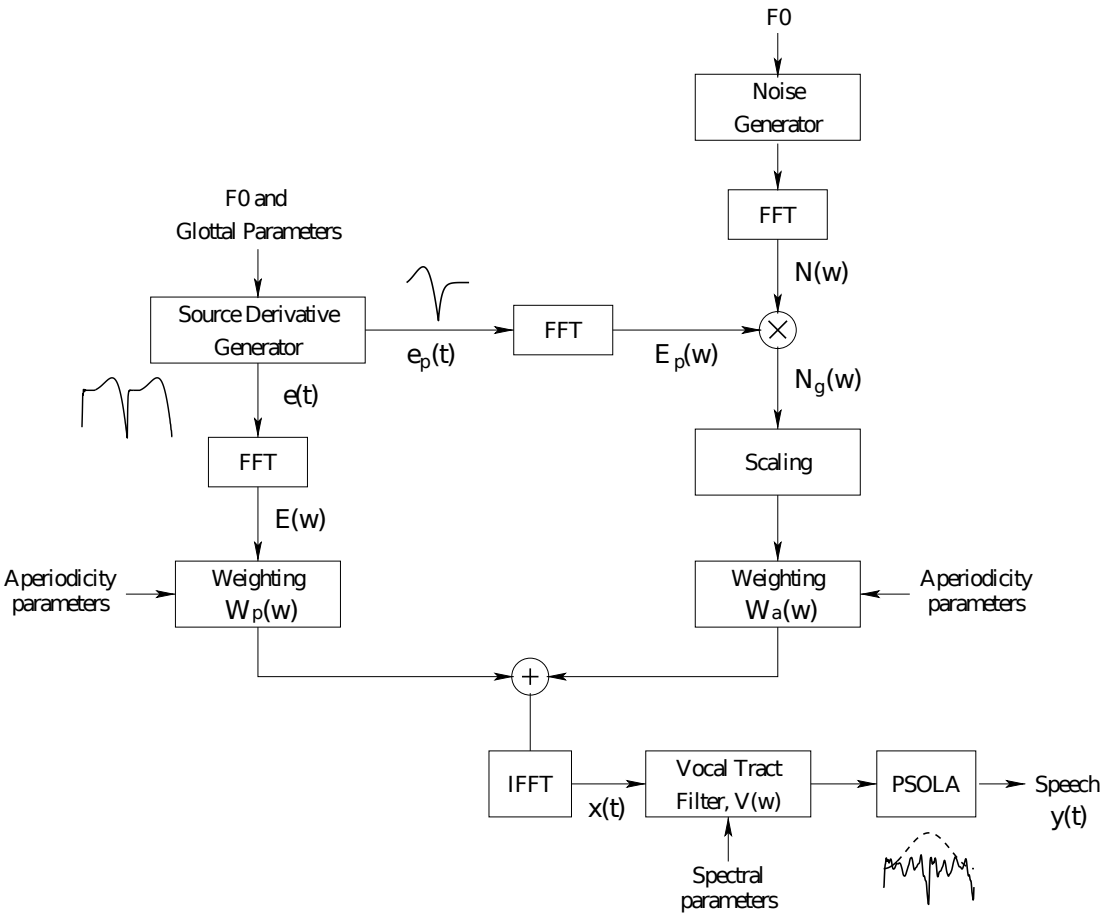

Figure 6.17: Block diagram of the speech synthesis method using the parameters estimated by the GSS method. The glottal source derivative waveform represented in this figure was obtained using the LF-model, as an example.

has the same power as the noise. The weighting operation has been explained in more detail in Section 6.2.3. The plots a) and b) of Figure 6.18 show the amplitude spectra of the two excitation components before and after the weighting, respectively.

In contrast to the delta pulse, the glottal source signal is not spectrally flat and its energy does not depend on the fundamental period only. In general, the shape of the glottal source waveform depends on all the glottal parameters and its energy varies with these parameters too. For this reason, either the glottal source signal or the white noise have to be transformed so that the weighting operation is performed correctly for synthesising speech using the GSS parameters. The solution proposed in this thesis to combine the glottal source signal with the STRAIGHT mixed excitation model is to shape the spectral envelope of the source derivative on the white noise before the weighting operation. The spectral envelope of the source can be described

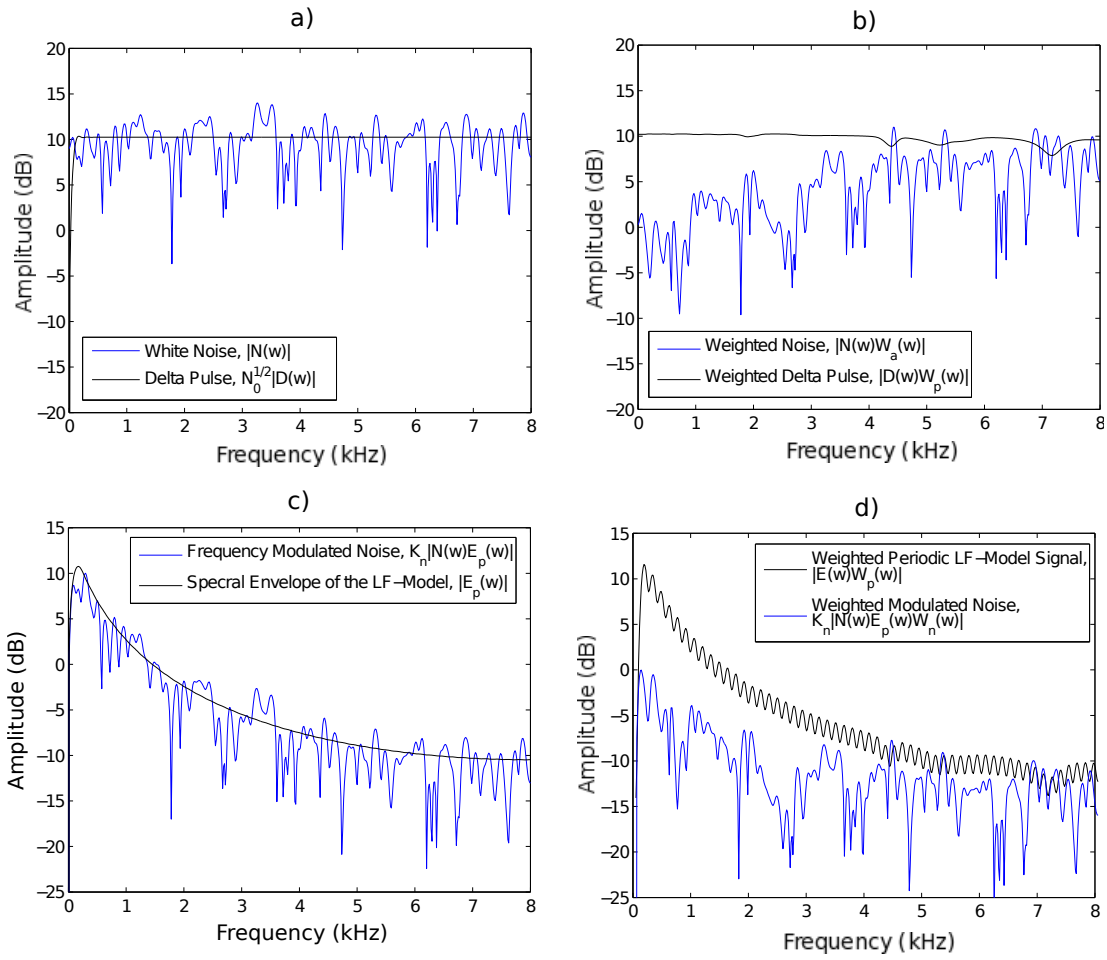

Figure 6.18: Weighting effect on the mixed excitation components using the STRAIGHT aperiodicity measurements and two types of periodic signals. In a) and b), the periodic component is represented by the delta pulse. In c) and d) the mixed excitation is generated using the LF-model: c) amplitude spectrum of white noise shaped by the spectral envelope of the LF-model, and d) effect of weighting on the modulated noise and LF-model periodic signal.

as the impulse response $D(w)E_p(w)$, in which $E_p(w)$ is the transfer function of one period of the glottal source signal. This technique can be represented by

$$N_g(w) = |E_p(w)|N(w), \tag{6.29}$$

where $N_g(w)$ is the *frequency modulated noise*. Figure 6.18 c) shows an example of $N_g(w)$, which was obtained using the LF-model signal as the *modulating signal*. Figure 6.18 d) shows the weighting effect on both the LF-model signal and the modulated noise. In this example, the amplitude spectrum of the LF-model component of the

excitation, $E(w)$, has harmonics because it consists of two cycles of the LF-model waveform.

Unlike the GPF method described in Section 6.3, which transforms a reference LF-model signal into a spectrally flat signal, this method transforms the noise and keeps the glottal source properties unchanged in the excitation. Therefore, the dynamic variations of the glottal characteristics, such as the source tilt and the glottal formant, can be modelled independently from the vocal tract spectrum.

6.4.3.3 Amplitude Scaling of the Noise

The periodic component of the excitation consists of two periods of the glottal source signal. The noise excitation has the same duration as the periodic excitation and it is scaled in amplitude for the two signals to have the same power. The white noise $N(w)$ (zero mean and variance one) has power equal to one, whereas the delta pulse train $P(w)$ has power $1/N_0$. The noise signal, $N_g(w)$, is multiplied by the scale factor $K_n = 1/\sqrt{N_0}$ so that is has the same power as the source signal, $P(w)E(w)$. It is important that the amplitude scaling is performed on the noise instead of the periodic component, in order to avoid the variation of amplitude parameters of the glottal source model. For example, if the LF-model is used to model the glottal source derivative and it is scaled in amplitude so that it matches the unit power of the delta pulse signal, then the amplitude of maximum excitation of the LF-model, E_e, is altered.

6.4.4 Voice Quality

The glottal parameters estimated using the GSS analysis method can be modified to transform characteristics of the glottal source signal used for generating speech. For example, by implementing the GSS method using the LF-model, the shape of the glottal source waveform used to synthesise speech can be easily modified. This method can be used to transform voice characteristics of the synthetic speech, as the glottal source parameters are strongly correlated with voice quality. Section 6.5.3 describes an application of the GSS method using the LF-model for voice transformation.

Also, speech synthesis using the GSS method does not have the limitations of the GPF method for voice transformation, as the glottal source waveform used by the GSS synthesis method is not transformed by the glottal post-filter.

6.4.5 GSS Compared with Other Analysis Methods

The separation and estimation of the glottal source and the vocal tract filter from the speech signal is a difficult problem to solve, as explained in Section 2.2.3. This is a blind separation problem, which is typically solved by making assumptions about the speech model. For example, the speech model is generally assumed to be linear, the vocal tract is often approximated by an all-pole filter and the glottal source by an acoustic glottal source model or pole-zero representation. However, interaction between the voice source and the vocal tract does also exist, which makes it even more complicated to accurately estimate the glottal source and vocal tract components.

The glottal source and the vocal tract filter are often estimated from the speech signal using an iterative method which estimates these signals jointly, such as the iterative inverse filtering and the glottal inverse filtering methods described in Sections 2.2.3.3 and 2.2.3.4 respectively. However, these methods typically use approximations which are not always valid or depend on the initial values and the convergence of optimisation algorithms. Another problem with methods which jointly estimate the source and the vocal tract is that errors in the estimated source signal affect the vocal tract estimation and vice-versa. As a consequence, the spectrograms of the vocal tract estimated using these methods are usually not as smooth as the spectrograms estimated by spectral envelope estimation methods like the one used by STRAIGHT.

The GSS analysis-synthesis method overcomes problems commonly found in the estimation of the source and the vocal tract filter by combining a method for glottal source estimation with a method for spectral envelope extraction. The main goal is to effectively separate the characteristics of a glottal source model from the speech signal and to obtain smooth parameter contours for both the vocal tract spectrum and the glottal source. It is important to correctly separate the glottal source and the spectrum parameters when they are modelled independently, e.g. by HMMs in a statistical speech synthesiser. Also, smooth parameter trajectories avoid distortion of the synthetic speech quality due to speech parameter discontinuities.

The GSS analysis method has several characteristics which are attractive for speech synthesis applications. The following aspects of this method could be advantageous compared with other methods which estimate the glottal source signal and the vocal tract filter from speech:

- Errors in the glottal parameter estimates can be reduced before the source-tract separation, e.g. using a smoothing technique.

- Vocal tract transfer function can be estimated using a spectral envelope extraction technique which permits a smooth spectrogram to be obtained, e.g. using STRAIGHT analysis.
- Smooth glottal parameter trajectories and smooth spectrogram can be recombined for synthesis of high-quality speech.
- Vocal tract estimation does not need to be pitch-synchronous.

GSS estimates the glottal parameters before the vocal tract is calculated. Therefore, errors in glottal parameter estimates can be attenuated before separating the source effects from the speech signal. In contrast, methods which calculate the source and the vocal tract parameters jointly, e.g. Alku et al. (1991); Fu and Murphy (2006), can only easily reduce discontinuities in the source parameter trajectories after the source-tract separation. Thus, the effect of glottal source estimation errors on the vocal tract estimation is difficult to avoid in these methods. These errors may cause discontinuities in the vocal tract parameter trajectories.

There are robust spectral envelope analysis methods which can produce a smooth spectrogram, e.g. the technique used by STRAIGHT. By using such a method in GSS, the estimated vocal tract spectrogram is expected to be smooth, under the assumption that the trajectories of the estimated glottal parameters are also smooth. Therefore, the extraction of smooth glottal source contours is a key factor to synthesise high quality speech using GSS.

Typically, accurate vocal tract estimation methods are pitch-synchronous or require the estimation of the closed phase. Such methods typically require a robust glottal epoch detector. For example, Wong et al. (1979) proposed to perform the LPC analysis on the closed phase in order to avoid errors caused by variations of the vocal tract during the pitch cycle or caused by source-tract interaction. Alku et al. (1991) also proposed to perform the LPC analysis over the pitch period in order to more accurately estimate the glottal source signal and the vocal tract. In contrast, the estimation of the spectral envelope does not usually require the detection of glottal instants. In this work, the GSS analysis is implemented using STRAIGHT to compute the spectral envelope. Although this vocoder uses a pitch-adaptive window to calculate the spectral envelope, it does not require glottal epoch detection. However, the implementations of the GSS method in this thesis use a pitch-synchronous technique to estimate the glottal parameters. The next section describes the first implementation of the GSS method using the LF-model and the STRAIGHT vocoder, which was performed in

this work. During this work, this implementation was also applied to a HMM-based speech synthesiser (as described in Section 7.3) and further improved (as described in Sections 8.2 and 8.2.2).

6.5 Application of GSS Using LF-model

6.5.1 Estimation of the LF-model and Vocal Tract

6.5.1.1 F_0 and Glottal Epochs

The fundamental frequency F_0 and the glottal epochs were estimated in the first stage of the GSS method, since they were used to estimate the glottal source derivative signal pitch-synchronously. The glottal epoch parameter corresponds to the maximal amplitude peak of the glottal flow derivative cycle (one period long), which is associated with the glottal closure instant. Therefore, the glottal epoch was also used as an estimate of the instant of maximum excitation of the LF-model, t_e.

Both F_0 and the glottal epoch were estimated using the F_0 and epoch detectors (Talkin and Rowley, 1990; Talkin, 1995) of the ESPS tools. F_0 values were calculated using the *get_f0* function, while the epochs were calculated using the *epochs* function and the estimated F_0 values. In this way, the extracted epochs were consistent with the F_0 values, i.e. epochs were only estimated for voiced speech ($F_0 > 0$). For unvoiced speech frames, the F_0 and epoch values were set equal to zero.

6.5.1.2 Glottal Source Signal Estimation

In this implementation of the GSS method, the inverse filtering technique with pre-emphasis was used for estimation of the glottal source derivative signal, $v(t)$. The inverse filter coefficients were estimated by LPC analysis of the pre-emphasised speech signal. This is a popular and simple method to obtain the LPC residual signal, which was described in Section 2.2.3.1.

Speech frames, $s^i(t)$, were sampled at 16 kHz and had duration equal to twice the fundamental period. The coefficients of the inverse filter were calculated pitch-synchronously (analysis window centered at the glottal epochs) from the pre-emphasised speech signal (α=0.97), using the autocorrelation method (order 18) and a Hanning window. Then, the *derivative of the glottal volume velocity* (DGVV), $v^i(t)$, was estimated by inverse filtering the short-time signal $s^i(t)$.

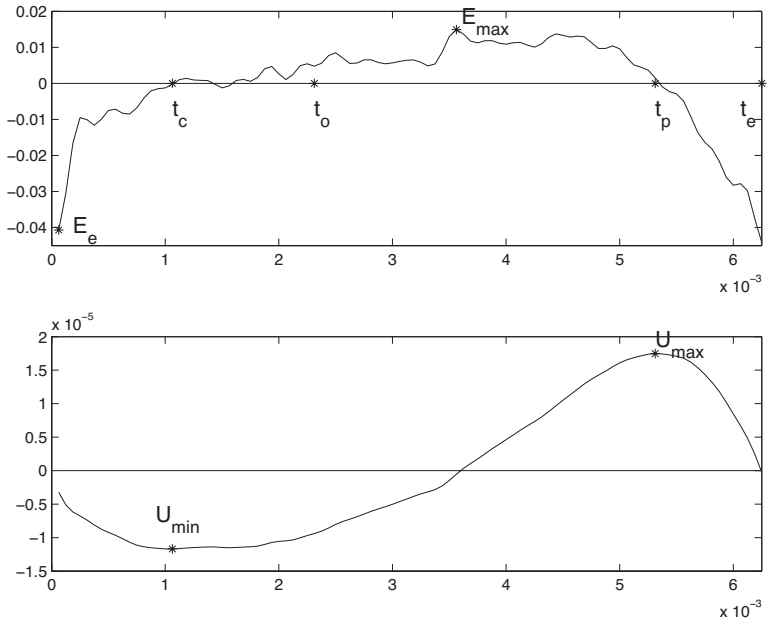

Figure 6.19: Estimates of t_c, t_o, and t_p. Top: a pitch cycle of the LPC residual; Bottom: integrated glottal source derivative signal (estimate of the glottal flow).

6.5.1.3 Initial Estimates of the LF-model Parameters

Initial estimates of the LF-model parameters, with the exception of t_e (estimated as the glottal epoch), were obtained by performing direct measurements on the estimated DGVV, $v^i(t)$. This short-time signal was one period long and delimited by two consecutive glottal epochs, which were indexed as $i-1$ and i, respectively. Afterwards, the estimated trajectories of the LF-parameters for each utterance were smoothed using the *median function*, in order to alleviate estimation errors. The E_e parameter was directly estimated from the residual signal as the absolute value of the amplitude of $v^i(t)$ at the glottal closure instant (glottal epoch $i-1$). Amplitude-derived measurements of the glottal flow and its derivative were also used to estimate the glottal opening, maximum flow and complete closure instants: t_o, t_p, and t_c parameters of the LF-model respectively.

The glottal flow signal, $g^i(t)$, was calculated by taking the integral of the short-time DGVV signal, $v^i(t)$. The DGVV signal was high-pass filtered by a linear phase FIR filter with cut-off frequency of 80 Hz prior to the integration, in order to reduce any effects of *low frequency amplitude fluctuation* that result from the integration. Next, the point of maximal flow amplitude U_{max} gave the instant t_p and the point of minimum flow amplitude U_{min} was the estimate of t_c. Figure 6.19 shows an example of the t_p and

t_c estimates. This method is based on techniques which estimate the glottal opening and closing instants from the EGG signal, e.g. Krishnamurthy and Childers (1986). The EGG signal is a measure of the vocal folds' conductivity during phonation. The closer the vocal folds are to each other, the higher the conductivity. The EGG signal has similar characteristics to the glottal flow signal and it is frequently used to estimate glottal source parameters.

Figure 6.19 shows an example of the t_o estimate. This parameter was calculated from U_{max}, U_{min}, and the maximal value of $v^i(t)$, E_{max}, using the following equation from Gobl and Ní Chasaide (2003):

$$t_o = \frac{2(U_{max} - U_{min})}{\pi E_{max}} \tag{6.30}$$

The short-time signal used to estimate the LF-model parameters started at the glottal epoch. Therefore, the instant of maximum excitation was assumed to be equal to zero, i.e. $t_e = 0$. However, the conventional LF-model signal starts at $t_o = 0$, instead of t_e. This was not a problem, because t_e was calculated from t_o as $t_e = T_0 - t_o$.

The estimated glottal source signals often have a noise component, e.g. caused by aspiration noise or ripple. Typically, the estimation of glottal parameters by direct measurements is affected by the noise of the source signal, as explained in Section 2.2.4. However, the amplitude-based measurements proposed in this section appeared to be robust to the noise characteristics of the glottal source derivative signal, such as aspiration noise and ripple.

The t_a parameter is defined as the time instant where the tangent (slope) to the decaying exponential function of the LF-model at $t = t_e$ hits the time axis. Figure 5.1 shows an example of the tangent in dashed line and the parameter t_a. This parameter usually is more difficult to estimate than the other LF-model parameters and few methods to directly estimate t_a from the glottal source signal can be found in the literature. Typically, it is estimated by fitting the LF-model signal to the DGVV signal. However, the performance of the optimisation algorithm also depends on a good initial estimate of this parameter. A simple method was developed in this work for estimating t_a. First, the derivative of the DGVV signal, $d\left[v^i(t)\right]/dt$, was calculated. Next, the peak of maximal amplitude of this signal over a relatively short-time interval starting at the glottal epoch was detected. Figure 6.20 shows an example of the estimated peak, which is represented by M. This amplitude M was the estimate of the decaying exponential slope at $t = t_e$, as this slope is maximum at $t = t_e$ (Figure 5.1 helps to visualise this property). Finally, T_a was estimated as $T_a = E_e/M$.

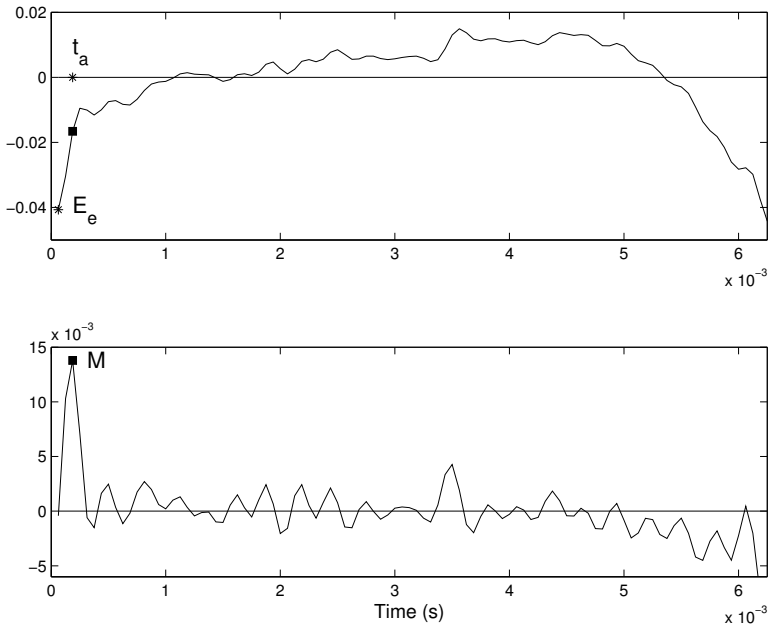

Figure 6.20: Estimation of t_a. Top: a pitch cycle of the LPC residual; Bottom: derivative of the residual signal.

Figure 6.21 a) shows the trajectories of the LF-model parameters estimated by the direct methods described in this section, for a segment of speech. The LF-model parameters are set equal to zero in the unvoiced regions, as they are not defined for unvoiced speech.

6.5.1.4 Optimisation of the LF-model Parameter Estimates

Methods based on fitting a voice source model to the data are often used to accurately estimate the glottal parameters. The t_o, t_p, and T_a parameters were estimated using an automatic method that fits the LF-model signal to the DGVV signal. In this application, the five parameter version of the LF-model (without t_c) was used, which is given by (5.4). Therefore, the instant of complete glottal closure t_c was not estimated.

The fitting method consisted of minimising the mean-squared error between the LF-model and the short-time signal, $v^i(t)$, using a non-linear optimisation algorithm. In this work, the *Levenberg-Marquardt algorithm* (Marquardt, 1963) was used to solve this optimisation problem. The initial estimates for this iterative method were the t_o, t_p, and T_a values estimated by direct methods (described in Section 6.5.1.3). The Levenberg-Marquardt method was implemented using the MATLAB function *lsqnonlin*. This function solves *non-linear least squares* problems of the form: $\min f^2(x)$,

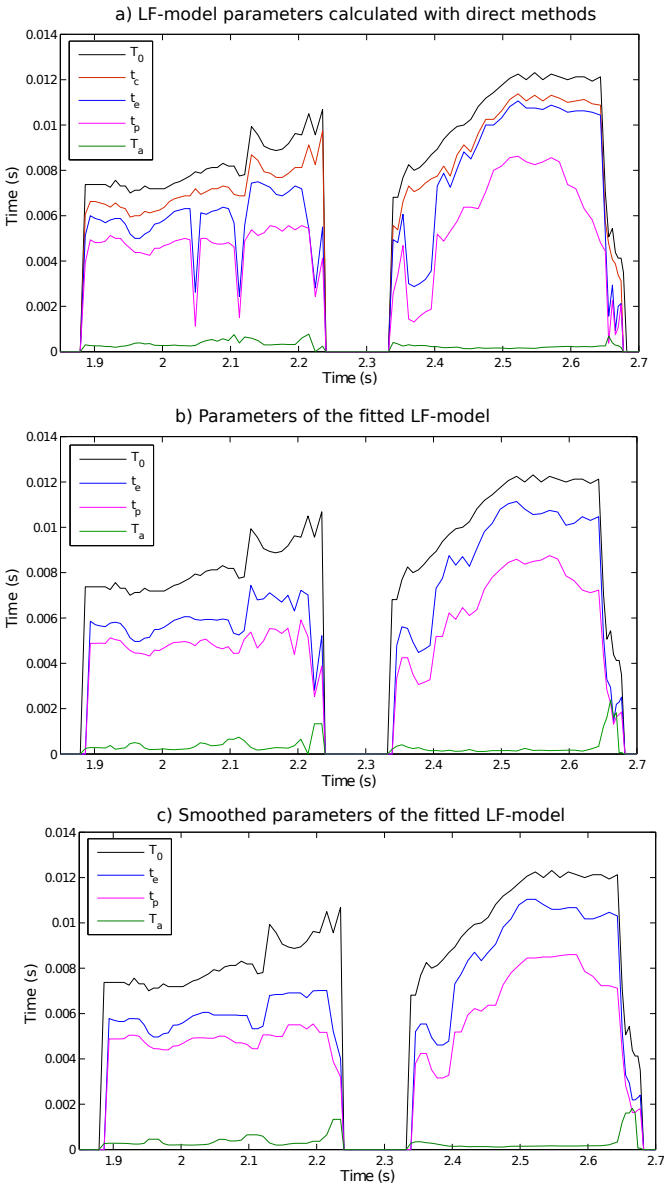

Figure 6.21: Trajectories of the LF-model parameters estimated for a segment of a recorded utterance. This segment corresponds to the words ”danger trail”, which is located approximately at the middle of the utterance. The T_0 contour is calculated using the F_0 detector of the ESPS tools. a) Trajectories estimated based on amplitude measurements of the glottal source derivative; b) Trajectories estimated by fitting the LF-model to the glottal source derivative; c) Smoothed trajectories of the parameters estimated by the fitting method.

where $f(x)$ is a *cost function*. In this work, the cost function used was $f(x) = e_{LF}(t) - v^i(t)$, where $e_{LF}(t)$ represents the LF-model signal. $e_{LF}(t)$ is a period of the LF-model which starts at $t = t_e$. This starting instant was chosen so that it coincides with the

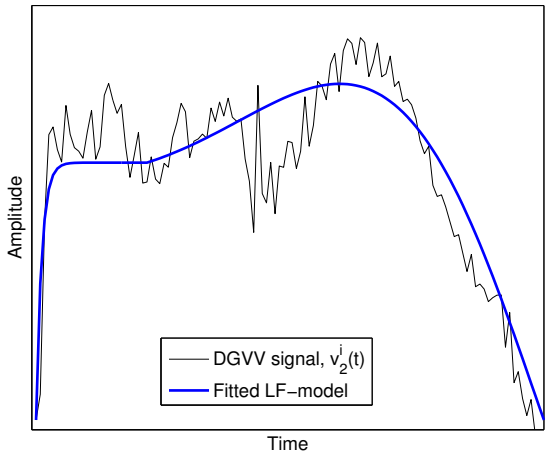

Figure 6.22: Example of the LF-model fitted to a short-time signal of the glottal source derivative signal.

glottal closure (glottal epoch) of the DGVV signal, $v^i(t)$. Note that t_e is different from the conventional starting instant of the LF-model, which is the glottal opening t_o. Figure 6.22 shows an example of the LF-model signal fitted to a DGVV short-time signal. After the fitting procedure, t_e was calculated as $t_e = T_0 - t_o$ (t_e is equal to the duration from the glottal opening instant to the instant of maximum excitation).

In the fitting method, the estimated instants of maximum excitation (epochs) were chosen as the starting and ending points of the LF-model waveform, because the glottal epochs were considered to be estimated more accurately than the other LF-model parameters (estimated using direct methods). As a consequence the LF-model parameter estimates using the fitting method depends on the performance of the glottal epoch detector. Nevertheless, the glottal epoch estimation method used in this experiment was assumed to be sufficiently robust and accurate.

Figure 6.21 b) shows LF-model parameter trajectories estimated for a segment of speech by using the fitting method. These trajectories are also smoothed by the median function. This operation reduces trajectory discontinuities caused by estimation errors. Figure 6.21 c) shows the smoothed trajectories of the LF-model parameters. The smoothed curve of the E_e parameter (amplitude of maximum excitation) is shown in Figure 6.23. In this example, E_e varies approximately in inverse proportion to T_0. This is consistent with the typical prosodic correlates of this LF-model parameter, which are described in Section 5.3.3.

From Figures 6.21 b) and c), a strong correlation between glottal parameters and T_0 can be observed, with the exception of the parameter T_a. The parameters t_e and

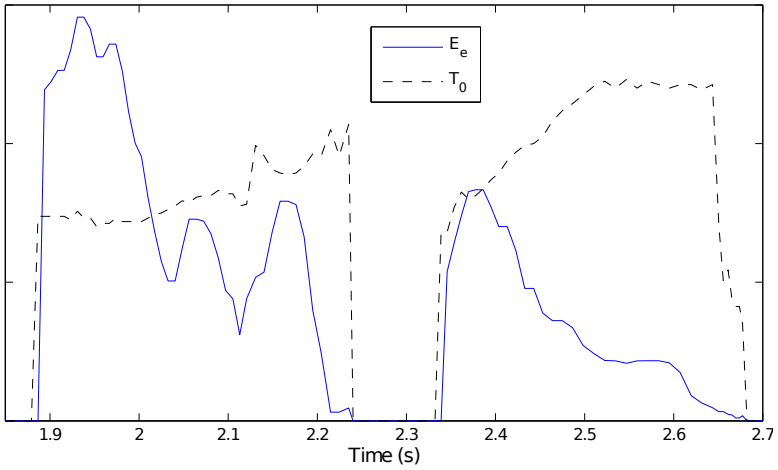

Figure 6.23: Trajectories of the E_e and T_0 parameters calculated for a segment of a recorded utterance. The segment corresponds to the words "danger trail", which is located approximately at the middle of the utterance. In this example, the T_0 parameter was scaled in amplitude by a constant factor for better comparing the T_0 and E_e trajectories.

t_p appear to vary in direct proportion to T_0, whereas the relationship between T_a and T_0 is not clear. There are also parts of the contours which show a different pattern of variation with T_0 than the linear. For example, a valley occurs on the trajectories of t_e and t_p from $t = 2.35$ to $t = 2.4$, whereas the T_0 contour has an approximately constant slope in this time interval. This might be related to variations of the glottal parameters related to prosodic aspects, such as syllable stress. These variations of the LF-model parameters with T_0 are expected, according to the prosody correlates of the LF-model parameters (discussed in Section 5.3.3).

6.5.1.5 Estimation of the Vocal Tract Spectrum

The spectral parameters were not estimated pitch-synchronously (using the glottal epochs). The speech signal was segmented at 5 ms frame rate into 40 ms long frames, $s^j(t)$, instead. This duration is equal to the default frame duration of STRAIGHT analysis. However, it was necessary to map each speech frame, $s^j(t)$, to a glottal epoch i, because the LF-model parameters were calculated pitch-synchronously for speech frames centered at the glottal epochs. This mapping was performed by finding the closest glottal epoch i to the center of each short-time signal $s^j(t)$. The set of LF-model parameter values associated with each selected epoch i was used to generate

one period of the LF-model signal, $e^i_{LF}(t)$, starting at the glottal opening instant t_o.

The vocal tract filter was estimated by removing the source model effects from the speech spectrum and calculating the spectral envelope of the resulting signal, as described in Section 6.4.2. Each speech frame $s^j(t)$ was multiplied by a Hamming window and zero-padded to have the length of 1024 samples, for the SFT analysis. The LF-model signal $e^i_{LF}(t)$ was also zero-padded to 1024 sample points. Next, the speech spectrum, $S^j(w)$, was divided by the amplitude spectrum of the LF-model signal, $\left|E^i_{LF}(w)\right|$, in order to remove the glottal source model effects. That is,

$$V^j(w) = S^j(w) / \left|E^i_{LF}(w)\right| \tag{6.31}$$

The spectral effects of $\left|E^i_{LF}(w)\right|$ are mainly related to the glottal formant and the spectral tilt characteristics of the LF-model. Finally, the STRAIGHT vocoder was used to calculate the spectral envelope of the signal $V^j(w)$. For unvoiced speech, the spectral parameters were estimated by computing the spectral envelope of $S^j(w)$ using STRAIGHT.

6.5.2 Copy-synthesis

The speech synthesis method using the parameters estimated by GSS was described in Section 6.4.3. Each voiced frame i of the excitation signal was generated by concatenating two periods of the LF-model waveform. They started at t_e and had durations T^i_0 and T^{i+1}_0, respectively. The first LF-model cycle was generated from the glottal parameters estimated for the frame i: t^i_e, t^i_p, T^i_a, and E^i_e. The t_e and t_p parameters of the second cycle were calculated under the assumption that the dimensionless parameters of the LF-model (OQ, SQ and RQ) were the same as the first cycle. That is, the glottal parameters are assumed to vary linearly with the fundamental period. For example, the t_p estimate for the second cycle was $\hat{t}_p = t^i_p T^{i+1}_0 / T^i_0$. This linear approximation for the variation of certain LF-model parameters is considered to be good because the variation of the dimensionless parameters between contiguous frames is generally not significant. The T_a and E_e parameters of the second cycle were set equal to the values of the first cycle respectively, as they did not show significant variation with T_0 from the analysis measurements. In this application of the GSS synthesis method, the LF-model signal is not mixed with the noise component. That is, the excitation of voiced speech consists of the periodic component only. The reason to exclude the effect of the noise component is to directly compare the LF-model signal against the impulse

train, because the noise component could reduce the buzziness of the synthetic speech caused by the impulse train and the LF-model signal. Moreover, the noise component is expected to have the same effect on the quality of speech synthesised using the STRAIGHT and GSS methods, because it is modelled using the same aperiodicity parameters.

The spectrum of the synthetic speech frame, $S^i(w)$, was calculated by multiplying the amplitude spectrum of the LF-model waveform by the vocal tract transfer function, which is given by the spectral parameters (FFT coefficients). In this process, the LF-model spectrum was calculated by performing the 1024 point FFT, using a Hamming window. The speech waveform was generated by computing the IFFT of $S^i(w)$ and removing the Hamming window effect from the resulting signal. Finally, the speech frames were concatenated using a pitch-synchronous overlap-and-add technique described in Section 6.4.3.1.

6.5.3 Voice Quality Transformation

In this application, the GSS analysis-synthesis method was used for voice transformation by modifying the LF-model parameters estimated for a speech signal and re-synthesising the speech signal using the new parameters. For synthesis, the F_0 and spectral parameters remained the same. Speech spoken with modal voice was transformed into breathy and tense voices by modifying the mean values of the OQ, SQ, and RQ parameters of the LF-model. This method is described in the following paragraphs.

First, the LF-model parameters were estimated for sentences spoken with three voice types: modal, breathy, and tense. Then, the mean values of the OQ, SQ, and RQ parameters of the LF-model were calculated for each utterance, by using the formulas given in Section 5.2.3. That is, these parameters were calculated for each speech frame i by:

$$OQ^i = \frac{t_e^i + T_a^i}{T_0^i} \tag{6.32}$$

$$SQ^i = \frac{t_p^i}{t_e^i - t_p^i} \tag{6.33}$$

$$RQ^i = \frac{t_a^i - t_e^i}{T_0^i} \tag{6.34}$$

The next step was to calculate the variations of the mean values of the dimensionless parameters between each voice quality and the modal voice. For example,

	Breathy			Tense		
	$\Delta\overline{OQ}$ (%)	$\Delta\overline{SQ}$ (%)	$\Delta\overline{RQ}$ (%)	$\Delta\overline{OQ}$ (%)	$\Delta\overline{SQ}$ (%)	$\Delta\overline{RQ}$ (%)
utt. 1	2.5	-7.7	17.6	-3.2	51.8	73.9
utt. 2	10.0	0.2	13.2	-8.3	30.4	39.2
utt. 3	-2.5	-23.8	10.1	-6.4	15.4	24.4
utt. 4	5.6	-16.4	51.2	-4.04	14.0	47.8
utt. 5	5.9	-16.8	62.7	-6.3	1.7	24.8

Table 6.1: Percentage variation of the mean values of the LF-model parameters between a sentence spoken with a voice quality (breathy or tense) and the same sentence spoken with modal voice. For example, the variation of the mean OQ for an utterance spoken with breathy voice is calculated as $\Delta\overline{OQ}_{breathy} = (\overline{OQ}_{breathy} - \overline{OQ}_{modal})/\overline{OQ}_{modal}$.

the variation of the mean value of the OQ for the breathy voice is $\Delta\overline{OQ}_{breathy} = E[OQ_{breathy}] - E[OQ_{modal}]$, where $E[x]$ represents the mean computed over the total number of speech frames of an utterance. Table 6.1 shows the variation of the mean values of the dimensionless parameters, which were calculated for five different utterances. These values are given in terms of percentage of the modal voice mean values. In general, the breathy voice had higher $\overline{OQ}$, lower $\overline{SQ}$, and higher $\overline{RQ}$ than the modal voice. This behaviour observed for the LF-model parameters is in agreement with the voice quality correlates of these parameters, which were described in Section 5.3.2. For the tense voice, the five utterances had lower $\overline{OQ}$, higher $\overline{SQ}$, and higher $\overline{RQ}$ than the modal voice. These results are also in acordance with the voice quality correlates, with the exception of the RQ parameter, which is typically lower for the tense voice compared with the modal voice. One possible explanation for this unexpected result is the limitation of inverse filtering using pre-emphasis (described in Section 2.2.3) to accurately estimate the DGVV signal. A major problem with this technique is that it does not correctly separate the spectral tilt of the glottal source from the speech signal. The RQ parameter is particularly affected by poor modelling of the spectral tilt by inverse filtering (using pre-emphasis), because this parameter is strongly correlated with

the spectral tilt. The OQ and SQ are less influenced by the spectral tilt, which might explain the expected variations of these parameters for the tense voice.

The range of the $\overline{OQ}$, $\overline{SQ}$, and $\overline{RQ}$ values in Table 6.1 (calculated for the five utterances) is relatively large for the two voice qualities (breathy and tense). One of the explanations for this result is that the values of the dimensionless parameters vary significantly along an utterance and across utterances because they also depend on prosodic factors, as explained in Section 5.3.3. Another factor is that it might be difficult for the speaker to reproduce the same type of voice quality along an utterance and for the different utterances. Estimation errors of the LF-model parameters could also contribute to the high variance values. Nevertheless, the voice quality transformations were performed for each utterance using the values of $\overline{OQ}$, $\overline{SQ}$, and $\overline{RQ}$ calculated for that utterance. For this reason, the variations of these parameters across utterances was not considered to be important. Also, the general trend of variation of these parameters (whether they increase or decrease) is similar for the different utterances as discussed in the previous paragraph.

The transformed trajectories of the LF-model parameters were obtained by multiplying the measurements of the glottal parameters of the modal voice by scale factors, so as to reproduce the target variation of the voice quality parameters (mean values of OQ, SQ, and RQ). The formulas used to calculate the scale factors were derived from the formulas of the voice quality parameters, given by (6.32) to (6.34), and from the deltas of the mean values of the voice quality parameters. For example, for transforming the voice quality of the speech frame i, from modal to breathy, the scale factors are given by

$$k_{T_a}^i = 1 + \frac{\Delta\overline{OQ}_{breathy}}{RQ^i} \tag{6.35}$$

$$k_{t_p}^i = \frac{t_e^i}{t_p^i}\frac{\Delta\overline{SQ}_{breathy} + SQ^i}{1 + \Delta\overline{SQ}_{breathy} + SQ^i} \tag{6.36}$$

$$k_{t_e}^i = \frac{T_0^i}{t_e^i}(\Delta\overline{OQ}_{breathy} + OQ^i) - \frac{k_{T_a}^i T_a^i}{t_e^i} \tag{6.37}$$

The scale factors used to transform a modal voice into a tense voice were also calculated using the previous equations, but the delta parameters derived for the tense voice ($\overline{OQ}_{tense}$, $\overline{SQ}_{tense}$, and $\overline{RQ}_{tense}$) were used instead of the breathy parameters. Figure 6.24 shows the estimated trajectories of the LF-parameters for a segment of speech spoken with modal voice and the transformed trajectories for synthesising that speech

segment with breathy voice. The main effect of scaling the LF-parameters using (6.35) to (6.37) is to change the mean component of the LF-parameter trajectories, while the dynamic component of the LF-parameter trajectories remain approximately unchanged. Thus, the local aspects of voice quality which are correlated with prosody are preserved, such as voice quality variations in stressed syllables. On the other hand, the mean values of the LF-model parameter trajectories which are expected to be related to the overall voice quality of the utterance are modified by the scaling operations.

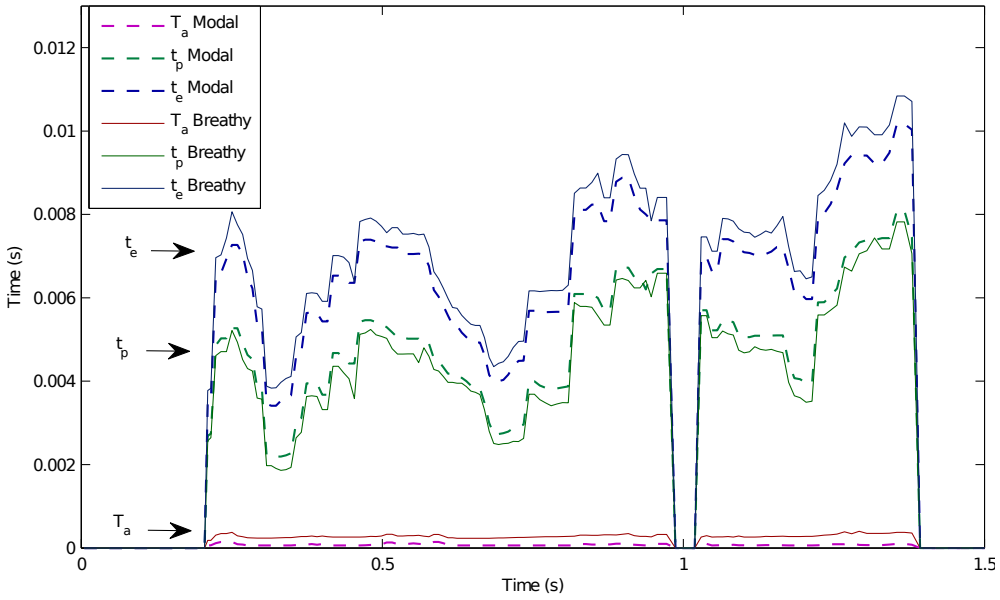

Figure 6.24: Estimated trajectories of the LF-parameters for an utterance spoken with a modal voice and the respective transformed trajectories which were calculated to synthesise speech with a breathy voice.

