# Multiple Acoustic Cues for Korean Stops and Automatic Speech Recognition

Weonhee Yun


Thesis submitted for the degree of Doctor of Philosophy

University of Edinburgh

2003

# Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Weonhee Yun

# Acknowledgements

I am deeply grateful to Steve Isard, my supervisor for his encouragement and help in finishing my course in the long run. Without him, I would have never arrived at the final destination of my journey. I thank Bob Ladd and Andrew Breen for their helpful comments. I would also like to thank Simon King, Robert Clark, Korin Richmond, Joe Frankel and all the other members of the Centre for Speech Technology Research for their advice on computing and various subjects. Especially, I would like to express my personal gratitude to Laurence Molloy. As an officemate, sitting by me for a long time, he always gave me a helping hand, whenever I was struggling with mathematical notations and concepts. I am also thankful to Tae-Yeoub Jang, whose work paved the way to pursue my topic. I also owe to his family for helping my wife and kids in various ways while they were here in Edinburgh. My thanks also go to Hyunsung Chung and his family for their enthusiastic welcome whenever my family visited them.

I would also like to thank Heecheol Yoon, Eunmi Park, Heekyung Choi, Hunjin Kim, Kiheung Kim, and many other members of the Edinburgh University Korean Society. Spending time with them has given great comfort to me. Particularly, I would like to express my special gratitude for two Korean couples in Edinburgh, Kihwang Lee and Youngsoo Jung, and Jinsang Hwang and Hayeon Cho. Their warm invitations on various memorable days and events have relieved me from homesickness.

Many thanks go to my friends, Sungwon Hong, Youngsoo Kim, Seungjae Lee, Minyeop Lim, Heungseok Oh, Seungkeun Oh, Ikjoong Youn, and many others, who are proud of having me as a friend. I am also honoured to have them as friends.

I would like to express my heartfelt thanks to Kook Chung, my MA supervisor. His introductory course to phonology a long time ago inspired me to get into linguistics, and later on, he guided me to the field of speech technology.

I cannot thank my wife, Myoungsook enough for understanding me no matter what and for believing in me patiently. Finally, I thank my boys, Sungpil and Sungwoo for continuing to grow fine sons without my presence.

I dedicate my thesis to my parents, whose educational passion and self-sacrificial devotion to their children are the source of my power to overcome difficulties in life.

# Abstract

The purpose of this thesis is to analyse acoustic characteristics of Korean stops by way of multivariate statistical tests, and to apply the results of the analysis in Automatic Speech Recognition (ASR) of Korean. Three acoustic cues that differentiate three types of Korean oral stops are closure duration, Voice Onset Time (VOT) and fundamental frequency (F0) of a vowel after a stop. We review the characteristics of these parameters previously reported in various phonetic studies and test their usefulness for differentiating the three types of stops on two databases, one with controlled contexts, as in other phonetic studies, and the other a continuous speech database designed for ASR. Statistical tests on both databases confirm that the three types of stops can be differentiated by the three acoustic parameters. In order to exploit these parameters for ASR, a context dependent Hidden Markov Model (HMM) based baseline system with a short pause model is built, which results in great improvement of performance compared to other systems. For modelling of the three acoustic parameters, an automatic segmentation technique for closure and VOT is developed. Samples of each acoustic parameter are modelled with univariate and multivariate probability distribution functions. Stop probability from these models is integrated by a post-processing technique. Our results show that integration of stop probability does not make much improvement over the results of a baseline system. However, the results suggest that stop probabilities will be useful in determining the correct hypothesis with a larger lexicon containing more minimal pairs of words that differ by the identity of just one stop.

# Contents

# List of Figures

# List of Tables

# CHAPTER 1

# Introduction

The aim of this study is to establish the characteristics of three acoustic cues for differentiating three types of Korean stops produced at the same place of articulation by way of multivariate statistical analyses, and to exploit them in Automatic Speech Recognition (ASR). Multivariate Analysis of Variance (MANOVA) tests on two databases: one with stops recorded in *carrier sentences* to single out core characteristics of the acoustic cues, and the other with contextual variability constructed for training and testing an ASR system. MANOVA tests show that the three acoustic cues, closure duration, Voice Onset Time (VOT) and fundamental frequency (F0) of a vowel after a stop, maintain their characteristics, as reported in other phonetic studies, in speech with contextual variability and that they work as a whole to differentiate three types of stops. In order to see the usefulness of the acoustic cues in ASR, we first develop a baseline system using context-dependent models with a short pause. After developing an automatic segmentation technique for labelling of closure duration and VOT, we model the three acoustic parameters externally with univariate and multivariate probability density functions (pdfs). By employing a post-processing technique, stop probabilities from the models are multiplied by probabilities of N-best hypotheses. Results after reordering the hypotheses show that the stop probability does not make much impact on the system. However, they do show that the three acoustic cues correct errors when the difference between two competing

hypotheses is a minimal pair of words differing by the identity of a stop. In fact, our database turns out not to contain as many minimal pairs as a more general dictionary with a large vocabulary.

## 1.1 Acoustic Cues of Phonation Types for Korean Stops

Different languages have different phonation types for stops that are produced at the same place of articulation. In English, for example, voiced stops differ from voiceless ones in their phonation types. French also has a two-way distinction for stops. A single acoustic cue responsible for the distinction is known as VOT. It is also been known that different languages use different VOT profiles (Lisker & Abramson 1964).

VOT is defined as the moment at which the voicing starts relative to the release of a closure (Ladefoged 1982:130). Voicing for voiced stops starts before the release of a closure, so VOT has a negative value whereas for voiceless stops, voicing starts after the release, which results in a positive VOT.

Unlike English or other languages that have a two-way distinction, Korean has a three-way distinction in phonation types of stops, as does Thai. Korean stops are often classified as lax, tense and aspirated stops. Lisker & Abramson (1964) attempted to differentiate the three types of stops using a single criterion, VOT, as they did in Thai. However, unlike Thai, Korean exhibits an overlap in VOT between the neighbouring types of stops. That means a single acoustic cue is not enough to differentiate three types of stops in Korean.

As opposed to the single acoustic cue for differentiating the three types of Korean stops, Kim (1965) proposed another criterion, *tensity*, which can be obtained by measuring vowel F0 after a stop. With these two acoustic cues, three types of stops can be differentiated in such a way that VOT distinguishes tense stops from lax and aspirated stops and vowel F0 distinguishes lax stops from tense and aspirated stops.

In addition to these two acoustic cues, closure duration was also suggested as an acoustic cue for differentiating tense stops from lax stops (Han 1996). Perception tests by increasing closure duration of a lax stop and decreasing that of a tense stop revealed that a stop with a long closure duration is perceived as a tense stop and vice versa.

Using multiple acoustic cues, three types of Korean stops can be safely differentiated.

## 1.2 Statistical Verification of Multiple Acoustic Cues

The three acoustic cues were independently reported in phonetic studies and the experiments confirming their status as cues for differentiating the three types of stops were done on speech spoken in *carrier sentences*. The amount of speech data these studies were based on was relatively small (Kim 1965, Han & Weitzman 1965, 1967, 1970, Park *et al.* 1982, Zhi *et al.* 1990, Silva 1992, Han 1996). If it is going to be possible to use these acoustic cues in ASR, they should be valid in speech with contextual variability because an ASR system cannot be confined to just one type of sentence.

Differently from the other phonetic studies, Jang (2000b) used a speech database constructed for an ASR system and revealed that vowel F0 after a stop showed a systematic difference depending on the three types of stops positioned in various contexts. However, closure duration, VOT and vowel F0 after a stop have not been tested all together.

Closure duration and VOT were statistically tested to verify their status as acoustic cues for type differentiation of stops (Lee 1998b), and they were also examined in relation to their positions in prosodic structure (Cho & Keating 2001). Each of these two acoustic cues was viewed as an independent variable. However, considering speech production through the human vocal apparatus, complete independence of productions by speech organs responsible for the acoustic cues is hardly imaginable. Assuming the acoustic cues not as independent variables, but as interacting variables is more appropriate. Given that variables are not independent, a correct statistical approach is a multivariate analysis.

We will attempt to analyse the acoustic cues in three different ways from the previous studies. First, a multivariate approach will be adopted. Second, analyses will be done on speech with contextual variability in order to test the validity of the acoustic cues in ASR. And finally, a large number of samples will be used for more proper statistical tests.

Through the multivariate statistical tests, we will confirm that the three types of stops are substantially different in the measures of closure duration, VOT and vowel F0 after a stop and that none of the three cues can exclusively be used as a single criterion for differentiation of three types of stops in speech with contextual variabilities.

## 1.3 Multiple Acoustic Cues and Speech Recognition

As noted in Rabiner & Juang (1993) and Koreman *et al.* (1997), consonants are harder to recognise automatically than vowels. Stops account for a large proportion of Korean consonants.

Since the three acoustic cues differentiate the three types of stops, exploitation of these cues in an ASR system should result in the improvement of system performance. A possible way of exploiting these acoustic cues in ASR is to let the system automatically learn characteristics of these cues. Since a stochastic approach to ASR using Hidden Markov Models (HMMs) is well established and is very successful, employing the HMM for our purpose looks natural. Jang (2000b) developed a system based on the HMM which successfully exploited F0. However, the other two cues we intend to use are durations and durational modelling by HMMs is known to be incorrect (Levinson 1986). Moreover, getting accurate closure duration and VOT from the signal for training using HMMs is also problematic.

We adopt an alternative way of using the acoustic cues obtaining the three parameters from the signal separately from the HMM and modelling each parameter externally with a parametric pdf. We then combine stop probabilities from the models of the three acoustic

parameters with probabilities of hypotheses from N-best recognition. By rescoring and reordering the N-best hypotheses, our system picks a new first-best hypothesis.

For comparison, we set up a baseline system using context-dependent triphone models with a short pause. This baseline system, before incorporation of stop probabilities, outperforms all other published systems trained on the same database. However, we did not make significant further advances with our final system in which the three acoustic parameters were externally modelled. That is partly because our speech database does not include a sufficient number of minimal pairs of words differing by a single stop. Nevertheless, our system will show the importance of employing acoustic cues for stops when two competing hypotheses contain words from minimal pairs.

## 1.4 Thesis Outline

The outline of this thesis is as follows.

In Chapter 2, we review phonetic studies of the three acoustic cues in relation to the differentiation of three types of Korean stops. We account for the factors that affect the magnitude of the acoustic parameters. The functional load of stops in Korean is also investigated in the literature where stops in a dictionary and text corpora were counted. The investigation shows the quantitative dominance of stops in Korean. Jang's (2000b) phone recognition experiment is reproduced to show stops are problematic in ASR.

In Chapter 3, to establish the status of the three acoustic parameters as acoustic cues, MANOVA tests were done, first on a database which had a lot of words with stops spoken in a carrier sentence, and second on a database which was continuous speech built for ASR experiments. The statistical tests on the second database reveal that acoustic parameters for differentiating three types of stops still play a role in spite of contextual variabilities. The tests will also show that the three acoustic cues work as a whole.

In Chapter 4, a baseline system is described. Various methods to improve the performance of the baseline system are also described. The performance of context-dependent triphone models with a short pause is better than the other systems which were developed on the same speech database (Choi *et al.* 1995, Yun *et al.* 1997, Jang 2000b).