6.6 Perceptual Evaluation of GSS Using LF-model

6.6.1 Overview

A forced-choice (AB) perceptual evaluation was conducted in order to compare the LF-model with the impulse train, with respect to speech naturalness and parametric flexibility for voice quality transformations. Speech was generated by copy-synthesis using the GSS implementation with the LF-model, which was described in Section 6.5.2. This method is suitable for comparing the LF-model with the impulse train because the spectral parameters used to synthesise speech with the two excitation models can be calculated using the same spectral envelope estimation technique. In addition to the

comparison of these two excitation models, this experiment also permitted the evaluation of the performance of the GSS method for copy-synthesis and voice quality transformations. Table 6.2 summarises the characteristics of the methods evaluated in this evaluation.

	Analysis-Synthesis Methods	
	GSS with LF-model	Baseline Method
Analysis	Inv. Filt. Pre-emphasis: LF-param. ESPS tools: F_0, epochs GSS: vocal tract	ESPS tools: F_0, epochs STRAIGHT: spectral envelope
Excitation	LF-model	Impulse
Synthesis	GSS synthesis	FFT process. & OLA
Evaluation	Naturalness, Voice quality	

Table 6.2: Summary of the forced-choice (AB) perceptual test which was conducted to compare the LF-model with the impulse train.

6.6.2 Recorded Speech

A male English speaker was asked to read ten sentences with a modal voice and two different voice qualities: breathy and tense. He had listened to examples of tense and breathy speech beforehand, which were obtained from the following University of Stuttgart webpage: `http://www.ims.uni-stuttgart.de/phonetik/EGG/page10.htm`. The sentences contained only sonorant sounds, as the study concerned voiced speech. The use of other sounds, such as voiced fricatives and unvoiced speech could decrease the performance of the epochs detector and increase the errors in the estimated LF-parameters.

6.6.3 Synthetic Speech

Each utterance spoken with modal voice quality (neutral quality) was synthesised by copy-synthesis using the GSS method, as described in Section 6.5.2. This method uses

the LF-model to represent the glottal source derivative and the STRAIGHT vocoder to compute the spectral envelope. The modal voice utterances were also synthesised using the impulse train instead of the LF-model. The speech synthesis method using the impulse train was similar to the GSS method using the LF-model, with the exception that the LF-model waveform was replaced by a delta pulse and the spectral parameters represented the spectral envelope of speech (computed by STRAIGHT) instead of the vocal tract. The delta pulse was placed at the instant of maximum excitation t_e (approximately at the center of the excitation), and had amplitude equal to $\sqrt{T_0}$. The F_0 values were the same for the two speech synthesis methods (estimated using the ESPS tools).

Five sentences from the recorded speech corpus were also synthesised with breathy and tense voices respectively by transforming the LF-model parameter trajectories of the modal voice using the voice transformation method described in Section 6.5.3. These transformations were performed using the Δ values measured for these utterances which are given in Table 6.2. In this experiment, speech synthesised using the voice transformation method was compared to the resynthesised modal speech only, because the main objective of the experiment was to show that the LF-model provides more parametric flexibility for voice transformation than the impulse train. In the future, more experiments could be conducted to better evaluate the performance of the GSS method using the LF-model for voice transformation. For example, the transformation of modal speech to reproduce a certain non-modal quality (e.g. breathy) could be also compared to non-modal speech resynthesised using the GSS method or recorded speech spoken with the same non-modal voice.

6.6.4 Experiment

6.6.4.1 Lab Experiment

The experiment was first conducted in a quiet room of the CSTR lab, using headphones. Twenty three students, who were all English native speakers, were paid to participate in the test.

The listening test was divided into five parts. In the first, subjects were presented with 20 pairs of stimuli (10 utterances, randomly chosen and repeated twice with the order of the samples alternated). Each pair consisted of a sentence synthesised using the LF-model and the same sentence synthesised using the impulse train. For each pair, they had to select the version that sounded more natural. Each synthetic utterance

used in the test had been previously scaled in amplitude to have the absolute value of the maximal amplitude equal to that of the recorded utterance.

The second and third parts of the test were similar to the first, but the recorded speech was compared to speech synthesised using the impulse train and speech synthesised using the LF-model, respectively.

In the fourth part, listeners were first presented with two pairs of recorded utterances in order to show them the difference between modal and tense voices. This test consisted of 10 pairs, corresponding to 5 different sentences. Each pair contained a sentence synthesised with modal voice (by copy-synthesis) and the same sentence synthesised with the transformed trajectories of the LF-parameters which were calculated for the tense voice. Subjects had to select the speech sample that sounded most similar to the tense voice. Finally, the fifth part was similar to the fourth, with the difference that sentences synthesised with breathy voice were used instead of sentences synthesised with tense voice. In this part, listeners were asked to select the speech sample that sounded most similar to breathy voice.

6.6.4.2 Web Experiment

The same experiment was also conducted on the web, after the lab evaluation. Twelve listeners participated in the test, using headphones. The listening panel consisted of students and staff from the University of Edinburgh, including seven speech synthesis experts and ten native speakers. No payment was offered to the participants in this experiment.

For the web experiment, each synthesised utterance was multiplied by a scale factor so that the total speech power of the utterance was equal to the total power of the respective recorded utterance. This amplitude scaling was different from the one used in the lab test. The reason for this adjustment was to reduce the difference in loudness between the synthetic and the recorded utterances of each pair, which was found in the stimuli after the lab test had finished. The recorded utterances were systematically perceived as louder than the synthetic speech. By using the power normalisation that difference in loudness was reduced.

6.6.5 Results

The results obtained from the lab and web listening tests are shown in Figure 6.25. All the results are statistically significant with $p\text{-}value \leqslant 0.01$.

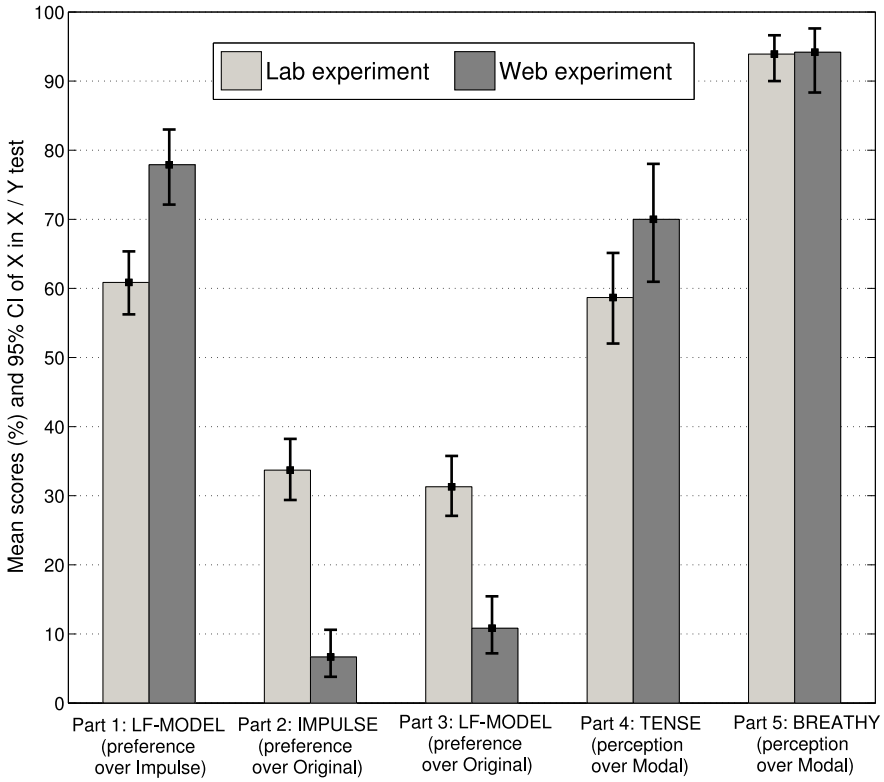

Figure 6.25: Preference rates and 95% confidence intervals obtained for each part of the forced-choice test.

In general, speech synthesised using the LF-model sounded more natural than speech synthesised using the impulse train. The preference for the LF-model was significantly higher in the web test than in the lab evaluation. In the web test, the participation of speech synthesis experts and the power normalisation of the speech samples are possible causes of the difference in results to the lab test. The results obtained in the two experiments were expected because the impulse train produces a buzzy speech quality, whereas that effect is attenuated by using the LF-model to represent the excitation.

Synthetic speech obtained higher scores than expected when compared to recorded speech, especially in the lab test. This result was unexpected, since the LF-model does not represent all the details of the true glottal source signal. For example, the LF-model cannot model certain voice effects such as aspiration noise, which is often perceived in voiced speech.

A detailed analysis of the lab test results showed that six listeners clearly preferred the synthetic speech to the recorded speech. The same listeners also clearly preferred speech synthesised using the impulse excitation to the LF-model. An ex-

planation might be that a small number of listeners (six out of ten) preferred speech spoken with a more buzzy voice quality over the natural voice of the speaker. Another explanation might be that the differences in loudness, which were observed between speech samples used in the lab test, influenced the perception of speech naturalness for some listeners. Further experiments are necessary to test these hypothesis. However, the differences between the results of the lab and web tests were not investigated because in both experiments the results showed a significant improvement of the speech quality by using the LF-model instead of the impulse train.

The unexpectedly good results obtained by synthetic speech in the comparisons against natural speech also indicate that the GSS synthesis method can produce high-quality speech by copy-synthesis, either using the impulse train or the LF-model.

Speech synthesised using the transformed LF-parameter trajectories to reproduce a breathy voice quality almost always sounded more breathy than speech synthesised using the estimated trajectories for modal voice. The results obtained for speech synthesised using the transformed LF-parameter trajectories to reproduce a tense voice quality were not as good as those obtained for breathy voice. A possible reason to explain this result is that speech features other than the LF-parameters are important to correctly model this voice quality, e.g. the F_0 parameter. Another factor which could have negatively affected the results for tense voice is related to possible errors in the estimation of the return phase parameter, T_a. This might be an important factor because the measured variation of the mean return quotient ($RQ = T_a/T_0$) between modal and tense speech was different from that expected, as explained in Section 6.5.3. The accuracy of the LF-model parameter estimation method was not evaluated in this work. However, the LF-model signal seemed to fit well to the estimated glottal source derivative signal in several utterances used in this experiment, from the informal analysis of these utterances by the author.

6.7 Conclusions

Three different analysis-synthesis methods have been described in this chapter. One of them is the STRAIGHT vocoder. The other two were developed in this work in order to use an acoustic glottal source model for synthesising speech, instead of the impulse train used by STRAIGHT. These methods are called Glottal Post-Filtering (GPF) and Glottal Spectral Separation (GSS) respectively. Table 6.3 summarises the main characteristics of these methods with respect to the type of parameters extracted

during analysis, the speech synthesis technique and the control over glottal source characteristics.

	Analysis		Synthesis	Control
	Source	Spectrum		Glot. Source
STRAIGHT	F_0	aperiod. meas., spectral envel.	Pulse phase proc., MBE gener., Min. phase filter.	None
Glottal Post-filter.	F_0, Post-filt. coeffic.	aperiod. meas., spectral envel.	Post-filt. LF-model, MBE gener., FFT proc. & PSOLA	LF-model variations
Glottal Spectral Separation	F_0, Glottal param.	aperiod. meas., spectral envel., vocal tract	Glottal source model, MBE gener., FFT proc. & PSOLA	glottal source model

Table 6.3: Summary of the characteristics of the analysis/synthesis methods.

One of the main advantages of STRAIGHT compared with other vocoders, such as the LPC vocoder, is that it calculates a smooth spectrogram of the speech signal, by effectively removing the periodicity characteristic of voiced speech. However, STRAIGHT uses a delta pulse to model the periodic component of the mixed excitation. This signal does not represent the glottal source characteristics and its spectrum has strong harmonics, which are typically associated with a buzzy speech quality. STRAIGHT reduces the buzziness effect by processing the phase of the delta pulse and by using a multi-band mixed excitation (MBE) model. This model consists of weighting the periodic and noise components of the excitation, in the frequency domain, and adding them together. Speech is synthesised by shaping the mixed excitation with the spectral envelope, which is described by a minimum-phase filter.

The GPF and GSS methods represent the periodic part of the excitation using a different signal from the impulse train. The periodic excitation signals used by these methods allow glottal source characteristics to be controlled and have a spectrum with

less periodicity than the impulse train. Although the GPF and GSS methods use a different excitation model from the model used by STRAIGHT, this vocoder can be easily incorporated into these methods to extract a smooth spectrogram.

The GPF method passes a chosen LF-model signal through a glottal post-filter, to transform the LF-model signal into a spectrally flat signal. This signal is used to model the periodic component of the mixed excitation. Speech is synthesised by shaping a spectrally flat excitation signal with the spectral envelope, as in STRAIGHT. However, instead of using the minimum-phase filtering technique of this vocoder, speech is synthesised multiplying the FFT parameters of the excitation with those of the spectral envelope. The resulting short-time speech signals are overlapped-and-added using windows centered at the instants of maximum excitation of the LF-model, t_e. A great advantage of this method, when compared with STRAIGHT, is that the input LF-model to the glottal post-filter can be changed during speech synthesis for voice transformation. Nevertheless, the control over the glottal source properties has some limitations because the effects of the LF-model parameter variations on the speech signal depend upon the glottal post-filter used.

The GSS method models speech as the convolution of the glottal source signal and the vocal tract filter. The vocal tract transfer function is estimated by separating the spectral effects of the glottal source from the speech signal and then calculating the spectral envelope of the resulting signal. This method uses a mixed excitation model which consists of mixing a glottal source signal with a random signal, in order to better model the noise characteristics of speech. For generating a short-term speech signal, the mixed excitation is convolved with the vocal tract filter (using FFT processing). Then, the short-term speech signals are concatenated using a pitch-synchronous overlap-and-add technique. The performance of the GSS analysis is mainly dependent on the glottal source estimation, as the spectral envelope can be estimated using a robust analysis method (e.g. the STRAIGHT method). Moreover, the effect of glottal parameter errors on the estimation of the vocal tract filter can be reduced, e.g. by performing a smoothing of the glottal parameter contours.

A forced-choice AB listening test was performed in order to compare the LF-model with the impulse signal, in terms of speech naturalness and voice quality transformation. In this evaluation, the GSS method was used to synthesise speech using the LF-model by copy-synthesis and to transform the voice quality of the synthetic speech. The GSS method performed well in the evaluation, which indicates that it can be used to produce high-quality speech. The results of this evaluation indicate that speech qual-

ity can be improved by using the LF-model instead of the impulse signal for synthesising speech. In this test, the LF-model was compared to the impulse signal, without using the noise component of the mixed excitation. These results also show that the LF-model offers higher parametric flexibility than the impulse train to model voice quality.

Chapter 7

HMM-based Speech Synthesiser Using LF-model: HTS-LF

7.1 Introduction

The LF-model was incorporated into a HMM-based speech synthesiser which uses the STRAIGHT analysis-synthesis method. This system is an implementation of the Nitech-HTS 2005 speech synthesiser (Zen et al., 2007a). Nitech-HTS 2005 is a very popular speaker-dependent HMM-based speech synthesiser, which performed very well against other speech synthesisers in the Blizzard Challenge 2005 (Zen et al., 2007a). The *Blizzard Challenge* is an annual event in which participants are provided with a speech corpus and have to synthesise a set of test utterances. Then, an overall evaluation of the synthesisers is conducted and the results can be examined in the *Blizzard Challenge Workshop*. Nitech-HTS 2005 has been used as the reference HMM-based speech synthesiser in the Blizzard Challenge since 2006. Another motivation for using a system similar to the Nitech-HTS 2005 is that it is an improved version of the HTS version 2.1 (Tokuda et al., 2009), which is publicly available for research purposes. More recent HMM-based speech synthesisers have been proposed, which obtained better results than the Nitech-HTS 2005 system, e.g. Yamagishi et al. (2007b). However, these systems are typically speaker-independent and are not publicly available. The speaker-independent approach is commonly used to synthesise multiple speakers' voices and typically requires a larger speech corpus (with speech from different speakers) than the speaker-dependent approach. In this work, the speaker-dependent approach was chosen because this research concerns the synthesis of a single speakers' voice. Furthermore, the speech quality obtained with a

speaker-dependent speech synthesiser is comparable to that obtained with a speaker-independent system, if the size of the speech corpus which is used to build the first type of synthesiser is large enough.

The HMM-based speech synthesiser using the LF-model which was developed during this thesis is called HTS-LF. This system uses the Glottal Spectral Separation (GSS) method (described in Section 6.4) for speech analysis and synthesis, instead of the STRAIGHT vocoder used by the *baseline system* (implementation of the Nitech-HTS 2005 system). The statistical modelling part of the baseline system was also modified to incorporate the LF-model parameters. This adjustment mainly concerned the structure of the statistical model, while the HMM training methods remained approximately the same.

This chapter first describes the baseline HMM-based speech synthesiser which uses STRAIGHT for analysis and synthesis. Then, the parts of the HTS-LF system which are different from the baseline system are described in Section 7.3.

7.2 Baseline System

The structure of the baseline HMM-based speech synthesiser which uses STRAIGHT is similar to that of the HTS system, which was described in Section 3.4.1. In this work, this baseline system is named HTS-STRAIGHT.

The HTS-STRAIGHT system analyses the text sentences of the speech corpus in order to extract the phonetic labels and contextual factors. In this process, the system generates *context-dependent labels* for English using the text analysis tools of the FESTIVAL unit-selection speech synthesiser (Black et al., 2004). The factors of the contextual labels are the same as those used in the conventional HTS system (Tokuda et al., 2002). Examples of these parameters can be found in Section 3.4.2. The HTS-STRAIGHT system also analyses the recorded speech to estimate the excitation and spectral parameters. The excitation parameters are F_0 and *aperiodicity weights* in five frequency bands, while the spectral envelope parameters are mel-cepstral coefficients. The aperiodicity measurements and the spectral envelope are computed using the STRAIGHT analysis method. The phonetic and speech parameters are then used to train the context-dependent HMMs and decision trees are used to cluster the trained statistical models. For speech synthesis, the parameter generation algorithm uses the statistical models to generate speech parameters from the input text. Finally, speech is generated from the excitation and spectral parameters using the STRAIGHT synthesis

method.

7.2.1 STRAIGHT Analysis and Synthesis

The baseline speech synthesiser uses the MATLAB programs of STRAIGHT for analysis and synthesis. This system requires higher computational time and memory for running these MATLAB programs, compared with the HTS system (version 2.1) which uses mel-cepstral analysis and MLSA filtering for synthesis. One reason for using STRAIGHT is that it is a high-quality speech vocoder which has been successfully implemented in the Nitech-HTS 2005 system. Also, the STRAIGHT analysis method (estimation of the spectral envelope and aperiodicity parameters) can be combined with the GSS analysis method to incorporate the LF-model into the HMM-based speech synthesiser. The implementation of the GSS method using STRAIGHT performed well in the copy-synthesis experiment presented in Section 6.6 and it is also expected to perform well when integrated into the baseline HMM-based speech synthesiser.

7.2.1.1 Analysis

The fundamental frequency, F_0, is estimated using the F_0 detector of the ESPS tools which is an implementation of the RAPT algorithm (Talkin, 1995). This method performs similarly to the fixed-point analysis method used by STRAIGHT (Kawahara et al., 1999b). However, the method of the ESPS tools was chosen in this work because it permitted the tuning of parameters of the F_0 detector in order to obtain a more accurate F_0 estimate.

STRAIGHT is used to calculate the aperiodicity measurements and the FFT coefficients of the spectral envelope of the short-time speech signal, as described in Section 6.2.2. These parameters are transformed to features which are more suitable for the statistical modelling. For the case of the spectral envelope, it is converted to a representation in terms of mel-cepstral coefficients. For the aperiodicity, *five weights* are obtained by averaging the aperiodicity amplitude spectrum in the five frequency bands: 0-1, 1-2, 2-4, 4-6, and 6-8 kHz.

7.2.1.2 Synthesis

The method used by STRAIGHT to synthesise speech was described in Section 6.2.3. For voiced speech, the excitation is obtained by *weighting* a pulse signal and white

noise and adding them together. The weighting functions are calculated from the aperiocidity parameters. STRAIGHT generates the pulse by processing the phase of a delta pulse, in order to reduce the strong periodicity of the impulse train signal and improve speech naturalness. For unvoiced speech, the excitation is modelled as white noise. Finally, the minimum-phase impulse response of the speech signal is calculated from the mel-cepstral coefficients and then the speech signal is generated by convolving this impulse response with the excitation signal.

7.2.2 Statistical Modelling

7.2.2.1 Statistical Model

The statistical model is a five-state left-to-right HMM. Each state output density function is modelled by a single Gaussian probability distribution. The state duration is also modelled by a Gaussian distribution. In this case, the HSMM structure described in Section 3.3.4 is used to explicitly model the duration.

Each observation feature vector at time t, $\boldsymbol{o}_t$, consists of five streams: spectrum, aperiodicity, $\log F_0$, Δ of $\log F_0$ and Δ^2 of $\log F_0$. The spectrum and aperiodicity parameters are modelled using a continuous probability distribution, while the last three streams are modelled using a continuous distribution for the voiced and a discrete distribution for the unvoiced space. A MSD-HMM (Tokuda et al., 1999) is used to model these parameters. The aperiodicity parameters consist of the five frequency band weights $\boldsymbol{v}_t$ and their delta (Δ) and delta-delta (Δ^2) parameters, whereas the spectral parameters are the static mel-cepstral coefficients $\boldsymbol{c}_t$, and their Δ and Δ^2 coefficients. In this work, the number of mel-cesptral coefficients used is 39. Figure 7.1 shows the structure of the speech parameter vector.

7.2.2.2 Context Clustering

There are many contextual factors (e.g. phonetic, prosodic and linguistic) that affect spectrum, F_0 and duration. Context-dependent HMMs are used to model these effects. However, it is difficult to cover all possible context-dependent units because the amount of training data that is usually available does not include all combinations of contextual factors. Similarly to the HTS system, HTS-STRAIGHT performs clustering of the trained HMMs using *decision trees*, which was described in Section 3.3.3. The spectral, F_0 and duration parameters are clustered independently because they have their own influential contextual factors. The HTS-STRAIGHT system uses the

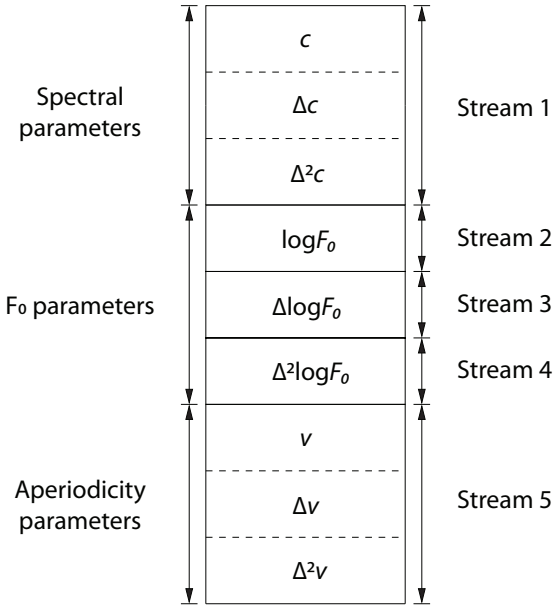

Figure 7.1: Multi-stream structure of the speech feature vector, in the HTS-STRAIGHT system.

minimum description length (MDL) criterion (Shinoda and Watanabe, 2000) for the tree-based clustering.

The HMMs associated with *leaf nodes* in the decision tree which have a common mean and variance are also tied in order to avoid data sparsity problems. For a set of models of tied leaf nodes, $U = \{U_1, U_2, ..., U_M\}$, the log-likelihood $\mathcal{L}(U)$ of U generating a set of T observation vectors, with $\boldsymbol{o}_t$ having dimension L, can be approximated by the following equation:

$$\begin{aligned} \mathcal{L}(U) &= \sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_t(m)\log\mathcal{N}_m(\boldsymbol{o}_t;\mu_m,\boldsymbol{V}_m) \\ &= -\frac{1}{2}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_t(m)\left(L+L\log(2\pi)+\log|\boldsymbol{V}_m|\right), \end{aligned} \tag{7.1}$$

where μ_m and $\boldsymbol{V}_m$ are the mean vector and the *diagonal covariance matrix* of the Gaussian probability distribution $\mathcal{N}_m$ at node S_m, respectively. In this equation, $\gamma_t(m)$ represents the probability of the observed frame $\boldsymbol{o}_t$ being generated by the node S_m.

The MDL principle uses the *description length* parameter, l, to find the optimal probabilistic models. The description length can be calculated as

$$l(\boldsymbol{U}) = -\mathcal{L}(U) + LM\log\left(\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_t(m)\right) + C, \tag{7.2}$$

where C is the *code length* (assumed to be a constant), required to choose the model. When a given node S_m associated with the model U is divided into two nodes, a new model U' is calculated for the *child nodes*. The difference between the description lengths before and after splitting, $\delta l = l(U') - l(U)$ is used as the *stopping criterion*. If $\delta l < 0$, then the node is divided. Otherwise, it is not divided.

7.2.2.3 HMM Parameter Estimation

The parameter calculation of a HMM λ with known phonetic transcription $\boldsymbol{Z}$ can be described by the following optimisation problem:

$$\lambda^* = \arg\max_{\lambda} P(\boldsymbol{O}|\boldsymbol{Z}, \lambda), \tag{7.3}$$

where $\boldsymbol{O}$ is the sequence of speech parameter vectors obtained during analysis. The HTS-STRAIGHT system uses the Baum-Welch algorithm, which was described in Section 3.2.5 to solve this problem. State duration probability density functions are estimated simultaneously with the other λ parameters, as they are modelled explicitly by a HSMM (same optimisation problem as for a HMM). The HTS-STRAIGHT system uses the the HTK-3.4 tools (Young et al., 2006) to implement the Baum-Welch algorithm and to perform the necessary operations to calculate the HSMM parameters. The main functions of the HTK tools (HTS versions of these tools) used for statistical modelling are summarised below:

- HCompV: calculation of the global speech parameter mean and covariance.
- HInit: calculation of initial estimates for the HMM parameters by using the speech parameters and the Viterbi alignment algorithm.
- HERest: calculation of the state duration probability density functions and Baum-Welch re-estimation of the parameters of a single HMM using a set of speech parameter vectors.
- HHEd: tying across selected HMMs and decision tree-based context clustering.

Figure 7.2 shows the block diagram of the HMM parameter estimation method used by the HTS-STRAIGHT system, which can be divided into two parts. The first is related to the HMM estimation without taking into account the context, i.e. training of *context-independent* (CI) HMMs or *monophone* HMMs. The second concerns the re-estimation of *context-dependent* (CD) HMMs, also called *full-context* HMMs.

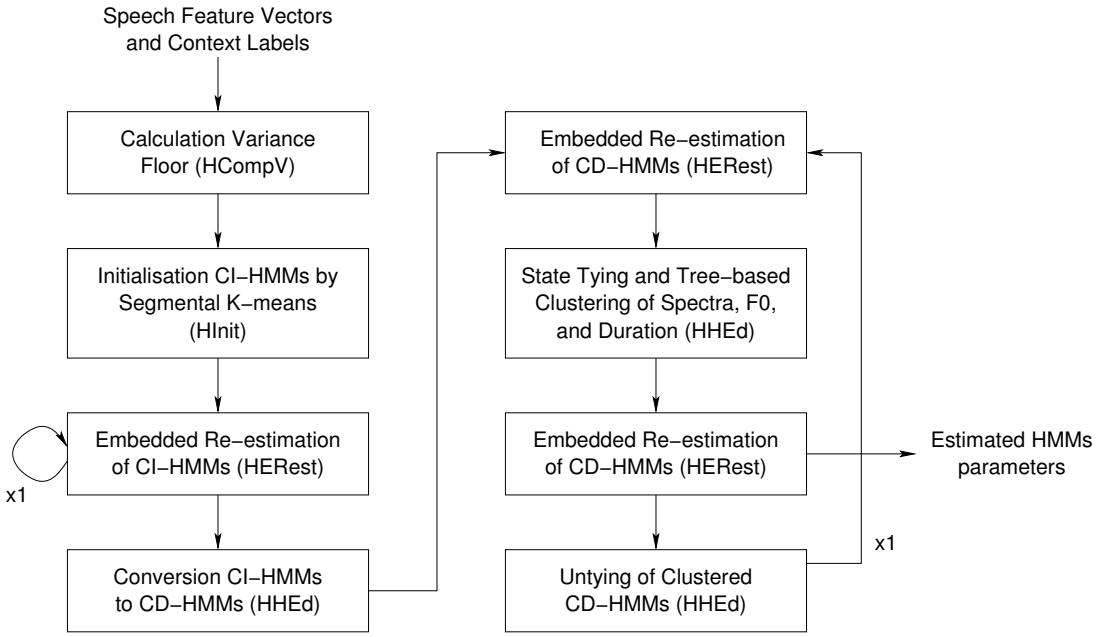

Figure 7.2: Block diagram of the training procedure in the HTS-STRAIGHT system.

In the CI-HMM estimation part, HCompV is used to initially calculate the *global speech variance* and a *variance floor* value. Next, HInit initialises the models by using the speech feature vectors, the monophone labels (labels of the phone model without context information) and the variance floor. This tool performs the *segmentation* of the training observations by recursively clustering the vectors in each segment using a *K-Means* based algorithm (Young et al., 2006) and using Viterbi alignment. The parameters of the CI-HMMs are re-estimated by the HERest tool. This tool uses the Baum-Welch algorithm (Baum et al., 1970; Young et al., 2006) to estimate the parameters of each HMM, from the phonetic transcriptions, the observation feature vectors and the initial estimates of the model parameters. The Baum-Welch re-estimation is performed more than once (twice in this case), in order to more accurately estimate the HMM parameters. The next step is to clone the CI-HMMs into context-dependent sets of models using HHEd and the labels with contextual information.

In the CD-HMM estimation part, the models are first re-estimated using HERest. In general, the amount of training data is not sufficiently large to accurately model all the contextual information by CD-HMMs and the more complex the model is (larger amount of contextual information), the more data are needed. In order to avoid this problem, HHEd is used to cluster the resulting CD-HMMs using decision trees and to perform tying of the clustered models. Tied models can share their data and param-

eters, which avoids the problem of data insufficiency. The mel-cepstral coefficients, $\log F_0$, aperiodicity parameters, and duration are clustered using different decision trees, respectively. The resulting models are re-estimated again. Finally, CD-HMM estimation is refined by performing another iteration of the re-estimation and context-clustering, after untying the clustered CD-HMMs.

7.2.3 Speech Parameter Generation

7.2.3.1 Algorithms

The problem of generating the speech parameter vector sequence $\boldsymbol{O}$ from the HMM λ, for a given word sequence $\boldsymbol{Z}$, is to maximise the output probability distribution with respect to $\boldsymbol{O}$, as follows:

$$\boldsymbol{O}^* = \arg\max_{\boldsymbol{O}} P(\boldsymbol{O}|\boldsymbol{Z},\lambda) \tag{7.4}$$

This problem can be solved using the recursive method based on the expectation-maximisation (EM) algorithm, which was described in Section 3.3.1.4. The HTS-STRAIGHT system implements the EM algorithm using the HMGenS tool, which is publicly available on-line as part of the HTS (version 2.1) program (Tokuda et al., 2009).

Similarly to the HTS system, HTS-STRAIGHT also provides a small run-time synthesis engine, called *hts_engine*, which generates speech parameters using a Viterbi-based method (Tokuda et al., 2000), described in Section 3.3.1.3. *hts_engine* works without the HTK/HTS libraries and it is faster than HMGenS. However, the HMGenS program was used in this work, because it uses an EM-based algorithm which is expected to more accurately generate the speech parameters than the *hts_engine* program.

7.2.3.2 Global Variance

Speech parameter trajectories obtained using the methods described in the previous section and using both static and dynamic features often are excessively smooth (Toda and Tokuda, 2007). This is an effect of the statistical modelling, as it does not capture details of the parameter trajectories of natural speech with sufficient accuracy.

Over-smoothing of the parameter trajectories causes the synthetic speech to sound muffled. Several methods have been proposed in order to reduce this problem. For example, Ling et al. (2006b) proposed a method to enhance the formants of the syn-

thesised speech by using the linear spectral pair (LSP) parameters, instead of mel-cepstral coefficients. The HTS-STRAIGHT system uses a parameter generation algorithm which considers the *global variance* of the generated feature trajectory to reduce the over-smoothing effect (Toda and Tokuda, 2007). This technique is described in the following paragraphs.

Toda and Tokuda (2007) observed that the global variance (GV) of the spectral parameters estimated by the conventional parameter generation algorithm (implemented using *hts_engine*) was smaller than the GV measured for the same utterance of natural speech. The generated trajectory was close to the mean vector sequence of the HMM. The solution proposed by Toda and Tokuda (2007) consists of compensating for this GV difference using a transformation of the feature trajectory.

The GV of a D-dimensional static feature vector $\boldsymbol{c}$, over a time sequence with duration T, is calculated as

$$\boldsymbol{v}(\boldsymbol{c}) = \{v(1), v(2), ..., v(D)\}^\top \tag{7.5}$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}(c_t(d) - \bar{c}(d))^2 \tag{7.6}$$

$$\bar{c}(d) = \frac{1}{T}\sum_{\tau=1}^{T} c_\tau(d), \tag{7.7}$$

where $\boldsymbol{c}_t = \{c_t(1), c_t(2), ..., c_t(D)\}^\top$ is the static feature vector at frame t and $\bar{c}(d)$ is the mean of the d-dimension of the static feature vector over the time sequence.
The parameter generation algorithm considering a Gaussian distribution λ_v for modelling the GV (Toda and Tokuda, 2007) maximises the following likelihood:

$$P(\boldsymbol{O}|\lambda, \lambda_v) = \sum_{\mathbf{q}} P(\boldsymbol{O}, \boldsymbol{q}|\lambda)^w P(\boldsymbol{v(c)}|\lambda_v), \tag{7.8}$$

where w is the weight for controlling the balance between the likelihood of the HMM model λ and the GV likelihood $P(\boldsymbol{v(c)}|\lambda_v)$. The probability density function of the GV is represented by a Gaussian distribution with mean μ_v and with a diagonal covariance matrix U_v, i.e.

$$P(\boldsymbol{v(c)}|\lambda_v) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{U}_v|}} \exp\left(-\frac{1}{2}(\boldsymbol{v(c)} - \mu_v)^T \boldsymbol{U}_v^{-1}(\boldsymbol{v(c)} - \mu_v)\right) \tag{7.9}$$

Toda and Tokuda (2007) set w equal to the ratio of the number of dimensions between the vectors $\boldsymbol{v(c)}$ and $\boldsymbol{O}$, that is, $w = 1/(3T)$. The Gaussian distributions of λ_v and

λ are independently trained from the speech corpus. The function of the likelihood $P(\boldsymbol{v(c)}|\lambda,\lambda_v)$ is to increase the GV by the adequate amount.

The conventional parameter generation algorithm (without considering GV) is used to generate an initial speech parameter trajectory. This algorithm calculates an optimal observation feature vector $\boldsymbol{O}^*$ and state sequence $\boldsymbol{q}^*$, e.g. by solving (3.43) and (3.42). Next, the following likelihood is maximised with respect to $\boldsymbol{c}$ under the condition that $\boldsymbol{q}^*$ is known:

$$L = \log\left[P(\boldsymbol{O}|\boldsymbol{q}^*,\lambda)^w P(\boldsymbol{v(c)}|\lambda_v)\right] \tag{7.10}$$

This optimisation is performed by using the iterative *Newton-Raphson method* (Kelley, 2003). The initial trajectory of this iterative algorithm, $\boldsymbol{c}'$, is obtained by the following linear transformation of the trajectory calculated with the conventional algorithm, $\boldsymbol{c}$:

$$c'_t(d) = \sqrt{\frac{\mu_v(d)}{v(d)}}\left(c_t(d) - \bar{c}(d)\right) + \bar{c}(d) \tag{7.11}$$

Toda and Tokuda (2007) indicate that $\boldsymbol{c}'$ usually gives a larger value of the likelihood L than $\boldsymbol{c}$, when $w = 1/(3T)$.

7.3 Incorporation of the LF-model

For the integration of the LF-model into the baseline system, it was necessary to modify the analysis-synthesis method (STRAIGHT method) and adjust the statistical modelling part. The GSS method is used to estimate the LF-model and the vocal tract transfer function parameters from speech and to generate the speech waveform instead of the STRAIGHT vocoder. The system which uses the GSS method and the LF-model is called HTS-LF.

7.3.1 GSS Analysis

The GSS method for estimation of the LF-model and the vocal tract parameters is implemented as in the copy-synthesis application, which was described in Section 6.5. This method is summarised as follows:

1. F_0 *and glottal epochs*: ESPS tools.

2. *Glottal source derivative*: inverse filtering of the pre-emphasised speech signal ($\alpha = 0.97$). In this operation, the inverse filter is obtained from the LPC coefficients which are computed pitch-synchronously using Hanning windows centered at the glottal epochs.

3. *LF-model parameters* (voiced speech): initial estimates using direct measurements on the period of the LPC residual (delimited by two consecutive epochs) and non-linear optimisation algorithm to fit the LF-model waveform to the residual signal. The resulting trajectories are smoothed in order to alleviate estimation errors.

4. *Vocal tract parameters (voiced speech)*: quotient between the speech spectrum and the amplitude spectrum of the LF-model signal (one period long) and spectral envelope computation of the resulting signal using STRAIGHT. The FFT parameters of the envelope are converted to mel-cepstral coefficients.

5. *Spectral envelope (unvoiced speech)*: STRAIGHT analysis and conversion of the resulting FFT coefficients to mel-cepstral coefficients.

6. *Aperiodicity parameters*: STRAIGHT analysis and conversion of aperiodicity measurements to weights in five frequency bands.

7.3.2 Statistical Modelling of the LF-parameters

7.3.2.1 Statistical Model

The structure of the statistical model of the HTS-LF system is similar to that of the HTS-STRAIGHT system. It is a five-state left-to-right HSMM and both the state output density function and the state duration are modelled by a single Gaussian distribution. However, there is a difference in the feature data streams: the F_0 parameter vectors (including dynamic features) of the HTS-STRAIGHT system are replaced by the LF-model parameter vectors in HTS-LF. That is, the feature vector of the HTS-LF system consists of five streams: spectrum, aperiodicity, LF-parameters, Δ of LF-parameters, and Δ^2 of LF-parameters. The LF-parameters are: $\log(1/t_e)$, $\log(1/t_p)$, $\log(1/T_a)$, $\log(E_e)$, and $\log(1/T_0) = \log(F_0)$. The spectrum and aperiodicity parameters are modelled by a continuous HMM with a diagonal covariance matrix, while the last three streams are modelled by a MSD-HMM (Tokuda et al., 1999). MSD-HMM is used to model the LF-parameters because they are not defined in the unvoiced regions.

The prototype HMM definition file of HTS-STRAIGHT was modified in order to take into account the LF-model parameters in HTS-LF. The length of the HTS-LF feature vector has fifteen more parameters than that of HTS-STRAIGHT.

The LF-model parameters, their Δ and Δ^2 are modelled in different streams (three streams are used). In each stream, the parameters are modelled by using a single Gaussian distribution with diagonal covariance matrix for the voiced space. For the unvoiced space, a single discrete distribution which outputs one symbol is used.

The LF-model and aperiodicity parameters are modelled in different streams because the periodic and noise components of the excitation are assumed to be independent. The LF-model does not take into account the noise component of speech, such as aspiration noise which is mainly produced during the open phase of the glottal cycle. This might be a limitation for voice transformation using the HTS-LF system. For example, accurate modelling of the aspiration noise is important to reproduce a breathy voice correctly. The covariance between the LF-model and the noise of the glottal source could be modelled by HMMs (using the same stream for both components) if the glottal source model represented the correlation between the periodic and noise components. In the opinion of the author, noise modelling in the HTS-LF system could be improved by using a time-domain model of the noise which was compatible with the LF-model. Such a model is further discussed in Section 10.3.1.2.

The clustering of the statistical models in the HTS-LF system is expected to result in smaller decision trees for the LF-model parameters than those obtained for F_0 in the HTS-STRAIGHT system, because the feature vector is larger in HTS-LF and the MDL is the same in the two systems. This effect related to the difference between the HMM structure of the HTS-LF and the HTS-STRAIGHT systems was confirmed experimentally for the voice built for these systems, which will be presented in Section 7.4.

7.3.2.2 HMM Parameter Estimation

The observation vector probability distributions of the HMM are calculated as in the HTS-STRAIGHT system. The HMM training method in the HTS-STRAIGHT system was described in Section 7.2.2.3. The decision trees used to cluster the LF-model parameters in the HTS-LF system are built using the contextual factors used to cluster F_0 in the HTS-STRAIGHT system. These contextual factors are assumed to perform well for the LF-model parameters because there is a strong correlation between F_0 and the other LF-model parameters. The stopping criterion used to build the decision trees is also the same for the two speech synthesisers.

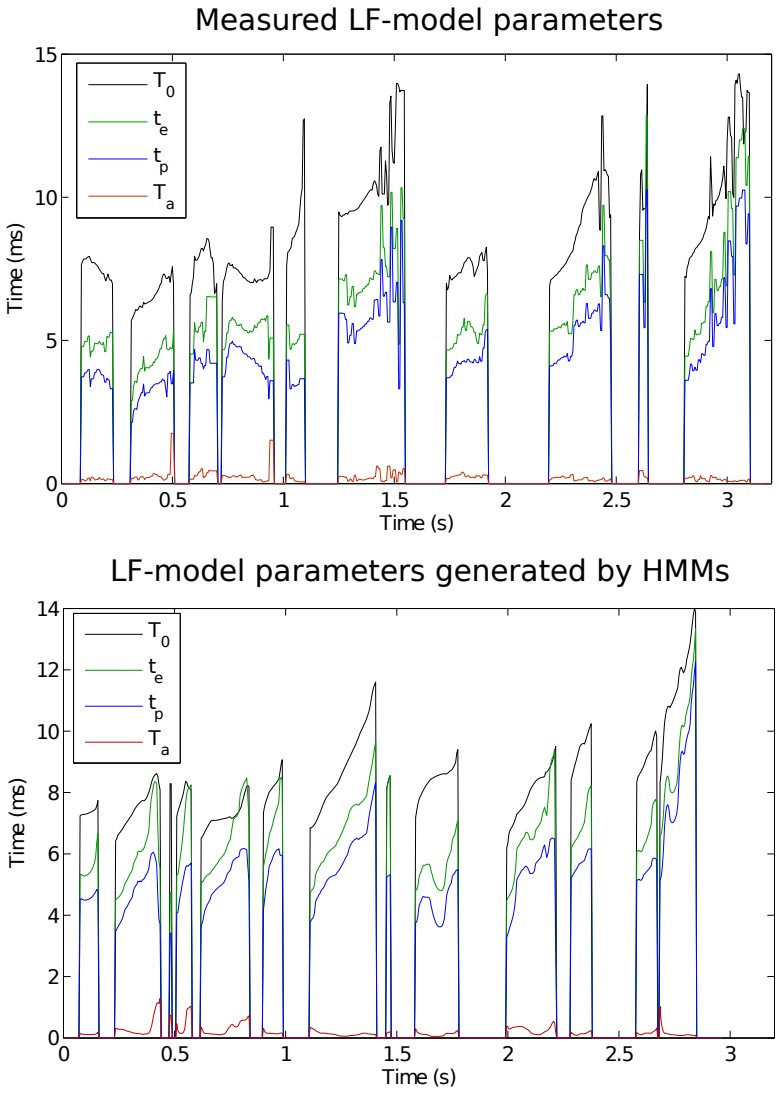

Figure 7.3: Trajectories of the LF-model parameters estimated for an utterance by the speech analysis method of the HTS-LF system and respective parameter trajectories generated by the system for the same utterance.