In Chapter 5, we explain how we integrate stop probabilities from the external models of three acoustic parameters into an ASR system by way of a post-processing technique. Automatic segmentation for closure and VOT to obtain better labels for accurate closure durations and VOT will be described. For modelling each acoustic parameter, univariate and multivariate parametric distributions are employed. Detailed calculation of stop probability and reordering procedure is also described. On the basis of our reordered first-best hypotheses, we analyse the results and explain how improvement would be possible, if a lot of minimal pairs of stops were involved in competing hypotheses.

Finally, in conclusion, what we have found in this study will be presented and further work in relation to using three acoustic parameters will be considered.

**Comment**

The investigation of acoustic characteristics of stops in the speech database was joint work with Jang, Tae-yeoub, who was a former Ph. D student in the Centre for Speech Technology Research, the University of Edinburgh. His main interest was in vowel F0 after a stop and I was working on stops' durational properties. Jang and I hand-labelled the database together. After he submitted his thesis, I began investigations combining the two sets of parameters. Since the characteristics of vowel F0 are included in this research, there are some overlaps with Jang's research in various parts of this thesis. In Chapter 2, discussions on 1. segmental F0, 2. vowel F0 in relation to prosodic units, 3. results of phone-level recognition, and 4. functional load of stops in several datasets will be taken from Jang (2000b). In Chapter 3, I will rely on his descriptions of our joint hand-labelling criteria and on normalisation of F0 for speech data with contextual variation. In Chapter 4, his lexicon and language model will be adapted for developing my baseline system.

My own contribution will be all of the discussions on closure duration and VOT, all of the statistical tests presented in Chapter 3, a new baseline system in Chapter 4 and the post-processing ASR system in Chapter 5. When necessary, I will reproduce figures and tables in Jang's thesis with proper citation.

# CHAPTER 2

# Overview of Korean Stops and their Acoustic Cues

## 2.1 Introduction

The main purpose of this chapter is to provide background knowledge for the research by describing characteristics of Korean oral stops and acoustic parameters for the stops. We also provide justification of targeting stop sounds in ASR by presenting a result of a phone recognition experiment done by Jang (2000b), and by examining frequencies of stops in a dictionary and text corpus.

Since we are focusing on three phonation types of stops produced at the same place of articulation, the acoustic parameters we are investigating are cues for differentiating the three types of stops. The differentiation of the three types of stops was attempted in Lisker & Abramson (1964) using VOT but Kim (1965) claimed that a proper differentiation of the three types of stops cannot be made without consideration of vowel F0 after a stop. Closure duration was also proposed as a cue for the differentiation through perception tests (Zhi *et al.* 1990). Characteristics of these parameters will be described in detail.

We explore some of the factors that may affect the characteristics of the acoustic parameters. In particular, a stop's position in the prosodic structure is an important factor that

affects the magnitude of each parameter. Speech rate is also investigated in relation to durational variation.

We show problems of stop recognition in ASR. The result of a phone recognition experiment, which was done by Jang (2000b), shows a low recognition rate for stops. The low recognition rate for stops may cause the overall word recognition rate to be pulled down, if the stops occur frequently. In order to confirm this, the functional load of stops is also examined. It can be measured in terms of frequencies of stops in a dictionary and text corpus. Most of the work was done in Jang (2000b) and in addition to it, we provide an independent source in which the frequencies of stops were investigated (Jin 1992).

In this chapter, there are three main parts. In the first part, description of Korean oral stops and acoustic parameters is given. In the second, factors that may cause the acoustic parameters to vary are described. Finally in the third, some reasons are reviewed for thinking that Korean ASR would especially benefit from improved stop recognition.

## 2.2 Classification of Korean Stops

Phonemically, Korean has 9 oral stops among 19 consonants. Table 2.1 shows all the consonants that are used as phonemes in Korean. One of the conspicuous characteristics of Korean consonants is that oral stops produced at the same place of articulation have three way contrasts in their phonation types[1]. (2.1) gives examples illustrating this three way contrast.

| | | | | | | |
|---|---|---|---|---|---|---|
| (2.1) | pul | 'fire' | p'ul | 'horn' | $p^h$ul | 'grass' |
| | tal | 'moon' | t'al | 'daughter' | $t^h$al | 'mask' |
| | ki | 'energy' | k'i | 'meal' | $k^h$i | 'rudder' |

[1]Previously in Jang (2000b), three types of stops are described as differentiation in the manner of articulation. However, the three types of stops are related to glottal shape and movement, so it seems to be more appropriate to view them as having different phonation types.

| | bilabial | alveolar | palatal | velar | glottal |
|---|---|---|---|---|---|
| Stop | p<br>p'<br>$p^h$ | t<br>t'<br>$t^h$ | | k<br>k'<br>$k^h$ | |
| Affricate | | | c<br>c'<br>$c^h$ | | |
| Fricative | | s<br>s' | | | h |
| Nasal | m | n | | ŋ | |
| Liquid | | l | | | |

Table 2.1: Phoneme inventory of Korean obstruents. (Reproduction from Jang (2000b:10))

For classification of different phonation types of stops, stops in the first row in (2.1) are classified as lax or slightly aspirated stops. Stops in the second one are tense or unaspirated stops, and those in the third are (heavily) aspirated stops. All of the stops in (2.1) are in word-initial syllable onset position. Possible Korean syllable structures are:

(2.2) CV, CVC, VC, V [2]

Korean syllable structure allows only one consonant in onset or coda position phonetically. In the coda position, all stops are neutralised to their homorganic lax stops. The neutralised lax stops are unreleased stops, which are represented as [p˺], [t˺], and [k˺], respectively. However, it should be noted that if a word is composed of more than two syllables, a word-medial coda stop is realised as a released stop by way of a resyllabification rule when the following syllable has no onset consonant. For example, the possible syllable structures of two consecutive syllables in a word are:

(2.3) 1. (C)V.CV(C)

[2] 'V' and 'C' represent 'Vowel' and 'Consonant', respectively. Semivowels and a following vowel are treated as a vowel for the convenience of explanation.

2. (C)VC.V(C)
3. (C)VC.CV(C) [3]

The second syllable structure in (2.3) undergoes resyllabification. It is known as a general tendency in many languages that the consonant in this position is associated with the next syllable as an onset to avoid onsetless syllables (Kenstowicz 1994:280). Thus, the syllable is restructured as illustrated in (2.4). As mentioned above, the result of resyllabification is that if the intervocalic consonant is a stop, then it becomes phonetically a released stop because it is not in coda but in the onset position. The three-way contrast, which is neutralised syllable finally, is manifested in this context.

(2.4) (C)VC.V(C) $\longrightarrow$ (C)V.CV(C)

If the intervocalically positioned stop in (2.4) is lax, it is realised as a voiced stop phonetically, while the other types of stops in this position stay the same. More generally, when lax stops are surrounded by voiced sounds, they become voiced. Furthermore, intervocalic lax stops can also be realised as fricatives and approximants. In particular, bilabial and velar lax stops are optionally produced as [β], and [ɣ], respectively (Huh 1985). In fast or sloppy speech, they are sometimes realised as homorganic approximants(Lee *et al.* 1993, Zhi 1993)[4].

In the third syllable structure in (2.3), if the coda consonant of the first syllable is an obstruent, and at the same time, the onset consonant of the second syllable is a lax stop, then the second stop is tensified[5]. Another phonological rule related to the stops of the same syllable structure as used in the tensification rule is the aspiration rule. If /h/ occurs

[3]'.' represents a syllable boundary.

[4]Lindblom (1983) explains the change from the lax stop to voiced fricative or approximant using the notion, 'Undershoot'. When the speech rate is increased, and the power of articulation as well as the driving force remains the same, then time for articulators to make a proper configuration for a phone is reduced. Because of the shortage of the time, the active articulator does not reach its target, thus it 'undershoots' the proper target position.

[5]Tensification is not always conditioned by phonological restrictions. Arbitrary tensification can be applied in Sino Korean.

as one of a pair of consecutive consonants and the other consonant is one of /p, t, k/, the pair is coalesced to a homorganic aspirated stop.

These phonological rules can be applied across a word boundary when the stop is not positioned either at the right or left edge of a prosodic unit, such as an *Accentual Phrase* (Jun 1993). Since the prosodic structure is related to the scope of a phonological rules, we will discuss the rules with respect to prosodic units more in Section 2.4.

We have seen three way contrasts of stops and allophones of them depending on syllable positions. In the next section, we will explain how the three-way contrast of stops is characterised by acoustic cues.

## 2.3 Characteristics of Acoustic Cues for Three-way Contrast

The general articulatory characteristic of stops is that there is a temporary closure in supraglottal space, which results in a silence that can be measured as closure duration. The delay from the release of the closure to the onset of voicing for the following vowel is manifested as VOT. F0 of the following vowel also varies with the type of stop. We will see how these acoustic cues are associated with phonation types of stops.

### *2.3.1 Voice onset time*

The delay in voicing is produced by laryngeal activity. The vocal folds are abducted for voicelessness, which is a characteristic of all Korean stops, and then, after the release of the blockage in the vocal tract, the vocal folds are adducted for voicing of the following vowel. The delay from the release of the closure to the onset of voicing for the following segment is attributed to the laryngeal muscular movement to set the vocal cords in the right position for voicing.

Lisker & Abramson (1964) measured VOT of stops in syllable onset position cross-linguistically and concluded that VOT is a single underlying variable for differentiating

features such as voicing, aspiration and force of articulation. In their measurements of Korean stops, heavily aspirated stops have the longest VOT whereas tense or unaspirated stops have the shortest, with VOT for slightly aspirated stops in between. Kim (1970) found through the examination of a cineradiographic film of the laryngeal area that the difference in VOT of stops results from different amounts of glottal opening. A similar observation was made by Kagaya (1974). He measured glottal width for each type of stops in CV and VCV position. A schematic view of the relationships between VOT and glottal widths of Korean stops is shown in Figure 2.1. The elliptical shapes represent relative glottis sizes for each type of stop. The top portion of the figure represents the change in distance between the two articulators making the stop closure.

According to Laver (1994:133), a segment can be divided into three phases. In the onset phase, the articulator is moving towards its final target degree of stricture. In the medial phase, the articulator achieves its target. In the offset phase away from the current target, either the articulator moves to a resting position, or toward a new target, in which case we can refer to an overlapping phase.

Figure 2.1, shows a closure during the medial phase, which is indexed as (a), and the closure is released in the overlapping phase, which starts from the second vertical line. After the release, the articulators move away from each other to make a configuration for the following vowel sound.

A possible explanation for the observed correlation between glottal opening and VOT is that the the wider the vocal folds are opened for voicelessness, the greater the distance the vocal folds have to move to before voicing can begin, and hence the longer the delay in voicing onset.