7.3.3 Synthesis Using the LF-model

7.3.3.1 Speech Parameter Generation

The HTS-LF system uses the same parameter generation algorithm as HTS-STRAIGHT, which was described in Section 7.2.3. However, the settings of this algorithm in HTS-LF are adjusted to its HMM structure. For example, the dimension of the F_0 feature vector in the baseline system is lower than the dimension of the LF-model parameter vector in the HTS-LF system.

Figure 7.3 shows an example of the LF-model parameters estimated for an utter-

ance by the HTS-LF system during analysis and the trajectories of the same parameters generated by the synthesiser. In general, the parameter generation algorithm produces a smoother trajectory than the one obtained during speech analysis, mainly due to statistical modelling by the HMMs. One advantage of this smoothing effect is attenuation of parameter discontinuities due to estimation errors in speech analysis. However, parameter smoothing by HMM-based speech synthesisers is typically excessively high which causes synthetic speech to sound muffled. There are other types of errors which can be occasionally observed in parameter contours generated by the HTS-LF system. These errors are related to validity of the LF-model parameter constraints, given in Section 5.2.1. For example, t_e is higher than the period T_0 in a short speech region located around the 0.8 s mark, in Figure 7.3. This problem might be related to errors in the LF-model estimation and inaccurate modelling of glottal parameters by HMMs. An algorithm was developed to reduce estimation errors of LF-model parameters, which is described in Section 8.2.2. For accurately modelling glottal parameters by HMMs, one possible solution is to use a sufficiently high amount of speech data for training.

7.3.3.2 Speech Waveform Generation

The HTS-LF system employes the speech waveform generation method described in Section 6.5.2, which was used for the copy-synthesis application using GSS. This method uses the LF-model and vocal tract filter parameters to synthesise speech. Conversely, the baseline HMM-based speech synthesiser (the HTS-STRAIGHT system) uses the STRAIGHT synthesis method. The synthesis method used by the HTS-LF system is summarised in the next paragraphs.

The excitation frame for voiced speech, $g^i(t)$, is obtained by mixing a periodic and a noise component. The periodic signal consists of two periods of the LF-model waveform, centered at the instant of maximum excitation t_e, while the noise is a random sequence with the same duration as the periodic signal. The two components are weighted in the frequency domain using the aperiodicity parameters and then added together, as explained in Section 6.4.3.2. Next, the excitation is multiplied by a Hamming window and zero-padded to 1024 samples to calculate the FFT, $X^i(w)$.

Speech is synthesised by calculating the convolution of the excitation signal with the vocal tract transfer function. This operation is performed in the frequency domain by multiplying the spectrum of the excitation by the vocal tract spectrum, i.e. $S^i(w) = P^i(w)G^i(w)V^i(w)$, where $P^i(w)$ is the FT of a delta pulse train, $G^i(w)$ is the FT of $g^i(t)$ and $V^i(w)$ is the vocal tract filter which is obtained from the mel-cepstral coefficients.

The next step is to calculate the speech waveform $y^i(t)$ by IFFT of $Y^i(w)$. Next, the effect of the Hamming window used to calculate the excitation spectrum is removed from $y^i(t)$. Finally, speech frames are concatenated using overlap-and-add windows which are asymmetric and centered at the instants of maximum excitation.

7.4 Preliminary Evaluation of the HTS-LF System

7.4.1 AB Perceptual Test

A forced-choice AB listening test was carried out in order to evaluate the HTS-LF system, by comparison with the HTS-STRAIGHT system. Table 7.1 summarises the characteristics of the systems evaluated in this experiment.

	Systems	
	HTS-LF	HTS-STRAIGHT (baseline)
Analysis	Inv. Filt. Pre-emphasis: LF-param. ESPS tools: F_0, epochs GSS: vocal tract STRAIGHT aperiodicity	ESPS tools: F_0, epochs STRAIGHT: spectral envelope STRAIGHT aperiodicity
Excitation	Mixed LF-model & noise	Mixed impulse & noise
Synthesis	GSS synthesis	STRAIGHT
Evaluation	Speech Naturalness	

Table 7.1: Summary of the HMM-based speech synthesisers used in the perceptual experiment which was conducted to evaluate naturalness of the synthetic speech.

The US English BDL speech corpus (male speaker) of the CMU ARCTIC speech database (Kominek and Black, 2004) was used to build the voices of the HTS-LF and HTS-STRAIGHT systems, respectively. The size of the BDL speech corpus is approximately one hour.

The stimuli consisted of 36 pairs of utterances: 18 utterances synthesised by each system, randomly chosen and repeated twice with the order of the samples alternated.

The type of sentences used for synthesis was of conversational speech, for example "I would like to have a five star hotel".

The evaluation was conducted via the web. Subjects were asked to listen to the pairs of stimuli and for each pair they had to select the version (A or B) that sounded best. They were able to listen to the files in any order, and as many times as they wished. They were also instructed to make a random choice if they could not decide on the version they preferred.

The listening panel was composed of students and staff from the School of Informatics. Fourteen listeners participated in the test, of which six were native speakers of English.

7.4.2 Results

The results of this perceptual experiment are shown in Table 7.2. The difference between the scores obtained by the HTS-LF and the HTS-STRAIGHT systems are statistically significant with $p\text{-}value \leq 0.01$.

	HTS-STRAIGHT	HTS-LF
Mean preference (%)	44.4	55.6
95% Conf. Interv. (%)	[40.1 48.9]	[51.1 59.9]

Table 7.2: Mean scores and 95% confidence intervals obtained by the two HTS synthesisers in the AB forced-choice evaluation.

On average, the HTS-LF system obtained a higher rate of preference. However, the improvement in performance by HTS-LF when compared with the baseline system was lower than expected, based on the results of the previous evaluation described in Section 6.6.5. In this previous experiment, speech synthesised by copy-synthesis using the GSS method was significantly preferred (preference rate over 60%) over speech synthesised using the impulse train. Examples of speech synthesised by the two systems are accessible through the link `http://homepages.inf.ed.ac.uk/cgi/jscabral/hts-lf-model.html`. From the results of this experiment, it is difficult to explain why difference in speech quality between the HTS-LF and the HTS-STRAIGHT systems appears to be lower than the difference between the GSS method and the baseline method in the copy-synthesis experiment. Possible factors to explain these results are:

- Errors in the LF-model and vocal tract parameter estimation using GSS could deteriorate the performance of the HTS-LF system.
- Statistical modelling of the LF-model parameters in the HTS-LF system might be less accurate than statistical modelling of F_0 in HTS-STRAIGHT.
- Excitation model of voiced speech used in the copy-synthesis evaluation (periodic component only) is different from that used by the HMM-based speech synthesisers (multi-band mixed excitation) in this experiment.
- Method used to synthesise speech using the impulse train in the copy-synthesis evaluation was different from the synthesis method (STRAIGHT vocoder) used by the HTS-STRAIGHT system.

The importance of these factors is discussed in the next paragraphs. In order to reduce the effects of the factors which are considered to be the most important, improvements were made to the HTS-LF system which are presented in the next chapter. Although these possible causes of speech distortion were not directly tested, further experiments conducted in this work, which are presented in Sections 8.4 and 9.2, permitted to obtain more conclusions about the causes of speech distortion in the HTS-LF system.

Errors in the LF-model parameter estimation are expected to have influenced the performance of the HTS-LF system, because the method to estimate the glottal source derivative (inverse filtering with pre-emphasis) might not be sufficiently accurate. For example, Section 2.2.3 described more complex inverse filtering techniques which are more accurate compared with inverse filtering using pre-emphasis. LF-model parameter errors could also affect spectral parameter estimation, as vocal tract parameters are estimated by separating the LF-model from the speech spectrum in the GSS method. Moreover, speech parameter discontinuities caused by estimation errors are expected to have a more negative effect on the quality of speech obtained by HMM-based speech synthesis than by copy-synthesis. This difference is because resynthesised speech frames obtained by copy-synthesis are very similar to the original speech frames, whereas speech parameter discontinuities might degrade statistical modelling in HMM-based speech synthesis.

In order to improve the robustness of the GSS analysis a more accurate method for glottal source estimation was implemented into the HTS-LF system, than inverse filtering with pre-emphasis. Also, an algorithm for errors detection and correction of the estimated LF-model parameters was developed in order to overcome errors related

to parameter values outside their valid ranges. Figures 7.4 to 7.6 show examples of the effect of these types of errors on the LF-model signal. The improvements for the speech analysis are described in the next chapter.

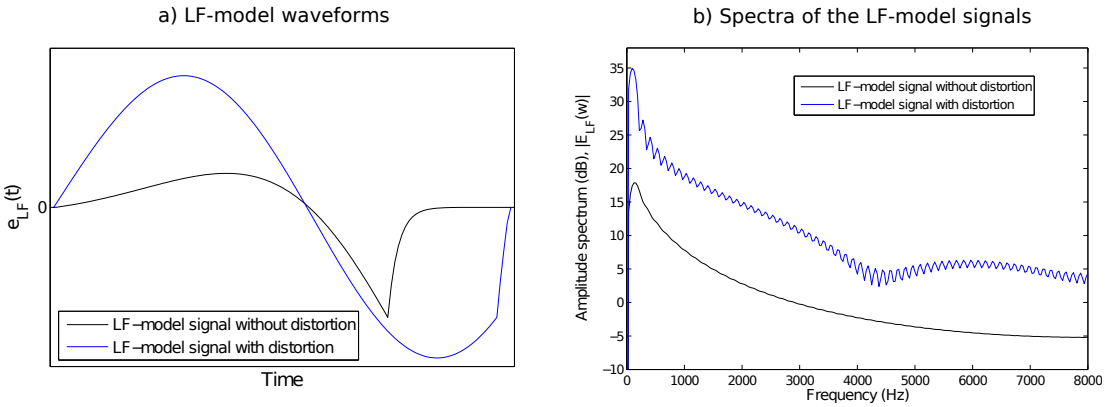

Figure 7.4: Example of a distorted LF-model signal for which the constraint $t_e \leq 3/2t_p$ is not satisfied. In this example, $t_e = 1.3t_p$ for the original LF-model signal. The parameter t_e was increased to obtain the distorted signal ($t_e = 7/4t_p$), while the other parameters remained the same (within their valid range of values).

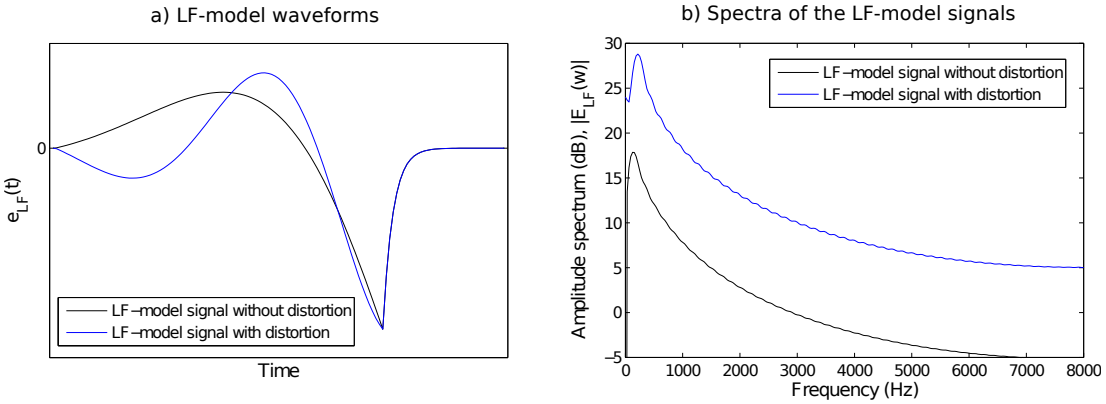

Figure 7.5: Example of a distorted LF-model signal for which the constraint $t_e \leq 3/2t_p$ is not satisfied. In this example, $t_e = 1.3t_p$ for the original LF-model signal. The parameter t_p was decreased to obtain the distorted signal ($t_p = 2/5t_e$), while the other parameters remained the same (within their valid range of values).

The trajectories of the LF-model parameters generated by HTS-LF seem to be smooth enough and similar to the trajectories measured on real speech, from visual comparisons made for several utterances. An example of these trajectories is given in Figure 7.3. Also, F_0 modelling in the HTS-LF system does not appear to be affected

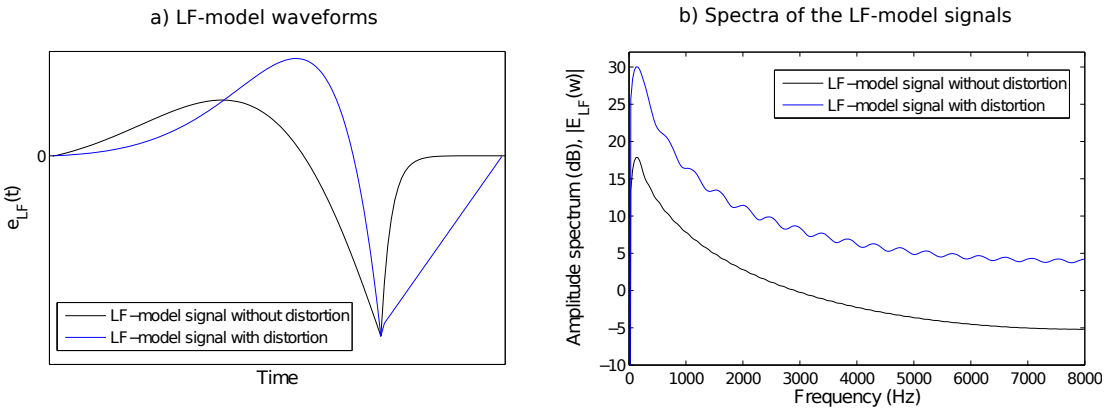

Figure 7.6: Example of a distorted LF-model signal for which the constraint $T_a \leq T_0 - t_e$ is not satisfied. In this example, $T_a = 0.18$ ms and $T_0 - t_e = 2.2$ ms for the original LF-model signal. The parameter T_a was increased to obtain the distorted signal ($T_a = 2.3$ ms), while the other parameters remained the same (within their valid range of values).

by using a vector feature stream for the LF-model parameters (F_0 is one of the parameters of this stream). For these reasons, statistical modelling of the LF-parameters in the HTS-LF system is assumed not to significantly cause speech quality degradation relative to the baseline system.

The noise component of the multi-band mixed excitation and the phase processing of the impulse signal by STRAIGHT reduces the buzziness produced by the impulse train. For this reason, the buzzy effect due to the impulse train is expected to be less relevant in the comparison between the HTS-LF and HTS-STRAIGHT systems than in the copy-synthesis experiment (voiced speech was synthesised using the conventional impulse train signal without mixing it with noise).

From the author's informal comparison of the speech synthesised by the two HMM-based speech synthesisers in this perceptual evaluation, they sounded different for most utterances. For some speech samples, the "buzzy" or "metallic" quality produced by the HTS-STRAIGHT system was clearly higher, when compared with the HTS-LF system. In other cases, speech synthesised with HTS-LF contained speech artefacts which could be more perceptually important than the buzziness characteristic of the HTS-STRAIGHT system. The most common and relevant speech artefacts perceived for the HTS-LF system were related to an excessively high energy of the noise or audible clicks in speech segments around the instants of *voicing transition* (voiced-to-unvoiced and unvoiced-to-voiced). The high energy variations which were occasionally observed in the synthetic speech are expected to be related to parameter modelling

problems in the voicing transition regions. A method for reducing energy variations between synthetic speech frames at voicing transitions was developed in order to avoid this problem. However, this technique requires modelling the power parameter in the HTS-LF system. This method is described in the next chapter, as an improvement performed to the HTS-LF system.

7.5 Conclusion

The HMM-based speech synthesiser with LF-model (HTS-LF) is based on a high-quality HMM-based speech synthesiser which uses the STRAIGHT vocoder (which is referred in this work as HTS-STRAIGHT system). The GSS analysis-synthesis method was integrated into the baseline system (HTS-STRAIGHT) so that the HTS-LF system is able to use the LF-model. Table 7.3 summarises the characteristics of these systems. They use a five-state left-to-right HMM with explicit duration modelling (HSMM). The HMM training is performed using the typical EM algorithm for the HMM parameter re-estimation and decision tree state tying clustering. A parameter generation algorithm considering global variance is used in order to reduce the problem of over-smoothed parameter trajectories.

The main differences between the HTS-LF and the HTS-STRAIGHT systems are the multi-stream structure of the speech parameter vector and the analysis-synthesis methods. HTS-STRAIGHT models F_0, its Δ and Δ^2 parameters by using a stream for each of these parameters, whereas HTS-LF models the five LF-model parameters, their Δ, and Δ^2 also using three streams. Both systems use the F_0 detector of the ESPS tools (Talkin, 1995) to estimate F_0. The HTS-STRAIGHT system uses the STRAIGHT analysis method to estimate the spectral envelope of the speech signal and the aperiodicity parameters. Meanwhile, the HTS-LF system uses the GSS method to extract the LF-model and vocal tract filter parameters. The GSS method estimates the LF-model parameters from the LPC residual (calculated by performing pre-emphasis inverse filtering on the speech signal) and the spectral parameters are obtained by removing the spectral effects of the LF-model from the speech signal and computing the spectral envelope of the resulting signal using STRAIGHT. The aperiodicity parameters are also calculated using STRAIGHT. The HTS-STRAIGHT system uses the MATLAB version of STRAIGHT to generate the speech signal, while the HTS-LF system uses the GSS synthesis method which generates speech from the LF-model and the vocal tract filter parameters. Both methods use a multi-band mixed excitation which is ob-

	HTS-STRAIGHT	HTS-LF
Analysis	STRAIGHT ESPS tools: F_0, epochs	LF-model estimation GSS & STRAIGHT ESPS tools: F_0, epochs
Synthesis	STRAIGHT	GSS synthesis
HMM Feature Vectors	39 mel-sp.coef., 39 Δ, 39 Δ^2 5 aperiodicity, 5 Δ, 5 Δ^2 log F_0, Δ, Δ^2	39 mel-sp.coef., 39 Δ, 39 Δ^2 5 aperiodicity, 5 Δ, 5 Δ^2 5 log LF-param., 5 Δ, 5 Δ^2
HMM struc.	5 states left-to-right; HSMM; MSD-HMM	
Prob. Distr.	Gauss. / Multi-space (F_0)	Gaussian / Multi-space (LF-param.)
Training	EM algorithm and Tree-based clustering with MDL criterion	
Par. Gener.	Maximum Likelihood criterion with GV	

Table 7.3: General characteristics of the HTS-STRAIGHT and HTS-LF systems.

tained by weighting the periodic and noise signals using the aperiodicity parameters and adding them together.

An AB listening test was conducted in order to evaluate the speech naturalness of the HTS-LF system, compared with the HTS-STRAIGHT system. From the results, speech synthesised with HTS-LF was slightly preferred on average over speech synthesised with HTS-STRAIGHT. However, the results for the HTS-LF system were expected to be better, as speech synthesised using the LF-model (by copy-synthesis) was significantly preferred over speech synthesised using the impulse train, in the perceptual evaluation presented in Section 6.6. A potential factor of speech distortion in HTS-LF is the effect of peaks observed in the energy envelope of the synthetic speech at voicing transitions, which were often associated with audible artefacts. Parameter estimation errors during speech analysis could also be a cause of speech distortion in the system. The next chapter describes improvements which were made to the HTS-LF system in order to increase the robustness of the GSS analysis method and in order to

avoid the energy peaks which occur in voicing transition regions.

Chapter 8

Improvements to the HTS-LF System

8.1 Introduction

The HTS-LF system described in the previous chapter was implemented using a simple method for estimating the glottal source derivative from the speech signal, the inverse filtering with pre-emphasis method. Although this is a simple technique, it does not accurately separate the glottal source effects from the vocal tract filter, especially the spectral tilt associated with the source. Inaccurate estimation of the glottal source could contribute to errors in LF-model parameterisation, because such errors could produce irregularities in the glottal source derivative waveform which are not represented by the LF-model. For example, LF-model parameters must satisfy certain constraints and inaccurate estimation of the glottal signal could result in a set of estimated glottal parameters which are not valid and could produce a distorted LF-model waveform. Also, problems in glottal source estimation could result in poor modelling of the source characteristics by HMMs in the HTS-LF system. This chapter describes the iterative inverse filtering method which was implemented into the HTS-LF system in order to improve the accuracy of the glottal source derivative estimation. Also, an algorithm to detect and correct LF-model parameter errors which was developed in this work for improving the HTS-LF system will be described. In addition, a method to correct energy envelope distortion in the speech frames around *voicing transitions* was also developed in this thesis and integrated into the HTS-LF system.

The last part of this chapter presents a perceptual listening test which was conducted in order to evaluate the HMM-based speech synthesisers developed during the work of this thesis, which use the LF-model. The synthetic speech was evaluated in terms of speech naturalness, intelligibility, and similarity of the synthetic voice to the

original speaker's voice. This experiment included the HTS-LF system which incorporated the improvements described in this chapter, the HTS-STRAIGHT system (baseline system) described in Section 7.2, a modified version of this system which used the Glottal Post-Filtering (GPF) method for synthesis, and other versions of the HTS-STRAIGHT and the HTS-LF systems which were used to evaluate aspects related to the excitation model and speech waveform generation technique.

8.2 Speech Analysis Improvements

8.2.1 Iterative Adaptive Inverse Filtering

The Iterative Adaptive Inverse Filtering (IAIF) method (Alku et al., 1991) was implemented to calculate the glottal source derivative signal, in the GSS analysis stage of the HTS-LF system. This method has also been used in the HMM-based speech synthesiser proposed by Raitio et al. (2008), which models the excitation of voiced speech using a glottal inverse filtered signal. Figure 8.1 shows the block diagram of the IAIF technique. This method was introduced in Section 4.5.2.2 and its implementation in the HTS-LF system is described in the following paragraphs.

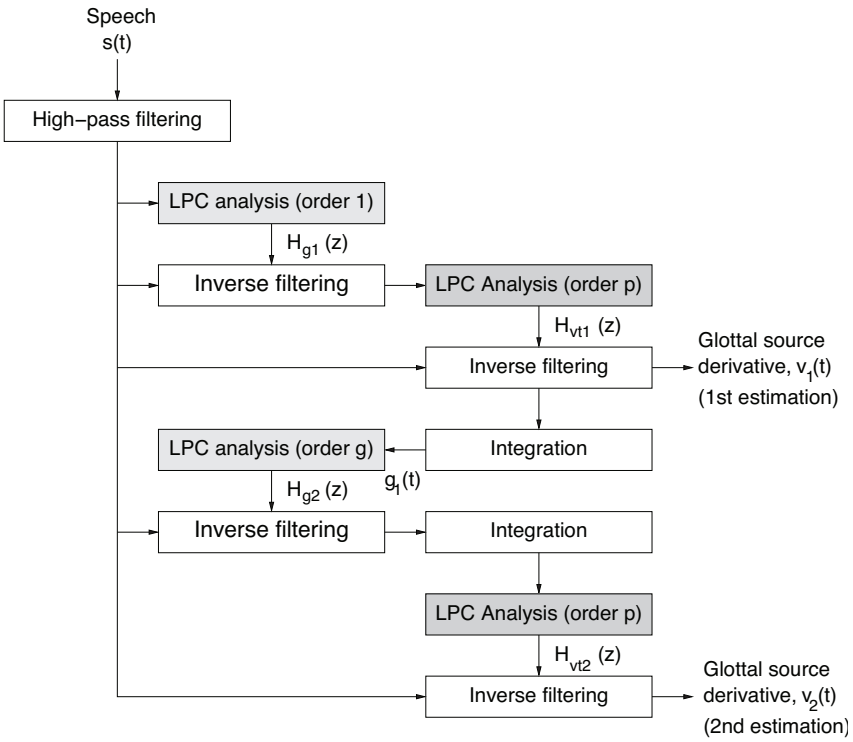

Figure 8.1: Flowchart of the IAIF method.

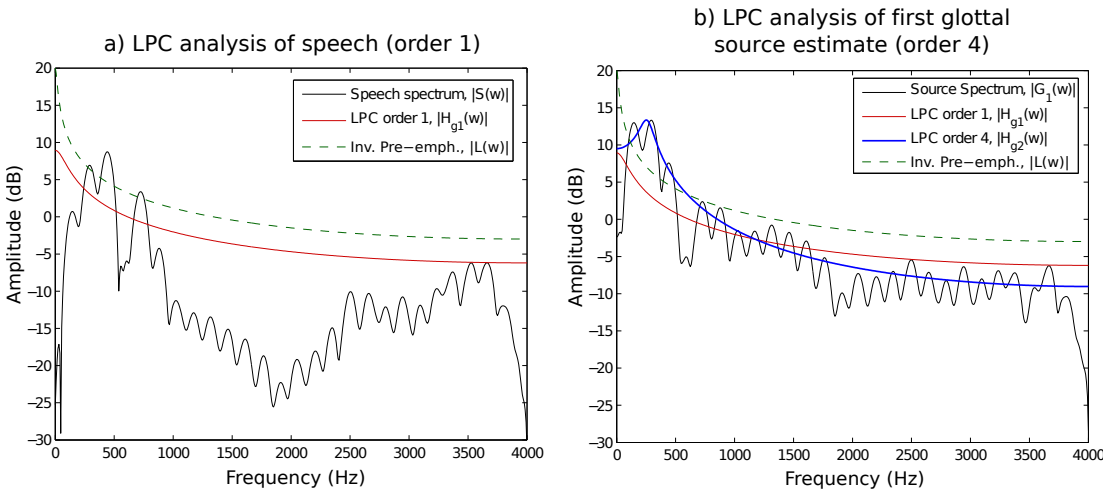

Figure 8.2: Example of LPC analysis in the IAIF method: a) LPC analysis of the speech signal (order one) for the first estimation of the spectral tilt associated with the glottal source and the radiation; b) LPC analysis of the initial estimate of the glottal source signal $g_1(t)$.

The IAIF method performs recursive LPC analysis pitch-synchronously. In the HTS-LF system, each short-time speech signal $s^i(t)$ is centered at the glottal epoch i, has duration equal to two fundamental periods (delimited by the glottal epochs $i-1$ and $i+1$) and is multiplied by a Hamming window with the same duration. The duration of the speech frame is constrained to the interval of 20 ms to 30 ms, in order to obtain a good time-frequency resolution in LPC analysis. The glottal epochs are estimated using the ESPS tools (Talkin, 1995). Each short-time signal is high-pass filtered at 50 Hz in order to remove low-frequency fluctuations and is down-sampled to 8 kHz, which is the same sampling frequency used by Alku et al. (1991).

The first inverse filtering operation of the IAIF method is comparable to a pre-emphasis filtering operation. It removes from the speech signal a rough estimate of the spectral tilt associated with the glottal source and the lip radiation. However, pre-emphasis inverse filtering is typically performed by a time-invariant filter, whereas the inverse filter in IAIF is calculated by first-order LPC analysis of the speech signal. The IAIF method is expected to more accurately model the spectral tilt than pre-emphasis inverse filtering, because in the IAIF method the spectral tilt is adapted to the input speech signal. Figure 8.2 a) shows an example of the amplitude spectra obtained by LPC analysis of order one, $|H_{g1}(z)|$, and the inverse of the pre-emphasis transfer function, $|L(z)|$. The pre-emphasis is modelled by $M(z) = 1 - \alpha z^{-1} = 1/L(z)$, with

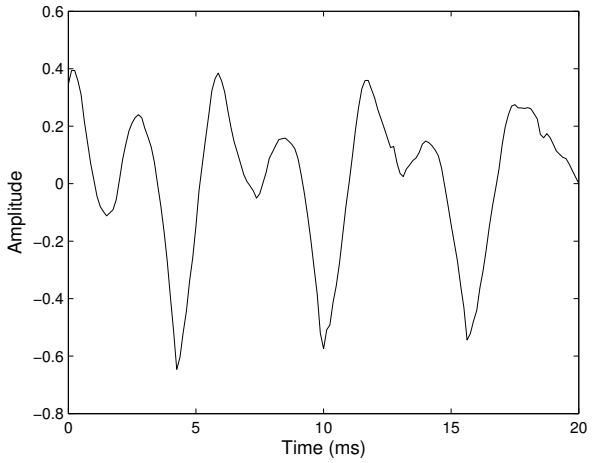

Figure 8.3: Segment of the glottal source derivative signal $v_2(t)$ calculated using the IAIF method.

$\alpha = 0.97$. The initial estimate of the vocal tract, $H_{vt1}(z)$, is calculated by performing LPC analysis of order $p = 10$ on the signal obtained by the initial inverse filtering (LPC order one). The initial estimate of the glottal source derivative, $v_1(t)$, is calculated by inverse filtering the speech signal with $H_{vt1}(z)$. After cancelling the lip radiation through integration, the all-pole model of the glottal source signal, $H_{g2}(z)$, is calculated by LPC analysis of order $g = 4$. Figure 8.2 b) shows an example of the amplitude spectrum of $H_{g2}(z)$. The spectral effect of the glottal source (represented by $H_{g2}(z)$) and the lip radiation are canceled from the speech signal through inverse filtering and integration, respectively. The second vocal tract estimate, $H_{vt2}(z)$, is obtained by performing another LPC analysis of order $p = 10$ to the output of the inverse filter. The final estimate of the glottal flow derivative, $v_2(t)$, is obtained by canceling the spectral effect of the vocal tract, $H_{vt2}(z)$. The signal $v_2(t)$ is up-sampled to 16 kHz, in order to obtain a good time resolution in estimation of the glottal time instants. Figure 8.3 shows an example of the glottal source derivative signal, $v_2(t)$.

8.2.2 Error Reduction in LF-model Parameters

LF-model parameters are constrained to the values indicated in Section 5.2.1, so that the LF-model waveform can be calculated and does not have distortion. It is important that the estimated LF-parameter values satisfy these constraints, in order to avoid parameter estimation errors by the GSS method and statistical modelling problems, in the

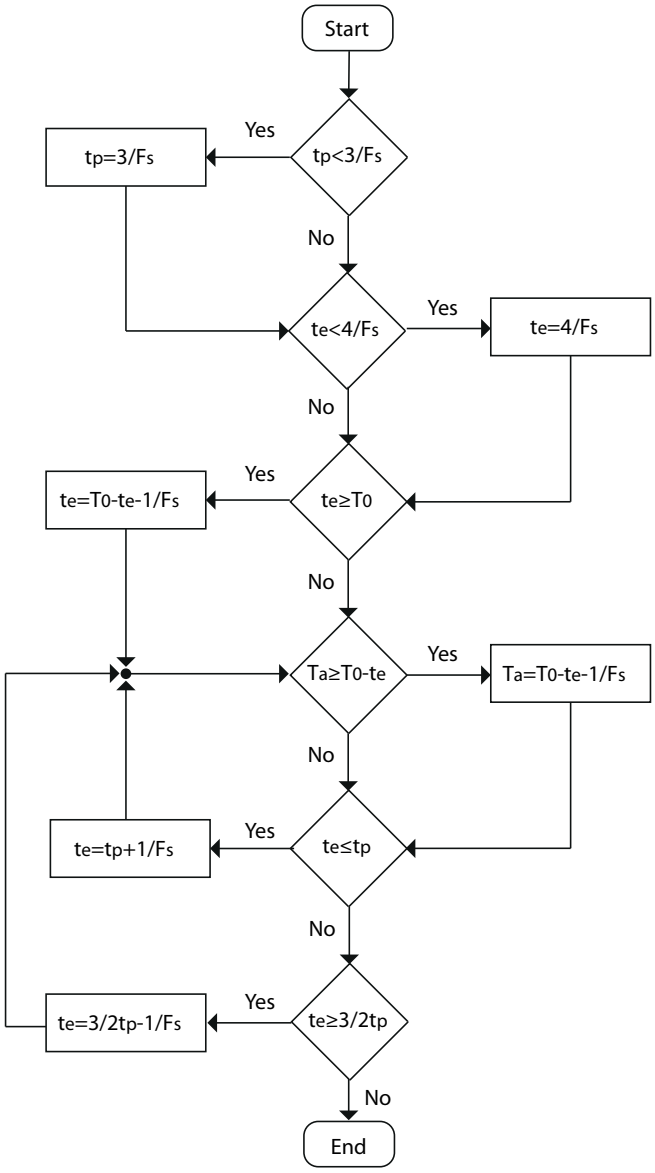

Figure 8.4: Block diagram of the algorithm used to correct LF-model parameter values, in order to avoid distortion in the LF-model waveform.

HTS-LF system. For example, a distorted LF-model signal would produce errors in the vocal tract transfer function because the spectrum of the LF-model is used to estimate the vocal tract filter. An algorithm to detect and correct errors in LF-model parameter estimation was developed during this thesis, in order to improve the robustness of the analysis and synthesis methods used by HTS-LF.

Figure 8.4 shows the algorithm developed to detect if the LF-model parameters satisfy several constraints and to correct them, in order to avoid distortions in the LF-model waveform. In the GSS method used by the HTS-LF system, the LF-parameters

are estimated pitch-synchronously using frames of the glottal source derivative delimited by glottal epochs (instants of maximum excitation). This method estimates the instant of glottal opening, t_o, the return phase parameter T_a, the instant of maximum flow, t_p, and the amplitude of maximum excitation, E_e. The t_e parameter is calculated as $t_e = T_0 - t_o$. First, t_p is evaluated in order to find if it is lower than its minimum value of $3/F_S$. If t_p does not satisfy this condition then it is set equal to $3/F_S$. The same test and correction operation is used for the parameter t_e but using a minimum value of $4/F_S$ instead of $3/F_S$. These minimum values of t_p and t_e were chosen empirically so that they were sufficiently low based on typical range values of these parameters. Subsequently, more constraints on the LF-model parameters are sequentially tested. If a parameter does not satisfy a given constraint it is set equal to the closest value within the possible interval of values for that parameter. In addition, if t_e is corrected, then the constraint $T_a \geqslant T_0 - t_e$ must be tested again. The error correction algorithm is not used to improve the accuracy of the LF-model parameter estimation, but to adjust the estimated LF-parameters so that they satisfy their constraints. This algorithm improves the robustness of the LF-model parameter estimation, because an invalid set of LF-parameter values could produce a significantly distorted LF-model waveform (as shown in Figures 7.4 to 7.6). Figure 8.5 shows an example of a distorted LF-model signal (does not satisfy one of the LF-parameter constraints) and the resulting LF-model signal after applying the error correction algorithm.

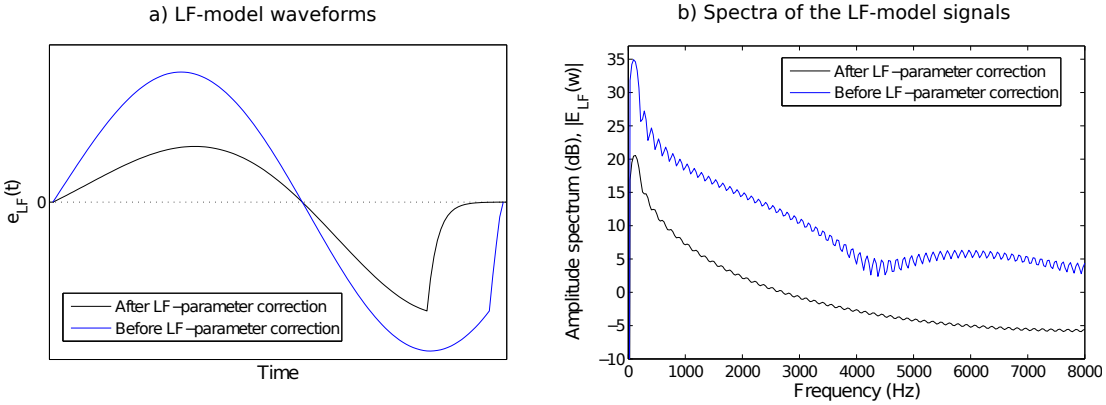

Figure 8.5: Example of a distorted LF-model signal which does not satisfy the constraint $t_e \leq 3/2t_p$, i.e. $t_e = 7/4t_p$ for this signal. The error reduction algorithm corrected this signal by setting t_e equal to $3/2t_p$, while the other parameters remained the same (within their valid ranges).

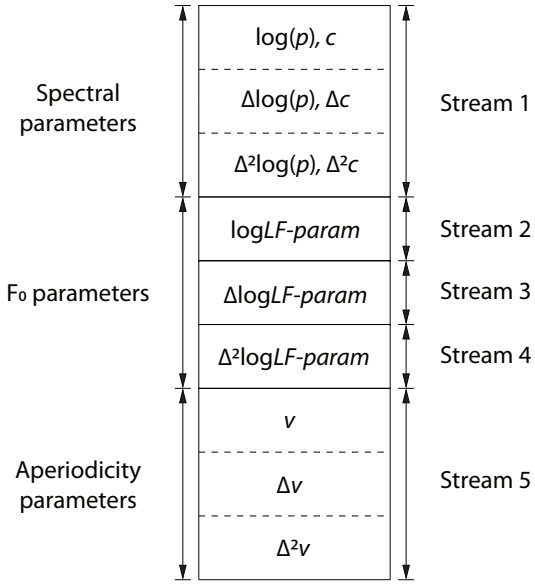

Figure 8.6: Multi-stream structure of the speech feature vector that includes the power parameter p, in the HTS-LF system.

8.3 Energy Adjustments of the Synthetic Speech

A method to adjust the energy of synthetic speech frames in the HTS-LF system was developed in this work, in order to reduce discontinuities in the energy envelope of the speech signal. For using this method in HTS-LF, the power parameter is estimated from recorded speech and it is modelled by the HMMs.

8.3.1 Statistical Modelling of the Power

The power parameter of the speech frame $s^i(n)$ is calculated as

$$p = \frac{1}{N}\sum_{n=1}^{N}\left(s^i(n)\right)^2, \tag{8.1}$$

where N is the number of samples of the speech signal. This parameter is then modelled in the same stream as the mel-cepstral coefficients which represent the vocal tract transfer function. The power and spectral parameters are expected to be correctly modelled in the same stream, as the power parameter is closely related to the c_0 mel-cepstral parameter. Figure 8.6 shows the structure of the speech parameter vector. The spectral parameter vector consists of the logarithm of the power ($\log p$) and the mel-cepstral coefficients, $\boldsymbol{c}$. The LF-model parameters, their dynamic features, and the aperiodicity features, $\boldsymbol{v}$, are modelled by different streams.

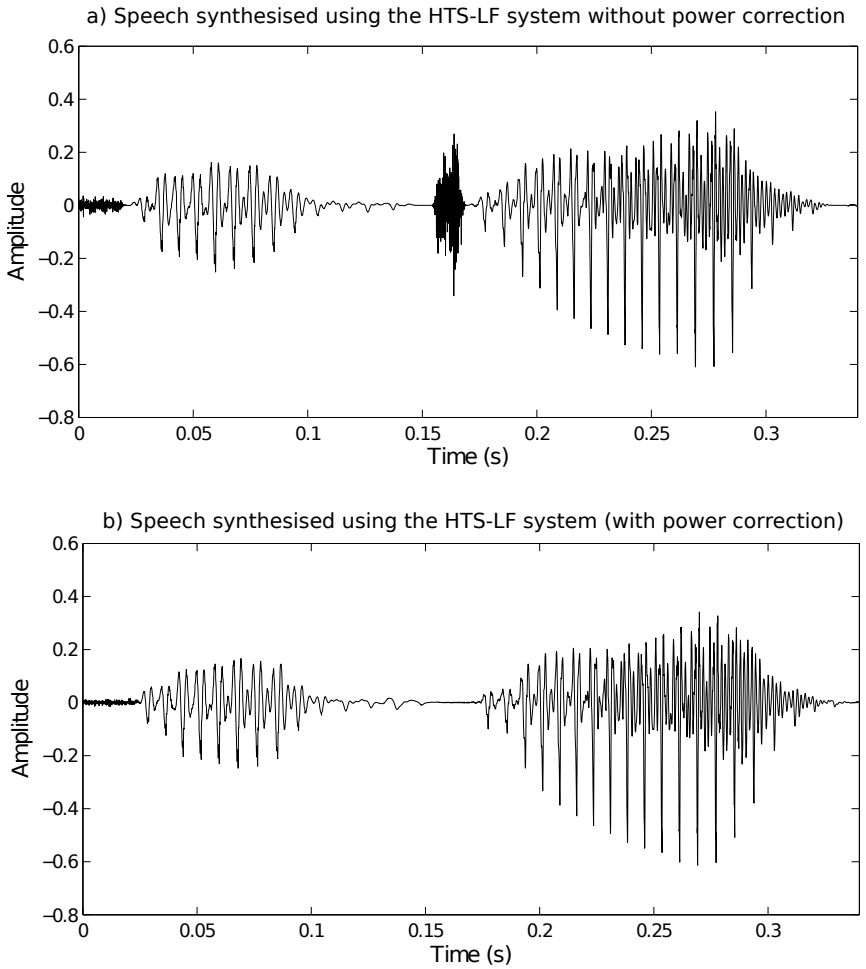

Figure 8.7: Speech segment of an utterance synthesised using the HTS-LF system. a) Speech synthesised without using the power correction algorithm of the HTS-LF system; b) Speech synthesised using the power correction algorithm.

8.3.2 Synthesis Using Power Correction

The HTS-LF system performs a power adjustment of synthesised speech frames before the overlap-and-add operation. The *power correction method* is used to reduce speech quality degradation caused by excessively high energy variations around voicing transitions. Figure 8.7 a) shows an example of excessively high energy noise produced by the HTS-LF system (without performing power correction) just before a transition between unvoiced (silence) and voiced speech (around the 0.16 s mark), which causes speech quality deterioration.

Two possible ways to perform the energy correction of the synthetic speech using the power parameter have been considered in this work. One way is to transform

the power of each speech frame so that it matches the power value generated by the system for that frame (obtained using the HMMs). The other way is to only correct the energy of the speech frames which are in the neighbourhood of voicing transitions. The second solution was chosen because it produced better results than the first.

The time intervals T_{v1} and T_{v2} were derived heuristically from experiments in order to correct the power before and after a voicing transition, respectively. For example, if there is a voicing transition from the frame i to the frame $i+1$, then the power correction is applied to the frames within the *voicing transition interval* $\left[t^i - T_{v1}, t^{i+1} + T_{v2}\right]$. t_i and t_{i+1} are the time instants of the central points of the frames i and $i+1$, respectively. The power correction algorithm is described in the following paragraphs.

If a synthetic speech frame $y^j(n)$ is within a voicing transition interval it is scaled in amplitude by a scale factor k_p^j, that is,

$$\tilde{y}^j(n) = k_p^j y^j(n) = \frac{e_t^j}{e_s^j} y^j(n), \tag{8.2}$$

where e_s^j is the energy of the synthetic speech signal $y^j(n)$ and e_t^j is the *target speech energy*. The energy of the signal $y^j(n)$ with length N is calculated as

$$e_s^j = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(y^j(n)\right)^2} \tag{8.3}$$

The target energy is calculated using the power contour generated by the synthesiser and the last synthetic speech frame $\tilde{y}_p^{j-1}(n)$ that was corrected in power, as follows:

$$e_t^j = \frac{p_j}{p_{j-1}}\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\tilde{y}^{j-1}(n)\right)^2}, \tag{8.4}$$

where p_{j-1} and p_j are the values of the power parameter generated by the synthesiser for the frames $j-1$ and j, respectively.

When a synthetic speech frame $y^j(n)$ is not within a voicing transition interval, it is scaled in amplitude by a scale factor k_g^j, that is,

$$\tilde{y}^j(n) = k_g^j y^j(n) = \frac{e_t'^j}{e_s^j} y^j(n), \tag{8.5}$$

where the target energy $e_t'^j$ is now calculated as

$$e_t'^j = \frac{e_s^j}{e_s^{j-1}}\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\tilde{y}^{j-1}(n)\right)^2} \tag{8.6}$$

In this equation, e_s^j represents the energy of the synthetic speech frame $y^j(n)$ and e_s^{j-1} is the energy of the last speech frame (not corrected in power). This amplitude scaling of the synthetic speech frames which are not in the voicing transition regions is performed to obtain a smooth energy variation between the last frame of the voicing transition region and the first frame of the non-transition region. This operation produces an energy contour in the non-transition regions of voicing which is the same as if the power correction was not performed in these regions, apart from a scale factor.

The amplitude scaling of the synthetic speech frames avoids the discontinuities of the energy contour in voicing transition regions. However, the amplitude scaling generally modifies the power of the speech frames in the non-transition region. This effect is because the target energy of the speech frame e_t^j is calculated from the energy of the previous frame. For example, the energy correction of the last speech frame $y^{i-1}(n)$ in a voicing transition region affects the energy of the first speech frame $y^i(n)$ in the next non-voicing transition region. If $y^i(n)$ is scaled by the factor k_p^i, then all the frames in the same non-voicing transition region are scaled by the same amount. This problem is overcome by scaling the whole voiced or unvoiced speech segment, v, by a factor k_t^v just after its last frame is synthesised. This *global scale factor* is calculated for each voiced and unvoiced segment as to match the energy of the segment if no power correction was performed, as follows:

$$k_t^v = \sqrt{\frac{\sum_{n=1}^{L} z_p(n)}{\sum_{n=1}^{L} z_o(n)}}, \tag{8.7}$$

where N is the length of the voiced/unvoiced segment, $z_p(n)$ is the voiced/unvoiced segment of speech synthesised with power correction and $z_o(n)$ is the voiced/unvoiced segment of speech synthesised without power correction.

The description of the algorithm for power correction is summarised in the following lists of steps.

Voiced-Unvoiced transitions:

1. amplitude scaling of each voiced frame j within $\left[t^i - T_{v1}, t^i\right]$ by k_p^j.
2. amplitude scaling of each unvoiced frame j within $\left[t^{i+1}, t^{i+1} + T_{v2}\right]$ by k_p^j.
3. amplitude scaling of each remaining frame j of unvoiced segment v, by k_g^j.
4. amplitude scaling of the whole unvoiced segment v by k_t^v.

Unvoiced-Voiced transitions:

1. amplitude scaling of each unvoiced frame j within $= \left[t^i - T_{v1}, t^i\right]$ by k_p^j.
2. amplitude scaling of each voiced frame j within $\left[t^{i+1}, t^{i+1} + T_{v2}\right]$ by k_p^j.
3. amplitude scaling of each remaining frame j of the voiced segment v, by k_g^j.
4. amplitude scaling of the whole voiced segment v by k_t^v.

Figure 8.7 shows an example of the effect of the power correction algorithm on the reduction of speech distortion in the HTS-LF system. In Figure 8.7 a), the speech segment synthesised without the power correction algorithm contains noise with excessively high energy just before the transition between unvoiced (silence) and voiced speech (around the 0.16 s mark). Figure 8.7 b) shows that the HTS-LF system using power correction does not produce this speech artefact (high energy noise).

8.4 Evaluation of HMM-based Speech Synthesisers Using LF-model

A subjective speech synthesis experiment was conducted in order to evaluate the HTS-LF system which incorporates the improvements described in the previous sections, an HMM-based speech synthesiser which incorporates the GPF method to generate the speech waveform and other statistical speech synthesisers which are variations of the HTS-STRAIGHT and the HTS-LF systems. The perceptual evaluation is based on the *Blizzard listening test setup*, which was conceived by Black and Tokuda (2005). This type of test was used mainly because it is adequate for evaluation of a relatively large number of speech synthesisers and it was designed to evaluate different speech quality aspects, such as speech naturalness, intelligibility, and similarity of the synthetic voice to the original speaker's voice.

The perceptual experiment conducted during this thesis is divided into four types of test: evaluation of *voice similarity*, evaluation of speech naturalness by *mean opinion scores* (MOS), evaluation of speech naturalness by *forced-choice pairwise comparison*, and evaluation of intelligibility. The Blizzard test was adjusted in order to incorporate the forced-choice part, since it did not originally include this type of speech naturalness evaluation. This evaluation is more complex and much more complete than the AB

listening test which was described in Section 7.4, in which the HTS-LF and HTS-STRAIGHT systems were compared in terms of speech naturalness only. Moreover, the experiment presented in this chapter includes the improved HTS-LF system and a larger number of systems.

8.4.1 Systems

Table 8.1 gives an overview of the principal systems used in the perceptual evaluation. The main goal of this experiment is to evaluate the speech quality of the following three systems: HTS-STRAIGHT, HTS-STRAIGHT using Glottal Post-Filtering (named HTS-GPF system) and HTS-LF. The baseline system (HTS-STRAIGHT) was described in Section 7.2. The difference between the HTS-STRAIGHT and HTS-GPF systems is during synthesis only. The first represents the excitation by mixing an impulse train with a noise signal (multi-band mixed excitation model), while the second uses a different signal to model the periodic component of the mixed excitation than the impulse train used by HTS-STRAIGHT. This signal is obtained by *whitening* the spectrum of a LF-model signal using a glottal post-filter, as described in Section 6.3.

	Systems		
	HTS-LF	HTS-STRAIGHT	HTS-GPF
Analysis	**IAIF: LF-parameters** ESPS tools: F_0, epochs GSS & STRAIGHT: v. tract STRAIGHT aperiodicity	ESPS tools: F_0, epochs STRAIGHT: spec. envelope STRAIGHT aperiodicity	
Excitation	Mix LF-model & noise	Mix imp. & noise	**Mix GPF & noise**
Synthesis	GSS synthesis	STRAIGHT	**FFT & OLA**
Evaluation	Naturalness, Intelligibility, Voice similarity		

Table 8.1: Summary of the characteristics of the HTS-LF, HTS-GPF, and HTS-STRAIGHT systems which were used in the perceptual evaluation (based on the Blizzard test setup).

Three other systems were also included in the experiment, which are variations of the HTS-STRAIGHT and HTS-LF systems. Table 8.2 summarises the characteristics of these systems. Two of them are versions of these synthesisers which do not use the noise component of the multi-band mixed excitation: HTS-STR-PR and HTS-LF-PR, respectively. These systems are used in order to study the effect of the noise component of the excitation on speech quality. The remaining system is a modified version of the HTS-STRAIGHT system, which uses a speech generation technique similar to that of HTS-LF instead of STRAIGHT. It allows us to compare the HTS-STRAIGHT and HTS-LF systems, avoiding any influence of the STRAIGHT vocoder on speech quality.

	Systems		
	HTS-LF-PR	HTS-STR-PR	HTS-FFT
Analysis	IAIF: LF-parameters ESPS tools: F_0, epochs GSS & STRAIGHT: v. tract STRAIGHT aperiodicity	ESPS tools: F_0, epochs STRAIGHT: spec. envelope STRAIGHT aperiodicity	
Excitation	**LF-model**	**Imp.**	Mix imp. & noise
Synthesis	GSS synthesis	STRAIGHT	**FFT & OLA**
Evaluation	Naturalness, Intelligibility, Voice similarity		

Table 8.2: Summary of the characteristics of the HTS-LF-PR, HTS-STR-PR, and HTS-STR-PR systems (the first is a variation of the HTS-LF system, while the others are variations of the HTS-STRAIGHT system) which were used in the perceptual evaluation (based on the Blizzard test setup).

8.4.1.1 HTS-LF

The HTS-LF system evaluated in this experiment incorporates the improvements described in Sections 8.2 and 8.3. That is, it uses the IAIF method to estimate the glottal source derivative signal, it uses an algorithm which corrects errors of the estimated

LF-model parameters and it uses a technique to adjust the energy of synthetic speech frames in voicing transition regions.

8.4.1.2 HTS-LF Without Noise Component of Excitation (HTS-LF-PR)

A version of the HTS-LF system which does not mix the LF-model signal with noise was also included in the experiment. The goal of using this system is to evaluate the importance of the noise component of the mixed excitation model of HTS-LF on speech quality.

8.4.1.3 HTS-STRAIGHT

The HTS-STRAIGHT system was described in detail in Section 7.2. It uses MATLAB STRAIGHT for analysis and synthesis. For speech synthesis, STRAIGHT processes the phase of the impulse signal by using the group delay function, as described in Section 4.3.3.2.

The original HTS-STRAIGHT system was modified in order to model the power parameter of speech by the HMMs. This parameter and its dynamic features (Δ and Δ^2) were added to the stream feature vector of the spectral parameters, which were the 39^{th} order mel-cepstral coefficients ant their dynamic features. Thus, the power parameter is modelled the same way as in the HTS-LF system. However, the power parameter is not used for speech synthesis by HTS-STRAIGHT. The purpose of modelling the power by HTS-STRAIGHT is to ensure that the difference in performance between the HTS-LF and HTS-STRAIGHT systems is not influenced by the effect of modelling the power parameter by HTS-LF. Although the power parameter modelling could affect the acoustic modelling of the spectral parameters (they are both in the same data stream), its effect is not expected to be significant. This assumption is based on the fact that the power parameter is closely related to the first mel-cepstral coefficient.

8.4.1.4 HTS-STRAIGHT Without Noise Component of Excitation (HTS-PR)

A variation of the HTS-STRAIGHT system which does not use the noise component of the mixed excitation was also used in the evaluation. This system has the same characteristics as the original HTS-STRAIGHT system, with the exception that the STRAIGHT synthesis program was modified so that it uses only the phase-manipulated impulse signal as the voiced excitation. That is, neither the spectrum of the impulse

signal is weighted using the aperiodicity parameters nor it is mixed with a noise component. This system, which is named HTS-PR, was used in this experiment in order to evaluate the effect of the noise component of the mixed excitation on speech quality.

8.4.1.5 HTS-STRAIGHT Using FFT-based Synthesis (HTS-FFT)

Another variation of the HTS-STRAIGHT system which was included in this evaluation, called HTS-FFT, uses an FFT-based processing technique to synthesise speech instead of STRAIGHT. This speech generation technique is similar to that used by the HTS-LF system.

The HTS-FFT system generates the excitation signal of voiced speech similar to the HTS-STRAIGHT system, by mixing a pulse signal (centered within a 1024 sample length frame to calculate the FFT) with a noise component. In this process, the two components are weighted in the frequency domain using functions defined by the aperiodicity parameters and added together. However, the phase of the pulse is not processed by HTS-FFT, in contrast to HTS-STRAIGHT. Next, the amplitude spectrum of the excitation is multiplied by the spectral envelope to obtain the speech spectrum. Finally, the speech signal is obtained by IFFT of the spectrum and then it is pitch-synchronously overlapped-and-added using a window centered at the pulse position. The main differences between this synthesis method and the STRAIGHT synthesis method are that STRAIGHT represents the speech spectrum by the minimum-phase impulse response (which is calculated from the spectral parameters) and it does not use the OLA technique.

The HTS-FFT system was used to compare the excitation model between the HTS-STRAIGHT and the HTS-LF systems, avoiding any influence of the STRAIGHT speech generation technique. Another reason for using the HTS-FFT system was to evaluate the speech waveform generation technique of the GSS method, compared with the STRAIGHT synthesis method (by comparing the HTS-STRAIGHT system against the HTS-FFT system).

8.4.1.6 HMM-based Speech Synthesiser Using Glottal Post-Filtering (HTS-GPF)

A version of the HTS-STRAIGHT system which synthesises speech using the GPF method was also developed. GPF was described in Section 6.3. Basically, it consists of using a glottal post-filter to transform a LF-model waveform into a spectrally flat signal. This signal is used to generate the mixed multi-band mixed excitation, instead

of the delta pulse signal.

The HMM-based speech synthesiser using GPF (HTS-GPF) uses the MATLAB STRAIGHT program to estimate the spectral envelope and aperiodicity parameters from the speech signal, as the original HTS-STRAIGHT system. The glottal post-filter is calculated by using the method described in Section 6.3.2.2. The way it was derived in this experiment is described in the following paragraphs.

The first process in the glottal post-filter calculation was to measure the LF-model parameters. The measurements were performed on eight utterances of the speech corpus. For the estimation of the LF-parameters T_a, t_p, and t_e, the LF-model was fitted pitch-synchronously to the glottal source derivative signal, by using a non-linear optimisation algorithm. This LF-model estimation method was the same as that used in the HTS-LF system, which was described in Section 7.3.1.

The LF-model measurements were used to calculate the mean values of the *dimensionless parameters*: OQ, SQ, and RQ. An estimate of the maximum F_0 of the speaker was also calculated. The t_e parameter of the *reference LF-model* was set approximately equal to the minimum T_0 of the speaker. Next, the other time parameters of the LF-model (t_p and t_a) were calculated by using the mean values of the dimensionless parameters and (5.12) to (5.14). In this way, the LF-model signal was short enough so as to avoid the problem of synthesising high-pitched speech (explained in Section 6.3.3.2) and the dimensionless parameter values were equal to the mean values obtained from the measurements.

Finally, the parameters of the glottal post-filter (the frequencies F_g and F_c) were calculated from the mean values of the LF-model parameters, using the method described in Section 6.3.2.2. The glottal post-filter was implemented as a 300^{th}-order FIR filter.

8.4.2 Speech Data

8.4.2.1 Speech Corpus

Two UK English speech databases were used to build the synthetic voices. They were provided by the Centre for Speech Technology Research. One is about ten hours of speech spoken by a male speaker which was obtained from the data released for the Blizzard Challenge 2009 (King and Karaiskos, 2009). The second contained about four hours of speech spoken by a female speaker. The male speech data is divided into two different subsets. One consists of a smaller set of phonetically-balanced sentences

taken from the CMU ARCTIC database (Kominek and Black, 2004), which is approximately one hour long. The second corresponds to sentences selected from news texts. Meanwhile, the female data corresponds to sentences selected from the news articles and from a novel.

8.4.2.2 Synthetic Voices

Three synthetic voices were built for the HTS-STRAIGHT and the HTS-LF, respectively, by using the speech databases. They were the following:

- Voice A: full voice from the male database.
- Voice B: voice from the ARCTIC subset of the male database.
- Voice C: female voice.

The acoustic models built for the HTS-STRAIGHT system were also used by the modified versions of this system (HTS-FFT, HTS-STR-PR, and HTS-GPF). This was possible because these modified systems differ from HTS-STRAIGHT only in terms of the method used to generate the speech waveform. For the same reason, the statistical models built for HTS-LF were used by the HTS-LF-PR system.

The phonetic labels of the speech data consisted of Festival utterance files created using the Unilex lexicon (Fitt and Isard, 1999). In addition to the phonetic transcription, they included contextual information such as segment, syllable, word and phrase level information.

8.4.3 Experiment

8.4.3.1 Speech Samples

For the male voices (voices A and B), the test sentences were the same as those of the Blizzard Challenge 2009 (King and Karaiskos, 2009), excluding the subsets which corresponded to the Blizzard Challenge 2007 and 2008 test sentences. The selected sentences were grouped in the following genres:

- 200 *news* sentences.
- 100 *novel* sentences.
- 100 *Semantically Unpredictable* (SU) sentences.

The novel and news sentences of the Blizzard Challenge 2009 were not used for the female speaker evaluation, because there were no recordings of the female speaker reading these test sentences. Instead, the test sentences were selected from a subset of sentences of the female speaker corpus, which was not used for voice building. This subset consisted of 100 news sentences. Test sentences of the genre "novel" were not used in the female voice listening test. However, the 100 SU sentences of the Blizzard Challenge 2009 were used as test sentences for the female voice evaluation, as recorded speech was not used to evaluate the SU sentences for this voice.

Each test sentence was synthesised by the six systems described in Section 8.4.1. From these sentences, the required number of sentences was randomly selected. For the full and ARCTIC male voices, the subset of sentences used for each of them consisted of:

- 42 news sentences.
- 35 novel sentences.
- 21 SU sentences.

For the female voice, only news and SU sentences were used. The randomly selected sentences consisted of 77 news sentences and 21 SU sentences.

8.4.3.2 Interface

The evaluation was conducted in a supervised perceptual lab at the University of Edinburgh. This lab was equipped with several rooms, which were especially designed for perceptual evaluations of audio. Each participant performed the evaluation in one of these rooms by using a computer interface and headphones. The estimated duration of the evaluation was 35 minutes.

The listening evaluation interface was based on the interface used for the Blizzard Challenge 2009 (King and Karaiskos, 2009). However, some adjustments were made to the original listening evaluation design. In the evaluation conducted in this thesis, there were five sections and each section was divided into a certain number of parts. The registration page contained instructions of the listening evaluation. Sections 1, 3, 4, and 5 of the test were very similar in design to sections of the Blizzard Challenge 2009, whereas Section 2 was designed specifically for this experiment. This section was designed in order to evaluate speech naturalness using an *ABX test*. This type of test has never been used in the Blizzard Challenge evaluations, apparently because it

would require too many utterances and listeners (the Blizzard Challenge evaluation usually includes a much higher number of systems). In this evaluation, the number of systems is not as high as the typical number of systems evaluated in the Blizzard Challenge, so the problem of a limited number of samples or listeners was not considered to be significant.

The listener tasks in each section of the test are described as follows:

- Section 1: Similarity (SIM) task. In each trial, listeners could play four reference samples of the original speaker and one synthetic sample. They were instructed to choose a response that represented how similar the synthetic voice sounded to the voice in the reference samples on a scale from 1 (“Sounds like a totally different person”) to 5 (“Sounds like exactly the same person”).

- Section 2: ABX task. In each trial, listeners heard one sample from each of two systems (A and B samples) most of the time. The exception was when listeners heard the same two samples (A and B were the same), which occurred once for each system ordering of the data set (explained later in Section 8.4.3.4). The samples of each pair corresponded to the same text sentence. For each pair of samples A-B, they then chose one of the three possible possible responses: (“A sounds more natural than B”), (“B sounds more natural than A”), and (“A and B sound equally natural”).

- Section 3: Mean Opinion Score (MOS) part, with speech samples from the *news* domain. In each trial, listeners heard one utterance and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 (“Completely Unnatural”) to 5 (“Completely Natural”).

- Section 4: Similar to MOS part of Section 3, but uses speech samples from the *novel* genre instead of news domain.

- Section 5: Intelligibility task, using SU sentences. Listeners were instructed to listen to one utterance in each trial and type what they heard. In the full male voice, the computer interface allowed the subject to play the sample more than once. The interface was modified for the evaluations of the ARCTIC and female voices, so that the subject was able to only listen to the utterance once.

8.4.3.3 Listeners

Ninety six undergraduate students from the University of Edinburgh were recruited to participate in the evaluation. They were all native speakers of UK English, aged 18-25 and received monetary compensation for their participation.

8.4.3.4 Listener Groups and System Orderings

Subjects were equally distributed among the three evaluations associated with the different voices. For each evaluation, each listener was assigned to a group. The number of groups was determined by the total number of systems in the evaluation, that is, the number of groups was 7 (equal to 6 systems plus the original speaker). Since the total number of listeners who participated in each voice evaluation was 32, there were three listener groups with 4 listeners and four other groups with 5 listeners.

For Sections 1, 3, 4 and 5 of the test, system orderings were systematically varied by using the *Latin square* design of the Blizzard setup (Fraser and King, 2007). Distinct Latin squares were constructed for all sections. The same Latin squares were used to evaluate the three voices, as each listener could only participate on the experiment once. The rows of a Latin square correspond to the listener groups and the columns correspond to the sentences. Then, each cell (i, j) of a 7 x 7 square represented the system that listener group i heard the sentence j. The sentence order was maintained across listener groups but the system order varied. Also, the position of a system in the Latin square of Section 3 (MOS news) was always different from its position in the Latin square of Section 4 (MOS novel). That is, the order of each system was never the same across the MOS sections. Moreover, the Latin Squares were designed so as to minimise possible ordering effects.

The Latin square associated with Section 5 of the evaluation was adjusted specifically for the female voice evaluation because there were no recordings of the female speaker reading SU sentences, unlike for the male speaker. This modification was similar to that described by Bennett (2005), which consisted of adding a row to an order 6 Latin square. The extra row was taken from another Latin Square of the same order. As a consequence, a row was repeated in each Latin square.

Section 2 of the evaluation was designed similarly to the *Multi-dimensional Scaling* (MDS) section of the Blizzard Challenge 2009 (King and Karaiskos, 2009). In this section, each listener group was assigned to 7 of the total 49 possible pairings of systems (including the original speaker). A *Graeco-Latin square* design was used to

distribute the pairs across the listener groups so that each pair was only repeated once in a different order, i.e. each system appeared once as the first and once as the second of a distinct pair, in each row of the square.

The test sentences used in each section of the evaluation were divided into different groups. Each sentence group was assigned to a Latin square (which determined the system orderings). The test sentences were also different between all the groups, with the exception of Section 2 of the test. In this section all the groups had the same set of sentences, to obtain a sufficiently high number of data points for each pairwise comparison. Table 8.3 shows the number of sentence groups that composed each section of the evaluation and the total number of sentences used in each section. The number of sentence groups of each section of the evaluation was chosen based on the importance which was given to each task, as the statistical significance of the results is strongly dependent on the number of samples of the test. The effect of glottal source modelling on speech naturalness was considered to be the most important aspect to be evaluated in this experiment. Therefore, the ABX and the MOS sections were given a higher number of sentence groups.

	Number of Groups	Total Number of Sentences
Section 1 (SIM)	3	21
Section 2 (ABX)	5	7
Section 3 (MOS)	4	28
Section 4 (MOS)	4	28
Section 5 (SUS)	3	21

Table 8.3: Number of sentence groups and total number of sentences of each section of the evaluation.

8.4.4 Results

The results of the perceptual evaluation are presented individually for each part of the listening test, in the next sections.

8.4.4.1 Similarity

In the first section of the Blizzard setup evaluation, listeners rated the similarity of a speech sample to the original speaker's voice by using a five point scale, which is an *ordinal scale*. The similarity results were analysed in terms of *medians*, as they are statistically meaningful for such scale (Clark et al., 2007a). Unlike the median, it is inappropriate to compare means on this type of scales.

Figure 8.8 shows the boxplot of the similarity scores between systems (including the natural speech) and the original speaker, for the three evaluations: full male voice, ARCTIC subset of the male voice, and female voice. The systems are ordered in descending order of the MOS means, although the ordering is not a ranking (the means are used to make the graphs more intuitive). The value of n in Figure 8.8 indicates the number of data points, which is the same for all systems. The median is represented by a solid bar across a box showing the quartiles. Whiskers extend to 1.5 times the inter-quartile range and outliers beyond this are represented as circles. *Pairwise Wilcoxon signed rank tests* between systems were calculated ($\alpha = 0.01$), in order to determine significant differences between systems.

Table 8.4 indicates the pairwise significance at 1% level ($p\text{-}value \leqslant 0.01$), for the different evaluations (full male voice, ARCTIC male voice, and female voice). Besides the median, the *median absolute deviation* (MAD), mean, and *standard deviation* (SD) values are also presented in Tables A.1 to A.2 (in Appendix A.1). The *p-values* calculated for the pairwise Wilcoxon signed rank tests are given in Tables A.3 to A.5 (in Appendix A.1).

Natural speech is significantly more similar to the original speaker ($p\text{-}value \simeq 0$) than all other systems. This result was expected, because natural speech was spoken by the original speaker. From Figure 8.8 and Table 8.4, the HTS-STRAIGHT system, the HTS-FFT system, and the system using the GPF method (HTS-GPF) scored the same, for all voices. These systems obtained a median score of 3 and are significantly more similar to the original speaker than the systems which use glottal source modelling (HTS-LF and HTS-LF-PR). The HTS-STR-PR system (HTS-STRAIGHT version which uses simple excitation) obtained the same score as the other versions of the

HTS-STRAIGHT system (HTS-STRAIGHT, HTS-FFT, and HTS-GPF), for the full male voice. However, HTS-STR-PR scored significantly lower in similarity compared with the same systems, for the female voice. The HTS-LF and HTS-LF-PR (HTS-LF without noise component) systems are equally similar to the original speaker. Finally, HTS-GPF is the only system which scored significantly higher than the HTS-STR-PR system, for the ARCTIC voice.

	S1	S2	S3	S4	S5
HTS-STRAIGHT (S1)					
HTS-GPF (S2)	*none*				
HTS-FFT (S3)	*none*	*none*			
HTS-STR-PR (S4)	***Fem.***	***Fem.***, ***ARCTIC***	***Fem.***		
HTS-LF (S5)	***All***	***All***	***All***	***All***	
HTS-LF-PR (S6)	***All***	***All***	***All***	***All***	*none*

Table 8.4: Significance difference of similarity scores between systems ($p < 0.01$), for the three voices: male full voice, ARCTIC subset of male voice and female voice. "*none*" means that the result is not significant for any voice and "***All***" means that it is significant for all the voices.

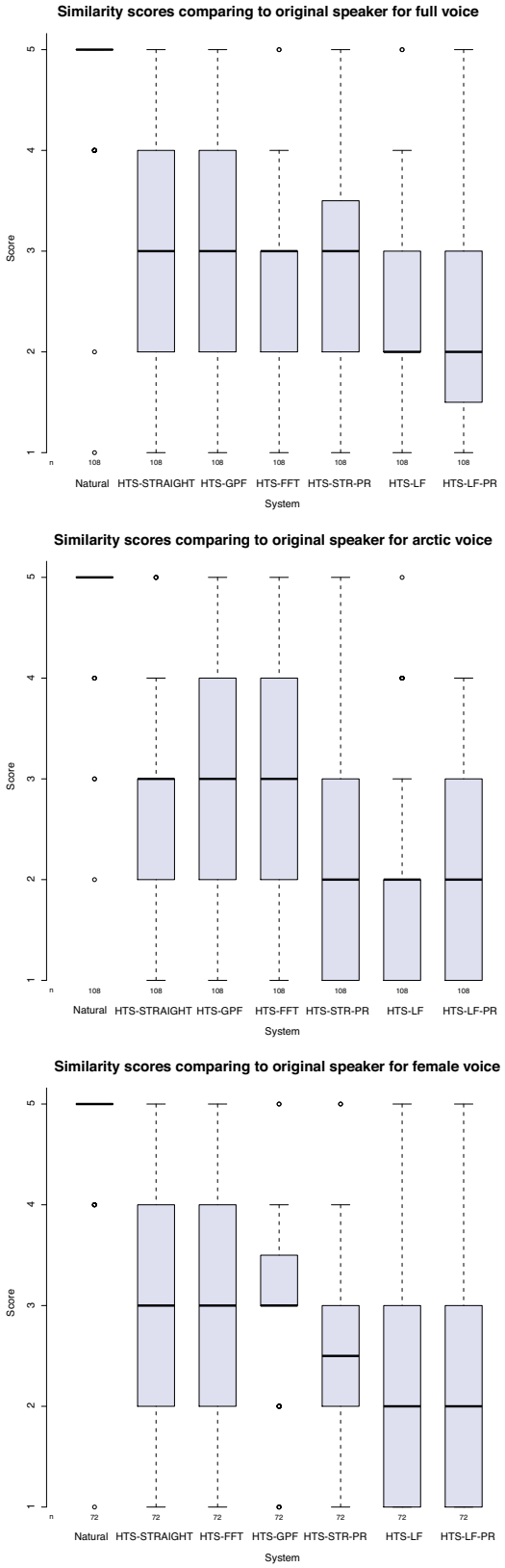

Figure 8.8: Similarity scores between systems and the original speaker (natural speech) for the three voices: full male voice, ARCTIC subset of the male voice, and female voice.

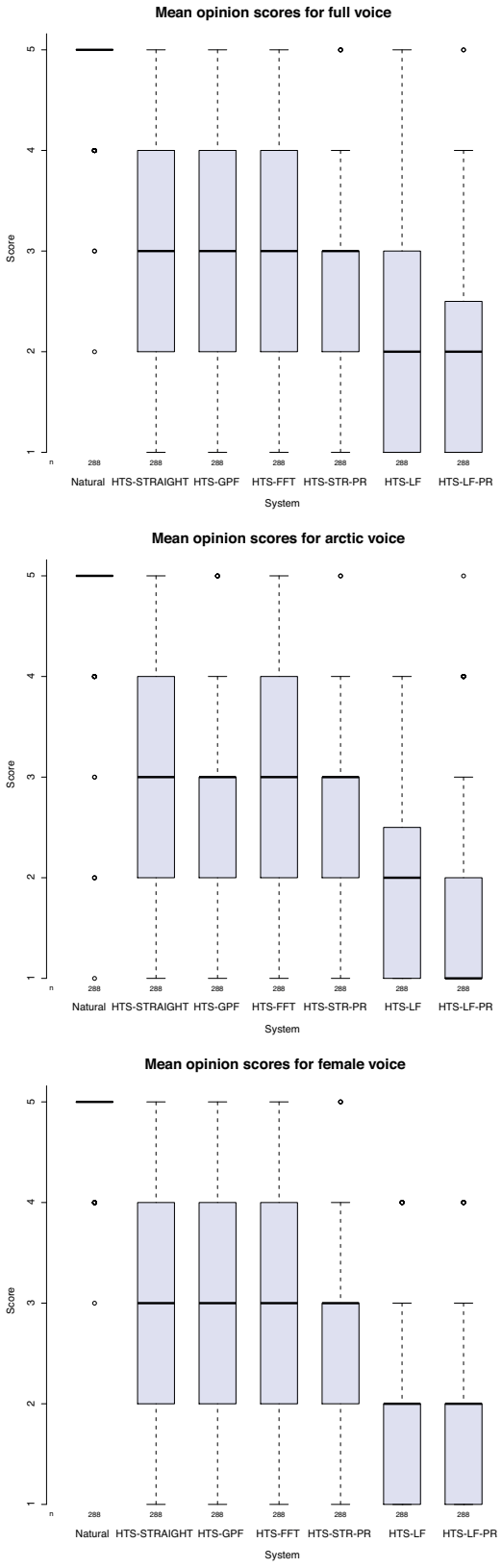

Figure 8.9: Mean opinion scores calculated for the news and novel sentences (Section 3 and 4 of the evaluation) and the three voices: full male voice, ARCTIC subset of the male voice, and female voice.

8.4.4.2 Naturalness - MOS

Mean Opinion Scores (MOS) were rated on a five point scale by the listeners, as in the similarity task. Therefore, median scores were again used for comparison. The MOS results were calculated for the Sections 3 and 4 of the evaluation together, which correspond to the news and novel test sentences respectively. The analysis was not performed for the news and novel test sentences separately, as to maintain the highest possible number of listener responses in the MOS evaluation. Figure 8.9 presents the MOS results for the three voices, by using a boxplot which is the same type as that used to show the similarity scores in Figure 8.8. Results of the pairwise Wilcoxon signed rank significance tests are shown in Table 8.5. Values of the median, MAD, mean, SD, and *p-values* of the significance tests are also presented in Tables A.6 and A.7 (in Appendix A.2). The *p-values* calculated are given in Tables A.8 to A.10 (in Appendix A.2).

	S1	S2	S3	S4	S5
HTS-STRAI. (S1)					
HTS-GPF (S2)	***Full***, ***Arctic***				
HTS-FFT (S3)	***Full***	*none*			
HTS-STR-PR (S4)	***All***	***Fem.***, ***Full***	***Fem.***, ***Arctic***		
HTS-LF (S5)	***All***	***All***	***All***	***All***	
HTS-LF-PR (S6)	***All***	***All***	***All***	***All***	***Fem.***, ***Arctic***

Table 8.5: Significance difference of MOS scores between systems ($p < 0.01$), for the three voices: male full voice, ARCTIC subset of male voice and female voice. "*none*" means that the result is not significant for any voice and "***All***" means that it is significant for all the voices.

From the results, natural speech is always significantly more natural than the synthetic speech for every HMM-based speech synthesiser, with *p-value*$\simeq 0$.

From Figure 8.8 and Table 8.5, the HTS-STRAIGHT system and its variations (HTS-GPF, HTS-FFT, and HTS-STR-PR) are equally natural as each other and they are all significantly more natural than the synthesisers which use glottal source mod-

elling (HTS-LF and HTS-LF-PR), for all the voices.

Finally, the HTS-LF system is as natural as HTS-LF-PR (without noise excitation) for the full and ARCTIC male voices. However, HTS-LF is significantly more natural than HTS-LF-PR, for the female voice.

8.4.4.3 ABX - Naturalness

In the ABX task of the evaluation, subjects were presented with pairs of utterances from different systems (the same sentence for each pair A-B), and were asked which utterance sounded more natural (A or B). They also had the option to answer that both utterances sounded equally natural (option X). Since this is a pairwise comparison test, the results are presented in terms of preference rates of a system (including natural speech) compared against a different system. The results of the preference rates and significance tests (*p-value*) obtained for every system in the three evaluations (full male voice, ARCTIC voice, and female voice) are presented in Appendix A.3. Table 8.6 summarises the statistical significance of the pairwise comparisons. The results which are statistically significant are described in the following paragraphs.

Natural speech was significantly preferred (*p-value*$\ll 0.01$) over all systems with preference rates higher than 90%, for the different voices.

	S1	S2	S3	S4	S5
HTS-STRAIGHT (S1)					
HTS-GPF (S2)	*none*				
HTS-FFT (S3)	*none*	*none*			
HTS-STR-PR (S4)	*none*	*none*	*none*		
HTS-LF (S5)	***Full***, ***Fem.***	***Full***, ***Fem.***	***All***	***Fem.***, ***Arctic***	
HTS-LF-PR (S6)	***All***	***All***	***All***	***All***	*none*

Table 8.6: Significance difference of ABX pairwise comparisons between systems ($p < 0.01$), for the three voices: male full voice, ARCTIC subset of male voice and female voice. "*none*" means that the result is not significant for any voice and "***All***" means that it is significant for all the voices.

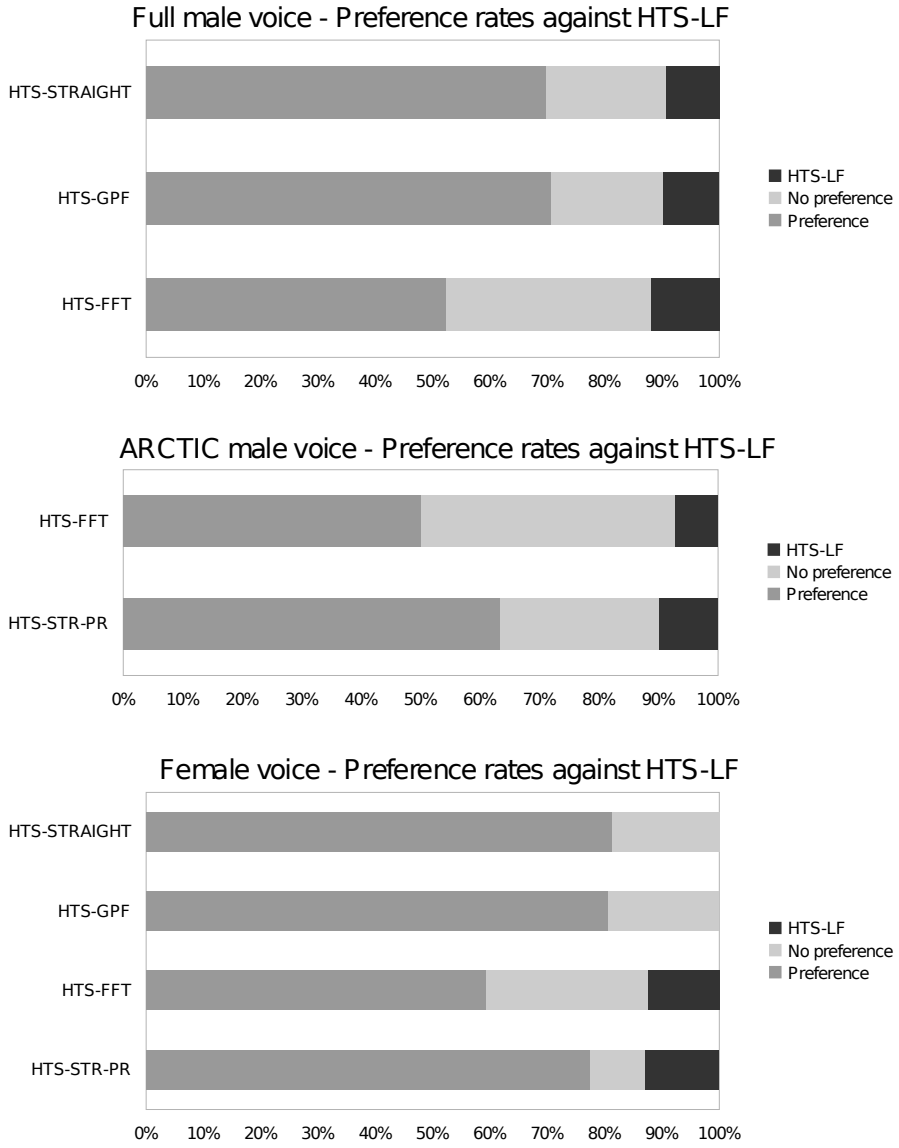

Figure 8.10: Preference rates obtained from the ABX comparisons (which were statistically significant), for the comparisons of the different systems againts the HTS-LF system. The results are presented for the different voices: full male voice, ARCTIC subset of the male voice, and female voice.

Similar to the MOS results, the HTS-STRAIGHT system and its variations (HTS-STR-PR, HTS-GPF, and HTS-FFT) are not significantly different between each other in terms of speech naturalness. However, they are significantly more natural than the synthesisers which use glottal source modelling (HTS-LF and HTS-LF-PR), in general. The two types of HTS-LF systems are also equally natural as each other.

Figure 8.10 shows the significant preference rates obtained from the pairwise comparisons of the systems against HTS-LF, for the three voices. The HTS-STRAIGHT

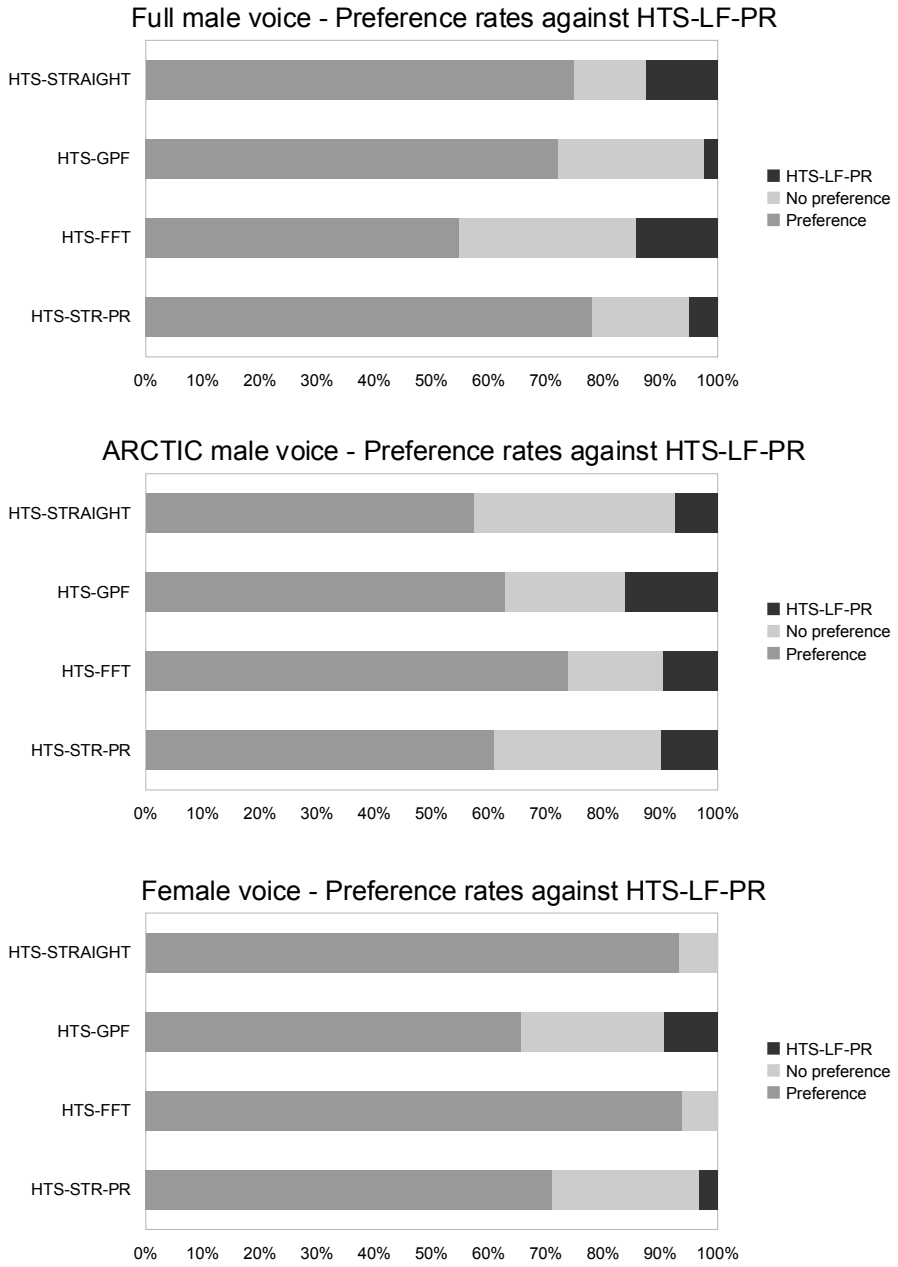

Figure 8.11: Preference rates obtained from the ABX comparisons (which were statistically significant) for the comparisons of the different systems against the HTS-LF-PR system. These results are presented for the different voices: full male voice, ARCTIC subset of the male voice, and female voice.

based systems were generally preferred over the HTS-LF system at least 50% of the time, whereas the highest preference rates obtained by the HTS-LF system against these systems is about 12%. The highest preference rates obtained by the group of HTS-STRAIGHT systems were around 80%, for the female voice. An exception to

these results is that HTS-STRAIGHT was not significantly different from the HTS-LF system for the ARCTIC male voice. In addition, the HTS-FFT and the HTS-STRAIGHT-PR systems were significantly more natural than the HTS-LF system, for the ARCTIC voice.

Figure 8.11 shows the significant preference rates obtained for the pairwise comparisons of the systems against the HTS-LF-PR system. From these results, all the HTS-STRAIGHT based systems were significantly more natural than the HTS-LF-PR system, with preference rates ranging from 50% to 90%, while the highest HTS-LF-PR score was about 14%.

The pairwise comparison between a speech synthesiser which uses mixed excitation against the same synthesiser using simple excitation was never significant at $p\text{-}value < 0.01$ That is, HTS-LF and HTS-STRAIGHT were not significantly different from the HTS-LF-PR and HTS-STR-PR systems, respectively.

8.4.4.4 Intelligibility

In Section 5 of the evaluation, subjects were presented with a SU sentence in each trial and were asked to type in what they heard. A *word error rate* (WER) score for each sample was calculated. This scale is an interval, so it is appropriate to compare WER results in terms of the means. For the male voices, natural speech was also included in the intelligibility evaluation. However, natural speech was not part of the stimuli of the female voice, because no recorded SU sentences were available for this voice. Figure 8.12 shows bar charts which represent the mean word error rates for the different systems (obtained for the three voices). The statistical significance of these results is shown in Table 8.7. The mean WER, SD values, and *p-values* of the significance tests can be found in Appendix A.4.

The trends found in the similarity and naturalness results continue for intelligibility. Natural speech is significantly more intelligible than the speech samples synthesised by every system. Also, the HTS-STRAIGHT system and its variations (HTS-GPF, HTS-FFT, and HTS-STR-PR) cannot be differentiated from one another in terms of intelligibility. However, they are significantly more intelligible than the systems which use glottal source modelling (HTS-LF and HTS-LF-PR). Moreover, HTS-LF and HTS-LF-PR are equally intelligible.

The intelligibility of the systems was not compared across the different types of voice (full male voice, ARCTIC voice and female voice), because certain evaluation factors were different between them. For example, all the listeners were different be-

tween the evaluations of the different voices. Furthermore, the intelligibility part of the evaluation was slightly different between the full male voice and the other two voices. In the first, subjects were instructed to listen to a speech sample once, although the interface allowed them to play it more than once. The evaluation interface of the two other voices was modififed so that listeners could only hear each sample once.