Even though VOT is a primary cue for differentiating phonation types of stops, the VOT range of one type of stop can be overlapped with another. In other words, VOT, as an acoustic cue, can be misleading in cases where a tense stop has longer VOT than a lax stop. In Lisker & Abramson (1964:35), the minimum value of VOT for lax stops is

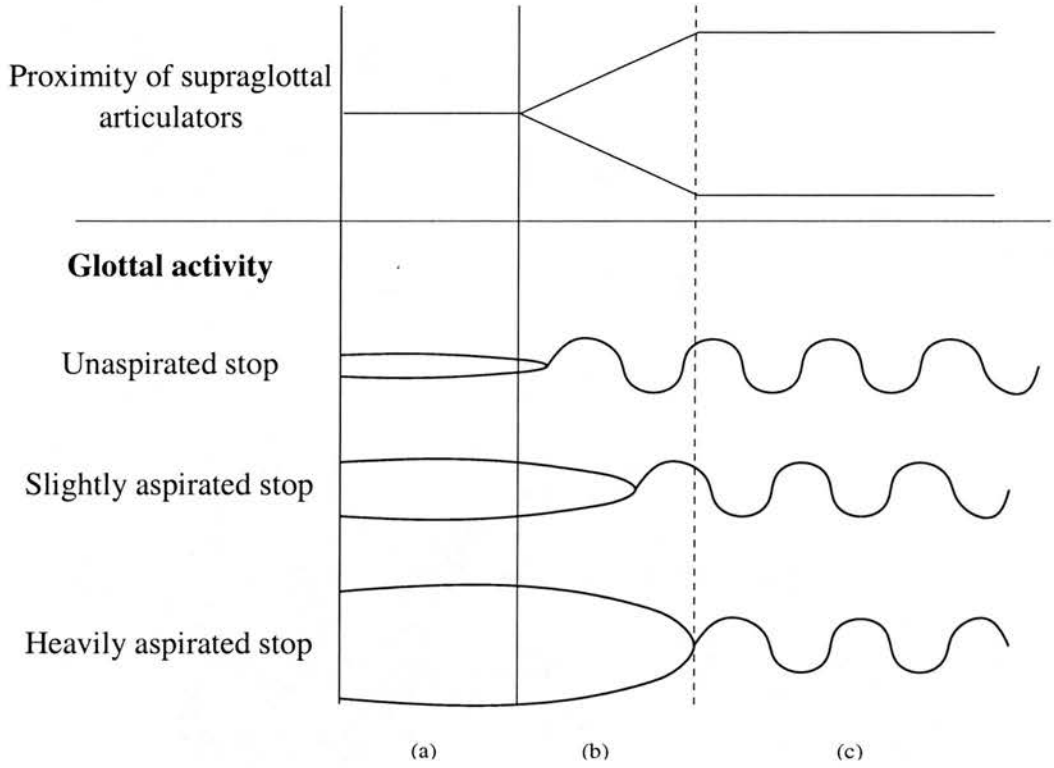



Figure 2.1: Simplified representation of proximity of subglottal articulators, VOTs and glottal widths for Korean stops. (a) is a medial phase of a stop. (b) is an overlapping phase of the stop and a following vowel. (c) represents a medial phase of the vowel. Ellipses represent width of open glottis and wavy lines stand for voicing. Source: Catford (1988) and Kagaya (1974)

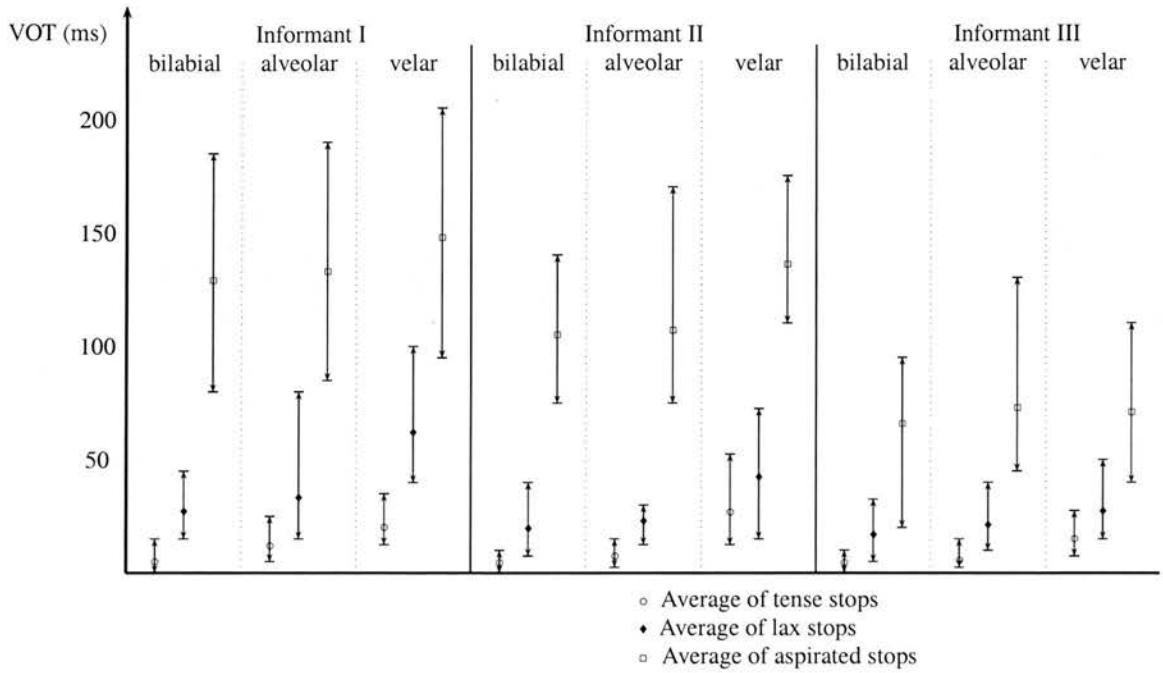



Figure 2.2: VOT ranges of each type of stop for Informant, I, II and III, respectively. Source: Han & Weitzman (1970:115)

less than the maximum VOT for tense stops at all places of articulation. But lax stops and aspirated stops do not have any overlapping in VOT. In Kim (1965:347), the VOT distribution for each type of stop in each place of articulation clearly shows that there is some overlapping VOT between tense stops and lax stops. In case of alveolar stops in the same study, even the lax stops are overlapped with the aspirated stops in VOT. More overlap in VOT between different types of stops can be found in Han & Weitzman (1970) as plotted in Figure 2.2.

Kagaya (1974:166) observed different glottal behaviour for stops in VCV contexts as compared in CV contexts. The glottal width of aspirated stops in VCV is slightly narrower than in the case of CV, which results in shorter VOT in VCV contexts. For lax

stops in CV contexts, the glottis begins to close gradually before the articulatory release, and it is rapidly adducted for the following vowel voicing whereas in VCV, vocal folds are observed to be adducted, which is in in accordance with the observation that the intervocalic lax stops are voiced. For tense stops, not much difference is found in the two contexts. The glottal width starts to be wider near the articulatory release and gradually decreases. Even though tense stops have a wide glottal width before the release of the articulators, the greatest width is still less than the maximum glottal width of aspirated stops.

VOTs in word initial and non-initial positions do not vary much in Lisker & Abramson (1964:50). But in Han & Weitzman (1965:36), aspirated stops in word medial position have one-third or one half of the VOTs in word initial position. VOT depending on prosodic position will be investigated in Section 2.4 more in detail.

### *2.3.2 Segmental F0 effect*

As already mentioned in the above section and shown in the Figure 2.2, VOT alone cannot completely separate three types of stops. Against Lisker & Abramson's (1964) view that the different types of stops can be categorised by VOT, Kim (1965) noted that overlap in VOT among different types of stops can be a problem when VOT is used as a single cue for differentiating the three types of stops. In particular, lax and tense stops produced at every place of articulation have overlaps in VOT. Kim observed that the length of the first glottal pulse for lax stop is longer than for the other types. In other words, vowels after lax stops have lower F0s than those after aspirated or tense stops. He proposed that an independent feature, *tensity* is needed along with voicing to correctly separate Korean stops.

Vowel F0 after a stop is not the only evidence for the feature, *tensity*. Kim related several things to the *tensity* feature: 1. energy distribution during aspiration, 2. duration of increased pressure, 3. amplitude at the beginning of the following vowel, 4. amount of airflow following the release, 5. contact area between the tongue and the roof of the

mouth for the occlusion, 6. lip muscle activity and 7. sound symbolism as a reflection of "native speaker's feeling towards the quality" of the stop sounds (Kim 1965:355-356). The vowel F0 is one of the physical evidence of the feature, *tensity*.

Han & Weitzman (1967, 1970), Hardcastle (1973), Kagaya (1974), and Jun (1996) also reported that vowels after aspirated and tense stops have higher F0 than after lax stops. But between tense and aspirated stops, F0 did not show consistent results. For example as shown in Figure 2.3, the average F0 value after alveolar tense stops for the informant III is higher than the average F0 after aspirated stops. Duration of the initial glottal cycles for the vowels after each type of stops were measured in Hardcastle (1973), which is similar to the method used in Kim (1965). In his experiment, tense stops showed the shortest average duration. Kagaya (1974:169) also showed tense stops shows higher F0 than the other types.

F0 measurement of vowels after stops was done with a considerable amount of data by Jang (2000a). Automatic labelling using forced alignment with the Hidden Markov Model Toolkit (HTK) (Young *et al.* 1996) and the pitch tracking program of ESPS (Entropic 1998) were used to get annotations of each phone region and F0 of vowels, respectively. He found that F0 after aspirated stops is consistently higher than after tense stops[6]. Since the vowel F0 can help in identifying different types of stops, this can be considered as an acoustic cue. Jang called the F0 effect on the identity of the preceding stop the *segmental F0 effect*. The characteristic of vowel F0 after a stop that we can draw from these studies is that tense and aspirated stops induce higher vowel F0 than lax stops.

### *2.3.3 Closure duration*

Lisker's (1957) experiments revealed that the voiced or voiceless stop distinction in intervocalic position is made by closure duration in English. Voiced stops have shorter closure duration than voiceless stops. Moreover, closure duration was reported to be affected by

[6]Affricates in Korean also have a three-way contrast and Jang (2000a) also found that they have a similar characteristic of vowel F0 to that of stops. That is, lax affricates have lower F0 whereas the other two types of affricates have higher F0.

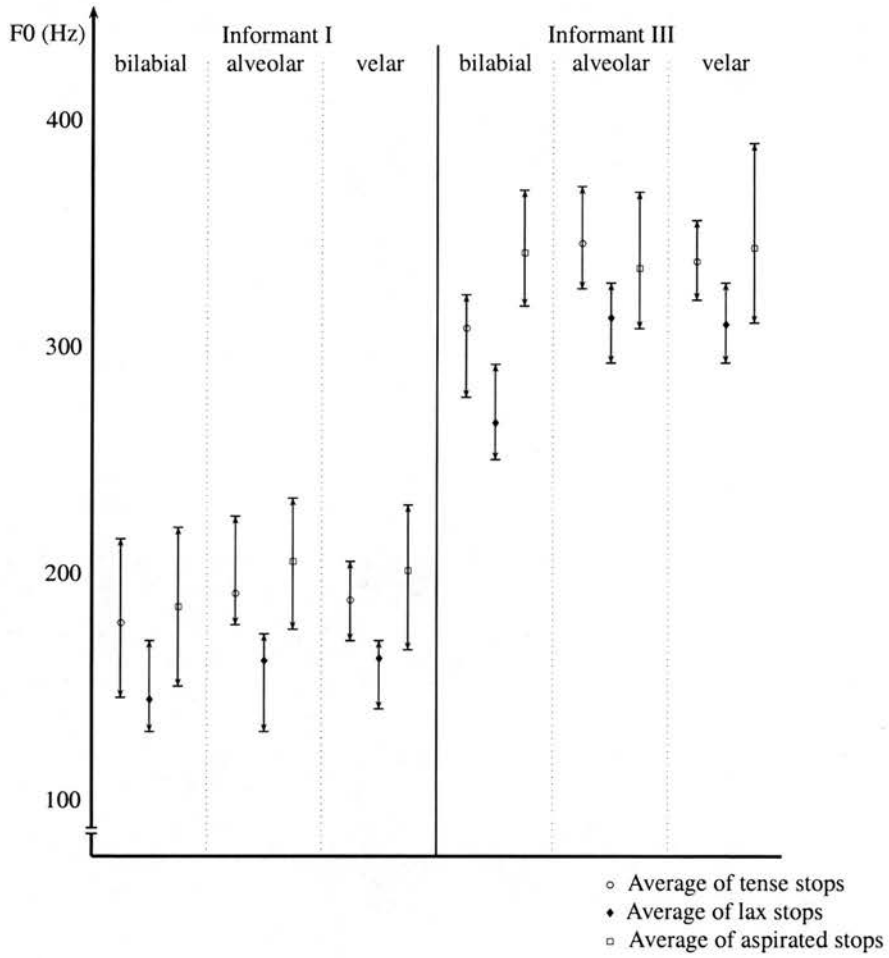



Figure 2.3: F0 ranges of each type of stop for Informant, I and III, respectively. Source: Han & Weitzman (1970:117)

| Prosody | Phrase Edge | | Word Edge | | Word Internal | |
|---|---|---|---|---|---|---|
| Preceding | Vowel | Nasal | Vowel | Nasal | Vowel | Nasal |
| p | 66 | 54 | 50 | 32 | 48 | 25 |
| $p^h$ | 85 | 73 | 77 | 53 | 84 | 57 |
| p' | 104 | 98 | 81 | 60 | 123 | 83 |
| Mean | 85 | 75 | 69 | 48 | 85 | 56 |

Table 2.2: Closure durations of bilabial stops in different prosodic contexts. When the preceding phone is a nasal, closure durations are shortened compared to those with a preceding vowel. Source: Silva (1992:120)

place of articulation in Repp (1984). Closure duration tends to be longer for labial stops than the stops in the other places of articulation.

Closure duration can also be used to distinguish phonation types of Korean stops. In early work by Park *et al.* (1982), closure durations were simply measured and illustrated. On the basis of their measurements, we can draw a tentative conclusion that there is a tendency for closure durations of tense and aspirated stops to be longer than those of lax stops.

Silva (1992) recorded stops with different preceding phone contexts in three prosodic positions, namely, phrase-edge, word-edge and word-internal. He measured closure durations of the three types of bilabial stops and observed that lax stops had the shortest closure durations, tense stops had the longest, and aspirated stops fell between the two other types. An Analysis of Variance (ANOVA) test showed that closure durations of each type of stop were statistically different from those of the other types.

Silva also found, as seen in Table 2.2, that preceding phone context affects closure durations of each type. Further investigation of the effect of preceding phone contexts on the acoustic characteristics of stops was done by Pae *et al.* (1999). As in Silva (1992),

vowels, nasals, and liquids were used as preceding phone contexts[7]. After nasals and liquids, closure durations are much shorter than after vowels, which is similar to Silva's (1992) results.

Perception tests have also been done to show that closure duration is an acoustic cue for differentiating tense and lax stops in intervocalic positions (Zhi *et al.* 1990, Han 1996). Zhi *et al.* (1990) investigated how perception of intervocalically positioned alveolar tense and lax stops is changed depending on voicing intensity and a duration ratio defined as:

(2.5)

$$\text{duration ratio} = \frac{\text{duration of preceding vowel}}{(\text{duration of preceding vowel} + \text{closure duration})}$$

They performed manipulations on both lax and tense stops. Because intervocalic lax stops are realised as voiced, there was voicing in the lax stop closure. First, the voicing amplitude of the lax closure period was reduced to zero in four steps. Then, each amplitude manipulated item was used for the further manipulation of duration ratio. Ten stimuli were prepared by repeatedly deleting a single period of the preceding vowel and inserting the same amount of closure duration. For the tense stop, the duration ratio was increased in eight steps. Since tense stops are always voiceless, regardless of phonological environment, no amplitude manipulation was necessary.

Zhi *et al.* found that decreasing the duration ratio of voiced stops made them more likely to be perceived as tense, and increasing the ratio for tense stops made them more likely to be perceived as lax. Voicing amplitude also had an effect. For the stimuli with the same duration ratio, those with lower amplitude were more likely to be perceived as tense.

This perception test does not confirm that closure duration on its own is a perceptual cue for tense stops because the preceding vowel duration was varied along with the closure durations. Considering the variability of vowel duration in fast or slow speech, using the

[7]One of the preceding phone contexts, liquids were recorded in Silva (1992) but the data was not used in the study.(Silva 1992:161)

preceding vowel duration as a variable of the perceptual cue does not look to be ideal for ASR .

A similar perception test was done in Han (1996). But in her experiment, vowel duration was not varied. Only the closure duration was manipulated for the test. Intervocalic lax stops with stepwise lengthening of closure durations by ten milliseconds (ms) each time were prepared to see whether lax stops with lengthened closure duration would be perceived as tense stops. Also, closure durations of tense stops were reduced in ten ms steps. The results showed that closure duration on its own had an influence on the perception of the two types of stops. The 50% crossover point for original lax stops to be perceived as tense stops was located between 85 and 90 ms of closure duration.

Closure duration has a drawback as a cue for differentiation of phonation types of stops in that it is not available when a stop is positioned sentence initially or after a pause. For the case of pauses, the boundary between the pause and the stop closure cannot be determined. Since pauses are believed to be indicators of prosodic structure, we will review the prosodic structure of Korean and studies of its effect on closure duration and the other two acoustic cues we have seen already in previous sections.

## 2.4 Acoustic Characteristics Depending on Prosody

In this section, the characteristics of acoustic cues that differentiate the three types of Korean stops will be reviewed in relation to the prosodic structure proposed by Jun (1993, 1998) and other suprasegmental factors.

As stated at the end of 2.3.1, the VOT for each type of stop is different depending on the position of the stop in a word. Word-initial and Word-medial position were generally used in measuring VOT and closure duration. These positions can be redefined as positions in the prosodic structure and systematic durational difference can be found due to the positions.

Prosodic structure alters the segmental F0 effect but it does not obliterate the segmental F0 effect. Jang (2000b) has already reviewed the relationship between prosodic units, which are constituents of prosodic structure, and the segmental F0 effect and concluded that prosodic influences manifest themselves at locations where the segmental F0 effect is most valuable (Jang 2000b:27).

There are other factors, such as intrinsic vowel F0, lexical stress, focus, declination, and speech rate that may affect the characteristics of acoustic cues for Korean stop differentiation. Except for speech rate, these other factors have been dealt with by Jang (2000b) and will be summarised on the basis of his review. In addition, speech rate as a factor for varying stop durations and for influencing the placement of prosodic boundaries will also be discussed.

### *2.4.1 Prosodic structure of Korean*

Prosodic structure is a hierarchical organisation of prosodic units. According to Nespor & Vogel (1986), a language universal rule, called *Prosodic Constituent Construction*, is applied to a process of grouping prosodic categories and the process obeys the *Strict Layer Hypothesis* (SLH), which is defined as:

> 1. A given nonterminal unit of the prosodic hierarchy, $X^p$, is composed of one of more units of the immediately lower category $X^{p-1}$.
> 2. A unit of a given level of the hierarchy is exhaustively contained in the superordinate unit of which it is a part (Nespor & Vogel 1986:7).

As Shattuck-Hufnagel & Turk (1996:207) put it, a prosodic category at one level is exhaustively parsed into constituents of the next-lower level, and those next-lower-level constituents are all of the same type.

According to Jun (1993), Korean is organised with prosodic units as follows:

(2.6) Utterance
Intonational Phrase
Accentual Phrase
Phonological/Prosodic Word
Syllable

For example, an *Utterance* is composed of one or more *Intonational Phrases* (IP) and an IP is also composed of one or more immediately lower units, which are *Accentual Phrases* (AP). The same relationship applies down to the *Syllable* level.

As Jun (1998) explained, her prosodic structure is different from that of Selkirk (1984) and Nespor & Vogel (1986), in that Jun's prosodic units above the Word level, such as AP and IP, were defined on the basis of variation of F0 contour and lengthening of vowels, whereas the prosodic structure of Selkirk and Nespor & Vogel was defined by indirect derivation of the prosodic units from syntactic structure. An advantage of intonationally motivated APs and IPs over syntactically derived prosodic units is that phonological rules such as *Post-obstruent Tensing*, *Vowel Shortening* and *Lenis Stop Voicing*, can be better described with the former than with the latter. Jun (1993:72) reported experimental findings that *Post-obstruent Tensing*, for example, was applied simply within the AP, while the same phonological rule had different domains in a syntactic approach, in which the domain was either within the Phonological Word by Cho (1987), or within the Phonological Phrase by Kang (1992), which is similar to the AP level.

Another potential advantage of Jun's prosodic structure is that we might attempt to detect the prosodic boundaries from the signal in an ASR system[8], by way of looking at the variation of F0. Furthermore, the detection of prosodic boundaries would give information about the prosodic positions of stop sounds, which will help in differentiating stops since stop durations differ depending on their prosodic positions. This would be much more

[8] An attempt has been made by Lee & Song (2000) for Korean.

difficult with a characterisation of syntactically derived prosodic units because syntax is not directly represented in the speech signal.

For each prosodic unit, we will list the prosodic rules which take that unit as their domain and explain the associated acoustic effect. Under the assumption of Jun's prosodic structure, Cho & Keating (2001) observed a *domain-initial strengthening effect*, in which initial stops of higher prosodic units are "articulatorily stronger" than those of lower ones, where strong articulation of stops is characterised by longer closure duration and VOT. No effects on F0 other than those associated with the tonal pattern are assigned to the prosodic unit. Since we already reviewed the allophonic effects of syllable position in 2.2, we will directly start with the Prosodic Word.

**Phonological/Prosodic Word**

A Phonological Word, or Prosodic Word (PW), is a constituent of a higher prosodic unit, the AP. Since the prosodic hierarchy follows the SLH, the PW is the same as the smallest AP. Jun (1993:225) defined the PW as "the minimal sequence of segments which can be produced as one AP" and carried on to state "the minimal sequence of segments produceable as an AP is a stem and its affixes (prefixes, inflectional or derivational affixes, postpositions and clitics)." An example is illustrated below.

(2.7) *[na-nin]$_{PW}$ [nə-lil]$_{PW}$ [coa-hæ]$_{PW}$*
I-TOP you-ACC like-END
'I like you'[9]

TOP: Topic marker
ACC: Accusative case marker
END: Verbal ending

Jun (1993) found that the PW is a domain of VOT lenition. VOTs of aspirated stops in PW medial position were observed to be shorter than the ones in PW initial position.