	S1	S2	S3	S4	S5
HTS-STRAIGHT (S1)					
HTS-GPF (S2)	*none*				
HTS-FFT (S3)	*none*	*none*			
HTS-STR-PR (S4)	*none*	*none*	*none*		
HTS-LF (S5)	***All***	***All***	***All***	***All***	
HTS-LF-PR (S6)	***All***	***All***	***All***	***All***	*none*

Table 8.7: Significance tests of WER ($p < 0.01$), for the three voices: male full voice, ARCTIC subset of male voice and female voice. "*none*" means that the result is not significant for any voice and "***All***" means that it is significant for all the voices.

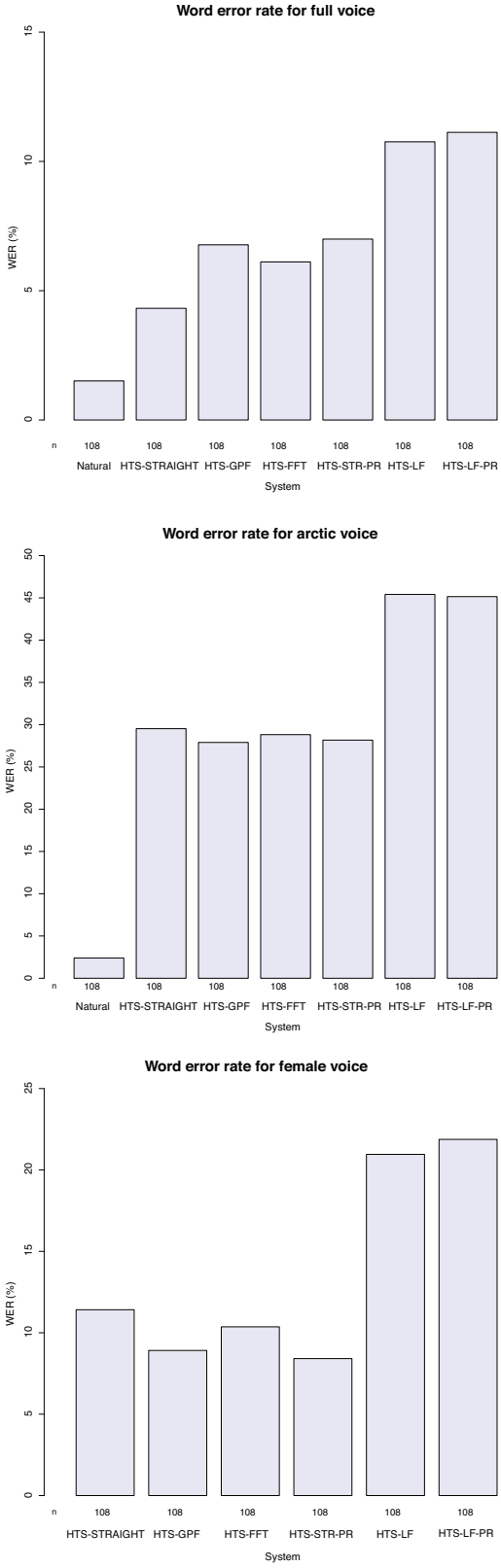

Figure 8.12: Word error rates of the systems and natural speech for the three voices: full male voice, ARCTIC subset of the male voice, and female voice. The natural speech was not evaluated in the intelligibility task for the female voice.

8.4.5 Discussion

8.4.5.1 HTS-STRAIGHT and HTS-LF Groups of Systems

Natural speech was always rated significantly better than synthetic speech, in all sections (SIM, ABX, MOS, and WER). Results show a clear difference in performance between two groups of systems, for the three voices. The first group consists of HTS-STRAIGHT and the different versions of this system (HTS-GPF, HTS-FFT, and HTS-STR-PR). The second consists of the systems using glottal source modelling (HTS-LF and HTS-LF-PR). In general, the systems in the HTS-STRAIGHT group scored significantly higher than the systems in the HTS-LF group in terms of similarity to the original speaker, naturalness, and intelligibility. However, it is not possible to state which system of the HTS-STRAIGHT group is the best because there is not a system which is significantly better than the others in any of the evaluation sections. For most of the cases, the systems in the HTS-LF group are also equally natural, intelligible, and similar to the original speaker.

8.4.5.2 HTS-GPF

The synthesiser which uses the GPF method (HTS-GPF) performed as well as the other systems which use STRAIGHT and a multi-band mixed excitation (HTS-STRAIGHT and HTS-FFT). This result indicates that the use of a flattened LF-model signal for the excitation does not affect significantly the speech quality of the synthesiser, when compared to the impulse signal. Nevertheless, HTS-GPF was expected to outperform HTS-STRAIGHT, because the LF-model obtained better results than the impulse, when speech was synthesised using the GSS waveform generation method in the perceptual experiment in Section 6.6. Nevertheless, the HTS-GPF system was expected to outperform HTS-STRAIGHT, because the impulse train was assumed to have a stronger harmonic structure than the spectrally flattened LF-model signal (strong harmonic structure is a cause of buzziness), as explained in Section 6.3.3.3. Possibly, the perceptual difference between the two signals (periodic components of the mixed excitation) is not significant because both were mixed with noise.

8.4.5.3 GSS Synthesis Method

The type of synthesis method used by the HTS-STRAIGHT synthesiser, either the STRAIGHT vocoder or the waveform generation technique of the HTS-FFT system,

did not significantly affect the speech quality. Therefore, the synthesis method used by the HTS-FFT system is competitive to the STRAIGHT method.

8.4.5.4 Mulit-Band Mixed Excitation Model

The results can also be analysed in terms of mixed excitation model versus simple excitation model (which does not use the noise component of the mixed excitation).

The similarity to the original speaker is not significantly affected by the noise component of the excitation for the male voice. However, for the female voice the HTS-STRAIGHT system using simple excitation (HTS-STR-PR) is lower in similarity than the HTS-STRAIGHT system using mixed excitation.

Using mixed excitation was shown to improve speech naturalness compared to simple excitation, for the female voice, although this factor was significant for the HTS-LF systems only. On the contrary, there was not a significant difference in naturalness between the HTS-STRAIGHT system its version using simple excitationand HTS ARCTIC subset of the male voice and the female voice.

8.4.5.5 Hypothesis to Explain the HTS-LF Results

The HTS-LF system was expected to produce higher speech quality than the HTS-STRAIGHT system, because it uses a more accurate model of the glottal source than the impulse train used by the HTS-STRAIGHT system. This hypothesis was supported by the results obtained in the AB perceptual test (presented in Section 7.4.2), which was conducted to evaluate the HTS-LF system using inverse filtering. Moreover, it was expected that by using the IAIF method instead of inverse filtering in HTS-LF, the preliminary results in Section 7.4.2 could be improved even further. However, HTS-LF was outperformed not only by HTS-STRAIGHT but also by the HTS-FFT system (which uses a speech waveform generation method similar to that of the HTS-LF system instead of STRAIGHT).

The use of the GSS method for analysis and synthesis is not expected to be an important reason for the speech quality degradation in the HTS-LF system, because the GSS method performed well in the copy-synthesis experiment presented in Section 6.6. However, the method used by GSS to estimate the glottal source derivative signal from speech appears to have influenced the performance of the synthesiser. It is assumed that when the IAIF method was used to more accurately estimate the glottal source derivative signal, the statistical modelling or the parameter generation parts of

the synthesiser unexpectedly performed worse. Errors in the LF-model parameter estimation are also not expected to be an important cause of speech distortion, because the LF-model waveform seemed to fit to the estimated glottal source derivative signal well. The performance of the fitting method was considered acceptable based on the visual comparison conducted by the author between the fitted LF-model signal and the respective glottal source signal, for every speech frames of several utterances. Also, the estimated LF-model parameters did not satisfy the LF-model constraints for a relatively low number of speech frames per utterance, on average. This number of frames was usually less than ten (an utterance typically had hundreds of voiced frames). Moreover, the detected LF-parameter errors were always successfully corrected by the parameter correction algorithm. Future experiments could be conducted for better evaluating the robustness and accuracy of the LF-model parameter estimation method developed in this work.

The hypothesis to explain the results of the HTS-LF system is that the vocal tract representation of the voiced speech spectrum used by this speech synthesiser negatively affects statistical modelling of the spectral parameters. The reason for this is that the vocal tract representation is different from the spectral envelope used to model unvoiced speech, which results in significant variations of the spectral parameters (estimated from recorded speech) between contiguous frames at voicing transitions. On one hand, this spectral discontinuity could contribute to degradation of statistical modelling. On the other hand, the speech parameter generation algorithm of the synthesiser is not appropriate for reproducing abrupt variations at voicing transitions, as it was developed to generate a smooth parameter contour by using both the static and delta parameters. Therefore, errors in the generated spectral parameter contours around voicing transitions are assumed to be sufficiently high to deteriorate speech quality.

Speech energy distortion was observed in speech synthesised by the HTS-LF system using inverse filtering (also observed for HTS-LF using IAIF), as described earlier in Section 7.4.2. This type of distortion can also be explained by the hypothesis that the spectrum is not modelled at voicing transitions correctly. Both the statistical modelling and the speech generation algorithm attenuate the spectral variation between the spectral envelope of unvoiced speech and the vocal tract representation of voiced speech at voicing transitions. If the variations between the two types of spectral representation is significant, then the error between the spectrum estimated from the original speech and the spectrum generated by the system could produce the excessively high energy variations observed in the synthetic speech.

Speech distortions due to energy discontinuities are not a known problem in the HTS-STRAIGHT system. This is explained by the fact that the energy of the synthetic speech is determined by the spectral envelope of the speech signal in both unvoiced and voiced regions, which is sufficiently smooth to be accurately modelled at voicing transitions by the HMMs.

8.5 Conclusion

Several transformations performed on the HTS-LF system in order to improve the quality of the synthetic speech were described in this chapter. The IAIF method was implemented into this system, since this method can more accurately estimate the glottal source derivative than inverse filtering using pre-emphasis. The objective of this modification was to improve the estimation of the LF-model parameters and modelling of the glottal source signal and vocal tract transfer function. In order to improve the robustness of the LF-model parameter estimation, an algorithm to validate the constraints of the LF-parameters and correct this type of errors was also developed and employed in the system for speech analysis and synthesis. Another modification made to the system was to extend the spectral parameter vector used by HMMs to include the speech power and to adjust the energy of the synthetic speech frames using this parameter. This method was used in order to overcome the energy distortion problem around transitions of voiced-unvoiced and unvoiced-voiced sounds, in the HTS-LF system.

A perceptual evaluation based on the Blizzard test setup was conducted in order to evaluate the performance of the two HMM-based speech synthesisers which use the LF-model, developed in this work. One was the HTS-LF system (with improvements) and the other was the HTS-GPF system, which is a variation of the HTS-STRAIGHT system that uses the GPF method for synthesis. The baseline system was the HTS-STRAIGHT system. Moreover, variations of these system were also included in the evaluation. They were the HTS-LF system without using the noise component of the excitation (the HTS-LF-PR system), the HTS-STRAIGHT system using simple excitation (the HTS-STR-PR system), and a version of the HTS-STRAIGHT system which used a speech waveform generation method similar to that of the HTS-LF system (the HTS-FFT system), instead of the STRAIGHT synthesis method. The HTS-LF-PR and HTS-STR-PR systems were both used in order to study the effect of the noise component of the excitation on speech quality and the HTS-LF-FFT system was used in order to evaluate the effect of the STRAIGHT synthesis method on speech quality, compared

with the waveform generation method used by HTS-LF (GSS synthesis method).

The results of the perceptual evaluation based on the Blizzard test setup are summarised as follows:

- HTS-STRAIGHT based systems (HTS-STRAIGHT, HTS-GPF, HTS-FFT, and HTS-STR-PR) outperformed the HTS-LF based systems.
- HTS-GPF performed as well as HTS-STRAIGHT.
- the HTS-STRAIGHT system (with mixed excitation) was significantly better than HTS-STR-PR (with simple excitation), in the speech naturalness test for the full male voice and in the similarity test for the female voice. For the rest of the results, there was no significant difference between the two systems.
- HTS-LF with mixed excitation performed better than HTS-LF without noise component of the excitation, only in terms of speech naturalness for the female voice. For all other parts of the evaluation, the performance was the same.

Part of the results of the perceptual evaluation were expected. The HTS-GPF system was expected to perform at least as well as the HTS-STRAIGHT system, because the spectrally flattened LF-model signal was expected to reduce the buzziness compared to the impulse train. Also, the good performance of the waveform generation technique of the GSS synthesis method, when compared to STRAIGHT, is supported by the good results obtained by the GSS method in the copy-synthesis experiment presented in Section 6.6.5. The positive results obtained by the mixed excitation compared to the simple excitation for the female voice were also expected. They are in agreement with other results reported in the literature, e.g. Yoshimura et al. (2001), which show that the mixed excitation model improves speech naturalness in HMM-based speech synthesis. The perceptual evaluation conducted in this work shows that the mixed multi-band excitation can also be important to voice similarity, particularly for the female voice. However, the mixed excitation model did not always improve naturalness and similarity to the original speaker's voice.

The results obtained by the HTS-LF system were expected to be at least as good as those obtained by the HTS-STRAIGHT system. The preliminary evaluation of the HTS-LF system (with inverse filtering instead of IAIF) indicated that this system could outperform the HTS-STRAIGHT system. The explanation for the low scores obtained by HTS-LF is that there is a problem in modelling the speech spectrum around voicing

transitions. This problem is assumed to be caused by rapid fluctuation of the spectrum at voicing transitions, related to the fact that spectral parameters represent the spectral envelope for unvoiced speech, whereas they represent the vocal tract filter for voiced speech. The deterioration in speech quality due to this problem appeared to be higher in this perceptual evaluation (the HTS-LF system used the IAIF method) than in the previous evaluation in which the HTS-LF system used inverse filtering (presented in Section 7.4). The interpretation of this result is that the IAIF method estimates the glottal source derivative signal more accurately than the inverse filtering technique, which results in increased differences between the vocal tract representation of voiced speech and the spectral envelope of unvoiced speech. For example, the spectral tilt of the glottal source derivative estimated by IAIF is usually higher than the spectral tilt of the residual calculated by inverse filtering with pre-emphasis.

Chapter 9

Analysis of Speech Distortion in the HTS-LF System

9.1 Introduction

The preliminary evaluation of the HTS-LF system presented in Section 7.4 indicated that this system was at least as good as the baseline, the HTS-STRAIGHT system. The HTS-LF system was then modified in order to improve its speech analysis and in order to reduce the speech distortion which was observed in the energy contour of the synthetic speech around *voicing transition instants* (voiced-unvoiced and voiced-unvoiced speech frame transitions). These improvements were described in Section 8.2. However, the results of the perceptual evaluation presented in Section 8.4 showed that the upgraded HTS-LF system was significantly outperformed by the HTS-STRAIGHT system.

The *objective measurement experiment* presented in this chapter was conducted in order to investigate the causes of the unexpected poor speech quality of the HTS-LF system. In this experiment, several speech properties were compared between the synthetic speech produced by the HTS-LF and HTS-STRAIGHT systems. The general aspects of the HTS-LF system which differentiate it from the HTS-STRAIGHT system are summarised as follows:

- Speech analysis: LF-model and spectral parameter estimation.
- Statistical modelling: additional LF-parameters.
- Speech waveform generation method.

The differences between the two systems and the hypothesis for the lower speech quality of the HTS-LF system are discussed in the following paragraphs.

The HTS-STRAIGHT system uses the STRAIGHT vocoder to estimate spectral envelope parameters during speech analysis, whereas the HTS-LF system uses the GSS analysis method. In this method, LF-model parameters are estimated from the speech signal and they are then used to remove the spectral effects of the LF-model signal from the speech spectrum. The vocal tract transfer function is estimated by computing the spectral envelope of the resulting signal using STRAIGHT. Both systems compute the F_0 and aperiodicity parameters by using the RAPT algorithm (Talkin and Rowley, 1990) and STRAIGHT respectively. Based on the comparison of the analysis methods used in the two systems, it is assumed that any problems during the analysis part of the HTS-LF system which could explain the poor performance of this system are related to the *LF-model estimation method* and the *voiced/unvoiced classification*. The relevant problem related to voicing classification in HTS-LF is that when a speech frame is wrongly classified as voiced, the IAIF method incorrectly estimates the glottal source derivative, since the excitation of unvoiced speech has the characteristics of white noise. Consequently, LF-model parameter estimation errors will occur for those speech frames. LF-model parameter errors also affect spectral parameter estimation in the GSS method, because this method uses the amplitude spectrum of the estimated LF-model waveform to separate the spectral characteristics of the the glottal source (the spectral tilt and the "glottal formant") from the speech signal. The case of a speech frame being wrongly classified as unvoiced is not considered to be important, as the effect of this error is the same as in the HTS-STRAIGHT system. That is, the spectral parameters of unvoiced speech (represent the spectral envelope) and the F_0 estimate are equal between the two systems.

The second point which differs between the two systems is the statistical modelling. There are two factors which could deteriorate speech parameter modelling in the HTS-LF system, when compared with the HTS-STRAIGHT system. One factor is that *errors in the LF-model parameter estimation* degrade the modelling of the speech features by HMMs. These errors might deteriorate not only the statistical modelling of the LF-model parameters but also the spectral parameters which are calculated using the GSS method. F_0 modelling could also be affected by LF-model parameter errors, as the F_0 and the other glottal source parameters are modelled in the same feature vector stream. The other factor which could deteriorate speech modelling is that the HTS-LF system uses a *different representation of the spectrum* for voiced and un-

voiced speech: the vocal tract transfer function and the spectral envelope, respectively. In contrast, the HTS-STRAIGHT system represents the spectrum of voiced and unvoiced speech by the spectral envelope. As result, for the HTS-LF system there is a higher variation of the spectral parameters between contiguous frames at a voicing transition than for the HTS-STRAIGHT system. The HMMs are not expected to accurately model this rapid fluctuation of the spectrum, due to the averaging characteristic of statistical modelling. Also, high spectral parameter discontinuities might degrade the modelling of this type of parameter by continuous HMMs. For example, even if the unvoiced and voiced speech frames of a voicing transition were modelled by different HMMs states, discontinuities of the dynamic features of the spectrum (Δ and Δ^2) could occur due to the spectral mismatch in the voicing transition frames. In addition, the feature generation algorithm of the HTS-LF system does not take into account the abrupt fluctuations of the spectrum at voicing transitions, as the algorithm attempts to generate smooth trajectories. The problem of correctly modelling spectral parameters at voicing transitions in the HTS-LF system could explain the speech distortions which were sometimes observed in speech synthesised by this system. As explained in Section 7.4.2, these distortions were the excessively high energy of noise in unvoiced frames next to voicing transitions and amplitude peaks in voiced frames next to voicing transitions. The power correction algorithm used in the HTS-LF system was developed in order to reduce these errors in the energy contour of the synthetic speech. However, it might not solve this problem completely. Furthermore, the power correction cannot solve possible spectral distortions of the synthetic speech, which are associated with the limitation of the synthesiser to model rapid fluctuations of the spectral parameters at voicing transitions.

The third difference between the HTS-LF and HTS-STRAIGHT systems is the *waveform generation technique*. The first system uses the GSS synthesis method developed in this work, whereas HTS-STRAIGHT uses the STRAIGHT vocoder. However, the speech generation method is not expected to have contributed to the degradation of speech quality in the HTS-LF system. This assumption is based on the results of the perceptual evaluation presented in Section 8.4.3, which showed that the HTS-STRAIGHT system performed similarly when it used the original STRAIGHT vocoder and when it used the same waveform generation method as that used by the HTS-LF system.

The possible reasons for speech quality degradation in the HTS-LF system which are considered in this experiment are:

- Problem in modelling the spectrum in voicing transition regions by HMMs.
- LF-model parameter estimation errors.
- Voiced/unvoiced classification errors.

The statistical modelling problem, which is due to the mismatch between the vocal tract representation and the spectral envelope at voicing transitions, is expected to be the most significant cause of speech quality deterioration in the HTS-LF system. Errors in the LF-model parameter estimation are assumed to be less important than the statistical modelling problem, because the HTS-LF system performed reasonably well in the preliminary evaluation presented in Section 7.4 (before the improvements to the LF-parameter estimation were implemented in the synthesiser). Also, from informal analysis of the F_0 and LF-model parameter contours, they appear to be smooth and similar to the contours obtained from the analysis of natural speech.

The next section describes the objective measurement experiment. In the subsequent sections, the methods used to measure each type of acoustic measurement are described and the respective results are presented. The correlation coefficients between the objective measurements and the perceptual test scores were also calculated and the results are presented in Section 9.6. This chapter ends with the overall discussion of the results and the conclusions.

9.2 Experiment

9.2.1 Overview

The objective measurement experiment described in this chapter consisted of measuring acoustic differences between the synthetic speech signals generated by the HTS-LF and HTS-STRAIGHT systems. Several types of acoustic characteristics, which are related to the speech energy, the spectral envelope of the speech signal and the glottal source, were analysed in order to investigate the causes of speech distortion in the HTS-LF system.

The HTS-LF and the HTS-STRAIGHT systems used in this experiment were the same as those used in the perceptual evaluation presented in Section 8.4. This permitted to examine if there was a correlation between the results of the acoustic measurements and the perceived speech quality. One method used to analyse this correlation consisted of plotting the results of the objective measurements in terms of the utterance

number, in which utterances were sorted in ascending order of the respective perceptual test scores. The other method consisted of calculating the correlation coefficients between the objective measurements and perceptual test scores.

The objective measurements were also analysed by comparing the degree of acoustic differences between *voicing transition and non-transition regions of speech*. The reasons for performing this analysis were to test the hypothesis that the main problem in the HTS-LF system is poor modelling of the spectrum in voicing transition regions and to evaluate the performance of the energy correction technique of the HTS-LF system. If the acoustic differences are higher in the voicing transition regions, then the hypothesis that the main cause of speech distortion is the spectrum modelling problem at voicing transition is reinforced. This condition is based on the assumption that the limitations of the LF-model parameter estimation method are expected to affect the speech frames associated with the different classes of voiced sounds approximately the same. For example, the LF-model estimation technique is assumed to perform similarly for voiced speech frames near the voicing transitions and speech frames away from transition regions (ignoring the effect of the LF-model errors due to incorrect voiced speech classification). On the contrary, the hypothesised voicing detection errors and the spectrum modelling problem at voicing transitions are assumed to be more relevant for the unvoiced and voiced speech frames near the voicing transitions.

9.2.2 Speech parameters

The following types of speech parameters were studied in this experiment:

- Energy.
- Mel-cepstral coefficients of the spectral envelope.
- FFT representation of the spectral envelope.
- First and second formants (F_1 and F_2 respectively).
- Spectral tilt.
- Difference in amplitude of the first two harmonics (H1-H2).
- Signal-to-noise ratio (SNR).

These parameter representations were chosen because they are perceptually important to speech quality and they enabled to investigate different types of speech characteristics. Energy discontinuities and its distance measurements were important to study the energy distortions which were observed in the synthetic speech of the HTS-LF system. Distance measurements of the spectral envelope were assumed to be the most relevant measurements to evaluate the spectral errors. The F1 and F2 parameters were considered to be phonetically important and relevant for speech intelligibility. Finally, the SNR, spectral tilt, and H1-H2 parameters were used because they are correlated with the LF-model parameters and the last two are also measures of the speech spectrum.

9.2.3 Systems

The systems used in the objective measurement experiment were the HTS-LF and HTS-STRAIGHT synthesisers, which were built for the full male voice used in the perceptual evaluation presented in Section 8.4. Although this perceptual evaluation was also conducted for the female voice and the ARCTIC subset of the male voice, only the full male voice was used in the objective measurement experiment. The reason for this choice is that the difference in performance between the HTS-LF and HTS-STRAIGHT systems was in general similar for the three voices. Therefore, it is assumed that the main factors which contribute to speech quality degradation in HTS-LF are approximately the same for the three voices. For the energy measurements, a version of the HTS-LF system which did not use the power correction algorithm was also used. This was done in order to evaluate the performance of the power correction method in reducing energy discontinuities in the synthetic speech.

The HTS-LF and HTS-STRAIGHT systems generate similar duration parameters for the same test sentence, since duration is modelled in the same way by the two synthesisers and the same speech data was used to build the full male voice for the two systems. Nevertheless, the duration models of HTS-STRAIGHT were replaced by the HTS-LF models. This was done to ensure the speech utterances synthesised by the two systems were phonetically aligned. The alignment of the pair of synthesised utterances was important in order to calculate the acoustic distance between the two speech signals consistently.

Distance measurements between synthetic speech and recorded speech were not performed in this evaluation. For carrying out this type of analysis it would be necessary to perform the phonetic alignment of the recorded speech to the synthetic speech

(e.g. performing a Viterbi alignment).

Some of the objective measurements, which were indicated in Section 9.2.2, were obtained directly from the parameter values generated by the HTS-STRAIGHT and HTS-LF systems, e.g. from the mel-cepstral coefficients. Other speech parameters, such as the energy, were calculated from the synthetic speech waveform. For the HTS-LF system, the synthetic speech was the same as the stimuli of this system which was used in the perceptual evaluation presented in Section 8.4.3. However, the speech synthesised by the HTS-STRAIGHT system which was used for the objective measurement experiment was not the same as the speech synthesised by the HTS-STRAIGHT system which was used in that perceptual evaluation. This difference was because the duration models built for the HTS-STRAIGHT system were replaced by those of the HTS-LF system. The alteration which was made to the duration models of the HTS-STRAIGHT system in this experiment is assumed not to have an important effect on speech quality, when compared with the utterances used in the subjective evaluation presented in Section 8.4.3. This approximation is considered to be valid because the duration model modification produces small variations in the phone and pause durations. Also this was informally verified for several sentences by listening to the utterances synthesised by the two HTS-STRAIGHT versions. Based on the previous assumption, the results of the perceptual evaluation presented in Section 8.4.3 were used to evaluate the correlation of the objective measurement results with the perceptual speech quality for the HTS-STRAIGHT system.

9.2.4 Test Sentences

The test sentences used in this experiment were those of the MOS, SIM and SU parts of the perceptual evaluation, which were described in Section 8.4.3. The sentences of the ABX part of the perceptual evaluation were not considered in the objective measurements because the ranking of the test sentences in terms of speech naturalness was more difficult to perform with the ABX scores than with the MOS scores. This is associated with the fact that the results from the ABX part are given as a forced-choice preference rate of a system, when it is compared with another system (pair-wise comparison).

The objective measurements were performed for the different set of sentences used in the different parts of the listening test: SIM, MOS, and SU sentences. The results of the correlation between objective measurements and perceptual test scores are pre-

sented for the SIM, MOS and SU sentence genres in Section 9.6. However, the speech distortion results for each type of objective measurement are presented only for the sentence genres which were considered to be the most important for that measurement type (such as energy or spectral envelope). One reason for making this simplification was that studying the results in terms of the type of sentence was not considered to be important for this work. Besides which, the results obtained for each objective measure were generally similar between the different sets of sentences. The results of the different objective measures were plotted for the following type of sentences:

- Energy and spectral envelope measures: news domain sentences of MOS part.
- F1 and F2 formant distance: news domain of MOS part and SU sentences.
- H1-H2, spectral tilt, and SNR distances: news domain of MOS part, and SIM sentences.

The results for the news sentences used in the MOS part of the perceptual evaluation were used to compare the speech naturalness scores with the acoustic analysis results, for all types of objective measurements. Also, the news domain group was selected because it includes higher number of sentences compared with the novel genre.

Formant frequencies are particularly important to speech intelligibility. For this reason, the results of formant distances were also plotted for the SU sentences (used in the intelligibility evaluation part of the perceptual evaluation). Meanwhile, H1-H2, spectral tilt, and SNR are typically more relevant for voice quality. For this reason, the results of the objective measurements obtained for these parameters were plotted for the sentences of the SIM part of the perceptual evaluation (associated with the voice similarity test), instead of the SU sentences.

9.2.5 Voiced/Unvoiced Speech Classification

The objective measurements were performed on synthetic speech frames with duration 40 ms and segmented at a 5 ms frame rate. This frame shift was appropriate for the classification of the analysis frames into voiced or unvoiced, because the synthesisers generated speech at a 5 ms frame rate (there is a correspondence between analysis frames and F_0 values generated by the speech synthesiser).

Transitions between voiced and unvoiced speech frames (both voiced to unvoiced and unvoiced to voiced transitions) were calculated for each test sentence using the

F_0 values that were generated by the speech synthesisers. That is, a speech frame was classified as unvoiced if the respective F_0 value was equal to zero and voiced otherwise. An unvoiced-voiced transition was assigned to an unvoiced frame (with $F_0 = 0$) that was right before a voiced frame ($F_0 > 0$), whereas the voiced-unvoiced transition was considered to be the voiced frame which preceded an unvoiced frame. All speech frames within a 50 ms time interval around a voicing transition were considered to be in a voicing transition region. This duration of the voicing transition region is equal to that which was derived heuristically for the power correction algorithm of the HTS-LF system (described in Section 8.3.2).

Silence regions of the speech signal cannot be detected using F_0. Although the speech analysis can be performed in these regions, the estimated parameters values are not relevant for this work and they affect the average values of the distance measures calculated for the unvoiced speech frames which are not in the voicing transition regions. This effect was reduced by performing the speech analysis on the region which starts 30 ms before the first unvoiced-voiced transition and ends 30 ms after the last voiced-unvoiced transition. This technique might discard some frames of unvoiced speech at the start and end regions of an utterance. The advantage is that relatively long segments of silence could be removed. By default, the parameters generated by the HTS-LF and HTS-STRAIGHT systems do not include the phone durations. However, it is possible to modify the systems in order to obtain the phone durations. This could be another solution to remove the silence regions.

9.3 Energy Distortion

The first version of the HTS-LF system, which was described in Chapter 7, occasionally produced speech artefacts related to high amplitude peaks in the energy envelope of the synthetic speech. This type of distortion was observed around voicing transition points and was perceived by the author's informal evaluation as audible "clicks" in voiced speech segments and excessive noise in unvoiced segments. In order to overcome this problem, the HTS-LF system was modified so as to model the power parameter of speech and so that it uses this parameter to adjust the energy of the synthetic speech frames in the voicing transition regions. This power correction method, which was described in Section 8.3, appeared to reduce the effect of the speech artefacts.

The number of energy discontinuities detected in speech synthesised by the HTS-LF system was used as a measure of energy distortion and to verify the effect of the

power correction algorithm on the reduction of these discontinuities. An energy distance measure was also used in order to evaluate the global effect of the power correction on the energy contour of the synthetic speech and to compare the energy distortion between the transition and non-transitions regions of the speech signal.

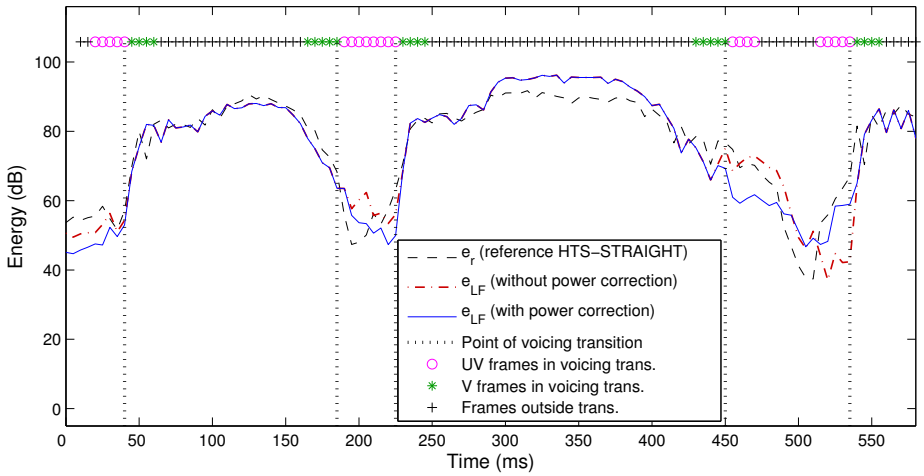

Figure 9.1: Energy contours (in dB) for part of a test utterance, which were calculated from speech synthesised by the HMM-based speech synthesisers. These systems were the HTS-STRAIGHT, the HTS-LF, and a version of the HTS-LF system which did not use the algorithm for energy correction in the voicing transition regions.

9.3.1 Energy Discontinuities

Energy discontinuity detection was performed using a *threshold-based method*. First, the energy parameter was calculated for the speech frames of each utterance synthesised by the HTS-LF system, the HTS-LF system without power correction, and the HTS-STRAIGHT system. Then, the energy contour of the synthetic speech produced by the HTS-STRAIGHT system was scaled in amplitude so that the mean value of the energy was equal to that of the utterance synthesised by the HTS-LF system. This scaling operation was performed so that the range of energy values of the utterances synthesised by the two systems was similar. Figure 9.1 shows the energy contours obtained for the three systems, over a part of a test sentence. The voicing transition regions are also represented in this figure. The next step of the energy discontinuity detection was the calculation of *delta values* from the energy absolute values, as $\Delta e_j = e_j - e_{j-1}$, where $j = 2,...,N$ and N is the total number of frames. Δe_j is a

measure of the *speed of energy variation* between contiguous frames. Energy discontinuities were detected using the Δe_j values (in dB) and by choosing an appropriate threshold. Although the mean of the energy contour obtained for the HTS-STRAIGHT system was adjusted, the range of energy values may differ between this system and the two versions of the HTS-LF system. For this reason, the use of the same $\log \Delta e_j$ threshold to detect discontinuities for the HTS-STRAIGHT and the HTS-LF systems might not be reasonable. However, if energy thresholds were determined for the HTS-STRAIGHT system and the other two systems separately, the comparison of the results of the two systems might also be incoherent. In order to overcome this problem, the energy discontinuities in the speech signals synthesised by each of the HTS-LF systems were estimated by calculating the difference between the Δe_j of these systems and the Δe_j of the HTS-STRAIGHT system, respectively. For each of the HTS-LF systems, these parameters were calculated as

$$\Upsilon_j = 10\log_{10}(\Delta e_j)_{LF} - 10\log_{10}(\Delta e_j)_r, \tag{9.1}$$

where $(\Delta e_j)_{LF}$ is the Δe_j calculated for the speech frame j synthesised by one of the HTS-LF systems and $(\Delta e_j)_r$ is the Δe_j calculated for the speech frame j synthesised by the HTS-STRAIGHT system. Finally, an energy discontinuity is detected in the speech synthesised by one of the HTS-LF systems when $\Upsilon_j > \Gamma$ or $\Upsilon_j < -\Gamma$, where Γ is the *amplitude threshold*. The first condition corresponds to a *"positive" discontinuity*, which represents a rapid increase in energy. Conversely, the second condition corresponds to a *"negative" discontinuity*, which is associated with a deep decrease in energy. These two types of discontinuities are distinguished here because they are assumed to have different perceptual effects on speech quality. A sudden increase in energy is expected to be more perceptually important than a decrease, because the louder a speech artefact is, the higher the chance that it is perceived as unnatural. From experiments, $\Gamma = 10$ dB was found to be an appropriate value to be used in this experiment. Figure 9.2 shows the Δe contours calculated over a part of a test sentence, for the three systems. The estimated "positive" and "negative" discontinuities are also represented in this figure. The effect of the power correction algorithm is clear in the voicing transition region around 200 ms. Two "positive" discontinuities were detected in this part for speech synthesised without power correction. This number was reduced to one when the power correction was used. However, the power correction has the opposite effect on the "negative" discontinuities estimated in the transition region around 500 ms. By attenuating high energy peaks, the power correction is expected to

reduce audible speech artefacts. Meanwhile, the increase in the number of "negative" discontinuities indicates that the algorithm might also produce an over-smoothing of the energy in the voicing transition regions.

Finally, using HTS-STRAIGHT as the reference system to calculate Υ_j depends on the assumption that energy discontinuities in synthetic speech are not a problem for this system. This assumption is supported by the better results obtained for the HTS-STRAIGHT system compared to the HTS-LF system in the perceptual evaluation presented in Section 8.4.

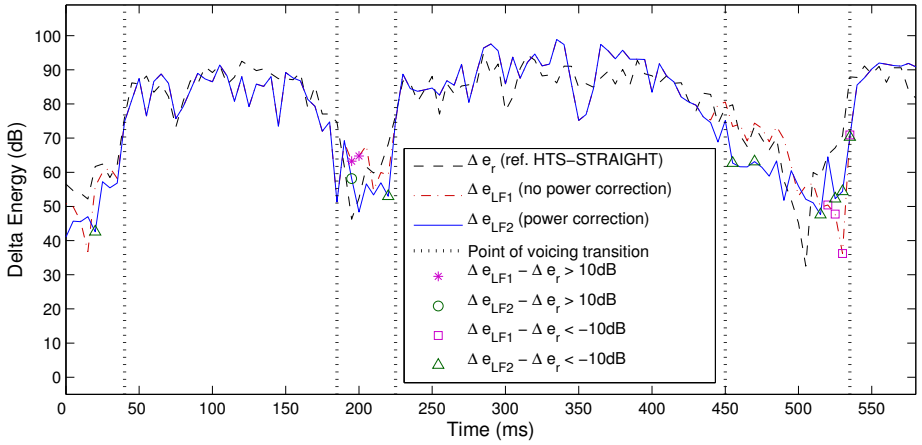

Figure 9.2: Delta energy (in dB) estimated over a part of a test sentence for the three systems: HTS-STRAIGHT and the two HTS-LF systems (versions with and without using power correction respectively). The points of energy discontinuity estimated for the HTS-LF systems are also represented. They were obtained using the thresholds $\Gamma = 10$ dB and $\Gamma = -10$ dB for the difference between their Δe and the Δe_r of the HTS-STRAIGHT system.

9.3.2 Euclidean Distance

The Euclidean distance, D_E, was also used as an objective measurement of energy distortion. The D_E parameter between two feature vectors, X and Y, is calculated as

$$D_E(X,Y) = \left(\sum_{i=1}^{n} (X_i - Y_i)^2 \right)^{1/2} \tag{9.2}$$

This distance was calculated between the Δe feature vectors of HTS-LF and the corresponding feature vectors of the HTS-STRAIGHT system. The results were then used

to calculate the mean value of this distance for the speech frames in the voicing transition regions of an utterance. The mean value of Δe was also calculated over all speech frames which were not in the voicing transition segments of an utterance. These measurements were repeated for speech synthesised using the HTS-LF system without the power correction technique.

9.3.3 Results

Energy discontinuities were classified as "positive" or "negative", using the method described in Section 9.3.1. They were detected by the conditions $\Upsilon_j > 10$ dB and $\Upsilon_j < -10$ dB, respectively. In these equations, Υ_j represents the difference between the energy delta Δe_j of the HTS-LF system and that of HTS-STRAIGHT, for the speech frame j.

Figure 9.3 a) shows the mean number of "positive" discontinuities obtained for the news domain test sentences of the MOS part of the perceptual test. The test sentences are sorted in ascending order of their MOS scores. In this figure, it is clear that the number of discontinuities in the voiced transition regions is substantially reduced by using the power correction algorithm in the HTS-LF system. Although the power correction does not have a significant effect on the reduction of the number of discontinuities for some sentences, it does not appear to have the opposite effect of increasing that number either. These results were expected and give support to the assumption that the power correction reduces the speech distortion associated with excessively high energy variations in voicing transition regions.

Figure 9.3 b) shows that the number of discontinuities detected in the non-transition regions is lower than in the voicing transition regions for most of the utterances. This result is in accordance with the hypothesis that there is more energy distortion in the voicing transition regions, due to the spectral modelling problem around the speech transition frames synthesised by the HTS-LF system. The number of discontinuities detected in a speech region (either voicing transition or non-transition) could be positively correlated with the number of speech frames analysed in that region. Nevertheless, this assumption strengthens the hypothesis considered above, as the number of speech frames in the voicing transition parts is around 30 to 40% of the total number of frames.

Figure 9.4 shows the results obtained for the detection of "negative" discontinuities in the energy contours of the synthetic speech. The number of discontinuities is

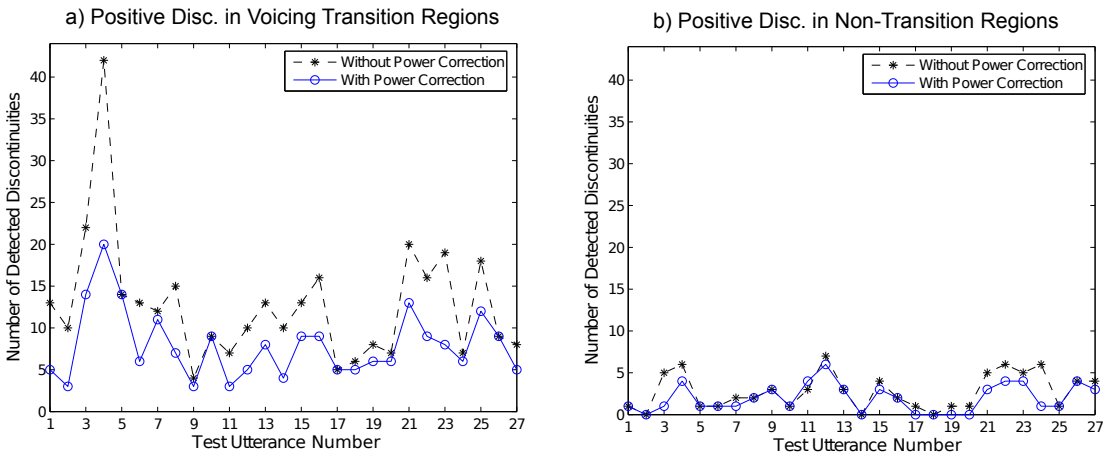

Figure 9.3: Number of "positive" discontinuities detected in the energy contour, which were estimated for the news domain sentences of the MOS part of the perceptual evaluation. "Positive" discontinuities were detected by using the threshold condition $\Upsilon_j > 10$ dB. a) Discontinuities detected in voicing transition regions; b) Discontinuities detected in non-transition regions. In both plots, the test sentences are sorted in ascending order of their respective perceptual test scores.

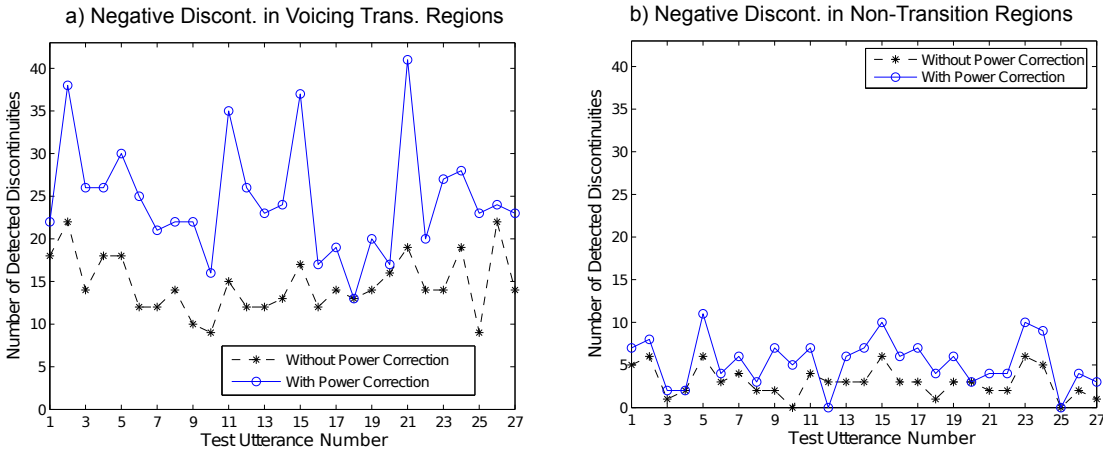

Figure 9.4: Number of "negative" discontinuities detected in the energy contour, which were estimated for the news domain sentences of the MOS part of the perceptual evaluation. "Negative" discontinuities were detected by using the threshold condition given by $\Upsilon_j < -10$ dB. a) Discontinuities detected in voicing transition regions; b) Discontinuities detected in non-transition regions. In the two plots, the test sentences are sorted in ascending order of their respective perceptual test scores.

higher around voicing transitions, as for the case of "positive" discontinuities. However, the power correction algorithm seems to have the opposite effect on the number of "negative" discontinuities, compared with the effect on "positive" discontinuities.