[9]This is the same example as Jang (2000b:24) used as an example of Korean accentual phrasing where he labelled each angled bracket as an AP. Since it is the smallest AP, it is actually the same as a PW.

| Number of syllables | Tonal patterns |
|---|---|
| 1-2 | XH |
| 3 | XH, XLH, XHH |
| 4 or more | XHLH |
| | (where X is L or H) |

Table 2.3: Types of phonetically realised AP from Jun (1998). The first tone (shown as X) is determined depending upon the type of onset consonant in the first syllable of AP (ie., segmental F0 effect). (Reproduction from Jang (2000b:24))

In Jun's prosodic structure, APs and IPs are defined by F0 contour but PWs are not, as mentioned in her definition of the PW. So, no tonal pattern is associated with the PW level, and thus, no factor that may affect the segmental F0 perturbation is expected.

For closure duration, Cho & Keating (2001) found that aspirated and tense alveolar stops in PW medial position have longer closure durations than those in PW initial position. They reported PW medial closure durations are longer by no more than 20% than PW initial ones. A similar result of PW medial closure duration lengthening for aspirated and tense bilabial stops can be found in Silva's (1992) experiment, which was represented in Table 2.2 on page 19. Word-edge and Word-internal positions correspond to PW initial and PW medial positions, respectively.

**Accentual Phrase**

As stated above, the AP along with the IP, is defined by its intonational pattern. The AP is also the domain of some phonological rules, such as *Post-obstruent Tensing* and *Lenis Stop Voicing*. *Post-obstruent Tensing* is a rule by which a lax stop after an obstruent becomes tensed unless the sequence is divided by an AP demarcation. A lax stop that is not preceded by an obstruent is subject to the *Lenis Stop Voicing* rule, which turns a lax stop into its voiced counterpart.

The tonal patterns associated with APs are well summarised by Jang (2000b:24) and reproduced here in Table 2.3. Jun (1998) states that the underlying tonal pattern of the AP in Seoul Korean is Low-High-Low-High (LHLH) or High-High-Low-High (HHLH). The tone of the first syllable is determined by the first segment of the syllable. When it starts with either an aspirated or a tense stop, the tone of the syllable is H. Otherwise, it gets L. When an AP has more than four syllables, the syllables between the third and the antepenult of the AP get their surface tones by interpolation between the H tone on the second syllable and L tone on the penult. The tonal patterns of APs which have fewer than four syllables are represented in Table 2.3.

Since the tonal pattern of an AP's first syllable is, in fact, an expression of the segmental F0 effect, the internal tonal pattern does not compete with the segmental F0 effect in this position. However, Lee (1998a) found that the segmental F0 effect was diminished after AP initial position. Jang (2000b:85) also observed a similar result in our stop-rich database[10]. He found that the ranges of F0 after aspirated and tense stops in the non-AP initial position were remarkably similar and did not show distinctions among three types of stops.

Jun's (1993) experiment conducted for VOT lenition showed that VOTs of aspirated stops in AP initial position are longer than those in AP medial position. AP medial position can be either a PW initial or a PW medial position. The VOT difference between PW initial and PW medial positions is even greater when the stop is AP initial as well as PW initial.

For closure duration, Cho & Keating (2001) showed that aspirated and tense stops in AP initial position have longer closure durations than those in PW initial position. As already noted, closure durations in PW medial position are longer than those in PW initial position but it is not clear whether closure durations for these stops in AP initial position are still longer than in PW medial position. Without comparing closure durations in these

[10]The description of the database will follow in Section 3.2 on page 54.

two positions, Cho & Keating (2001) suggested that the failure to observe a hierarchical lengthening effect might be explained by the PW medial tense stops being geminates.

**Intonational Phrase**

The IP is the prosodic unit immediately above the AP. The intonational contour for the IP is defined as having tonal patterns of one or more APs and an IP boundary tone which indicates the pragmatic meaning of the phrase. The IP boundary tone is associated with the final syllable of the IP. The IP boundary tone preempts the tonal pattern of the AP final syllable. Apart from the IP boundary tone in the phrase final syllable, the duration of the syllable is also lengthened (Jun 1993, 1998).

Jun claimed that the IP is a phonological rule domain for *Obstruent Nasalisation* and *Spirantisation*[11]. Both rules are related to coda stops, which are not released and never realised as onset stops after the application of these rules. Another characteristic of the IP is that it is often preceded by a pause, which can affect closure durations of stops in the IP initial but not in the Utterance initial position.

A syllable at the left edge of an IP is not associated with any special tonal pattern, while at the right edge there is an IP boundary tone. So, the segmental F0 effect on the syllable at the left edge, which is also AP initial position, does not interact with the IP boundary tone. Thus, the segmental F0 effect can be expected to be seen in this position.

Cho & Keating (2001) reported that IP initial lax and aspirated alveolar stops had longer VOTs than those of lower prosodic units. They did not find any systematic differences for the tense alveolar stop. They also reported that closure duration in IP initial position is longer than in the lower prosodic units.

Since the IP is usually preceded by a pause, it is impossible to separate the pause from the closure duration of an IP initial stop. Considering that a pause is not detected between

[11] */s/-Palatalisation* is also a rule that has an IP as a rule domain but it does not apply to stops. The rule says /s/ becomes [ʃ] before /y, i/.

the two APs, longer closure duration for a stop in IP initial position can be expected. The measurement of IP initial, but not Utterance initial, closure durations by Cho and Keating was meant to be acoustic silence between the two segments across IPs. So the pause must have been included in the closure duration. Even if the pause is a part of closure duration of their measurement, the result seems to be in accordance with the characteristics we have seen in 2.3.3. The result was also supported by "seal duration", which was obtained by the electropalatography[12]. Cho and Keating found that the seal duration of tense stops was longest whereas that of lax stops was the shortest, which is analogous to what measurements of closure duration show in other positions. Thus, it is reasonable to extend the findings for closure durations to IP initial positions.

**Utterance**

The Utterance is the largest prosodic unit in the prosodic hierarchy. Nespor & Vogel (1986:222) defined a domain of the Utterance as consisting of all the IPs corresponding to $X^n$ in the syntactic tree. The Utterance is also the largest phonological rule domain. For example in American English, the flapping rule has the Utterance as its rule domain. As far as Korean is concerned, no phonological rule that applies in an Utterance domain has been found.

The Utterance does not have any tonal pattern specific to it. So, the segmental F0 effect does not interact with Utterance position. Cho & Keating (2001) reported that VOTs and seal durations were longer in Utterance initial position than initially in lower prosodic units, but they did not measure closure duration for the same reason they did not measure closure durations for IP.

[12]The electropalatography is a recording of activated electrodes in the pseudo-palate. There are ninety six electrodes in total. Two measurements was made from the electropalatography by Cho & Keating (2001). One was the linguopalatal contact and the other was the seal duration. They defined the seal duration as "the duration between the first and the last frames in which the oral cavity was completely sealed."

### 2.4.2 *Interaction of prosodic structure and acoustic cues*

The *domain-initial strengthening effect* has two aspects. First, phones, including stops, that are prosodic domain initial are articulatorily stronger than those in domain medial positions. What is meant by strong articulation is to have more linguopalatal contact, as measured by electropalatography. Acoustically, stronger articulation gives rise to lengthening of phone duration. Second, the strengthening effect is greater for higher prosodic domains. In other words, the higher the prosodic domain is, the more linguopalatal contact is observed (Fougeron & Keating 1997).

Cho & Keating's (2001) work clearly shows that there is a *domain-initial strengthening effect* for Korean stops. The acoustic outcome of the articulatory strengthening of stop sounds is a greater magnitude of VOT and closure duration. The hierarchical relationship between the lengthening and prosodic units is also found except for PW initial position. In fact, closure duration of PW medial position is greater than that of PW initial position.

In (2.8), the hierarchical relationship is graphically represented and the way acoustic characteristics of Korean stops depend on the prosodic hierarchy is summarised on the basis of the work by Cho & Keating (2001).

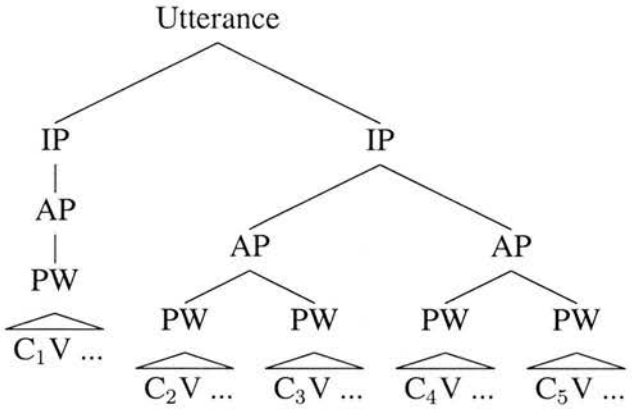
Cho & Keating (2001:176) show that there is a strong correlation between linguopalatal contact[13] and both VOT and closure duration. They also found strong correlation between linguopalatal contact and seal duration, which can be regarded as articulatory representation of closure duration.

The *domain-initial strengthening effect* will be investigated in our databases[14]. We will show the effect can be found in one of the databases. This will be discussed more later in Chapter 3.

[13]Each data frame from the electropalatography represents how much contact is made between the tongue and the palate. The value for the linguopalatal contact is a percentage of the contact in a frame that shows the peak contact.

[14]The databases will be described in Section 3.2 on page 54.

(2.8)




VOT

| | | | | | |
|---|---|---|---|---|---|
| Aspirated stops: | $C_1$ | > | $C_2$ | > | $C_4$ > $C_3$ (or $C_5$) |
| Lax stops: | $C_1$ | > | $C_2$ | > | $C_4$ > $C_3$ (or $C_5$) |
| Tense stops: | No systematic difference | | | | |

Closure duration

| | | | | | |
|---|---|---|---|---|---|
| Aspirated stops: | | | $C_2$ | > | $C_4$ > $C_3$ (or $C_5$) |
| Lax stops: | | | $C_2$ | > | $C_4$ > $C_3$ (or $C_5$) |
| Tense stops: | | | $C_2$ | > | $C_4$ > $C_3$ (or $C_5$) |

### *2.4.3 Other factors: intrinsic vowel F0, stress, focus, declination and speech rate*

Apart from prosodic structure, there are some other factors that may affect the acoustic characteristics of stops. In Jang (2000b), four factors, intrinsic vowel F0, stress, focus, and declination have been examined in relation to their influence on the segmental F0 effect. These factors will be briefly summarised here and speech rate will be also examined in relation to stops' durations and placement of prosodic boundaries.

*Intrinsic vowel F0*

One of the universal characteristics of vowel quality is that vowels have different average of F0 (Whalen & Levitt 1995). Ohala (1978:29) states that the average F0 of vowels is systematically related to vowel height. In other words, high vowels like [i], [u] have a higher average F0 while low vowels like [a], [æ] have a lower one.

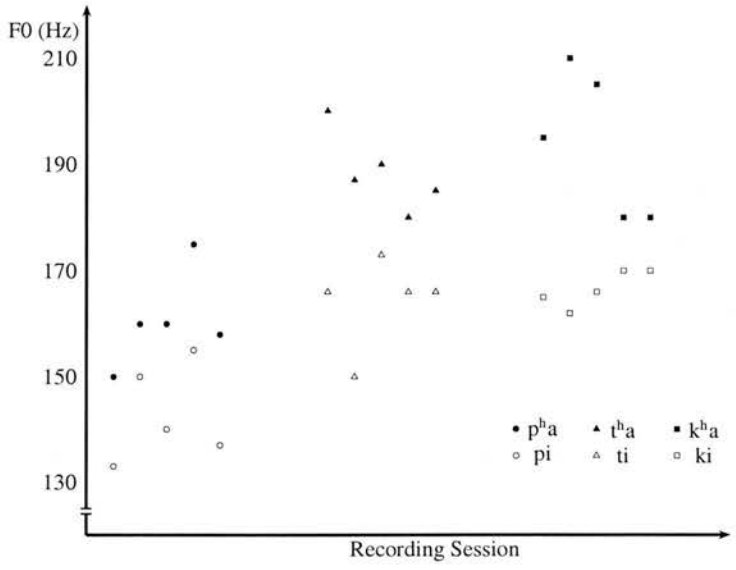



Figure 2.4: Comparison of interaction between intrinsic F0 effect and segmental F0 effect of a single speaker. Jang took the F0 values from Han & Weitzman (1967:16). Each pair of vertically aligned tokens were uttered in the same recording session. It is shown that, in each session, the F0 value of the syllable composed of [aspirated stop + low vowel] is always greater than that of the syllable composed of [lax stop + high vowel], which means that the intrinsic vowel F0 effect never completely preempts the segmental F0 effect. (Reproduction from Jang (2000b:27))

For Korean Han & Weitzman (1967) measured the F0 of vowels after various stops. The average F0s of vowels of different heights show that regardless of the type of preceding stop, higher vowels have higher F0 and lower vowels have lower F0. Jang (2000b:28) considered the possibility that the intrinsic vowel F0 effect could interfere with the segmental F0 effect. Since vowel F0 after an aspirated stop is generally higher than after a lax stop, vowel F0 after an aspirated stop followed by a lower vowel might be lower than after a lax stop followed by a high vowel due to the intrinsic vowel F0. However, he showed that the intrinsic vowel F0 never completely cancelled out the segmental F0 effect in the data used in Han & Weitzman (1967) by plotting each F0 value of a single utterance token in a graph which is reproduced in Figure 2.4. Jang also mentioned that the data in Han and Weitzman's survey was not good enough for estimating the average

size of the difference between high and low vowels because the range of the variation between speakers was too large (Jang 2000b:89).

To see whether the intrinsic F0 effect exists in other data, he used two databases, one of which was recorded with carrier sentences and the other of which was the KAIST continuous speech database[15] and concluded that the intrinsic F0 effect is consistently observed[16]. Nevertheless, the intrinsic F0 effect does not eliminate the segmental F0 effect. Put differently, average F0s of vowels after aspirated and tense stops are always higher than after lax stops regardless of vowel height.

*Lexical stress*

Increasing F0 is a way of realising prominence in some languages but in Korean syllable prominence is mainly achieved by lengthening the stressed vowel duration (Lee 1973) and increasing the amplitude of the syllable (Huh 1985). In brief, stress in Korean is completely predictable from syllable structure and its realisation is generally related to duration and intensity, not pitch. Thus, the segmental F0 effect is not affected by lexical stress in Korean.

*Focus*

The effect of focus in Korean is realised by restructuring the prosodic organisation of the utterance. If a word is focused in an utterance, a new AP starts with that word (Jun 1993).

Since AP initial is the position where the segmental F0 effect is well distinguished, focus does not inhibit the segmental F0 effect. Rather it reinforces the effect.

Apart from the structural change in prosodic organisation, focus brings about a change in vowel duration. Oh (1999) reported that vowel duration of the first syllable of a focused

[15]The description of the database will follow in Section 3.2.

[16]The intrinsic F0 effect failed to be observed in one case. In controlled speech data vowels after aspirated stops in non AP initial position did not show a statistically significant difference between high and low vowels (Jang 2000b:90).

word was significantly longer than that of a non-focused word. There has been no study on the durational change of VOT or closure duration of stops with respect to focus. However, since the main effect of a change in vowel duration is speech rate, we assume that durational change of VOT and closure duration by focus is subsumed by effect of speech rate on VOT and closure duration.

*Declination*

Declination is a process in which F0 tends to become progressively lower through the course of an utterance (Laver 1994:155). The possible result of declination would be a large variation of F0 values for vowels after stops within an utterance. For example, an aspirated stop at the beginning of an utterance would have higher F0 than the same type of stop at the end, even though the general tendency that aspirated stops have higher F0 than lax stops would not be violated. Considering long utterances in continuous speech, a more appropriate description of the segmental F0 effect is that aspirated stops at a given position of an utterance have higher F0 than lax stops positioned at a similar distance from the start of an utterance.

*Speech rate*

Two things will be considered in this section concerning speech rate in Korean. First, as Jun (1993) pointed out, speech rate changes the internal organisation of prosodic structure of an utterance and second, speech rate changes segmental duration (Zhi 1993).

Jun (1993) found that the number of voiced stops was increased as speech rate got faster. Since the rule of *Lenis Stop Voicing* applies to non-AP initial stops, she claimed that more voiced stops represented a smaller number of APs. (2.9) gives a phonemic representation of an utterance and (2.10) shows various organisations of APs, each of which can have a different phonetic representation.

| | | | | | |
|---|---|---|---|---|---|
| (2.9) | *na-nin* | *kinyə-ka* | *kili-n* | *kilim-il* | *po-nta* |
| | I-TOP | she-NOM | draw-REL | picture-ACC | see-END |

'I see a picture that she drew'

TOP: Topic marker
NOM: Nominative case marker
REL: Relativiser
ACC: Accusative case marker
END: Declarative ending

If each PW becomes an AP in this utterance, there are two stops that have undergone the *Lenis Stop Voicing* as in (2.10 *a*). When the number of APs is decreased from five to two, the number of voiced stop is increased from two to five as in (2.10 *d*).

(2.10) *a.* [na-nɨn]$_{AP}$ [kɨnyə-ga]$_{AP}$ [kɨli-n]$_{AP}$ [kɨlim-ɨl]$_{AP}$ [po-nda]$_{AP}$

*b.* [na-nɨn]$_{AP}$ [kɨnyə-ga gili-n]$_{AP}$ [kɨlim-ɨl]$_{AP}$ [po-nda]$_{AP}$

*c.* [na-nɨn]$_{AP}$ [kɨnyə-ga gili-n gilim-ɨl]$_{AP}$ [po-nda]$_{AP}$

*d.* [na-nɨn]$_{AP}$ [kɨnyə-ga gili-n gilim-ɨl bo-nda]$_{AP}$

*e.* [na-nɨn ginyə-ga gili-n gilim-ɨl bo-nda]$_{AP}$

When a phonemically lax stop is realised as phonetically voiced, the acoustic cues we are in pursuit of are not going to be available in the speech signal. Even though the F0 of vowels after voiced stops stays relatively low in comparison to the other two types, closure duration and VOT of voiced stops will not have the same characteristics as those of lax stops. Thus, voiced stops are not going to be target phones in this research, especially for building the ASR system.

Such rephrasing of the AP structure also results in a reorganisation of the IP structure. However, rules that have the IP as their rule domain, such as *Obstruent Nasalisation* and *Spirantisation* do not have phonetically realised stops as their output. In other words, speech rate also restructures the IP of the utterance but the phonetically realised stops are not affected by it.

While there have been some studies on the relationship between vowel and syllable duration for Korean, it is hard to find a phonetic study on the effect of speech rate on VOT and closure duration in Korean. An increase in speech rate can be expected to decrease VOT and closure duration but detailed studies for Korean stops have not been done yet. Only general information on how speech rate affects durations of lax stops is available.

Zhi (1993) recorded a phrase with three speech rates, slow, normal, and fast. He observed a 37% reduction of the total duration between normal and fast readings and a 53% increase between normal and slow. His phrase included two velar lax stops. One of the stops was underlyingly an onset stop and the other one was underlyingly in a coda but was realised as an onset through resyllabification. The duration of the former in normal speech was 30 ms. The stop had the same duration in fast speech but in slow speech it was 95 ms. The latter was 50 ms in normal speech and 65 ms in slow, and 25 ms in fast speech, respectively.

In Zhi's study, there seems to be a tendency for consonant duration to be less reduced than vowel duration[17]. Because his study did not include the variation of the other types of stops, and because it was not a statistical study, it is of limited help in understanding how speech rate affects stop durations. However, we can make predictions for Korean from studies done with data in English and some other languages.

Amerman & Parnell (1981) investigated the effect of speech rate on stop consonant perception in English. They showed that mean duration of VOT for voiceless stops in English was reduced by 41% in fast speech and was increased by 17% in slow speech, compared to the VOT of a moderate rate. Their perception test showed that even though the speech rate affected stop durations, stops spoken at the three different rates were perceived equally correctly in their original contexts and when excised and inserted in normal rate carrier sentences.

[17]It is hard to judge whether the lax stop in slow speech is actually an unvoiced lax stop or a voiced allophone just by examining spectrograms he presented, but it appears to be a voiced stop in normal and fast speech.

Kessinger & Blumstein (1997) observed changes in VOT depending on speech rate in three languages, Thai, French, and English. They classified VOT in each language into three categories, pre-voiced, short lag and long lag. Thai is similar to Korean in that it also has three types of stops, voiced, voiceless unaspirated and voiceless aspirated stops. So, Thai has all three categories whereas French has pre-voiced and short lag, and English has short lag and long lag categories.

Kessinger & Blumstein (1997) found that speech rate shifted distributions of VOT for pre-voiced and long lag stops in the direction of reducing VOT in fast speech but they did not find noticeable change of VOT for short lag stops. In Thai, they observed an overlap between short lag and long lag stops in fast speech but no overlap was found between pre-voiced and short lag stops. French and English do not have a three-way distinction in stops but they showed a similar result to Thai. In French, pre-voiced stops were much more affected in fast speech than short lag stops, and in English, changes of VOT for long lag stops were greater than for short lag stops. They concluded that speech rate affected the VOT in such a way that the phonemic distinctions on the continuum of VOT were still preserved (Kessinger & Blumstein 1997:165).

In Korean, VOT distinguishes the tense stop from the other two types. The order of VOT magnitude is tense, lax and then aspirated stops. However, we have already seen that unlike in Thai, VOT does not completely separate these three types. Nevertheless, tense stops seem to correspond to the short lag category of Kessinger & Blumstein (1997), because they have the shortest VOT among the three types. Some other reasons as well as the shortest VOT make us believe that the tense stops may be resistant to the effect of speech rate, as the short lag category is in Thai, French and English.

First, Han & Weitzman's (1970) measurements of VOT, which were illustrated in Figure 2.2, show that the range of VOT for tense stops is narrower than that for lax or aspirated stops. The data they used were recorded at a single speech rate, which was normal speed. Even at a single speech rate, VOT of lax and aspirated stops shows considerable variation. However, the data for tense stops support Cho & Keating's (2001) observation that "VOT

[for tense stops] is always quite short and there is little variation." Han and Weitzman's study controlled the other linguistic factors that might affect VOT. Even in this controlled environment, the other two types of stops have much more variation in themselves than the tense stop. So, speech rate might have more influence on lax and aspirated stops than on tense stops.