That is, the number of "negative" discontinuities increases when the power correction technique is used. This could be related to an over-smoothing of the energy in the transitions, when compared with the same regions of speech synthesised by the HTS-STRAIGHT system. This excessive reduction of the delta energy could deteriorate speech quality. Nevertheless, the perceptual effect of "positive" discontinuities is assumed to be more perceptually important to speech distortion than the energy over-smoothing effect.

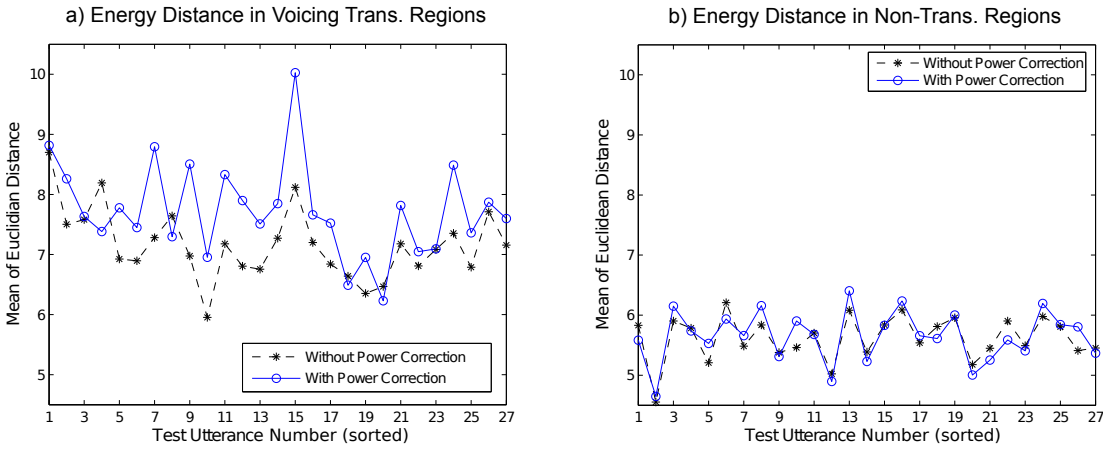

Figure 9.5: Mean values of the Euclidean distance between the energy of speech frames synthesised by the HTS-LF and HTS-STRAIGHT systems, for each test sentence. The sentences are sorted in ascending order of their respective perceptual test scores. a) Calculated for frames in voicing transition regions; b) Calculated for frames in non-transition regions.

The Euclidean distance (D_E), which was calculated between the energies of speech frames synthesised by the HTS-LF and HTS-STRAIGHT systems, was averaged over all frames associated with each test sentence. Figure 9.5 shows the mean values of D_E obtained for each sentence, when either the power correction technique was used or not. The results for voicing transition regions are shown in Figure 9.5 a). Unexpectedly, the energy distance is generally higher in these regions, when the power correction is used. The interpretation of this result is that the increase on the number of "negative" discontinuities has a stronger effect on the energy distance than the reduction of "positive" discontinuities on average, when the power correction is used.

By comparing Figures 9.5 a) and b), the energy distance is generally lower in regions which are away from voicing transition regions. Again, this result supports the hypothesis that speech distortion is higher in voicing transition regions due to the statistical modelling problem. This could be a factor which contributes to the lower per-

formance of the HTS-LF system, compared with HTS-STRAIGHT.

From Figure 9.5, the correlation between the mean energy distances and the MOS scores is not clear. Nevertheless, Figure 9.5 a) indicates that the test sentences with lowest scores are associated with relatively high distances.

9.4 Spectral Envelope Distortion

Spectral distance measures are commonly used in different fields of speech research. For example, they have been used in speech recognition for evaluation of feature representations (Gray and Markel, 1976). They have also been employed in speech coding for the study of perceptual effects of speech distortions (Quackenbush et al., 1988) and in unit-selection speech synthesis for prediction of audible discontinuities (Klabbers and Veldhuis, 2001; Stylianou and Syrdal, 2001; Vepa et al., 2002).

9.4.1 Spectral Envelope

9.4.1.1 Distance Measurements

The Euclidean distance, given by (9.2), on mel-cepstral coefficients and the *Kullback-Leibler distance* (Kullback and Leibler, 1951), D_{KL}, on power spectra are two distances widely used in speech synthesis, due to their good correlation with perceptually relevant characteristics of speech quality, e.g. Klabbers and Veldhuis (1998); Wouters and Macon (1998); van Santen (1997). In this experiment, the spectral envelope distances are calculated using these two metrics. The D_{KL} is a statistical measure, which consists of calculating the distance between two probability distributions $f(x)$ and $g(x)$, as follows:

$$D_{KL}(f,g) = \int (f(x) - g(x)) \log \left(\frac{f(x)}{g(x)} \right) dx \tag{9.3}$$

For the calculation of the spectral distance, $f(x)$ and $g(x)$ represent the spectral density. The spectral density $f(x)$ can be represented by:

$$f(x) = \sum_{n=-\infty}^{\infty} r(n) e^{jnx} \tag{9.4}$$

$$r(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x) e^{jnx} dx, \tag{9.5}$$

where $r(n)$ are the Fourier coefficients. In this experiment, the functions $f(x)$ and $g(x)$ are defined by the FFT coefficients which represent the spectral envelope of the speech signals synthesised by the HTS-LF and HTS-STRAIGHT systems, respectively. Meanwhile, the Euclidean distance, D_E, was calculated between the feature vectors of mel-cepstral coefficients.

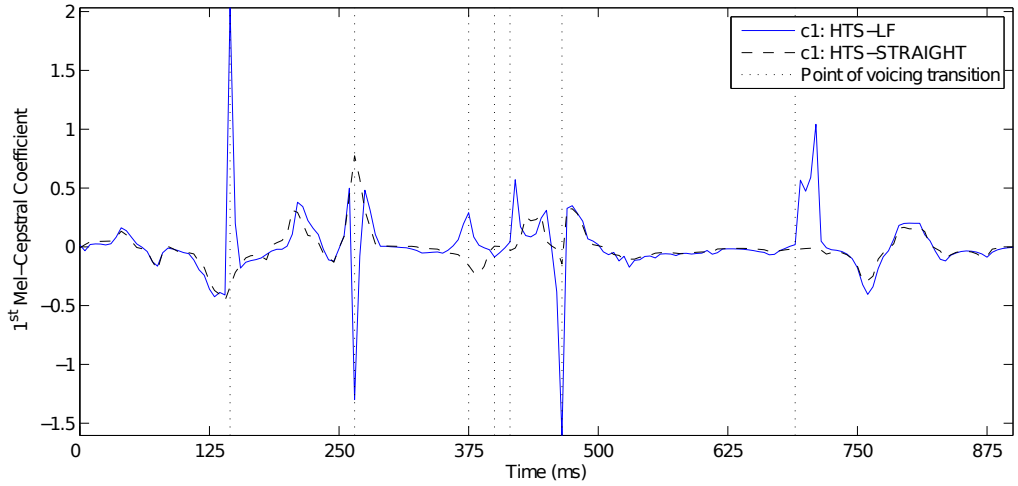

Figure 9.6: Trajectory of the 1^{st} mel-cepstral coefficient (c_1), which was obtained for the HTS-LF and HTS-STRAIGHT systems (over a part of a test sentence)

.

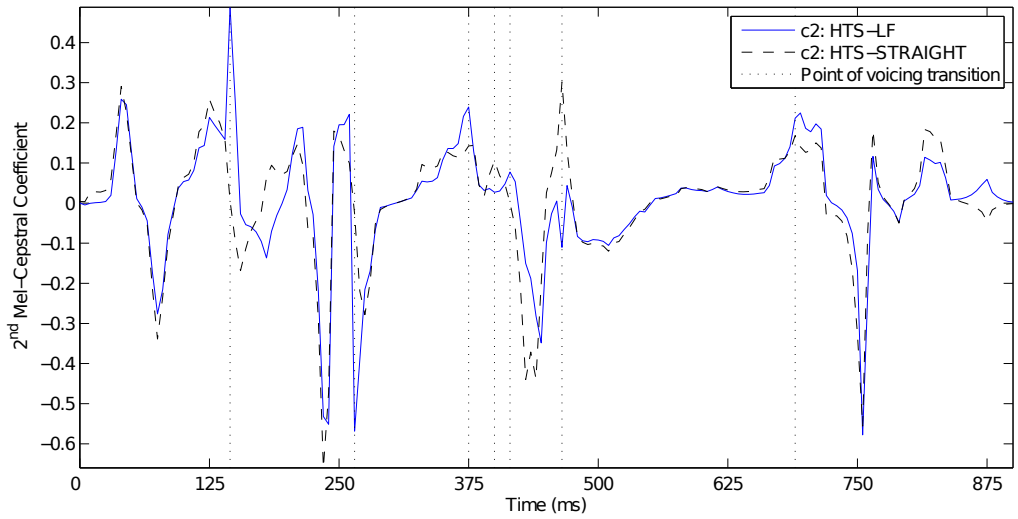

Figure 9.7: Trajectory of the 2^{nd} mel-cepstral coefficient (c_2), which was obtained for the HTS-LF and HTS-STRAIGHT systems (over a part of a test sentence)

.

The mel-cepstral coefficients generated by the HTS-STRAIGHT system represent the spectral envelope of the synthetic speech. The same parameters generated by the HTS-LF system represent the vocal tract instead. This system generates speech by convolving the excitation signal with the vocal tract spectrum, in which the periodic component of the excitation consists of two periods of the LF-model waveform. For calculating the spectral envelope of speech synthesised by the HTS-LF system, the mel-cepstral coefficients of the vocal tract were converted to FFT parameters and then they were multiplied by the amplitude spectrum of a single LF-model cycle (without adding noise). The resulting spectrum does not contain harmonics and represents the spectral envelope of the synthetic speech frame.

For computing the D_{KL} measure, the mel-cepstral coefficients generated by the HTS-STRAIGHT system were transformed to FFT coefficients. Meanwhile, the FFT parameters which represent the spectral envelope in the HTS-LF system were converted to mel-cepstral coefficients so as to compute the D_E measure. Each FFT feature vector of the two systems consisted of 512 coefficients and it was normalised in amplitude by dividing the coefficients by their sum. The mel-cepstral coefficient vector was defined by 38 elements (delta parameters were not used) and they were not normalised. The normalisation was not required because the first mel-cepstral coefficient, which is correlated with the energy of the signal, was not used to calculate the distance.

Figures 9.6 and 9.7 show the trajectories of the 1^{st} and 2^{nd} order mel-cepstral coefficients respectively, for a part of a test sentence. These examples show that there are high amplitude discontinuities of these parameters around the voicing transitions, for the HTS-LF system. In general, these discontinuities are not observed, or are less significant, in the parameter trajectories obtained using the HTS-STRAIGHT system.

9.4.1.2 Results

Figure 9.8 a) shows the mean Euclidean distances between the feature vector of mel-cepstral coefficients of HTS-LF (using power correction) and the vector of mel-cepstral coefficients of HTS-STRAIGHT. The distance is generally higher in the voicing transition regions, than in the voiced and unvoiced regions which do not include the voicing transitions parts. Also, the results obtained for these unvoiced and voiced regions are similar to each other.

Figure 9.8 b) shows the mean Euclidean distances between the delta values of the mel-cepstral coefficients. For the delta parameters, the mean distances calculated for the voicing transition regions are also higher than the distances calculated for the

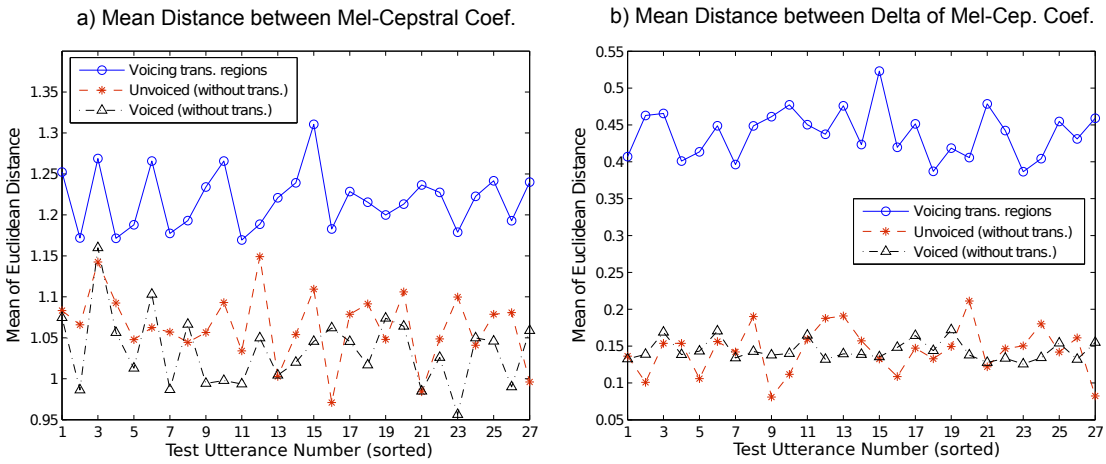

Figure 9.8: Mean Euclidean distance between mel-cepstral feature vectors (represent the spectral envelope) of the HTS-LF system and those of the HTS-STRAIGHT system, for the test sentences (sorted in ascending order of their respective perceptual scores). The results were calculated separately for three types of speech frames: in voicing transition regions, voiced speech away from transition regions and unvoiced speech away from transition regions. a) Mel-cepstral coefficients; b) Deltas of mel-cepstral coefficients.

voiced and unvoiced speech regions which are not included in the voicing transition parts. Moreover, the results obtained for the delta parameters show a clearer difference between the voicing transition regions and the other regions, than the results obtained for static mel-cepstral coefficients. The explanation for this is that the delta parameter is more affected by the rapid fluctuations of the spectral envelope at voicing transitions, as it represents the variation of the static parameter between consecutive frames.

The results obtained for the Kullback-Leibler distances between the FFT parameter vector (represents the spectral envelope) of the HTS-LF and HTS-STRAIGHT systems are shown in Figure 9.9. They are in accordance with the results obtained for the D_E measure.

The results obtained for the two distance measures of the spectral envelope show that there is a significant difference between the spectrum of speech synthesised by HTS-LF and that of the HTS-STRAIGHT system, in the voicing transitions parts. In contrast, the distance between the spectral envelopes of the two systems is much smaller in the non-transition regions. This result gives support to the hypothesis that the HTS-LF system produces lower speech quality than the HTS-STRAIGHT system due to the spectral modelling problem at voicing transitions.

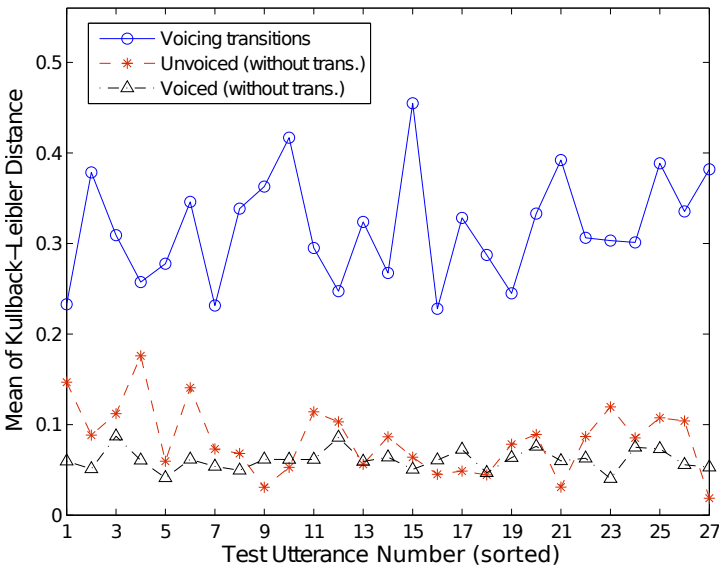

Figure 9.9: Mean Kullback-Leibler distance between FFT coefficient vectors (representing the spectral envelope) of the HTS-LF system and those of the HTS-STRAIGHT system, for each test sentence (sorted in ascending order of their respective perceptual scores). The results were calculated separately for three types of speech frames: in voicing transition regions, voiced speech away from transition regions and unvoiced speech away from transition regions.

9.4.2 Formants

9.4.2.1 Distance Measurement

The formant distance measure is often used in the phonetics field for studying coarticulation, e.g. van den Heuvel et al. (1996). It has also been used in speech synthesis, e.g. as an objective measure of *spectral discontinuity* by Klabbers and Veldhuis (2001). Formant errors may affect speech intelligibility, because the formants are important to phone differentiation. The HTS-LF system was outperformed by the HTS-STRAIGHT system in the intelligibility part of the perceptual test presented in Section 8.4. This is one of the reasons why the distance between the vectors defined by the first two formants (F_1 and F_2) of speech synthesised by the HTS-LF and HTS-STRAIGHT systems was included in the objective evaluation. Since formants are estimated from the spectral envelope, they were also used as an indicator of spectral envelope distortion.

The formant frequencies, F_1 and F_2, were calculated using the *formant tracker* of the ESPS/waves+ program, which employs a F_0 tracking algorithm based on the method of Talkin and Rowley (1990). The estimated formant frequencies were transformed to a *Mel-scale* as proposed by Klabbers and Veldhuis (2001). This Mel trans-

formation is given by

$$F^* = 2595 \log\left(1 + \frac{F}{700}\right) \tag{9.6}$$

Then, the Euclidean distance between the two dimensional feature vectors (each vector consisting of the F_1^* and F_2^* values) of speech synthesised by the HTS-LF and HTS-STRAIGHT systems was computed using (9.2).

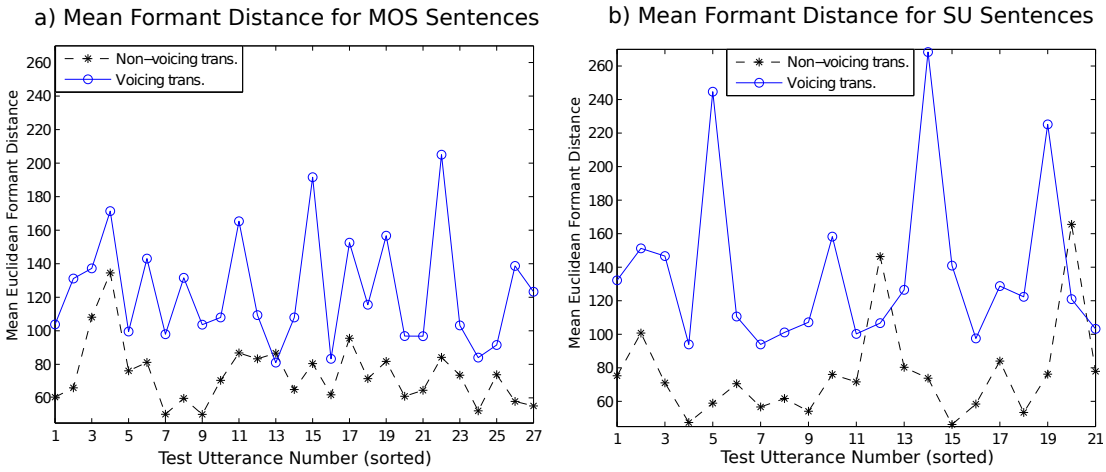

Figure 9.10: Mean Euclidean distance between the feature vectors defined by the first two formant frequencies of speech synthesised by the HTS-LF and HTS-STRAIGHT systems respectively, for the following test sentences used in the perceptual evaluation: a) Sentences of news domain used in the MOS part; b) SU sentences.

9.4.2.2 Results

Figure 9.10 shows the plots of the mean values of the Euclidean distance between the pairs of formants (F_1^* and F_2^*) of speech synthesised by the HTS-LF (using power correction) and HTS-STRAIGHT systems respectively. The results shown in Figures 9.10 a) and b) were obtained for the subset of news domain sentences of the MOS part of the perceptual evaluation and the SU sentences (intelligibility part) respectively. For both groups of sentences, the mean formant distances are higher in the voicing transition parts of voiced speech than in the remaining voiced regions on average. However, there is no apparent correlation between the formant distance and the perceptual test scores (MOS of speech naturalness for news sentences, and word error rates for SU sentences). Also, the variation of mean formant distance between utterances is relatively high. This effect might also be related to problems in the formant estimation, as it is difficult to accurately estimate formants.

It is expected that the higher values of mean formant distance observed in the voicing transition regions are related to the lower scores in speech naturalness and intelligibility for the HTS-LF system, compared with the HTS-STRAIGHT system in the perceptual evaluation.

9.5 Distortion of Speech Related to the Glottal Source

9.5.1 Spectral Tilt

The spectral tilt property of speech considered in this experiment refers to the decaying spectral characteristic of voiced speech, which is a perceptually important aspect of speech. It is mainly associated with voice quality, as there is a strong correlation between spectral tilt and voice source characteristics. The relationship between spectral tilt and the LF-model was described in Section 5.3.1. It is mainly correlated with the return phase parameter, T_a, of this glottal source model. One reason for measuring the spectral tilt was to study the effect of modelling LF-model parameters on the spectral distortion, in speech synthesised by the HTS-LF system. The spectral tilt distance was also used as an indicator of spectral envelope distortion in HTS-LF.

9.5.1.1 Distance Measurement

The spectral tilt of the speech signal was estimated using the method proposed by Murphy (2001). It consists of the ratio of the power energy below a frequency F_t to the energy above that frequency. Murphy (2001) calculated two spectral tilt measures, R_{14} and R_{24}, from the estimated *periodogram* (power spectral density) of the speech signal. R_{14} represented the ratio between the energies from 0 to 1 kHz to the energy from 1 to 4 kHz. Meanwhile, R_{24} was the level difference between the energies below and above $F_t = 2$ kHz (up to 4 kHz). R_{14} and R_{24} were calculated in the work of this thesis by using the FFT coefficients of the normalised spectral envelope (divided by the sum of the elements), H_k, as follows:

$$R_{14} = \frac{\sum_{k=1}^{N/N_y} |H_k|^2}{\sum_{k=N/8}^{N/2} |H_k|^2} \tag{9.7}$$

$$R_{24} = \frac{\sum_{k=1}^{N/N_y} |H_k|^2}{\sum_{k=N/8}^{N/2} |H_k|^2}, \tag{9.8}$$

where $N = 512$ is the total number of FFT coefficients and $N_y = 8$ kHz is the *Nyquist frequency* (equal to half the sampling rate of the speech signal). For example, $H_{N/8}$ corresponds to the frequency component at $f = 1$ kHz. The FFT coefficients representing the spectral envelope of the synthetic speech were obtained similarly as in the spectral envelope measurements described in Section 9.4, for the HTS-LF and HTS-STRAIGHT systems. Finally, the Euclidian distance between the spectral tilt parameters of speech synthesised by the HTS-LF and HTS-STRAIGHT systems was calculated using (9.2) for R_{14} and R_{24} respectively.

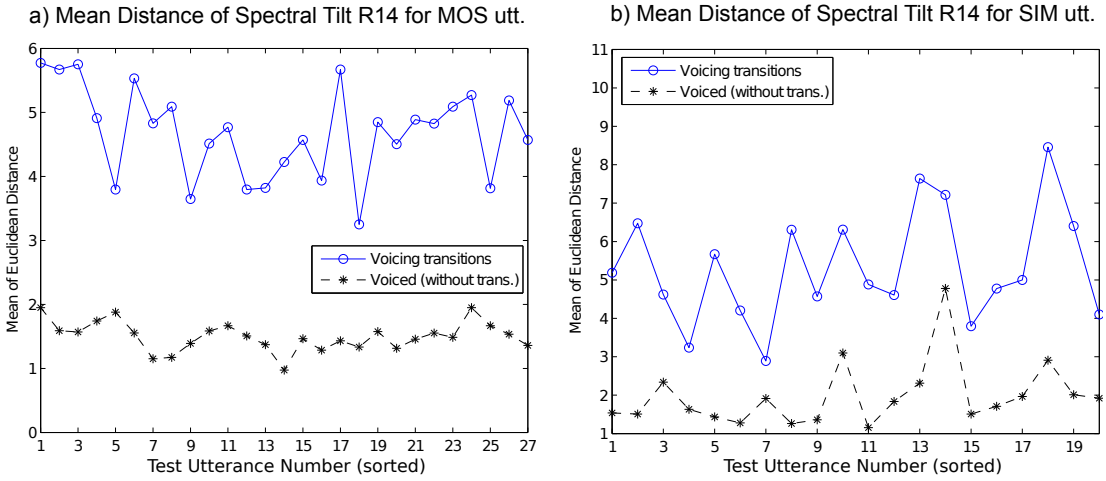

Figure 9.11: Mean distance of the spectral tilt measure R_{14} between the HTS-LF and the HTS-STRAIGHT systems, for the following sentences (sorted in ascending order of their respective perceptual scores): a) Sentences of news domain used in the MOS part of the perceptual evaluation, b) Sentences used in the voice similarity part (SIM) of the perceptual evaluation.

9.5.1.2 Results

Figure 9.11 shows the mean values of the Euclidean distances between the spectral tilt parameters, R_{14} and R_{24}, of the HTS-LF and HTS-STRAIGHT systems. This figure shows the results for the SIM and MOS sentences (news domain sentences). The sentences are sorted in ascending order of the scores obtained for the voice similarity and naturalness test sections of the perceptual evaluation respectively. The results shown in Figure 9.11 indicate that the spectral tilt distance R_{14} is higher in the voicing transition regions of voiced speech than in the remaining voiced parts. From this figure, it is difficult to find a correlation between the perceptual test scores and the distance measures, for both MOS and SIM sentences. Nevertheless, the three utterances with

lowest MOS scores have the highest tilt distances, for the results obtained for voicing transition regions.

The results obtained for R_{24} are shown in Figure 9.12. They are similar to those of R_{14}. Thus, the two spectral tilt distances are consistent with each other and with the results of the previous spectral distance measures (spectral envelope and formant frequencies) which were also higher in the voicing transition regions.

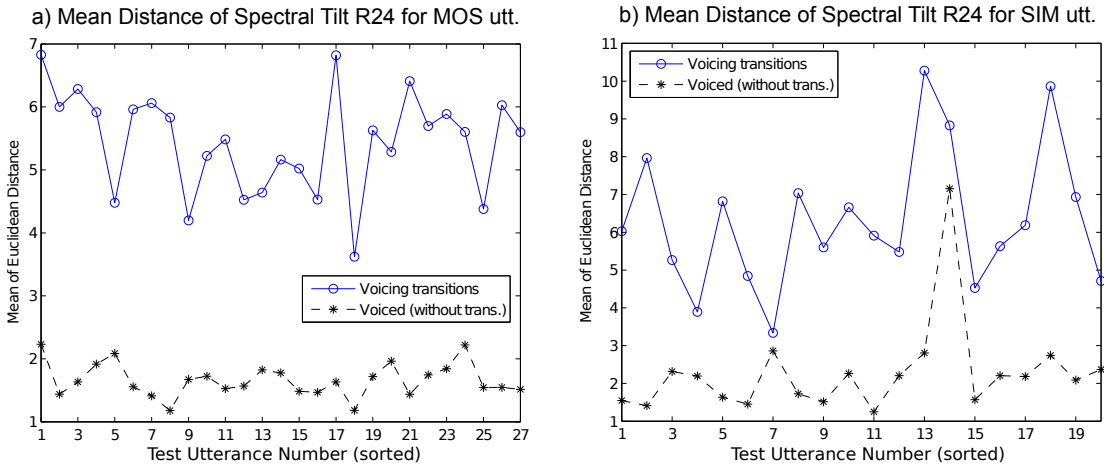

Figure 9.12: Mean distance of the spectral tilt measure R_{24} between the HTS-LF and the HTS-STRAIGHT systems. a) Sentences of news domain used in the MOS part of the perceptual evaluation; b) Sentences used in the voice similarity part (SIM) of the perceptual evaluation.

9.5.2 H1-H2

The difference in amplitude between the first two harmonics of the speech signal has an important effect on voice quality and is correlated with the glottal source signal. The correlation between H1-H2 and the LF-model parameters was described in Section 5.3.1. This spectral parameter is mainly affected by the the amplitude peak ("glottal formant") of the spectrum of the LF-model in the lower frequencies. This peak is more influenced by the SQ and OQ parameters than the RQ parameter of the LF-model.

9.5.2.1 Distance Measurement

The amplitudes H1 and H2 were estimated using the F_0 contour generated by the HMM-based speech synthesisers and the spectral envelope of the synthetic speech. In this process, the frequency components associated with the harmonics H1 and H2

were estimated as the closest components to F_0 and $2F_0$, respectively. Next, H1-H2 was calculated as the difference between the amplitudes of the spectral envelope at the respective frequencies. The spectral envelope was obtained for the HTS-STRAIGHT and HTS-LF systems as described in Section 9.4. Finally, the Euclidian distance between the H1-H2 parameter of the two synthesisers was calculated.

9.5.2.2 Results

Figure 9.13 shows the mean values of the Euclidean distance between the H1-H2 values of the HTS-LF and HTS-STRAIGHT systems. Figure 9.13 a) shows the mean distances for the news domain sentences of the MOS part of the perceptual evaluation (speech naturalness part), while Figure 9.13 b) shows the results for the sentences of the SIM part (voice similarity part).

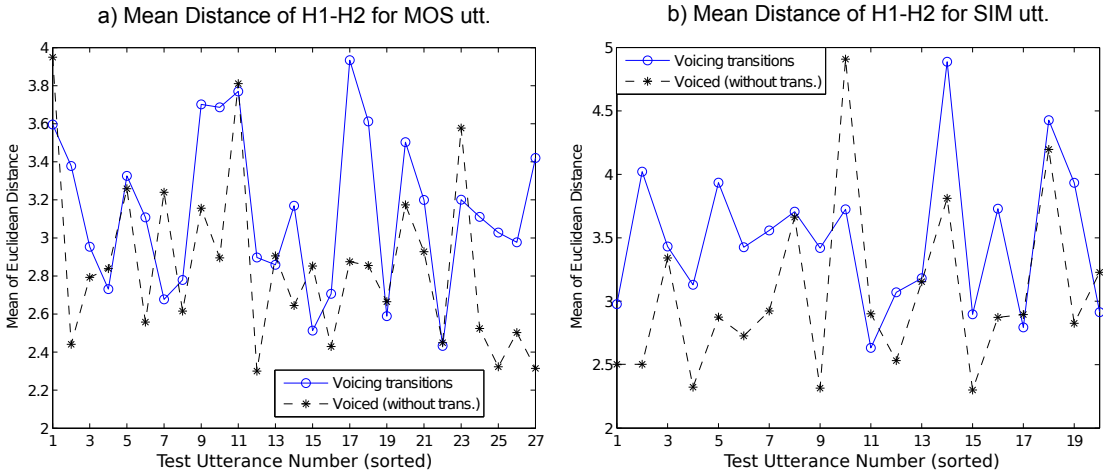

Figure 9.13: Mean H1-H2 distance between speech synthesised by the HTS-LF and HTS-STRAIGHT systems, for the following sentences (sorted in ascending order of their respective perceptual scores): a) News domain sentences used in the MOS part of the perceptual evaluation; b) Sentences used in the voice similarity part (SIM) of the perceptual evaluation.

The H1-H2 distances plotted in Figures 9.13 a) and b) do not seem to be a measure which differentiates the voicing transition regions of voiced speech from the other voiced regions. These results contrast with the previous results obtained for the other spectral distance measures (distance measures of the spectral envelope and spectral tilt). Also, a simple correlation between the distances and the perceptual test scores (test sentences are sorted in ascending order) is not observed from these figures.

The fact that the mean values of the H1-H2 distance are not higher in the voicing transitions regions on average indicates that LF-model errors due to problems in voiced/unvoiced classification are not significant. That is, if the voicing detection was an important problem, there would be sufficiently high LF-parameter estimation errors in the voicing transition segments that produced significant errors on the vocal tract spectrum estimated by the GSS method. Such LF-model errors would affect the spectral envelope of speech synthesised by the HTS-LF system, especially in terms of the H1-H2 and spectral tilt parameters. However, the H1-H2 distance measured on the synthetic speech does not seem to be dependent on the location of the speech frames with respect to voicing transition regions.

Figures 9.13 a) and b) also show that the H1-H2 distance between speech synthesised by the HTS-LF and HTS-STRAIGHT systems is relatively high for some utterances, compared with distances obtained for other test sentences. The high values of the H1-H2 distance can be explained by the differences in the mixed excitation model and spectrum representation between the HTS-LF and HTS-STRAIGHT systems. That is, HTS-LF uses the LF-model to generate the excitation and the vocal tract spectrum, whereas HTS-STRAIGHT uses the impulse train (processed in phase) to generate the excitation and the spectral envelope representation. Both the LF-model signal and the vocal tract transfer function used by the HTS-LF system affect the H1-H2 parameter. In contrast, H1-H2 is only influenced by the spectral envelope in HTS-STRAIGHT. However, it is not possible to determine what system models H1-H2 the best from these objective measurements. Other type of measurements, e.g. by comparing H1-H2 estimated from speech synthesised by the two systems to H1-H2 estimated from natural speech, could help to answer this question.

9.5.3 SNR

The SNR parameter is often used to evaluate speech quality in speech synthesis or speech coding, e.g. Sluijter et al. (1995). The HMM-based speech synthesisers used in this experiment model the noise component of voiced speech by mixing the periodic component of the excitation with a noise signal in different frequency bands (multi-band mixed excitation model). SNR is directly related to the aperiodicity parameters modelled by the HTS-LF and HTS-STRAIGHT systems, as these parameters represent the spectral weighting between the periodic and noise components of the excitation.

The SNR parameter has an important effect on speech naturalness of the speech

synthesisers. For example, the noise component of the multi-band mixed excitation model used in HTS-STRAIGHT generally has the effect of improving speech naturalness when compared with the simple excitation (excitation of voiced speech is modelled as an impulse train only). However, the higher the energy of the noise component the lower the SNR of the synthetic speech and an excessively low SNR might have the opposite effect on speech naturalness, by producing noisy sounding speech. Therefore, it is important to accurately model the SNR in the HMM-based speech synthesisers. SNR is also important to model the speaker's voice characteristics. For example, this parameter is expected to be lower for the breathy voice when compared with modal voice ("neutral" voice quality), due to the effect of *aspiration noise* in breathy voice.

9.5.3.1 Distance Measurements

To estimate the SNR, the powers of both the synthetic speech signal and the noise component of the speech signal were calculated. The noise signal was obtained by synthesising speech using the noise excitation only. Then, the SNR parameter was calculated as the ratio of the speech signal power (synthesised using mixed excitation) to the noise signal power. The distance between SNR values (in dB) of the two HMM-based speech synthesisers was calculated using the Euclidean distance measure given by (9.2).

9.5.3.2 Results

Figure 9.14 shows the mean Euclidean distance between the feature vectors which consisted of the SNR of speech frames synthesised by the HTS-LF and HTS-STRAIGHT systems respectively, for each test utterance. The plots a) and b) in Figure 9.14 represent the results for the MOS (news domain) and SIM sentences, respectively. Both distances obtained for MOS and SIM sentences are generally higher in the voicing transition regions of voiced speech than in the other voiced regions. However, the correlation between the distances and the perceptual test scores does not seem to be significant, in the two plots of Figure 9.14 (MOS and SIM sentences are sorted in ascending order of the scores).

The SNR parameter was expected to be approximately equal between the two systems, because the aperiodicity parameters were extracted from the speech signal and modelled using HMMs similarly by the two synthesisers. For explaining the relatively high SNR distances in the voicing transition parts, two possible causes are consid-

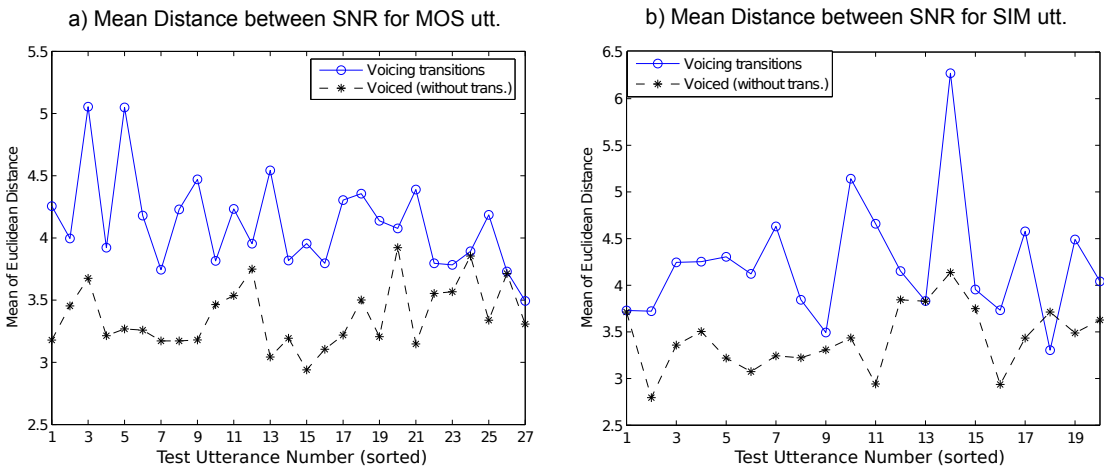

Figure 9.14: Mean of the Euclidean distances of SNR between speech synthesised by the HTS-LF and HTS-STRAIGHT systems, for each of the following sentences: a) Sentences of news domain used in the MOS part of the perceptual evaluation; b) Sentences used in the voice similarity part (SIM) of the perceptual evaluation.

ered. One reason is that poor modelling of the glottal source parameters in the voicing transition regions by the HTS-LF system could result in unnatural LF-model waveforms generated by the synthesiser in these regions. Since the HTS-LF system uses the LF-model signal to modulate the noise component of the excitation, the distortion associated with the generated LF-model parameters could also affect the SNR. This explanation of the results in the voicing transition regions is in accordance with the hypothesis that voiced/unvoiced classification errors could deteriorate the LF-model parameter estimation and affect the glottal source modelling, as explained in Section 9.1. The other factor is that the spectral envelope distortion in the voicing transition regions of the synthetic speech (which is indicated by the results in Section 9.4) could produce the higher SNR distance in the voicing transition regions. This hypothesis is based on the assumption that the SNR variation between the speech signals synthesised by the two systems depends on the variation on the respective spectral envelopes, because the SNR of voiced speech tends to be lower at the high-frequency part of the spectrum than at the low-frequency part. For example, a decrease in spectral tilt (increased ratio of energy at the higher-frequency part of the speech spectrum to the lower-frequency part) by HTS-LF when compared with the HTS-STRAIGHT system, could result in lower SNR of the speech synthesised by the HTS-STRAIGHT system. That is, the spectral tilt decrease emphasises the high-frequency part of the spectrum, which is expected to have lower SNR than the low-frequency part. In Section 9.4, the possi-

ble reasons which were used to explain the spectral envelope distortion in the voicing transition regions were the spectral discontinuities related to the spectral representation mismatch (spectral envelope for unvoiced and vocal tract for voiced speech) and voicing detection errors. In this case, these two problems are also considered to be the most important to explain the higher SNR in voicing transition regions.

9.6 Correlation Between Acoustic Distances and Speech Quality

The mean values of the objective measures calculated for each test utterance were also used to calculate the correlation coefficients between the objective results and the scores obtained for those utterances in the perceptual evaluation described in Section 8.4.

The population correlations between each of the objective measure results presented in the previous sections (energy, mel-cepstral coefficients, FFT coefficients, etc.) and the respective perceptual results were calculated separately for the test sentences associated with the similarity to the speaker's voice (SIM part), naturalness (MOS part) and intelligibility (SU part) sections of the perceptual evaluation.

The *population Pearson correlation* between the mean distances $X = \{x_1, x_2, ..., x_n\}$ and perceptual test scores $Y = \{y_1, y_2, ..., y_n\}$ was calculated by using the following formula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}, \tag{9.9}$$

where x_i and y_i are series of n measurements ($i = 1, 2, ..., n$), $\bar{x}$ and $\bar{y}$ are the means and s_x and s_y are the standard deviations of X and Y, respectively.

The correlation between perceptual test scores and objective distances was low ($|r_{xy}| < 0.5$) for all the cases. The correlation values can be found in Table B.1 (in Appendix B). These correlation results are discussed in Section 9.7.2.

9.7 Discussion

9.7.1 Speech Distortion

The results of the objective measurements showed that the acoustic distances between speech synthesised by the HTS-LF and the HTS-STRAIGHT systems were in general significantly higher for synthetic speech segments in voicing transition regions than for speech in the other regions. Based on this result, it is assumed that speech distortion in HTS-LF is significantly higher in the voicing transition regions. This result was expected according to the hypothesis that speech distortion in the HTS-LF system is related to discontinuities of the spectral parameters at voicing transitions. These parameter discontinuities were explained by two possible factors. One factor was the mismatch between the spectral representation of unvoiced and voiced speech (spectral envelope and vocal tract respectively) at voicing transitions. The other factor was the effect of possible voicing classification errors which resulted in poor estimation of the LF-model parameters for the speech frames around voicing transitions.

The following are the possible causes of speech distortion in the HTS-LF system, which were listed in Section 9.1:

- spectral modelling problem due to mismatch between spectral envelope and vocal tract at voicing transitions.
- problems during analysis of voiced speech segments in voicing transition regions due to voiced/unvoiced classification errors.
- *systematic* errors in estimation of the LF-model parameters.

The first two factors of the previous list are in accordance with the general results of the objective measurements, as the acoustic distances between speech synthesised by the HTS-LF and the HTS-STRAIGHT systems were significantly higher in voicing transition regions than in the other speech regions. The higher acoustic distance in voicing transition regions was observed for all the acoustic parameters analysed in the experiment, with the exception of the H1-H2 parameter. This parameter is expected to be particularly affected by errors in the LF-model parameter estimation, since the H1-H2 parameter is strongly correlated with the LF-model parameters. However, the H1-H2 distance measured on the synthetic speech frames did not seem to be dependent on the location of frames with respect to voicing transition regions. This result indicates that voicing decision errors might not be an important cause of speech distortion

in voicing transition regions. On the other hand, poor modelling of abrupt spectral fluctuations at voicing transitions mainly has the effect of smoothing these spectral variations. Such distortion is expected to affect in particular the overall spectral envelope characteristic (e.g. spectral tilt), as opposed to detailed aspects of the spectrum (such as the H1-H2 measure).

There is another point which suggests that LF-model estimation errors due to an eventual voicing detection problem are less relevant for the speech distortion than the spectral representation mismatch. It is the fact that LF-parameter errors cannot explain the higher acoustic distances of the spectral envelope and energy parameters in the unvoiced speech frames which belong to the voicing transition regions (LF-model parameters are not estimated and the spectral parameters represent the spectral envelope, for unvoiced speech frames). In contrast, the mismatch between the spectral representation of unvoiced and voiced frames can explain the higher acoustic distances obtained for the unvoiced frames in voicing transition regions.

The last cause of speech distortion, *systematic* errors of the glottal parameters, could be related to limitations of the method which was used in the HTS-LF system to estimate the glottal source derivative signal (the IAIF method) or limitations of the technique to estimate the LF-model parameters from the estimated source signal. For example, the LF-model may not accurately fit to every glottal source signal or the non-linear optimisation method which was used by the system to estimate the glottal parameters may not be sufficiently robust. This type of errors in LF-model parameter estimation is assumed to be *systematic*, i.e. it is expected to equally affect every speech frame classified as voice. This assumption may not be completely true. For example, the LF-model parameter estimation might not perform as well for voiced fricatives as for vowels, because the estimation of the glottal source signal by LPC inverse filtering is typically more difficult for voiced fricatives (these speech sounds are usually not as stationary and periodic as vowels). However, in this experiment the important point to consider is that the limitations of the LF-model parameter estimation method are not significantly dependent on whether a voiced speech frame (correctly classified as voiced) is near a voicing transition or not. It could happen that the speech frames of a voiced fricative were just after an unvoiced speech segment and the analysis of the LF-model parameters was poorer for these frames than others. Nevertheless, this effect is assumed not to be significant on average, as voiced sounds whose analysis is less accurate may also appear in non-transition regions of voicing. Given that the LF-model parameter estimation method used in the HTS-LF system is assumed to

perform similarly for the different voiced speech frames, its limitations are considered to be less significant than the other two problems (due to voicing decision errors and spectral representation mismatch) to explain the higher speech distortion in voicing transition parts.

Finally, the results of the objective measurements have indicated that the power correction technique, used in the HTS-LF system for synthesising speech, reduces the number of high energy peaks in the voicing transition regions of the synthetic speech. This effect contributes to reduce the speech energy distortion, although it seems not to solve this problem completely. In addition, power correction does not reduce the other types of speech distortion analysed in this experiment, in particular spectral distortion.

9.7.2 Correlation with Perceptual Test Scores

The acoustic differences between speech synthesised by the HTS-STRAIGHT and HTS-LF systems were analysed in terms of the mean value of those measurements calculated for a set of test sentences. Each set of sentences was associated with a given task (naturalness, intelligibility, or voice similarity tasks) of the perceptual evaluation described in Section 8.4. One way to study the correlation between the objective measurements and the perceptual test scores was to plot the results of the objective measurements in terms of the utterance number, in which utterances were sorted in ascending order of the respective scores. For example, it was expected that the spectral distance between the speech signals of the two systems was higher for test sentences which obtained lower perceptual test scores of speech naturalness (MOS). In this case, the mean value of the spectral distance was expected to decrease with the utterance number. However, for all objective measurements, it has not been apparent from the plots that the mean values of the acoustic measurements were correlated with the perceptual results.

Another method which was used in this experiment to investigate the relationship between the objective measurements obtained for the test utterances and the respective perceptual results was to calculate the correlation coefficients between the two results. However, we have found the correlation was also low for the different types of acoustic measurements.

The following are considered to be possible reasons for the low correlation values between objective and perceptual results:

- perceptual results obtained for each utterance are not adequate to study the per-

ceptual effects of speech distortions that occur in specific locations within the utterance.

- there is no direct relation between the degree of perceived distortion and the acoustic measurements.
- speech quality degradation depended on a combination of different types of acoustic distortion, but the analysis was done individually for each type of objective measurements.
- other speech aspects could be more important to speech quality degradation in the HTS-LF system than those that were analysed in this experiment.

These factors are explained in the following paragraphs.

The first factor listed above indicates that the utterance level is not adequate to study the correlation between the objective measurements and the perceived speech quality, in this experiment. This assumption is based on the fact that by taking the average of an objective measurement over all the analysis frames of an utterance, the effect of a higher distance value at a certain region along the utterance is given less emphasis than if the distance was averaged around that region. Also, the duration of the utterance might affect the correlation value. For example, in the hypothetical case that a speech frame k of an utterance has the highest distance value in that utterance, i.e. $D_i = D_{max}$ for $i = k$, and that the distance is much lower and approximately equal for the rest of the speech frames, i.e. $D_i \approx D_k$ for all $i \neq k$. Then, the longer the utterance (more speech frames), the closer will be the mean distance to the value D_k. That is, the longer the utterance, the lower is the effect of the point with a high distance value on the mean distance. However, if the speech distortion in that utterance is mainly associated with the highest distance value at a specific frame, the listener could judge the perceptual quality of the whole utterance mainly based on that speech distortion independently of the remaining speech frames of the utterance.

Another factor to explain the low correlation is that the acoustic differences between the HTS-LF and HTS-STRAIGHT systems might not be directly related to speech distortion. That is, the acoustic variations between the synthetic speech signals may or may not lead to speech distortion. Also the relationship between the measured acoustic differences and the perceptual test scores may not be linear, whereas the correlation coefficients calculated in this experiment measure a linear correlation between two variables.

The speech quality degradation in the HTS-LF system could also result from a combination of different types of acoustic parameter distortion. However, the correlation between objective measurements and perceptual test scores was analysed for each type of acoustic measure individually, because the analysis of the correlation in terms of the combination of different acoustic parameters is much more complex. In addition, there could be other types os acoustic measurements which were not analysed in this experiment which contributed significantly to speech distortion in the system.

Other types of experiments could be done in the future in order to study the correlation between perceptual results and objective measurements for short parts of an utterance. For example, a perceptual test could be conducted using vowels or words as the synthetic speech samples, instead of utterances. Additionally, other types of acoustic parameters and distance metrics could be investigated.

9.7.3 Future Improvements for the HTS-LF System

Through the analysis presented in this chapter, the main cause of speech distortion in the HTS-LF system is assumed to be the problem of modelling the spectral mismatch between the spectral envelope of unvoiced speech and the vocal tract of voiced speech at voicing transitions.

One way to reduce the speech distortion in the HTS-LF system could be to improve the statistical modelling of speech in voicing transition regions. For example, the spectral parameters could be modelled independently in the voiced and unvoiced regions, using MSD-HMMs, as for F_0. However, F_0 is modelled using a discrete probability distribution for unvoiced speech and a continuous probability density function for voiced speech. In the case of using a MSD-HMM for the spectral parameters, these parameters should be modelled using two continuous distributions, one for voiced and the other for unvoiced speech.

Another way to solve the spectrum modelling problem of the HTS-LF system is to avoid the discontinuities of the spectral parameters at voicing transitions. These discontinuities could be reduced by modelling the spectral envelope and the vocal tract separately using different streams. That is, one stream would be used to model the spectral envelope of unvoiced and voiced speech. The other stream would be used to model the vocal tract transfer function. In this stream, the vocal tract spectrum of unvoiced speech could be represented by the spectral envelope, as it is not possible to estimate the vocal tract for unvoiced speech using the GSS analysis method (the

spectrum of unvoiced speech is generally represented by the spectral envelope). In order to avoid an abrupt fluctuation of the spectrum at voicing transitions, a smoothing operation could be performed on the spectral envelope parameters of the unvoiced frames in the neighbourhood of voicing transitions, while the vocal tract parameters could remain the same. To synthesise unvoiced speech, the spectral parameters would be obtained from the spectral envelope stream and to synthesise voiced speech they would be obtained from the stream which contains vocal tract and spectral envelope parameters. The HMM-based speech synthesiser proposed by Raitio et al. (2008) uses a similar method to model the vocal tract spectrum of voiced speech and the spectral envelope of unvoiced speech (by using separate streams). It is not clear from that paper why the two types of spectra are modelled separately, but the reason could also be to avoid the spectral mismatch at the voicing transitions.

9.8 Conclusion

The results of the perceptual evaluation presented in Section 8.4 showed that the speech quality of the HTS-LF system is significantly lower than the quality of the HTS-STRAIGHT system. In order to investigate the reasons for such a difference in performance, the two systems were compared in terms of several acoustic characteristics: speech energy parameters, spectral envelope parameters, and parameters related to the glottal source (spectral tilt, H1-H2 and SNR). The conclusions of this experiment are summarised as follows:

- acoustic difference between an utterance synthesised by HTS-LF and the same utterance synthesised by HTS-STRAIGHT is generally higher for speech frames in voicing transition regions than for speech frames away from those regions.
- results suggest that the problem of modelling abrupt spectral parameter variations at voicing transitions by the HTS-LF system (due to mismatch between the spectral envelope and the vocal tract transfer function) is the most important factor of speech distortion.
- correlation between the mean acoustic difference of utterances synthesised by the two systems and the perceived speech quality of those utterances for the HTS-LF system was not found.

- the power correction technique used in the HTS-LF system reduces the number of high amplitude peaks (energy discontinuities) in the voicing transition regions, which seemed to be related to speech artefacts.

Apart from the spectral modelling problem at voicing transitions, errors in the LF-model parameter analysis are also a possible cause of the high acoustic differences observed between the speech synthesised by the HTS-LF and HTS-STRAIGHT systems. However, this second factor is considered to be less significant than the first, based on the results.

The high acoustic distances between the speech signals of the two HMM-based speech synthesisers, particularly around voicing transitions, is expected to be the main cause of speech quality degradation in the HTS-LF system. However, this hypothesis could not be verified in this experiment, as the correlation between the results of the acoustic analysis and the results of the perceptual evaluation was not significant. Since the speech distortion in the HTS-LF system has been found to be higher around the voicing transitions, future experiments could be conducted to study the correlation between acoustic measurements and perceived speech quality for speech segments shorter than the utterances used in this experiment. Such experiments could give more significant correlation results and prove the hypothesis that the speech distortion in HTS-LF is mainly caused by the acoustic differences in voicing transition regions. Further experiments using an improved HTS-LF system could also permit more conclusions to be obtained about the problems that caused the poor performance of the HTS-LF system in the perceptual evaluation presented in Section 8.4. For example, two methods were suggested for overcoming the problem of modelling the unvoiced and voiced spectra at voicing transitions, in Section 9.7.3.

Chapter 10

Conclusions

Most current HMM-based speech synthesisers use a parametric model of speech that consists of a spectrally flat excitation signal and a synthesis filter representing the spectral envelope of the speech signal. The simplest excitation model used by these systems consists of an impulse train for synthesising voiced speech and white noise for unvoiced speech. However, speech synthesised using an impulse train typically sounds robotic due to the strong periodicity characteristic of this signal. Recently, different types of excitation models have been applied to statistical speech synthesis in order to improve speech naturalness. In general, these models still assume that the excitation is a spectrally flat signal but they represent more characteristics of the voiced excitation in addition to the periodicity aspect, such as the noise and other non-periodic aspects. However, these models typically do not describe the important characteristics of the glottal source signal. In particular, this signal is characterised by a decaying spectrum instead of being spectrally flat. The major problem of using an excitation model that describes the glottal source is that this component of speech has to be separated from the synthesis filter. That is, the synthesis filter has to represent the vocal tract transfer function instead of the spectral envelope. In this work, the motivations for using the glottal source excitation in HMM-based speech synthesis were:

- to model glottal source aspects and the vocal tract parameters independently.
- to take into account the correlation between the F_0 and glottal parameters.
- to alleviate the robotic sound quality characteristic of the impulse train.
- to increase parametric flexibility for voice transformation.

One advantage of modelling the spectrum separately from the glottal source is that these components are assumed to be independent in the source-filter theory of speech production. By modelling them separately it is expected that the glottal source and spectrum modelling is improved. Also, the parameters of the glottal source model can be modelled together with the F_0 parameter so as to take into account the correlation which exists between them. When compared to the impulse train, the glottal source signal is less periodic and is a more realistic representation of the excitation of voiced speech in the speech production system (the energy of the excitation is spread along the period instead of being concentrated at one time instant). For this reason, the use of a glottal source model instead of the impulse train to represent the excitation is expected to reduce the robotic speech quality. Finally, the glottal source signal has several properties which are strongly correlated with voice quality (such as breathiness and creakiness). By using an excitation model that describes the glottal source, the voice quality of the synthetic speech can be better controlled.

10.1 Analysis-Synthesis Methods

Very little research work can be found in the literature about using glottal source parameters in statistical speech synthesis. Moreover, there is not currently a HMM-based speech synthesiser using glottal source modelling which permits glottal parameters correlated with voice quality to be directly controlled. This limitation is because the excitation parameters that have been modelled by current systems do not have a direct relation with the properties of the glottal pulse. Moreover, any of the current HMM-based speech synthesisers using glottal source modelling takes into account the correlation between F_0 and the glottal source parameters. The major contribution of this thesis is the integration of an acoustic glottal source model, the LF-model, into a baseline HMM-based speech synthesiser which is based on the HTS system. The LF-model parameters can be used to control several aspects of the voice source which are correlated with voice quality. Also, the correlation between the LF-model parameters and F_0 is taken into account in the speech synthesiser.

Two analysis-synthesis methods were developed in this work to integrate the LF-model into the baseline system. They are reviewed in the next paragraphs:

1. **Glottal Post-Filtering (GPF):**

    - Analysis: Calculation of the glottal post-filter from a chosen LF-model

signal (which is stored so as to be used for synthesis).

- Synthesis of voiced speech:
  i) Excitation of voiced sounds is generated by passing the stored LF-model signal through a glottal post-filter, which works as a whitening filter.
  ii) Convolution of the spectrally flat excitation with the spectral envelope, then overlap-and-add.
- Advantages:
  i) Stored LF-model signal can be modified so as to transform the voice characteristics of the synthetic speech.
  ii) Excitation contains phase information of the LF-model signal, which reduces buzziness.

2. **Glottal Spectral Separation (GSS):**

- Analysis of voiced speech:
  i) Glottal source parameters (LF-model parameters) are estimated from the recorded speech, e.g. using an inverse filtering technique for calculating the glottal source signal.
  ii) Spectral effects of the LF-model signal (generated using the LF-model parameters) are removed pitch-synchronously from the speech signal by dividing the amplitude spectrum of the speech by the amplitude spectrum of the LF-model signal.
  iii) Vocal tract spectrum is estimated by computing the spectral envelope of the signal obtained in ii).
- Synthesis of voiced speech:
  i) Generation of the excitation using the LF-model parameters.
  ii) Generation of the vocal tract spectrum from the spectral parameters.
  iii) Convolution of the excitation with the vocal tract spectrum, then overlap-and-add.
- Advantages:
  i) LF-model signal contains more phase information than the impulse train, which reduces buzziness.

ii) Glottal parameters can be used to control voice quality.

iii) Errors in the glottal parameter estimation can be alleviated before separating the glottal source aspects from the speech signal, e.g. by smoothing the glottal parameter contours.

iv) Vocal tract spectrum can be computed using a robust spectral envelope estimation method, as the glottal source and the vocal tract parameters are estimated independently. In contrast, the typical source-tract separation techniques estimate the glottal source and the vocal tract using the same model of speech, e.g by calculating the two components jointly or iteratively.

The GPF method was integrated into the baseline HMM-based speech synthesiser just by modifying the speech waveform generation technique of the system, as they use the same spectral envelope and excitation parameters to generate the speech waveform. The speech synthesiser using GPF was called HTS-GPF. The baseline system was also modified in order to incorporate the GSS analysis-synthesis method and in order to train the LF-model parameters by the HMMs. This system using glottal source modelling was called HTS-LF.

10.2 Summary of the Results

The first perceptual evaluation in this thesis was conducted to test the hypothesis that the use of the LF-model for speech synthesis improves speech naturalness and increases the degree of parametric flexibility to control voice quality aspects, when compared to the traditional impulse train. In this experiment, the GSS method was used to synthesise speech by copy-synthesis using the LF-model. In this way, any potential effect of statistical modelling on results was excluded. For synthesising speech using the impulse train, the same waveform generation and spectral envelope estimation techniques as those of the GSS method were used. The results showed that:

- Speech synthesised using the LF-model sounded significantly more natural than using the impulse train on average.
- Control over the LF-model parameters permitted to transform the voice quality of the synthetic speech. Conversely, the impulse train did not permit to perform the same voice transformations.

Another perceptual evaluation was conducted in order to test the hypothesis that the HTS-LF system produces more natural speech quality than the baseline system which uses the impulse train for generating the excitation. In this case, the HTS-LF system was compared against a synthesiser which used the STRAIGHT vocoder for analysis and synthesis (called HTS-STRAIGHT system), in an AB forced-choice experiment. For this experiment, the GSS method was implemented in the HTS-LF system using a simple inverse filtering technique to estimate the glottal source derivative signal and the STRAIGHT vocoder to compute the spectral envelope and aperiodicity parameters. These aperiodicity parameters are also used by the HTS-LF system to weight the spectra of the LF-model signal and the noise, in the generation of the mixed excitation signal. However, in the perceptual experiment the noise component of the excitation was not used, in order to exclude the effect of this component in the comparison of the LF-model against the impulse train signals. The characteristics of the implemented GSS analysis are summarised as follows:

- Inverse filtering with pre-emphasis for estimating the derivative of the glottal volume velocity (DGVV) signal.
- Estimation of the LF-model parameters by fitting the LF-model waveform pitch-synchronously to the DGVV signal, using a non-linear optimisation algorithm and initial estimates obtained by direct measurements on the DGVV signal.
- STRAIGHT analysis for computing the aperiodicity parameters and the spectral envelope of the signal obtained after removing the LF-model spectral effects from the speech signal.

The results showed that speech synthesised using the HTS-LF system sounded slightly more natural that speech synthesised using the HTS-STRAIGHT system. The results obtained for the HTS-LF system were expected to be better in terms of speech naturalness, because the LF-model clearly outperformed the impulse train in the copy-synthesis experiment. For this reason, it is assumed that the statistical modelling of the speech parameters in HTS-LF resulted in some speech quality degradation. This effect could be caused by errors in the LF-model parameter estimation or a problem in modelling the speech parameters by the HMMs. In particular, the author detected speech artefacts produced by the HTS-LF system, which were not characteristic of the HTS-STRAIGHT system. This type of distortion was related to high peaks observed

in the energy envelope of the synthetic speech around voicing transition instants. Subsequently, the HTS-LF system was modified in order to improve the estimation of the LF-model parameters and reduce speech artefacts. These improvements are listed below:

- Iterative adaptive inverse filtering method which is used to obtain a more accurate estimate of the DGVV signal than using pre-emphasis inverse filtering.
- Algorithm for detecting and correcting errors of the estimated LF-model parameters.
- Technique for adjusting the energy of the synthetic speech frames which are in the neighbourhood of voicing transitions using the power parameter.

A final perceptual evaluation was conducted in order to evaluate the HTS-GPF and the improved HTS-LF systems. The main results of this evaluation were:

- The HTS-GPF system performed as well as the baseline system which used the STRAIGHT vocoder for analysis and synthesis.
- The HTS-LF system did not perform as well as the baseline system in terms of speech naturalness, intelligibility and similarity to the original speaker's voice.
- The baseline system (HTS-STRAIGHT) performed similarly to a modified version of this synthesiser which used the GSS waveform generation technique (FFT processing and OLA instead of STRAIGHT synthesis).

In this evaluation the improved HTS-LF system was expected to outperform the baseline system, as the first HTS-LF version obtained positive results in the preliminary AB perceptual evaluation (the baseline system used in the two experiments was similar). The results indicate that the speech distortion in the HTS-LF system is most likely to be related to the speech analysis and the statistical modelling of the speech parameters, since the waveform generation technique in HTS-LF performed well when it was employed in the baseline system. The main difference between the speech analysis in the HTS-LF and baseline systems is the estimation of the LF-model parameters by HTS-LF. Possible causes of errors in the estimation of the LF-parameters are poor estimation of the DGVV signal, limitations of the LF-model parameterisation technique, and errors in the voiced/unvoiced classification of the speech frames. These LF-parameter errors could affect the statistical modelling of the speech parameters and explain the

poor speech quality of the HTS-LF system. However, speech distortion seemed to be particularly prevalent in voicing transition regions, from informal evaluation of several utterances used in the formal perceptual experiment. This effect could be explained by a problem in modelling rapid spectral parameter variations at voicing transitions, as the spectrum represents the vocal tract transfer function for voiced speech and the spectral envelope for unvoiced speech.

In order to investigate the causes of speech distortion in the HTS-LF system, objective measurements were performed on sentences synthesised by the HTS-LF and HTS-STRAIGHT systems (same sentences which were used in the perceptual evaluation). These measurements represented acoustic differences between the speech synthesised by the two systems. The results of this experiment showed that:

- Acoustic differences related to the energy and spectral envelope are significantly higher in the voicing transition regions than in the speech regions away from the voicing transitions in general.
- Acoustic differences related to the glottal source (spectral tilt, H1-H2, and SNR), were generally higher in the voicing transition regions, with the exception of the H1-H2 distance measure.
- Correlation between the mean values of the objective measurements calculated for the test sentences and the perceptual test scores obtained by the respective sentences was low.

Although correlation between the objective measurements and the perceptual speech quality was not found, it is assumed that the speech distortion in voicing transition regions is the most important factor of speech quality degradation in the HTS-LF system. In the experiment conducted in this work, it was not possible to verify this assumption, because the perceptual speech quality was evaluated for whole utterances. That is, the perceptual test scores could not be used to calculate the correlation between speech quality and the high acoustic differences observed in the voicing transition regions. The results of the objective measurements give support to the hypothesis that the mismatch between the spectral envelope and the vocal tract spectrum, at the voicing transitions, have a negative effect on the statistical modelling of the spectrum and that it is the most important factor causing speech distortion in the HTS-LF system. The robustness of the LF-model parameter estimation method is considered to be a less important factor causing speech distortion, because it is assumed that this factor is

not related to the higher acoustic differences observed in the voicing transition regions. One reason for this assumption is that the LF-model parameter errors are expected to affect all voiced speech frames in a similar manner on average. In addition, the LF-model parameters are strongly correlated with the H1-H2 parameter, but the results of the H1-H2 distance measure were similar for voiced speech in voicing transition regions and away from the voicing transitions.

The HTS-LF system uses the LF-parameters for training the HMMs and then the system uses these parameters to generate the excitation of voiced speech. This system allows the shape of the LF-model signal which is used to represent the excitation to be directly controlled, in order to transform voice characteristics of the synthetic speech. The HTS-GPF system does not use the LF-model parameters for training the HMMs. However, it passes a stored LF-model signal through a glottal post-filter for generating the excitation of voiced speech. This system also permits characteristics of the glottal source to be modified for voice transformation. However, the control over the glottal source characteristics is more limited than in the HTS-LF system, because the LF-model signal is used by the HTS-GPF system for generating a spectrally flat excitation, instead of being used directly to represent the excitation. The disadvantage of the HTS-LF system is that the speech quality is not as good as that of the HTS-GPF system. Nevertheless, there is scope for improvement of the HTS-LF system in future research.

10.3 Future Work

10.3.1 Synthetic Speech Quality

Two main factors which affect speech quality in state-of-the-art HMM-based speech synthesisers are the over-smoothing effect of speech parameter trajectories generated by the HMMs and the quality of the speech vocoder employed in these systems. Techniques have been proposed to reduce the excessive parameter smoothing, e.g. using a parameter generation algorithm considering global variance (Toda and Tokuda, 2007). The speech vocoding methods have also been improved in order to obtain more natural speech. For example, recent versions of the HTS system using the STRAIGHT vocoder produce significantly more natural speech than the traditional HTS system which generates voiced speech by passing an impulse train through the MLSA filter. Due to both factors, details of speech relevant for speech naturalness are somehow lost. Therefore, improving the statistical models to capture those speech details might

not be sufficient if those details are poorly represented by the speech parameters in the first place. Similarly, using better parametric models of speech might not have a significant impact on speech quality if the statistical models cannot correctly capture the increase in parameter detail. In this work, the main focus was to improve the parametric model of speech in HMM-based speech synthesis. In particular, the HTS-LF system proposed in this thesis represents voiced speech by passing the LF-model signal through the vocal tract filter. This system could be further improved in the future in terms of statistical modelling, the parametric model of speech and robustness of the speech parameter estimation methods.

10.3.1.1 Statistical Modelling

One of the findings in the work of this thesis was that the typical method to model the spectral parameters in HMM-based speech synthesis is not appropriate for the case of using the vocal tract and spectral envelope representations for voiced and unvoiced speech, respectively. In the opinion of the author, the use of the speech model which represents the vocal tract and the glottal source for voiced speech is the way forward to further improve speech quality. However, it is necessary to develop better methods for statistical modelling of the spectral parameters using this type of speech representation. In Section 9.7.3 two different methods were suggested to better model the abrupt variations of the spectral parameters at voicing transitions, in the HTS-LF system. They are compared in more detail in the next paragraphs.

One method for modelling rapid variations of the spectral parameters at voicing transitions consists of modelling the spectral envelope and vocal tract parameters in the same feature vector stream using a MSD-HMM. This model is defined by two spaces which are associated with continuous probability density functions (pdfs) for each state. The first space is used to model the spectral envelope for unvoiced speech and the second space is used to model the vocal tract for voiced speech. For each state, the probability of the first and second spaces are defined by the probability of the speech segments associated with that state being unvoiced or voiced respectively. These probabilities should be equal to those calculated to model F_0 using a MSD-HMM so that the F_0 values (voicing classification) is consistent with the spectral representation. When the probability of the unvoiced space is higher than a given threshold than the pdfs associated with the spectral envelope are used to generate the spectral parameters. Otherwise, the pdfs associated with the vocal tract representation are used. This method using a MSD-HMM to model the spectrum permits to model

the spectral envelope and the vocal tract independently and allows abrupt variations of the spectral parameters to occur at voicing transitions. The spectra of the first speech frames at a voicing transition are assumed to be independent from the spectra of the last speech frames before the voicing transition. However, the parameter generation algorithm uses dynamic feature constraints. Then, there is a problem in estimating the correct Δ and Δ^2 parameters for the first speech frames after a voicing transition. This problem also exists in modelling F_0 using a MSD-HMM.

The other method which could be used to better model the spectral parameters around voicing transitions consists of using two feature streams for the spectrum. One stream is used to model the spectral envelope for both unvoiced and voiced speech, as in the baseline HTS-STRAIGHT system described in Section 7.2. The second stream models the spectral envelope and the vocal tract parameters for unvoiced and voiced speech, respectively. This stream is similar to that used in the HTS-LF system to model the spectrum with the difference that a smoothing operation is performed on the spectral envelope of the unvoiced speech frames closest to a voicing transition, in order to produce a smoother transition between the spectral envelope and the vocal tract spectra. The stream which models the spectral envelope only is used to synthesise unvoiced speech. This spectral envelope is correctly modelled, because it was not transformed by any smoothing operation during analysis. The other stream is used to synthesise voiced speech using the vocal tract parameters. The vocal tract parameters are expected not to be affected by abrupt variations at voicing transitions, given that the spectral envelope was transformed during analysis to obtain smooth parameter variations at voicing transitions.

Both methods described in the previous paragraphs alternate the spectral parameters between those representing the spectral envelope for unvoiced regions and those representing the vocal tract for voiced regions. The method which uses a MSD-HMM has the advantage that it uses a single stream for the spectrum whereas the other method requires two spectrum streams. Also the second method depends on the performance of a spectral smoothing operation. For these reasons, the method using a MSD-HMM to model the spectrum might be a more effective and simple solution. A great limitation with these two methods is that they depend on the accuracy of the voiced/unvoiced speech classification. Voiced/unvoiced classification errors during analysis affect negatively the statistical modelling of the F_0 parameter. By imposing the constraint that the spectrum is represented by the spectral envelope and the vocal tract during unvoiced and voiced speech respectively, spectrum modelling is also affected by the

voiced/unvoiced decision. The ideal solution would be to model the vocal tract transfer function for both unvoiced and voiced speech. Such spectrum is expected to be sufficiently smooth as the articulators of the vocal tract system move relatively slow during speech production. However, it is very difficult to accurately estimate the vocal tract transfer function for unvoiced speech. Future work for improving the speech model in HMM-based speech synthesis is further discussed in the next section.

Future evaluations of the HTS-LF system using the previous methods could permit to conclude if these methods overcome the speech distortion at voicing transitions. The HTS-LF system could also be evaluated using voiced-only utterances in future experiments, such as in the evaluation of the GSS method presented in Section 6.6. The use of this type of sentences could reduce the effect of speech distortion in voicing transitions and permit to better evaluate the contribution of glottal source modelling for improving the quality of voiced speech.

10.3.1.2 Parametric Model of Speech

State-of-the-art speech vocoders, such as the STRAIGHT vocoder, produce speech which sounds very close to human speech. However, HMM-based speech synthesisers cannot synthesise speech which sounds as natural as vocoded speech. This degradation in speech quality compared to vocoded speech is expected as statistical modelling cannot capture all details of the speech signal, whereas such details can be reconstructed reasonable well using high-quality speech vocoders.

In speech coding the main challenge is to reduce the amount of speech parameters preserving the high-quality of the vocoded speech. The speech quality of an HMM-based speech synthesiser depends not only on the quality of the speech vocoder used by the system but also on the performance of the parametric representation of the speech signal for statistical modelling. For example, STRAIGHT is a high-quality speech vocoder which has been successfully used in HMM-based speech synthesis. One of the advantages of STRAIGHT for statistical modelling compared with other popular vocoders such as the LPC vocoder is that STRAIGHT extracts a smoother spectrogram. This characteristic is important because parameter discontinuities have a negative effect on acoustic modelling using HMMs. However, even using high-quality speech vocoders in HMM-based speech synthesis there is a clear gap between the quality produced using this method and vocoded speech. One way to further improve speech quality in HMM-based speech synthesis is to use a different speech representation than the typical spectral envelope of speech.

Another important aspect of the speech model in HMM-based speech synthesis is the separation of the spectrum characteristics that differentiate the speech units, e.g. phones, from the prosodic aspects. It is expected that this separation reduces the variability of the spectrum and consequently improves statistical modelling. This hypothesis motivated the representation of speech by the vocal tract and the glottal source components, in this work. Since the glottal source component is mainly related to prosodic and voice quality characteristics of speech, it is desirable to model the glottal source independently from the spectrum, i.e. the vocal tract transfer function. The speech model representing the glottal source and vocal tract is expected to improve the statistical modelling of the spectrum and prosody characteristics, compared to the spectral envelope representation of speech. The vocal tract transfer function is assumed to vary sufficiently slow to be well modelled by the HMMs. However, separating the glottal source from the vocal tract is a more complex problem than computing the spectral envelope of speech. Errors in vocal tract parameter estimation cause parameter discontinuities which affect negatively the statistical modelling by HMMs. In the opinion of the author, a direction to further improve speech quality in HMM-based speech synthesis is to develop methods to more accurately estimate the vocal tract and the glottal source components of speech. The ideal case is to represent speech using smooth and accurate parameter trajectories of the vocal tract transfer function and the glottal source signal. The GSS analysis/synthesis method developed in this work is a step forward in meeting this criteria as explained in Section 6.4.5. Basically, it attenuates parameter discontinuities of the vocal tract by performing a smoothing operation on the glottal source parameter trajectories and using STRAIGHT to compute a smooth spectrogram. If it was possible to accurately estimate the vocal tract filter during unvoiced sounds, speech could be represented using an uniform and continuous model of speech. Such a model is attractive for statistical modelling by HMMs and gives a close representation of the real speech production model.

The aperiodic component of speech is also important for speech naturalness. State-of-the-art HMM-based speech synthesisers model the noise component of speech in the frequency domain, e.g. using a MBE model or HNM. However, such models cannot represent well effects of the noise in the time-domain such as noise bursts or aspiration noise, which contribute to speech naturalness and are important to reproduce certain aspects of voice quality, such as breathiness (associated with aspiration noise). By using a sophisticated model of the noise in the time-domain, the quality of statistical speech synthesis could be improved. In particular, for the HTS-LF system it would be

desirable to use a time-domain model of the noise which could be combined with the LF-model. For example, aspiration noise is usually modelled as an amplitude modulated noise signal with its energy concentrated in the open phase of the glottal cycle. Therefore, aspiration noise could be combined with the LF-model signal by adding them pitch-syncronously and by using the glottal pulse for performing the amplitude modulation of the noise.

10.3.1.3 Hybrid Unit-selection/Statistical Speech Synthesis

The speech parameter trajectories generated by HMM-based speech synthesis can also be used to select the natural speech units to concatenate using the unit-selection method. The results of the evaluations conducted in the recent Blizzard Challenge 2010 (King and Karaiskos, 2010) indicate that this hybrid statistical/unit-selection approach can produce more natural speech than traditional unit-selection and HMM-based speech synthesis. The typical disadvantages of this hybrid method is the high computational complexity and memory requirements, which are not appropriate for several applications which require a low memory footprint. This method also provides low parametric flexibility for voice transformation. These reasons help to explain the high interest in improving the speech quality in HMM-based speech synthesis. That is, this method is suitable for a wider type of applications than unit-selection or hybrid statistical/unit-selection speech synthesis.

10.3.2 Applications

HMM-based speech synthesis using an acoustic glottal source model can be used for a wide range of applications. The following list indicates a set of topics where this synthesis method could be used:

- Voice transformation.
- Study of correlation between voice quality and glottal source parameters.
- Study of correlation between glottal source and prosody.
- Similarity to the speaker's voice (speaker's voice adaptation).
- Application to languages in which good glottal source modelling is considered to be important, such as Hindi.

10.3.2.1 Voice transformation

The GPF method can be used to transform voice characteristics of the synthetic speech, as explained in Section 6.3.4. Basically, it consists of using a different LF-model signal as input to the glottal post-filter than the stored LF-model waveform. In this work, an informal experiment was conducted in order to transform the voice quality of speech synthesised using the HTS-GPF system (which was described in Section 8.4.1.6). This experiment is described in Appendix C.

The HMM-based speech synthesiser using glottal source modelling which was developed during this thesis, the HTS-LF system, permits voice aspects of the synthetic speech to be transformed by modifying the LF-model parameters which are generated by the HMMs. This system allows the properties of the LF-model signal which is used to generate the excitation of voiced speech to be directly controlled, unlike the HTS-GPF system.

Formal perceptual experiments could be conducted in the future in order to evaluate the performance of the HTS-LF and the HTS-GPF systems in reproducing specific voice qualities. For example, the voice quality correlates of the LF-model parameters described in Section 5.3.2 could be used to synthesise speech with different voice qualities, e.g. breathy and tense.

In HMM-based speech synthesis, the parameters of the statistical models can be interpolated or adapted for transforming the voice characteristics of the synthetic speech. This transformation can be performed using a small amount of speech spoken with the target voice, e.g. the voice of a different speaker. However, the characteristics of the glottal source that are correlated with voice quality are typically incorporated into the spectral envelope and they might not be correctly transformed because the spectral envelope represents other speech characteristics in addition to the type of voice. The HTS-LF system has the advantage that these statistical parameter transformations can be performed independently for the glottal source and the spectral parameters, since they are modelled independently. Experiments could be conducted in the future in order to evaluate the performance of the HTS-LF system to transform the voice of the synthetic speech using the adaptation or interpolation techniques.

Also, voice transformation using the HTS-LF and the HTS-GPF systems could be improved by modelling other speech effects which are important to voice quality, e.g. aspiration noise and jitter.

10.3.2.2 Synthesis of Expressive Speech

The parametric flexibility to control glottal source aspects in the HMM-based speech synthesisers using the LF-model can also be used to synthesise expressive speech. For example, the techniques for voice transformation indicated in the previous section can also be applied to synthesis of speech with different emotions, or to produce certain speech effects (such as breathiness) which are difficult to synthesise without modelling glottal source parameters.

10.3.2.3 Application to Different Languages

In most languages, people control the movement of the glottis to produce voiced sounds with different pitches (determined by the rate of vibration of the vocal folds) and for the realisation of voiced and voiceless phonation. For examples, vowels are voiced sounds with a regular periodic pattern while unvoiced stop consonants are characterised by a voiceless phonation, such as the phone /k/ in English. However, the non-modal phonation, e.g. breathy or creaky, is also important to phonetic contrast in several languages. For example, in Hindi the contrast between breathy and modal voice is common in *obstruents* and *nasals* (Gordon and Ladefoged, 2001). The breathy voice is characterised by vocal folds that are highly abducted and by turbulent airflow through the glottis, as described in Section 5.3.2. The contrast between creaky and modal voice is also common in many languages (Gordon and Ladefoged, 2001). Creaky voice is commonly used as a marker of prosodic boundaries, such as in Finnish and English to mark vowel-initial words (Gordon and Ladefoged, 2001). In contrast to breathy voice, creaky voice is typically associated with high adduction of the vocal folds. More details about the characteristics of creaky phonation can be found in Section 5.3.2.

The contrast between modal voice and other voice qualities which is relevant in several languages could be more accurately modelled by a HMM-based speech synthesiser using glottal source modelling, such as the HTS-LF system. In this thesis, a Hindi voice was built using the HTS-STRAIGHT system (HMM-based speech synthesiser using STRAIGHT) which was described in Section 7.2. An example of speech synthesised with this Hindi voice can be found at:
`http://homepages.inf.ed.ac.uk/jyamagis/Demo-html/demo.html`

As future work, the Hindi voice could also be built using the HTS-LF system and evaluated against the HTS-STRAIGHT system, as the HTS-LF system is expected to

model the contrast between modal and breathy voice more accurately.

10.3.2.4 Correlation between Glottal Source Parameters and Prosody

Voice source dynamics are very important to prosodic aspects of speech, such as stressed syllables and intonation. However, the study of the correlations between glottal source parameters and speech prosody is usually performed on isolated speech sounds, such as vowels, to facilitate the analysis and to obtain more accurate results. The HTS-LF system could be used to extend the study of the prosodic correlates of the glottal source to a wide range of acoustic realisations and to study *supra-segmental* prosodic characteristics, by analysing the trajectories generated by the HMMs from input test sentences.

10.4 Final Remarks

In this work, two different HMM-based speech synthesisers were developed which incorporate an acoustic voice source model, the LF-model. One is the HTS-GPF system which represents the voiced excitation by passing the LF-model through a glottal post-filter and the spectrum by the spectral envelope of speech. The other is the HTS-LF system which uses the GSS method proposed in this thesis to estimate both the LF-model and vocal tract parameters. This system uses these parameters for training the HMMs and for generation of voiced speech in which the excitation is represented by the LF-model. Both the HTS-GPF and HTS-LF systems are competitive to a standard HMM-based speech synthesiser which uses the STRAIGHT vocoder.

The results of initial experiments conducted in this work to compare the LF-model to the traditional impulse train excitation were positive and showed that using the LF-model for speech synthesis could improve the speech quality. However, informal analysis of speech synthesised using the HTS-LF system indicated that there were some problems in this system which caused speech degradation. Even after performing improvements to the HTS-LF system in order to solve these problems, this system was not as successful as expected in a perceptual speech quality evaluation which was conducted in this work. Possible causes of speech distortion in the HTS-LF system were then investigated by conducting an objective measurement experiment. From the results of this experiment promising ideas to further improve the HTS-LF system have been proposed.

A great advantage of the HTS-LF and HTS-GPF systems compared to state-of-the-art HMM-based speech synthesisers is that they provide control over acoustic glottal parameters (the LF-model parameters). This is a valuable characteristic, especially because glottal parameters can be transformed for more correctly reproducing different voice qualities and synthesising more expressive speech, such as speech with vocal emotions.

This thesis proposed analysis/synthesis methods to incorporate an acoustic glottal source model into a HMM-based speech synthesiser. It also identified and investigated several difficulties encountered in developing a HMM-based speech synthesiser using glottal source modelling. Also, the speech synthesisers using the LF-model which were developed in this work can be useful in a wide variety of applications, which motivate future work. The expectations for glottal source modelling in HMM-based speech synthesis are high and this thesis contributed to the study of important aspects in this method. Moreover, we firmly believe the results of this work are very promising in this line of research.

Appendix A

Results of the Evaluation Based on the Blizzard Test Setup

A.1 SIM - Similarity

	Median		MAD		Mean		SD	
	full	arctic	full	arctic	full	arctic	full	arctic
Natural	5	5	0.0	0.0	4.8	4.9	0.57	0.44
HTS-STRAIGHT	3	3	1.5	1.5	3.0	2.6	1.03	1.13
HTS-GPF	3	3	1.5	1.5	2.9	2.7	0.93	1.10
HTS-FFT	3	3	1.5	1.5	2.7	2.6	1.02	1.12
HTS-STR-PR	3	2	1.5	1.5	2.8	2.3	1.02	1.06
HTS-LF	2	2	1.5	1.5	2.3	2.0	0.94	0.99
HTS-LF-PR	2	2	1.5	1.5	2.3	2.0	1.01	0.98

Table A.1: Similarity scores for the full male voice and ARCTIC subset of the male voice. Results are given for the different HMM-based speech synthesisers in terms of the median, median absolute deviation (MAD), mean, and standard deviation (SD).

	Median	MAD	Mean	SD
Natural	5	0.0	4.8	0.56
HTS-STRAIGHT	3	1.5	3.0	1.05
HTS-GPF	3	0.0	3.0	0.86
HTS-FFT	3	1.5	3.1	0.93
HTS-STR-PR	2.5	0.75	2.6	0.98
HTS-LF	2	1.5	2.1	0.87
HTS-LF-PR	2	1.5	2.0	1.06

Table A.2: Similarity scores for the female voice. Results are given for the different HMM-based speech synthesisers in terms of the median, median absolute deviation (MAD), mean, and standard deviation (SD).

	S1	S2	S3	S4	S5
HTS-STRAIGHT (S1)					
HTS-GPF (S2)	4.45E-1				
HTS-FFT (S3)	1.11E-2	3.09E-2			
HTS-STR-PR (S4)	9.77E-2	3.24E-1	1.85E-1		
HTS-LF (S5)	3.81E-8	8.03E-8	1.06E-3	6.04E-6	
HTS-LF-PR (S6)	6.38E-9	1.84E-7	2.18E-3	1.82E-5	9.59E-1

Table A.3: $P-values$ of similarity scores calculated for the HMM-based speech synthesisers, for the full male voice.

	S1	S2	S3	S4	S5
HTS-STRAIGHT (S1)					
HTS-GPF (S2)	5E-1				
HTS-FFT (S3)	8.5E-1	4.66E-1			
HTS-STR-PR (S4)	6.44E-2	9.85E-3	3.61E-2		
HTS-LF (S5)	1.20E-5	2.23E-7	4.40E-7	4.13E-3	
HTS-LF-PR (S6)	2.05E-5	1.15E-6	1.30E-6	8.12E-3	6.77E-1

Table A.4: $P-values$ of similarity scores calculated for the HMM-based speech synthesisers, for the ARCTIC subset of the male voice.

	S1	S2	S3	S4	S5
HTS-STRAIGHT (S1)					
HTS-GPF (S2)	8.56E-1				
HTS-FFT (S3)	3.58E-1	3.06E-1			
HTS-STR-PR (S4)	7.08E-3	2.77E-3	6.51E-4		
HTS-LF (S5)	2.22E-8	9.68E-8	4.19E-9	8.78E-6	
HTS-LF-PR (S6)	2.84E-7	2.12E-7	3.83E-9	2.53E-5	6.20E-1

Table A.5: $P-values$ of similarity scores calculated for the HMM-based speech synthesisers, for the female voice.

A.2 MOS - Naturalness

	Median		MAD		Mean		SD	
	full	arctic	full	arctic	full	arctic	full	arctic
Natural	5	5	0.0	0.0	4.9	4.9	0.38	0.46
HTS-STRAIGHT	3	3	1.5	1.5	3.2	2.9	1.03	1.06
HTS-GPF	3	3	1.5	1.5	3.0	2.6	1.01	1.01
HTS-FFT	3	3	1.5	1.5	2.8	2.7	1.05	1.04
HTS-STR-PR	3	3	1.5	1.5	2.7	2.5	0.97	0.98
HTS-LF	2	2	1.5	1.5	2.2	1.9	0.96	0.95
HTS-LF-PR	2	1	1.5	0.0	2.0	1.7	0.97	0.90

Table A.6: MOS scores obtained for the HMM-based speech synthesisers, for the full male voice and the ARCTIC subset of the male voice. Results are given in terms of the median, median absolute deviation (MAD), mean, and standard deviation (SD).

	Median	MAD	Mean	SD
Natural	5	0.0	4.9	0.26
HTS-STRAIGHT	3	1.5	3.1	0.94
HTS-GPF	3	1.5	2.9	0.91
HTS-FFT	3	1.5	3.0	0.95
HTS-STR-PR	3	1.5	2.6	0.97
HTS-LF	2	1.5	1.7	0.78
HTS-LF-PR	2	1.5	1.6	0.76

Table A.7: MOS scores obtained for the HMM-based speech synthesisers, for the female voice. Results are given in terms of the median, median absolute deviation (MAD), mean, and standard deviation (SD).

	S1	S2	S3	S4	S5
HTS-STRAIGHT (S1)					
HTS-GPF (S2)	4.79E-3				
HTS-FFT (S3)	1.2E-7	2.46E-2			
HTS-STR-PR (S4)	1.54E-10	2.22E-3	3.27E-1		
HTS-LF (S5)	3.46E-25	1.12E-20	5.33E-15	5.25E-10	
HTS-LF-PR (S6)	4.15E-30	1.29E-23	2.49E-17	2.07E-17	1.14E-2

Table A.8: $P-values$ of MOS scores calculated for the HMM-based speech synthesisers, for the full male voice.