Second, as noted in (2.8), VOT for tense stops is not much affected by the *domain-initial strengthening effect*, which means that VOT for tense stops is not subject to one form of systematic variation that affects lax and aspirated stops. Speech rate might also be related to strong or weak articulation by way of *undershoot*[18] In other words, a speaker when speaking fast cannot make a strong articulation due to time constraints. However, whether a tense stop is positioned in articulatorily strong or weak position, its VOT does not vary much. It seems likely that VOT for tense stops will not be affected much by the speech rate factor, either.

Third, vowel duration after aspirated stops in Cho & Keating's (2001) experiment shows a similar pattern to the one in Kessinger & Blumstein (1998). On the basis of the fact that syllable duration is increased as the speech rate gets slower, and that VOT for voiced stops in English has less variation than that for voiceless stops, as shown in their previous work (Kessinger & Blumstein 1997) and that of others (Miller *et al.* 1986, Volatis & Miller 1992), Kessinger and Blumstein investigated how voiceless stops took part in the increase of the syllable duration. What they found was that both VOT of voiceless stops and vowels participated in the increase of the syllable duration, and that VOT and vowel were almost equally increased, maintaining C/V ratio of the syllable across speech rates.

Cho & Keating (2001) observed that in Korean, vowel duration after aspirated stops in higher prosodic domain initial position had a tendency to get longer[19]. The higher the prosodic position, the longer the vowel duration. Since the *domain-initial strengthening*

[18]Refer to footnote 4 on page 11.

[19]The relationship does not apply to Utterance initial position. According to Cho and Keating, Korean vowels are less affected by durational variation than vowels in some other languages (Cho & Keating 2001:174)

*effect* applies to aspirated stops, VOT for aspirated stops also increased with domain boundary strength, resulting in the same tendency to maintain C/V ratio as seen with speech rate in English. If the tendency also holds up when vowels are lengthened by speech rate in Korean, VOT of aspirated stops will be more affected by speech rate than that of tense stops.

A similar durational pattern is seen in vowels after tense stops but their VOT does not increase. This also supports our hypothesis that tense stops correspond to the short lag category in Thai. In sum, speech rate may affect VOT for stops in Korean, but tense stops are probably less affected than the other two types.

Since our own data described in Section 3.2, was collected without considering variation of speech rate, it cannot be used to directly test our hypothesis. However, as Kessinger & Blumstein (1998) showed strong correlations between VOT and syllable duration, and between vowel and syllable duration, investigating correlation of them in Korean can give us a clue to whether the relationship among VOT, vowel, and syllable duration is similar in Korean and English. If a similar relation can be found in our data, we may conclude that some of the variation in syllable duration is probably due to speech rate. Furthermore, it can justify the classification of tense stops as short lag category as in Thai. They might then also behave like the short lag Thai stops in being invariant to speech rate. We will show that such a relationship holds in our data in Section 3.6.5 on page 80.

### *2.4.4 Summary of characteristics of acoustic cues*

Korean has three types of stops, which are all phonemically voiceless. They can be differentiated by three acoustic cues, VOT, closure duration, and vowel F0 after a stop.

VOT is shortest for tense stops and longest for aspirated ones. Closure duration (although it is not available in utterance initial position) is longest for tense stops and shortest for lax ones. F0 is highest after aspirated stops and lowest after lax ones.

Prosodic structure and other factors that might affect the acoustic cues have been investigated. Prosodic structure for example, affects the identity of the stop because phonological rules are applied within or across rule domains defined by prosodic structure. VOT and closure duration can be longer or shorter depending on prosodic position. Their acoustic variation is attributed to the *domain-initial strengthening effect*. Speech rate is another factor that affects VOT. However, its impact seems to be much less for VOT of tense stops.

Even though the above factors cause the statistical distributions of the affected acoustic cues to be more widely spread about their means, there are two things that do not seem to change. First, the order of overall magnitude of each acoustic cue for each type of stop is not changed. Second, the fact that in many cases, values of an acoustic cue have overlaps among the three types of stops is not changed, either. At the end of Section 2.5.3, we will briefly touch on the implication of these two facts in developing an ASR system using such acoustic cues for differentiating Korean stops.

## 2.5 Problems of Stop Recognition in ASR

In this section, we will summarise Jang's (2000b) experiment on phone unit recognition and his analysis of frequency of stops in three different data sets. The result of the experiment shows a low recognition rate for stop sounds in Korean, which is the motivation for the idea that improving recognition of stop sounds may result in a better ASR system. The analysis of stop frequency will also show how much an improvement in stop recognition can be expected to contribute to overall word recognition.

Since a phone recogniser does not include a language model, it may not be appropriate to relate the result of a phone recogniser directly to an ASR system integrated with a higher level grammar model. Nonetheless, it is reasonable to think that increasing the acoustic probability of a correct phone hypothesis will increase the probability of a word

that includes the phone, which will consequently, increase the chances that the word hypothesis will remain at the end of the recognition process.

Higher acoustic probability does not always guarantee highest overall probability, which is affected by other components such as the language model. However, candidates with high acoustic probabilities have much more chance of surviving as one of the N-best candidates. So, it is important to see whether there is a room for improvement in stop recognition.

### *2.5.1 Automatic phone recogniser*

Jang (2000b) built a phone recogniser similar to his baseline model for continuous speech recognition. Since we used a similar baseline model for our study, we will describe it fully in Chapter 4 and only discuss it briefly here.

The data Jang used was the KAIST speech database[20]. The vocabulary size is about 3000 words and it consists of 12 hours of continuously read speech. For training, 8790 utterance tokens by 89 speakers and for testing, 2073 utterance tokens by 21 speakers were used. The phone recogniser was built with HTK (Young *et al.* 1996). 37 phone models, including a silence model were trained. Each phone model was a 3 state, left to right continuous density HMM. The *Baum-Welch* and *viterbi* algorithms were used for training and decoding, respectively[21].

For parameterisation of speech waveforms into sequences of feature vectors, 25 ms windows were used with a frame shift of 10 ms. There were 39 feature vectors used. They are 12 Mel Frequency Cepstral Coefficients (MFCCs) plus energy, together with their first and second derivatives.

When using HTK, some sort of language model had to be given when decoding test data. The additional language model log probability is added to the acoustic log probability.

[20] More detailed description of the database will be in Section 3.2

[21] These algorithms are well explained in Rabiner & Juang (1993) and the HTK manual.

| Phone Class | | | No-gram | Unigram | Bigram |
|---|---|---|---|---|---|
| Vowel | | | 50.67 | 53.32 | 61.98 |
| Consonant | | | 59.61 | 59.31 | 64.95 |
| | Sonorant | | 62.57 | 60.25 | 66.22 |
| | Obstruent | | 57.36 | 58.59 | 63.98 |
| | | **stop** | **49.61** | **50.56** | **56.12** |
| | | affricate | 57.68 | 58.42 | 65.85 |
| | | fricative | 73.97 | 76.13 | 79.74 |
| Overall average | | | 56.97 | 58.81 | 62.70 |

Table 2.4: Summary of phone recognition accuracy(%) for natural classes. Accuracy is calculated through the formula: $((Correct - Insertion)/Number) \times 100$. Source: Jang (2000b:33)

Jang experimented with three language models. The first one is no-gram model, in which each phone is equally likely to appear. The second one is unigram model, in which each phone has a probability that represents its frequency in the corpus. Finally, the bigram model uses the probability that a phone, A is preceded by a phone B. Some of the tables Jang used in his thesis are reproduced here in order to show how well stops and other phone classes can be automatically recognised in the phone recogniser.

Table 2.4 shows that the stop sounds are the least correctly recognised phones across the three language models. As Jang noted, the unigram and bigram language models give only a modest improvement in phone recognition.

Considering the syllable structure of Korean, the phone sequence can in principle be a big help in phone recognition, which seems to go against Jang's explanation that "Especially in Korean, where no syllable initial or final consonant clusters are allowed, phonotactic constraints are not very useful for narrowing the search path of consonants" (Jang 2000b:33). Putting his remark differently, a language which has syllable initial or final consonant clusters such as English, could have a narrower search path. But he seems to have overlooked the simplicity of syllable structure in Korean.

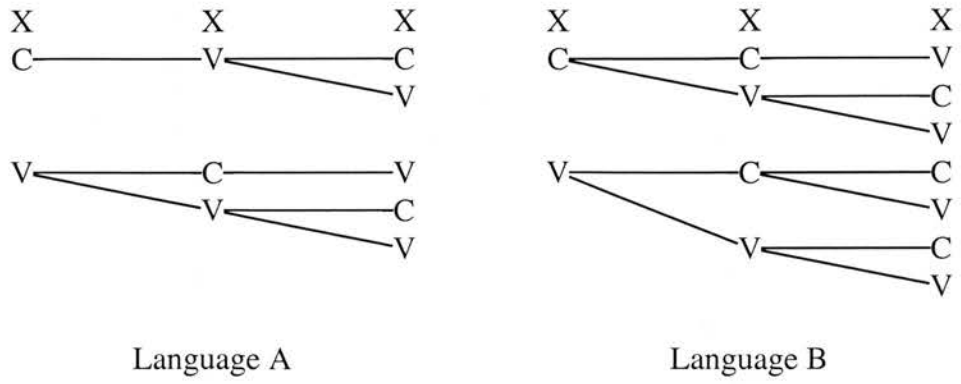



Figure 2.5: Phone Paths in two hypothetical languages. In Language A, syllable initial and final consonant clusters are not allowed whereas in Language B they are. Language B has more possible combinations of consonants and vowels. Semivowels and vowels are considered as V in this diagram.

Suppose we have three slots to fill in with vowels and consonants in two hypothetical languages. As shown in Figure 2.5, Language A has fewer paths than Language B because of its simpler syllable structure. If all other phonotactic conditions are the same except the syllable structure, Language B would require more work in searching all possible paths.

In English, for example, when we have a consonant in a phone string, it is hard to predict its syllable position because of the more complex syllable structure compared to Korean. A Korean consonant, on the other hand, can be in either an onset or a coda, which also gives an information of what could come before or after. And the coda consonants are restricted to seven phones, [p, t, k, l, n, m, ŋ], which means, if a consonant after vowel has higher acoustic probability than any of the seven coda consonant, it is probable that the next phone after this consonant is a vowel.

Although there is a benefit from the language models, particularly from the bigram model, the benefit seems to mostly be in vowel recognition. As seen in Table 2.4, com-

| Phone | Phonation confusion (%) | Place confusion (%) |
|---|---|---|
| p | 33.61 | 25.84 |
| p' | 71.43 | 26.67 |
| $p^h$ | 46.15 | 14.20 |
| t | 34.01 | 23.99 |
| t' | 42.86 | 24.49 |
| $t^h$ | 36.29 | 24.19 |
| k | 35.47 | 20.40 |
| k' | 71.14 | 11.94 |
| $k^h$ | 48.67 | 9.33 |
| Total | 37.98 | 21.10 |

Table 2.5: Proportion of substitution errors caused by phonation type confusion compared with those of place confusion. Values are the ratio of each type error to total substitution error. Phonation type confusion refers to the number of phones recognised as either of the other two different types of the same place category (eg., [p] as [p'] or [$p^h$]). The place confusion refers to the number of phones recognised as either of the other two different places with the same phonation type (eg., [p] as [t] or [k]). Source: Jang (2000b:36)

pared to the result with the no-gram model, the recognition rate of vowels is increased by 11.31% when bigram model is used. With the no-gram model, recognition rates of vowels and stops are not much different but with the bigram model, vowels are better recognised than stops. What the result of this experiment really shows is that the acoustic probabilities of stops are generally low. Recognition of vowels can be helped by the other source of information but for stops, the recognition rate is still low.