	S1	S2	S3	S4	S5
HTS-STRAIGHT (S1)					
HTS-GPF (S2)	1.04E-3				
HTS-FFT (S3)	2.75E-2	2.52E-1			
HTS-STR-PR (S4)	9.50E-7	1.13E-1	9.11E-3		
HTS-LF (S5)	6.13E-27	7.26E-18	9.86E-21	2.32E-14	
HTS-LF-PR (S6)	5.63E-34	1.62E-26	1.18E-28	4.33E-23	1.04E-3

Table A.9: $P-values$ of MOS scores calculated for the HMM-based speech synthesisers, for the ARCTIC subset of the male voice.

	S1	S2	S3	S4	S5
HTS-STRAIGHT (S1)					
HTS-GPF (S2)	5.33E-4				
HTS-FFT (S3)	6.53E-3	4.44E-1			
HTS-STR-PR (S4)	2.53E-10	1.57E-4	5.71E-6		
HTS-LF (S5)	1.93E-40	5.22E-33	9.04E-34	5.44E-27	
HTS-LF-PR (S6)	3.37E-39	1.18E-35	4.10E-36	1.79E-30	1.13E-1

Table A.10: $P-values$ of MOS scores calculated for the HMM-based speech synthesisers, for the female voice.

A.3 ABX - Naturalness

	S1	S2	S3	S4	S5	S6	S7
Natural (S1)		97.7	97.6	97.5	95.2	97.5	97.6
HTS-STRAIGHT (S2)	2.3		15.0	7.5	34.1	69.8	75
HTS-GPF (S3)	2.4	5.0		9.5	26.8	70.7	72.1
HTS-FFT (S4)	2.5	7.5	9.5		11.9	52.4	54.8
HTS-STR-PR (S5)	4.8	2.4	4.9	2.4		51.2	78.0
HTS-LF (S6)	2.5	9.3	9.8	11.9	14.6		12.2
HTS-LF-PR (S7)	2.4	12.5	2.3	14.3	4.9	2.4	

Table A.11: Preference rates (in percentage) from ABX comparisons obtained for the HMM-based speech synthesisers, for the full male voice.

	S1	S2	S3	S4	S5	S6	S7
Natural (S1)		97.7	97.6	95.0	92.9	97.5	100
HTS-STRAIGHT (S2)	0.0		7.5	12.5	22.0	53.5	57.5
HTS-GPF (S3)	2.4	10.0		4.8	9.8	48.8	62.8
HTS-FFT (S4)	5.0	7.5	2.4		4.8	50.0	73.8
HTS-STR-PR (S5)	4.8	2.4	4.9	4.8		63.4	61.0
HTS-LF (S6)	2.5	18.6	22.0	7.1	9.8		2.4
HTS-LF-PR (S7)	0.0	7.5	16.3	9.5	9.8	0.0	

Table A.12: Preference rates (in percentage) from ABX comparisons obtained for the HMM-based speech synthesisers, for the ARCTIC subset of the male voice.

	S1	S2	S3	S4	S5	S6	S7
Natural (S1)		100	96.8	96.7	100	100	100
HTS-STRAIGHT (S2)	0.0		26.7	16.7	19.4	81.3	93.3
HTS-GPF (S3)	0.0	3.3		0.0	19.4	80.6	65.6
HTS-FFT (S4)	3.3	13.3	12.9		29.0	59.4	93.8
HTS-STR-PR (S5)	0.0	0.0	19.4	0.0		77.4	71.0
HTS-LF (S6)	0.0	0.0	0.0	12.5	12.9		13.3
HTS-LF-PR (S7)	0.0	0.0	9.4	0.0	3.2	10.0	

Table A.13: Preference rates (in percentage) from ABX comparisons obtained for the HMM-based speech synthesisers, for the female voice.

	S1	S2	S3	S4	S5	S6
Natural (S1)						
HTS-STRAIGHT (S2)	5.1E-12					
HTS-GPF (S3)	3.8E-11	6.4E-1				
HTS-FFT (S4)	5.1E-12	1.0	1.0			
HTS-STR-PR (S5)	4.1E-10	6E-2	2.1E-1	6.4E-1		
HTS-LF (S6)	7.5E-11	4.2E-5	1.1E-4	7.9E-3	2.8E-2	
HTS-LF-PR (S7)	3.8E-11	4.2E-5	1.6E-6	7.9E-3	7.8E-7	5.3E-1

Table A.14: $P-values$ of preference rates calculated for the HMM-based speech synthesisers, for the full male voice.

	S1	S2	S3	S4	S5	S6
Natural (S1)						
HTS-STRAIGHT (S2)	2.0E-8					
HTS-GPF (S3)	1.5E-9	1.0				
HTS-FFT (S4)	1.1E-13	8.8E-1	1.0			
HTS-STR-PR (S5)	7.5E-11	2.1E-1	7.6E-1	1.0		
HTS-LF (S6)	3.8E-11	3.2E-2	1.2E-1	7.9E-3	7.6E-1	
HTS-LF-PR (S7)	9.1E-13	2.2E-3	1.9E-3	1.5E-5	1.5E-3	1.0

Table A.15: $P-values$ of preference rates calculated for the HMM-based speech synthesisers, for the ARCTIC subset of the male voice.

	S1	S2	S3	S4	S5	S6
Natural (S1)						
HTS-STRAIGHT (S2)	2.3E-10					
HTS-GPF (S3)	9.3E-10	2.0E-1				
HTS-FFT (S4)	5.8E-8	8.6E-1	4.7E-1			
HTS-STR-PR (S5)	9.3E-10	2.8E-1	1.0	1.5E-1		
HTS-LF (S6)	1.9E-9	2.6E-6	4.7E-6	7.0E-3	8.8E-4	
HTS-LF-PR (S7)	9.3E-10	5.8E-8	2.1E-3	1.5E-8	1.9E-4	8.6E-1

Table A.16: $P-values$ of preference rates calculated for the HMM-based speech synthesisers, for the female voice.

A.4 WER - Intelligibility

	Mean			SD		
	full	arctic	fem.	full	arctic	fem.
Natural	1.5	2.4	-	5.6	7.9	-
HTS-STRAIGHT	4.3	29.5	11.4	8.9	17.9	15
HTS-GPF	6.8	27.9	10.4	8.7	17.4	14
HTS-FFT	6.1	28.8	8.9	9.8	17.1	12
HTS-STR-PR	7.0	28.2	8.4	10.7	18.7	11
HTS-LF	10.0	45.4	21.0	13.4	19.1	18
HTS-LF-PR	11.1	45.1	21.9	13.1	20.2	19

Table A.17: Mean and standard deviation (SD) of the word error rates (in percentage) obtained for the HMM-based speech synthesisers, for the three voices.

	S1	S2	S3	S4	S5	S6
Natural (S1)						
HTS-STRAIGHT (S2)	1.1E-3					
HTS-GPF (S3)	2.7E-6	1.4E-1				
HTS-FFT (S4)	9.9E-6	8.6E-2	5.8E-1			
HTS-STR-PR (S5)	3.4E-6	7.0E-2	9.1E-1	7.8E-1		
HTS-LF (S6)	8.8E-9	2.0E-4	10.0E-3	2.2E-3	2.9E-2	
HTS-LF-PR (S7)	7.1E-10	1.3E-5	3.4E-3	1.2E-3	8.2E-3	9.0E-1

Table A.18: $P-values$ of the WER calculated for the HMM-based speech synthesisers, for the full male voice.

Appendix B

Objective Measurements

	MOS		SIM		SU	
	VT	All	VT	All	VT	All
Positive Energy Disc.	-0.03	0.016	0.22	0.21	-0.32	-0.23
Negative Energy Disc.	0.01	0.054	-0.01	0.03	0.12	0.05
D_E of Energy	0.04	0.017	0.11	0.44	0.05	0.25
D_E of mel-spec. coef.	0.13	0.054	0.26	0.23	0.32	0.47
D_E of Δ mel-spec. coef.	0.02	-0.083	0.17	0.06	-0.04	0.28
D_{KL} of FFT coef.	0.12	0.061	0.27	0.22	0.13	0.48
D_E of R_{14}	0.06	-0.027	0.17	0.22	0.39	0.49
D_E of R_{24}	0.14	-0.003	0.15	0.21	0.23	0.44
D_E of H1-H2	-0.11	-0.004	-0.02	0.18	0.05	0.49
D_E of SNR	-0.23	-0.079	0.07	0.27	0.12	0.21

Table B.1: Correlation coefficients between the objective measurements and the perceptual scores, calculated for the parts of the test sentences in voicing transition regions (VT) and for the whole test sentences (All).

Appendix C

Voice Transformation Experiment Using the HTS-GPF System

In the work of this thesis, an informal experiment was conducted in order to investigate the effect of modifying the parameters of the LF-model used by the HTS-GPF synthesiser (which was described in Section 8.4.1.6) on the synthetic speech signal generated by this system.

A small set of sentences were synthesised using the HTS-GPF system, for different shapes of the input LF-model waveform. Speech synthesised using the *reference* LF-model signal (stored LF-model signal), was considered to have neutral voice quality. This is the voice quality which is obtained by synthesising speech using a spectrally flat excitation in the HTS-GPF system, without performing any transformation to the excitation or the synthesis filter (which represents the spectral envelope). Then, the sentences were also synthesised with different voice characteristics by varying one of the dimensionless parameters of the LF-model: open quotient (OQ), speed quotient (SQ) and return quotient (RQ). These parameters were described in Section 5.2.3 and their voice quality correlates were explained in Section 5.3.2. Each parameter was decreased and increased by different degrees. For example, the OQ was multiplied by scale factors, which ranged from 0.2 to 1.8. Examples of the synthetic speech samples are accessible at `http://homepages.inf.ed.ac.uk/jscabral/hts-gpf.html`. The variation of voice characteristics with the degree of transformation of each LF-model parameter can be clearly perceived by listening to the synthetic speech signals. Moreover, each parameter appears to have a different effect on the voice quality. This result was expected, as the variation of each parameter has a different effect on the spectrum of the LF-model (Doval and d'Alessandro, 1999). The voice quality transformations

also seemed not to produce speech artefacts, even for relatively large degrees of transformation of the LF-parameters. Further experiments need to be conducted for finding the ranges of the LF-parameter variations which do not produce distortion in the synthetic speech.

Bibliography

Abdel-Hamid, O., Abdou, S., and Rashwan, M. (2006). Improving Arabic HMM based speech synthesis quality. In *Proc. of INTERSPEECH*, Pittsburgh, USA.

Acero, A. (1999). Formant analysis and synthesis using hidden Markov models. In *Proc. of EUROSPEECH*, Hungary.

Alku, P., Bäckström, T., and Vilkman, E. (2002). Normalized amplitude quotient for parameterization of the glottal flow. *J. Acoust. Soc. Am.*, 112:701–710.

Alku, P., Strik, H., and Vilkman, E. (1997). Parabolic spectral parameter: a new method for quantification of the glottal flow. *Speech Communication*, 22:67–79.

Alku, P. and Vilkman, E. (1994). Estimation of the glottal pulseform based on discrete all-pole modeling. In *Proc. of ICSLP*, Yokohama, Japan.

Alku, P. and Vilkman, E. (1996). Amplitude domain quotient of the glottal volume velocity waveform estimated by inverse filtering. *Speech Communication*, 18:131–138.

Alku, P., Vilkman, E., and Laine, U. K. (1991). Analysis of glottal waveform in different phonation types using the new IAIF method. In *Proc. of International Congress of Phonetic Sciences (ICPhS)*, Aix-en-Provence, France.

Allen, M., Hunnicutt, M. S., and Klatt, D. (1987). *From Text to Speech: The MITalk System*. Cambridge Univ. Press, Cambridge, UK.

Ananthapadmanabha, T. V. (1984). Acoustic analysis of voice source dynamics. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Ananthapadmanabha, T. V. and Fant, G. (1982). Calculation of true glottal flow and its components. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Arroabarren, I. and Carlosena, A. (2003). Glottal source parameterization: a comparative study. In *Proc. of ITRW (VOQUAL'03)*, pages 29–34, Geneva, Switzerland.

Avanzini, F., Alku, P., and Karjalainen, M. (2001). One-delayed-mass model for efficient synthesis of glottal flow. In *Proc. of EUROSPEECH*, Aalborg, Denmark.

Baker, J. (1975). The DRAGON system–An overview. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 23:24–29.

Banos, E., Erro, D., Bonafonte, A., and Moreno, A. (2008). Flexible harmonic/stochastic modeling for HMM-based speech synthesis. In *V Jornadas en Tecnologías del Habla*, Spain.

Barra-Chicote, R., Yamagishi, J., King, S., Montero, J., and Macias-Guarasa, J. (2010). Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication*, 52:394–404.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41:164–171.

Bennett, C. (2005). Large scale evaluation of corpus-based synthesizers: Results and lessons from the 2005 Blizzard Challenge. In *Proc. of INTERSPEECH*, Lisbon, Portugal.

Birkholz, P. (2010). VocalTractLab. Avaiable at http://www.vocaltractlab.de.

Black, A., Taylor, P., and Caley, R. (2004). The Festival Speech Synthesis System. http://www.festvox.org/festival.

Black, A. and Tokuda, K. (2005). Blizzard Challenge – 2005: Evaluating corpus-based speech synthesis on common datasets. In *Proc. of INTERSPEECH*, Lisbon, Portugal.

Black, A., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *Proc. of ICASSP*, Hawaii, USA.

Bozkurt, B. (2005). *Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals*. Ph.D. thesis, Faculté Polytechnique De Mons, Belgium.

Cabral, J. P. and Oliveira, L. C. (2005). Pitch-synchronous time-scaling for prosodic and voice quality transformations. In *Proc. of INTERSPEECH*, pages 1137–1140, Lisbon, Portugal.

Cahn, J. E. (1989). Generating Expression in Synthesized Speech. Master's Thesis, MIT.

Campbell, W. and Black, A. (1996). Prosody and the selection of source units for concatenative synthesis. In Heidelberg, editor, *Progress in Speech Synthesis*. Springer.

Carlson, R., Fant, G., Gobl, C., Granstrom, B., Karlsson, I., and Lin, Q.-G. (1989). Voice source rules for text-to-speech synthesis. In *Proc. of ICASSP*, Glasgow, Scotland.

Childers, D. and Lee, C. (1991). Vocal quality factors: Analysis, synthesis and perception. *J. Acoust. Soc. Am.*, 90:2394–2410.

Childers, D. G. and Ahn, C. (1995). Modeling the glottal volume-velocity waveform for three voice types. *J. Acoust. Soc. Am.*, 97:505–519.

Chung, J. and Schafer, R. (1990). Excitation modeling in a homomorphic vocoder. In *Proc. of ICASSP*, USA.

Clark, R., Podsiadlo, M., Fraser, M., Mayo, C., and King, S. (2007a). Statistical analysis of the Blizzard Challenge 2007 listening test results. In *Proc. of Blizzard Challenge Workshop*, Bonn, Germany.

Clark, R., Richmond, K., and King, S. (2007b). Multisyn: Open-domain unit selection for the Festival speech synthesis system. *Speech Communication*, 49:317–330.

d'Alessandro, C., D'Alessandro, N., Le Beux, S., and Doval, B. (2006). Comparing time-domain and spectral-domain voice source models for gesture controlled vocal instruments. In *Proc. of 5th International Conference on Voice Physiology and Biomechanics*, Tokyo, Japan.

de Cheveigné, A. (1996). Speech fundamental frequency estimation. TR-H-195, ATR Interpreting Telephony Laboratories, Osaka, Japan.

Deller, J. R., Proakis, J. G., and Hansen, J. H. (1993). *Discrete Time Processing of Speech Signals*. Macmillan, New York, USA.

Deng, L. (1998). A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. *Speech Communication*, 24:299–323.

Dines, J., Yamagishi, J., and King, S. (2009). Measuring the gap between HMM-based ASR and TTS. In *Proc. of INTERSPEECH*, Brighton, UK.

Ding, W., Kasuya, H., and Adachi, S. (1995). Simultaneous estimation of vocal tract and voice source parameters based on an ARX model. *IEICE Trans. Information and Systems*, E78-D:738–743.

Donovan, R. and Woodland, P. (1995). Automatic speech synthesiser parameter estimation using HMMs. In *Proc. of ICASSP*, Detroit, USA.

Doval, B. and d'Alessandro, C. (1997). Spectral correlates of glottal waveform models: an Analytic study. In *Proc. of ICASSP*, pages 1295–1298, Munich, Germany.

Doval, B. and d'Alessandro, C. (1999). The spectrum of glottal flow models. Notes et Documents LIMSI-CNRS (Notes and Documents of the Laboratory for Mechanics and Engineering Sciences), National Centre for Scientific Research, Stockholm, Sweden.

Doval, B., d'Alessandro, C., and Henrich, N. (2003). The voice source as a causal/anticausal linear filter. In *Proc. of ITRW (VOQUAL'03)*, Geneva, Switzerland.

Drugman, T., Bozkurt, B., and Dutoit, T. (2009a). Complex cepstrum-based decomposition of speech for glottal source estimation. In *Proc. of INTERSPEECH*, Brighton, UK.

Drugman, T. and Dutoit, T. (2009). Glottal closure and opening instant detection from speech signals. In *Proc. of INTERSPEECH*, Brighton, UK.

Drugman, T., Wilfart, G., and Dutoit, T. (2009b). A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *Proc. of INTERSPEECH*, Brighton, UK.

Drugman, T., Wilfart, G., Moinet, A., and Dutoit, T. (2009c). Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis. In *Proc. of ICASSP*, Taipei, Taiwan.

Dudley, H., Riesz, R., and Watkins, S. (1939). A synthetic speaker. *Journal of the Franklin Institute*, 227:739–764.

Dutoit, T. (1993). MBR-PSOLA : Text-to-speech synthesis based on an MBE re-synthesis of the segments database. *Speech Communication*, 13:435–440.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*. Klower Academic Publishers.

El-Jaroudi, A. and Makhoul, J. (1991). Discrete all-pole modelling. *IEEE Trans. Signal Processing*, 39:411–423.

Fant, G. (1953). Speech communication research. *IVA, Royal Swedish Academy of Engineering Sciences*, 24:331–337.

Fant, G. (1973). *Speech sound and features*. MIT press, Cambridge, MA.

Fant, G. (1979). Glottal source and excitation analysis. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Fant, G. (1981). The source filter concept in voice production. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Fant, G. (1982). The voice source - acoustic modeling. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Fant, G. (1993). Some problems in voice source analysis. *Speech Communication*, 13:7–22.

Fant, G. (1995). The LF-model revisited. Transformations and frequency domain analysis. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Fant, G. (1997). The voice source in connected speech. *Speech Communication*, 22:125–139.

Fant, G. and Kruckenberg, A. (1996). Voice source properties of the speech code. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Fant, G., Liljencrants, J., and Lin, Q. (1985). A four-parameter model of glottal flow. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Fant, G. and Lin, Q. (1988). Frequency domain interpretation and derivation. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Ferguson, J. D. (1980). Variable duration models for speech. In *In Proc. of Symp. Application of Hidden Markov Models to Text and Speech*, Princeton, NJ, USA.

Fitt, S. and Isard, S. (1999). Synthesis of regional English using a keyword lexicon. In *Proc. of EUROSPEECH*, Budapest, Hungary.

Flanagan, J. (1972). *Speech Analysis, Synthesis and Perception*. Springer-Verlag (2nd edition), New York.

Fraser, M. and King, S. (2007). The Blizzard Challenge 2007. In *Proc. of Blizzard Challenge Workshop*, Bonn, Germany.

Fröhlich, M., Michaelis, D., and Strube, H. (2001). SIM-simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *J. Acoust. Soc. Am.*, 110:479–488.

Fu, Q. and Murphy, P. J. (2006). Robust glottal source estimation based on joint source-filter model optimization. *IEEE Trans. Audio, Speech and Language Processing*, 14:492–501.

Fujisaki, H. and Ljungqvist, M. (1986). Proposal and evaluation of models for the glottal source waveform. In *Proc. of ICASSP*, Tokyo, Japan.

Fujisaki, H. and Ljungqvist, M. (1987). Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform. In *Proc. of ICASSP*, Dallas, USA.

Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. (1992). An adaptive algorithm for Mel-cepstral analysis of speech. In *Proc. of ICASSP*, San Francisco, USA.

Funaki, K. and Mitome, Y. (1990). A speech analysis method based on a glottal source model. In *Proc. of ICSLP*, Japan.

Funaki, K., Miyanaga, Y., and Tochinai, K. (1999). Recursive ARMAX speech analysis based on a glottal source model with phase compensation. *Proc. of ICSLP*, 74:279–295.

Gales, M. and Young, S. (2007). The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1:195–304.

Gardner, W. R. (1994). *Modeling and Quantization Techniques for Speech Compression Systems*. Ph.D. thesis, University of California, San Diego, USA.

Gauffin, J. and Sundberg, J. (1989). Spectral correlates of glottal voice source waveform characteristics. *Journal of Speech and Hearing Research*, 32:556–565.

Gobl, C. (1989). A preliminary study of acoustic voice quality correlates. STL-QPSR, Royal Institute of Technology, KTH, Stockholm, Sweden.

Gobl, C. (2006). Modelling aspiration noise during phonation using the LF voice source model. In *Proc. of INTERSPEECH*, Pittsburg, USA.

Gobl, C. and Ní Chasaide, A. (2003). Amplitude-based source parameters for measuring voice quality. In *Proc. of ITRW (VOQUAL'03)*, Geneva, Switzerland.

Gonzalvo, X., Socoro, J., Iriondo, I., Monzo, C., and Martinez, E. (2007). Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish. In *Proc. of 6th ISCA Speech Synthesis Workshop (SSW6)*, Bonn, Germany.

Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29:383–406.

Gray, H. A. and Markel, J. D. (1976). Distance measures for speech processing. *IEEE Trans. Acoustics, Speech and Signal Processing*, ASSP-24:380–391.

Guerchi, D. and Mermelstein, P. (2000). Low-rate quantization of spectral information in a 4 kb/s pitch-synchronous CELP coder. In *IEEE Workshop on Speech Coding*, Delavan, U.S.A.

Han, S., Jeong, S., and Hahn, M. (2009). Optimum MVF estimation-based two-band excitation for HMM-based speech synthesis. *ETRI Journal*, 31:457–459.

Hanquinet, J., Grenez, F., and Schoentgen, J. (2005). Synthesis of disordered speech. In *Proc. of INTERSPEECH*, Lisbon, Portugal.

Hanson, H. M. and Chuang, E. S. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *J. Acoust. Soc. Am.*, 106:1064–1077.

Hedelin, P. (1984). A glottal LPC-vocoder. In *Proc. of ICASSP*, USA.

Hemptinne, C. (2006). Integration of the Harmonic plus Noise Model (HNM) into the Hidden Markov Model-Based Speech Synthesis System (HTS). Master's Thesis, IDIAP Research Institute.

Hirtum, A., Lopez, I., Hirschberg, M., and Pelorson, X. (2003). On the relationship between input parameters in the two-mass vocal-fold model with acoustical coupling and signal parameters of the glottal flow. In *Proc. of ITRW (VOQUAL'03)*, Geneva, Switzerland.

Högberg, J. (1997). Data driven formant synthesis. In *Proc. of EUROSPEECH*, Greece.

Holmes, J. N. (1972). *Speech Synthesis*. Mills & Boon, London.

Hong, S., Kang, S., and Ann, S. (1994). Voice parameter estimation using sequential SVD and wave shaping filter bank. In *Proc. of ICSLP*, Yokohama, Japan.

Huang, X., Acero, A., and Hon, H. W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall, USA.

Iida, A., Campbell, N., Iga, S., Higuchi, F., and Yasumura, M. (2000). A speech synthesis system for assisting communication. In *Proc. of ITRW on Speech and Emotion*, pages 213–214, Belfast, N. Ireland.

Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale. In *Proc. of ICASSP*, Boston, USA.

Imai, S. and Furuichi, C. (1988). Unbiased estimator of log spectrum and its application to speech signal processing. In *Proc. of EURASIP*, Greece.

Isaksson, A. and Millnert, M. (1989). Inverse glottal filtering using a parameterized input model. *Signal Processing*, 18:435–445.

Iseli, M., Shue, Y., Epstein, M. A., Keating, P., Kreiman, J., and Alwan, A. (2006). Voice source correlates of prosodic features in American English: A pilot study. In *Proc. of INTERSPEECH*, Pittsburgh, USA.

Ishizaka, K. and Flanagan, J. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *The Bell System Technical Journal*, 52:1233–1268.

Jackson, L. B. (1996). *Digital filters and signal processing*. Kluwer Academics, USA.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer, New York.

Kaburagi, T. and Kawai, K. (2003). Analysis of voice source characteristics using a constrained polynomial model. In *Proc. of EUROSPEECH*, Geneva, Switzerland.

Kane, J., Kane, M., and Gobl, C. (2010). A spectral LF model based approach to voice source parameterisation. In *Proc. of INTERSPEECH*, Tokyo, Japan.

Karlsson, I. and Liljencrants, J. (1996). Diverse voice qualities: models and data. In *Proc. of Swedish Phonetics Conference (Fonetik)*, Nasslingen, Sweden.

Kawahara, H. (1997). STRAIGHT - TEMPO: A universal tool to manipulate linguistic and para-linguistic speech information. In *Proc. of IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Florida, USA.

Kawahara, H., Cheveigné, A., Banno, H., Takahashi, T., and Irino, T. (2005). Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. In *Proc. of INTERSPEECH*, Lisbon, Portugal.

Kawahara, H., Estill, J., and Fujimura, O. (2001). Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. In *Proc. of MAVEBA*, Firenze, Italy.

Kawahara, H., Katayose, H., Cheveigné, A., and Patterson, R. D. (1999a). Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In *Proc. of EUROSPEECH*, Budapest, Hungary.

Kawahara, H., Masuda-Katsuse, I., and Cheveigné, A. (1999b). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27:187–207.

Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. (2008). Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Proc. of ICASSP*, Las Vegas, USA.

Keller, E. (2005). The analysis of voice quality in speech processing. *Lecture notes in computer science*, 3445:54–73.

Kelley, C. T. (2003). *Solving Nonlinear Equations with Newton's Method*. SIAM books, USA.

Kim, S., Kim, J., and Hahn, M. (2006). HMM-based Korean speech synthesis system for hand-held devices. *IEEE Trans. Consumer Electronics*, 52:1384–1390.

Kim, S. J. and Hahn, M. (2007). Two-band excitation for HMM-based speech synthesis. *IEICE Trans. Information and Systems*, E90-D:378–381.

King, S. and Karaiskos, V. (2009). The Blizzard Challenge 2009. In *Proc. of Blizzard Challenge Workshop*, Edinburgh, UK.

King, S. and Karaiskos, V. (2010). The Blizzard Challenge 2010. In *Proc. of Blizzard Challenge Workshop*, Tokyo, Japan.

Klabbers, E. and Veldhuis, R. (1998). On the reduction of concatenation artefacts in diphone synthesis. In *Proc. of ICSLP*, Sydney, Australia.

Klabbers, E. and Veldhuis, R. (2001). Reducing audible spectral discontinuities. *IEEE Trans. Speech and Audio Processing*, 9:39–51.

Klatt, D. (1982). The Klattalk text-to-speech conversion system. In *Proc. of ICASSP*, Paris, France.

Klatt, D. (1987). Review of text-to-speech conversion for English. *J. Acoust. Soc. Am.*, 82:737–793.

Klatt, D. and Klatt, L. (1987). Analysis, synthesis and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, 2:820–857.

Kominek, J. and Black, A. (2004). The CMU Arctic speech databases. In *Proc. of 5th ISCA Speech Synthesis Workshop (SSW5)*, Pittsburgh, USA.

Krishnamurthy, A. (1992). Glottal source estimation using a sum-of-exponentials model. *IEEE Trans. Signal Processing*, 40:682–686.

Krishnamurthy, A. and Childers, D. (1986). Two-channel speech analysis. *IEEE Trans. Signal Processing*, 34:730–743.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86.

Lawrence, W. (1953). The synthesis of speech from signals which have a low information rate. In Jackson, W., editor, *Communication Theory*. Butterworth.

Levinson, S. (1986). Continuously variable duration hidden Markov models for automatic speech recognition. *Computer Speech and Language*, 1:29–45.

Liljencrants, J. (1968). The OVE III speech synthesizer. *IEEE Trans. Audio Electroacoustics*, AU-16:137–140.

Lin, W., Koh, S., and Lin, X. (2000). Mixed excitation linear prediction coding of wideband speech at 8 kbps. In *Proc. of ICASSP*, Washington, DC, USA.

Ling, Z. and Wang, R. (2006). HMM-based unit selection using frame sized speech segments. In *Proc. of INTERSPEECH*, Pittsburgh, USA.

Ling, Z., Wu, Y., Wang, Y., Qin, L., and Wang, R. (2006a). USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method. In *Proc. of Blizzard Challenge Workshop*, Bonn, Germnay.

Ling, Z., Wu, Y., Wang, Y., Qin, L., and Wang, R. (2006b). USTC system for Blizzard Challenge 2006 an improved HMM-based speech synthesis method. In *Proc. of Blizzard Challenge Workshop*, Pittsburgh, USA.

Lobo, A. P. (2001). Glottal flow derivative modeling with the wavelet smoothed excitation. In *Proc. of ICASSP*, USA.

Maia, R., Toda, T., Zen, H., Nankaku, Y., and Tokuda, K. (2007a). A trainable excitation model for HMM-based speech synthesis. In *Proc. of INTERSPEECH*, Antwerp, Belgium.

Maia, R., Toda, T., Zen, H., Nankaku, Y., and Tokuda, K. (2007b). An excitation model for HMM-based speech synthesis based on residual modeling. In *Proc. of 6th ISCA Speech Synthesis Workshop (SSW6)*, Bonn, Germany.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proc. of IEEE*, 63:561–580.

Markel, J. D. and Gray, A. H. (1976). *Linear Prediction of Speech*. Springer-Verlag, New York.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11:431–441.

Masuko, T. (2002). *HMM-based speech synthesis and its applications*. Ph.D. Thesis, Tokyo Institute of Technology, Tokyo, Japan.

McCree, A. and Barnwell III, T. (1995). A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Trans. Speech and Audio Processing*, 3:242–250.

McCree, A., Truong, K., George, E. B., Barnwell, T., and Viswanathan, V. (1996). A 2.4 kbit/s MELP coder candidate for the new U.S. federal standard. In *Proceedings of the ICASSP*, Atlanta, GA, USA.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. In Chen, C. H., editor, *Pattern Recognition and Artificial Intelligence*, pages 374–388. Academic.

Milenkovic, P. (1986). Glottal inverse filtering by joint estimation of an AR system with a linear input model. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 34:28–42.

Milenkovic, P. (1993). Voice source model for continuous control of pitch period. *J. Acoust. Soc. Am.*, 93:1087–1096.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones. *Speech Communication*, 9:453–476.

Murphy, P. J. (2001). Spectral tilt as a perturbation-free measurement of noise levels in voice signals. In *Proc. of EUROSPEECH*, Aalborg, Denmark.

Murray, I. R. and Edgington, M. D. (2000). Rule-based emotion synthesis using concatenated speech. In *Proc. of ITRW on Speech and Emotion*, pages 173–177, Belfast, N. Ireland.

Murthy, K., Prasanna, S., and Yegnanarayana, B. (2004). Speaker-specific information from residual phase. In *Internat. Conf. on Signal Processing and Communications*, Bangalore, India.

Ní Chasaide, A. and Gobl, C. (1993). Contextual variation of the vowel voice source as a function of adjacent consonants. *Language and Speech*, 36:303–330.

Ní Chasaide, A. and Gobl, C. (2004). Voice quality and F0 in prosody: Towards a holistic account. In *Proc. of Speech Prosody*, Nara, Japan.

Ó Cinnéide, A., Dorran, D., Gainza, M., and Coyle, E. (2010). Towards a method to determine the glottal formant parameters of voiced speech without time-domain references. In *Proc. of Irish Signals and Systems Conference (ISSC)*, Cork, Ireland.

Öhlin, D. and Carlson, R. (2004). Data-driven formant synthesis. In *Proc. of Swedish Phonetics Conference (Fonetik)*, Stockholm, Sweden.

Oliveira, L. (1993). Estimation of source parameters by frequency analysis. In *Proc. of EUROSPEECH*, Berlin, Germany.

Ostendorf, M. and Bulyko, I. (2002). The impact of speech recognition on speech synthesis. In *Proc. of IEEE Workshop on Speech Synthesis*, Santa Monica, USA.

Pelorson, X., Hirschberg, A., van Hassel, R., Wijnands, A., and Auregan, Y. (1994). Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model. *J. Acoust. Soc. Am.*, 96:3416–3431.

Plumpe, M., Acero, A., Hon, H., and Huang, X. (1998). HMM-based smoothing for concatenative speech synthesis. In *Proc. of ICSLP*, Australia.

Plumpe, M., Quatieri, T., and Reynolds, D. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Speech and Audio Processing*, 7:569–586.

Price, P. (1989). Male and female voice source characteristics: inverse filtering results. *Speech Communication*, 8:261–277.

Proakis, J. G. and Manolakis, D. G. (1996). *Digital Signal Processing: Principles, Algorithms, and Applications*. Prentice-Hall, Upper Saddle River, NJ.

Qi, Y. and Bi, N. (1994). Simplified approximation of the 4- parameter LF model of voice source. *J. Acoust. Soc. Am.*, 96:1182–1185.

Quackenbush, S. R., Barnwell III, T. P., and Clement, M. A. (1988). *Objective Measures of Speech Quality*. Prentice Hall, Englewood Cliffs, New Jersey.

Quatieri, T. F. (1979). Minimum and mixed phase speech analysis-synthesis by adaptive homomorphic deconvolution. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-27:328–335.

Quatieri, T. F. (2001). *Discrete-time speech processing: principles and practice*. Prentice Hall.

Rabiner, L. and Juang, B.-H. (1993). *Fundamentals of Speech Recognition*. PRT Prentice Hall, New York.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of IEEE*, 77:257–286.

Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. (2008). HMM-based Finnish text-to-speech system utilizing glottal inverse filtering. In *Proc. of INTERSPEECH*, Brisbane, Australia.

Ratnayake, N., Savic, M., and Sorensen, J. (1992). Use of semi-Markov models for speaker-independent phoneme recognition. In *Proc. of ICASSP*, San Francisco, USA.

Rosen, G. (1958). Dynamic analog speech synthesizer. *J. Acoust. Soc. Am.*, 30:201–209.

Rosenberg, A. (1971). Effect of glottal pulse shape on the quality of natural vowels. *J. Acoust. Soc. Am.*, 49:583–598.

Rothenberg, M., Carlson, R., Granström, B., and Gauffin, J. (1975). A three-parameter voice source for speech synthesis. *Speech Communication*, 2:235–243.

Rouibia, S. and Rosec, O. (2005). Unit selection for speech synthesis based on a new acoustic target cost. In *Proc. of INTERSPEECH*, Lisbon, Portugal.

Russell, M. (1993). A segmental HMM for speech pattern modeling. In *Proc. of ICASSP*, April.

Russell, M. J. and Moore, R. K. (1985). Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition. In *Proc. of ICASSP*, Tampa, USA.

Schnell, K. (2006). Pitch modification of speech residual based on parameterized glottal flow with consideration of approximation error. In *Proc. of ICASSP*, Toulouse, France.

Schoentgen, J. (1993). Modelling the glottal pulse with a self-excited threshold autoregressive model. In *Proc. of EUROSPEECH*, Berlin, Germany.

Schoentgen, J. (2002). Analysis and synthesis of the honatory excitation signal by means of a pair of polynomial shaping functions. In *Proc. of ICSLP*, Denver, USA.

Schoentgen, J. (2003). On the bandwidth of a shaping function model of the phonatory excitation signal. In *Proc. of ITRW on Non-Linear Speech Processing (NOLISP)*, France.

Sciamarella, D. and d'Alessandro (2002). A study of the two-mass model in terms of acoustic parameters. In *Proc. of INTERSPEECH*, USA.

Shinoda, K. and Watanabe, T. (2000). MDL-based context-dependent subword modeling for speech recognition. *Journal Acoustic Soc. Japan*, 21:79–86.

Sluijter, R. J., Wuppermann, F., Taori, R., and Kathmann, E. (1995). State of the art and trends in speech coding. *Philips Journal of Research*, 49:455–488.

Steiglitz, K. and Dickinson, B. (1977). Computation of the complex cepstrum by factorization of the z-transform. In *Proc. of ICASSP*, Hartford, USA.

Stevens, K. N. (1998). *Acoustic Phonetics*. MIT Press, Cambridge, MA.

Story, B. (2003). Physical modeling of voice and voice quality. In *Proc. of ITRW (VOQUAL'03)*, Switzerland.

Strik, H. (1998). Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *J. Acoust. Soc. Am.*, 103:2659–2669.

Strik, H. and Boves, L. (1992). On the relation between voice source parameters and prosodic features in connected speech. *Speech Communication*, 11:167–174.

Strik, H. and Boves, L. (1994). Automatic estimation of voice source parameters. In *Proc. of ICSLP*, Yokohama, Japan.

Stylianou, Y. (1996). *Harmonic plus Noise Model for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. Ph.D. thesis, Ecole Nationale Superieure des Telecommunications, Paris.

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. Speech and Audio Processing*, 9:21–29.

Stylianou, Y. and Syrdal, A. K. (2001). Perceptual and objective detection of discontinuities in concatenative speech synthesis. In *Proc. of ICASSP*, Salt Lake City, USA.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In Kleijn, W. B. and Paliwal, K. K., editors, *Speech Coding and Synthesis*, pages 495–518. Elsevier Science.

Talkin, D. and Rowley, J. (1990). Pitch-synchronous analysis and synthesis for TTS systems. In *Proc. of ESCA Workshop on Speech Synthesis*, Autrans, France.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (1998). Speaker Adaptation for HMM-based speech synthesis system using MLLR. In *Proc. of 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, Blue Mountains, Australia.

Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (2001). Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In *Proc. of ICASSP*, France.

Thomson, M. (1992). A new method for determining the vocal tract transfer function and its excitation from voiced speech. In *Proc. of ICASSP*, San Francisco, USA.

Titze, I. (1984). Parameterization of the glottal area, glottal flow, and vocal fold contact area. *J. Acoust. Soc. Am.*, 75:570–580.

Toda, T. and Tokuda, K. (2007). Speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans. on Information and Systems*, Vol. E90-D:816–824.

Tokuda, K., Kobayashi, T., and Imai, S. (1995a). Speech parameter generation from HMM using dynamic features. In *Proc. of ICASSP*, Detroit, USA.

Tokuda, K., Masuko, T., Miyazaki, N., and Kobayashi, T. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *Proc. of ICASSP*, Phoenix, USA.

Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T., and Imai, S. (1995b). An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In *Proc. of EUROSPEECH*, Madrid, Spain.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. of ICASSP*, Istanbul, Turkey.

Tokuda, K., Zen, H., and Black, A. (2002). An HMM-based speech synthesis system applied to English. In *Proc. of IEEE Workshop on Speech Synthesis*, Santa Monica, USA.

Tokuda, K., Zen, H., and Kitamura, T. (2004). Reformulating the HMM as a trajectory model. In *Proc. of Beyond HMM – Workshop on Statistical Modeling Approach for Speech Recognition*, Kyoto, Japan.

Tokuda, K., Zen, H., Yamagishi, J., Black, A., Masuko, T., and Sako, S. (2009). The HMM-based speech synthesis system (HTS) version 2.1. http://hts.sp.nitech.ac.jp/.

Tooher, M. and McKenna, J. (2003). Variation of the glottal LF parameters across F0, vowels, and phonetic environment. In *Proc. ITRW (VOQUAL'03)*, Geneva, Switzerland.

van den Heuvel, H., Cranen, B., and Rietveld, T. (1996). Speaker variability in the coarticulation of /a, i, u/. *Speech Communication*, 18:113–130.

van Santen, J. (1997). Prosodic modeling in text-to-speech synthesis. In *Proc. of EUROSPEECH*, Rhodos, Greece.

Vaseghi, S. V. (1995). State duration modelling in hidden Markov models. *Signal Processing*, 41:31–41.

Veldhuis, R. (1998). A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. *J. Acoust. Soc. of Am.*, 103:566–571.

Vepa, J., King, S., and Taylor, P. (2002). Objective distance measures for spectral discontinuities in concatenative speech synthesis. In *Proc. of ICSLP*, Denver, USA.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Information Theory*, IT-13:260–269.

Wong, D. Y., Markel, J., and Gray, J. A. H. (1979). Least squares glottal inverse filtering from acoustic speech waveform. *IEEE Trans. Acoustics, Speech and Signal Processing*, 27:350–355.

Wouters, J. and Macon, M. (1998). Perceptual evaluation of distance measures for concatenative speech synthesis. In *Proc. of ICSLP*, Sydney, Australia.

Yamagishi, J. (2006). *Average-voice-based speech synthesis*. Ph.D. Thesis, Tokyo Institute of Technology, Tokyo, Japan.

Yamagishi, J. and Kobayashi, T. (2005). Adaptive training for hidden semi-Markov model. In *Proc. of ICASSP*, Philadelphia, USA.

Yamagishi, J. and Kobayashi, T. (2007). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. *IEICE Trans. Information and Systems*, E90-D:533–543.

Yamagishi, J., Kobayashi, T., Tachibana, M., Ogata, K., and Nakano, Y. (2007a). Model adaptation approach to speech synthesis with diverse voices and styles. In *Proc. of ICASSP*, USA.

Yamagishi, J., Onishi, K., Masuko, T., and Kobayashi, T. (2003). Modeling of various speaking styles and emotions for HMM-based speech synthesis. In *Proc. of EUROSPEECH*, Geneva, Switzerland.

Yamagishi, J., Zen, H., Toda, T., and Tokuda, K. (2007b). Speaker-independent HMM-based speech synthesis System - HTS-2007 system for the Blizzard Challenge 2007. In *Proc. of Blizzard Challenge Workshop*, Pittsburgh, USA.

Yan, Z., Qian, Y., and Soong, F. (2009). Rich context modeling for high quality HMM-based TTS. In *Proc. of INTERSPEECH*, Brighton, UK.

Yoshimura, T. (2002). *Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-To-Speech Synthesis*. Ph.D. Thesis, Nagoya Institute of Technology, Nagoya, Japan.

Yoshimura, T., Masuko, T., Tokuda, K., Kobayashi, T., and Kitamura, T. (1997). Speaker interpolation in HMM-based speech synthesis system. In *Proc. of EUROSPEECH*, Rhodes, Greece.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *IEICE Trans. Information and Systems*, J83-D-II:2099–2107.

Yoshimura, T., Tokuda, K., Masukom, T.and Kobayashi, T., and Kitamura, T. (2001). Mixed excitation for HMM-based speech synthesis. In *Proc. of EUROSPEECH*, Aalborg, Denmark.

Young, S., Evermann, G., Gales, M., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P. (2006). The HTK book version 3.4. http://htk.eng.cam.ac.uk/.

Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, 174:215–243.

Zen, H., Oura, K., Nose, T., Yamagishi, J., Sako, S., Toda, T., Masuko, T., Black, A., and Tokuda, K. (2009). Recent development of the HMM-based speech synthesis system. In *Proc. of Asia-Pacific Signal and Information Processing Association (APSIPA)*, Sapporo, Japan.

Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007a). Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. *IEICE Trans. Information and Systems*, E90-D:325–333.

Zen, H., Tokuda, K., and Kitamura, T. (2007b). Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech and Language*, 21:153–173.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2004). Hidden semi-Markov model based speech synthesis. In *Proc. of INTERSPEECH*, Jeju Island, Korea.

Zhang, L. (2009). *Modelling Speech Dynamics with Trajectory-HMMs*. Ph.D. Thesis, The University of Edinburgh, Edinburgh, UK.