Table 2.5 shows another ground for concentrating on improving stop recognition for ASR. The low recognition rate of stops is partly due to confusion among the different types at the same place of articulation. Because different types of stops at the same place of articulation are phonetically classified by VOT, closure duration and F0 after stops, these acoustic cues are not well modelled with conventional HMMs. The next section will explain more about this problem.

In sum, the acoustic probability of stops is generally low and the classification of types of stops is also difficult. More confusion in phonation types than in place differentiation suggests that implementing acoustic cues for differentiation of phonation types based on phonetic studies might improve stop recognition.

### 2.5.2 *Quantitative contribution of stops in Korean words and syllables*

We have seen that recognition of stops in ASR is not as good as that of other phone types and that differentiation of phonation types is the most likely source of errors in recognition of stops. However, if the stops in Korean were not much used in words, the negative effect caused by the error would be negligible. Thus, it is necessary to find out how much functional load Korean stops bear in words (Jang 2000b:37).

In order to calculate frequencies of stops in Korean, Jang used three datasets. The first one is 2350 syllables which are used for all Korean words written orthographically[22]. The second data set is a word dictionary created and distributed by Natural Language Processing Laboratory, Pohang University of Science and Technology (POSTECH). The third dataset is a transcription of three hours recorded telephone dialogues between an hotel operator a guest concerning hotel reservation. The dataset has 5561 utterances in which 13063 linguistically meaningful words are used among 5802 unique words[23]. The transcription did not follow orthographical rules but what a transcriber actually heard.

The third dataset is better than the other two for the purpose of investigating the functional load of stops in a couple of reasons. First, as Jang mentioned, it is spontaneous speech. The other two datasets simply represent how stops are distributed in a dictionary. But in real speech, some words, especially function words, will be frequently repeated and this makes some phones appear more than others, which affects the frequencies of the

[22] According to Huh (1985:229), the number of possible syllables in Korean is 3520. The difference between the two numbers results from the fact that many syllables are never used in orthographically correct Korean words. Orthographically unused syllables can appear in pronunciation.

[23] Jang (2000b:39) states that the total word count is 16568 which includes vocatives, interjections and disfluency.

| Phone | (a) Syllables | (b) Dictionary Words | (c) Dialogue Words | (d) Textbook Syllables |
|---|---|---|---|---|
| p | 5.5 | 8.21 | 6.25 | 3.70 |
| t | 5.4 | 5.22 | 5.22 | 7.76 |
| k | 7.3 | 12.73 | 15.78 | 12.04 |
| **LAX Total** | **18.20** | **26.19** | **27.20** | **23.50** |
| p' | 3.1 | 1.49 | 0.16 | 0.51 |
| t' | 3.7 | 1.19 | 1.69 | 2.27 |
| k' | 5.1 | 2.54 | 0.37 | 2.44 |
| **TENSE Total** | **11.90** | **5.22** | **2.22** | **5.22** |
| $p^h$ | 4.4 | 3.10 | 3.19 | 0.85 |
| $t^h$ | 4.5 | 2.29 | 1.51 | 1.02 |
| $k^h$ | 4.5 | 1.33 | 0.99 | 1.10 |
| **ASPIRATED Total** | **13.4** | **6.72** | **5.69** | **2.97** |
| **Total** | **43.50** | **38.13** | **35.11** | **31.69** |

Table 2.6: Quantitative analysis of stops appearing in four different data sources. Each number represents the percentage of words or syllables containing the phone in each data source. (a) from 2350 syllables, (b) from 113194 dictionary words, (c) from 82 natural continuous speech dialogues with 13063 linguistically meaningful words, and (d) from various articles with 101637 syllables. Source: Jang (2000b:40) and Jin (1992:86)

phones. Second, since the transcription is based on actual pronunciations, it is more accurate than the others that are based on automatic conversion of an orthographic string into a phonetic form. So, the calculation of frequency is also more accurate. However, there could be a problem with depending on such a dataset. The topic is hotel reservations and some content words related to the subject could give rise to biased information on phone distribution.

Table 2.6 is a summary of Jang's investigation of stop frequencies in the three datasets. In addition to his analysis, another phone distribution using an independent dataset is illustrated in the table. Jin (1992) also examined frequencies of phones in Korean, based on primary school Korean language textbooks and scripts used for broadcasting. All sentences were transcribed in compliance with orthographical rules and then spaces between

words were deleted when the two neighbouring words could be read as one unit. Finally, the revised string was converted to a phonetic string using phonological rules. The result should approximate a standard pronunciation of the texts. 101637 syllables were found in the data.

Jin's (1992) results are in line with Jang's. The difference between the results of the second and third datasets and that of Jin's (1992) data is that the second and third datasets count words starting with one of the stops, while Jin counts syllables starting with the stops. In either case, more than 30% of the words or syllables counted begin with a stop.

### *2.5.3 Multiple acoustic cues for classification of Korean stops*

As pointed out in Holmes & Huckvale (1994), the more information is available, the better ASR can be developed. For Korean stops, we have seen three acoustic cues phonetically and perceptually attested. F0 has been successfully used in developing an ASR system by Jang (2000b), but the remaining two cues have not been included in his system. He gave four reasons why temporal cues such as VOT and closure duration are not attractive as useful speech parameters for ASR.

First, Jang said that temporal cues were highly dependent upon the context of the consonants (Jang 2000b:42). For example, closure duration after a pause cannot be measured because a silence for a pause and a closure are not acoustically separable. He also argued that even in a position where a closure duration is available, like in intervocalic position, lax stops become voiced so that closure durations will not occur. However, using multiple acoustic cues will sort these problems out. Admittedly, a closure duration after a pause cannot be used because of its unavailability but VOT and F0 are still available. For lax stops that are realised as voiced intervocalically, it is also true that closure durations of voiced stops cannot be measurable because voiced stops do not have an acoustically silent part for closure by definition. However, voiced stops are prosodically positioned AP medially because AP is a prosodic domain for *Lenis Stop Voicing* and more importantly, the segmental F0 effect is diminished after AP initial position. That is, as we

| Prosodic positions | Utt. ini. | IP ini. | AP ini. | PW ini. | PW med. |
|---|---|---|---|---|---|
| Acoustic cue | **VOT** | | | | |
| LAX | √ | √ | √ | | |
| TENSE | √ | √ | √ | √ | √ |
| ASP | √ | √ | √ | √ | √ |
| | **Closure duration** | | | | |
| LAX | | √ | √ | | |
| TENSE | | √ | √ | √ | √ |
| ASP | | √ | √ | √ | √ |
| | **F0** | | | | |
| LAX | √ | √ | √ | | |
| TENSE | √ | √ | √ | | |
| ASP | √ | √ | √ | | |

Table 2.7: Availability of acoutic cues depending on prosodic positions

noted on page 26, Jang himself found that F0 is not effective in non-AP position (Jang 2000b:85).

Contrary to Jang's claim, in non-AP position as shown in Table 2.7, it is not F0 but closure duration and VOT that give us better information in distinguishing three types of stops in Korean. Even though closure duration and VOT are not measurable for voiced stops, no availability, by itself, tells us that the stop in question is voiced, thus a lax stop. On the other hand, since F0 is low in voiced stops and high in tense and aspirated stops, it can only distinguish between lax and the others.

Jang's (2000b) second argument against using temporal cues is that these cues are mainly effective for stop sounds but not for affricates (Jang 2000b:44). This seems to be legitimate because closure duration and VOT have been suggested as acoustic cues for stops

not for affricates. But, why would anyone like to use these temporal cues for differentiating affricates in the first place? Neither closure duration nor VOT has been reported as an acoustic cue for differentiating the three types of affricates. Raising the second issue could be justifiable only if the temporal cues were reported as acoustic cues for affricates.

Our study is focused on stop sounds and the main objective of building an ASR system is to see whether three acoustic cues can improve the recognition of stops which may result in overall system improvement, not to see whether affricates can be classified by temporal cues. For the purpose of comparison, when constructing an ASR system, F0, excluding closure duration and VOT, will be used for the affricates.

Thirdly, Jang notes that the traditional HMM is not capable of introducing duration of units properly (Jang 2000b:47). Since this is a structural problem of HMMs, it is not limited to Korean stop durations. This should be overcome whenever any duration of any unit is modelled with the traditional HMM. Revised HMM models for durations by Ferguson (1980) and Levinson (1986) still have a problem in modelling VOT and closure duration because durations are expressed as numbers of analysis frames, typically 10 ms long. If any temporal cue is shorter than the frame shift of the analysis, revised models are of no use. To avoid this problem, VOT and closure duration will be modelled in a different way as explained in Chapter 5.

Finally, Jang mentions that the reliable extraction of the durational characteristics is difficult (Jang 2000b:47). Since the characteristics of temporal cues are highly variable depending on speech rate, getting consistent characteristics from spontaneously spoken speech is hard. However, as we have already seen in the above Section 2.4.3, VOT, for example, varies depending on speech rate, but the order of VOT magnitudes is not changed. This means the characteristics of VOT are consistent even in a highly variable situation.

The advantage of having consistent characteristics across speech rates when modelling VOT of each type of stop is that it is possible to pool VOT data from utterances regardless of their speech rate. The order of mean VOTs for each type will not change. However,

there is a drawback to pooling the data. Pooling the data always brings about enlarged overlapping areas between the two neighbouring types' distributions.

There could be two ways of increasing the overlapping area between two neighbouring statistical distributions. As shown in Figure 2.6, if the standard deviation of one type gets greater due to data pooling, the overlapping area is increased. Another possibility is that the means of the two distributions get closer. In this case, even if the standard deviations are not changed, the overlapping area is always increased.

Since the new distribution by pooling the data should cover all the samples in the context a and b in Figure 2.6, the standard deviation gets greater. And the wider distribution makes the mean somewhere between the two means of original distributions. Consequently, the mean of the new distribution is closer to the mean of the distribution of Type B.

These overlapping areas will be a source of confusion in determining which type. However, multiple acoustic cues will help to relieve the confusion. If we have only one acoustic cue for classifying Korean stops, we have to decide which is more advantageous, pooling the data in order to overcome data sparsity, or not pooling the data in order to reduce the overlapping area which causes confusion. Since we have other acoustic cues, we have multiple criteria that can possibly clarify the confusion.

More detailed explanation of the system implementation will be discussed in Chapter 5.

## 2.6 Summary

So far, we have explored Korean stops and their acoustic characteristics. Korean stops at the same place of articulation have three different phonation types. Differentiation of the three phonation types of stops at each place of articulation has been reported in various phonetic studies to be dependent on three acoustic cues, VOT, closure duration and vowel F0 after the stop.

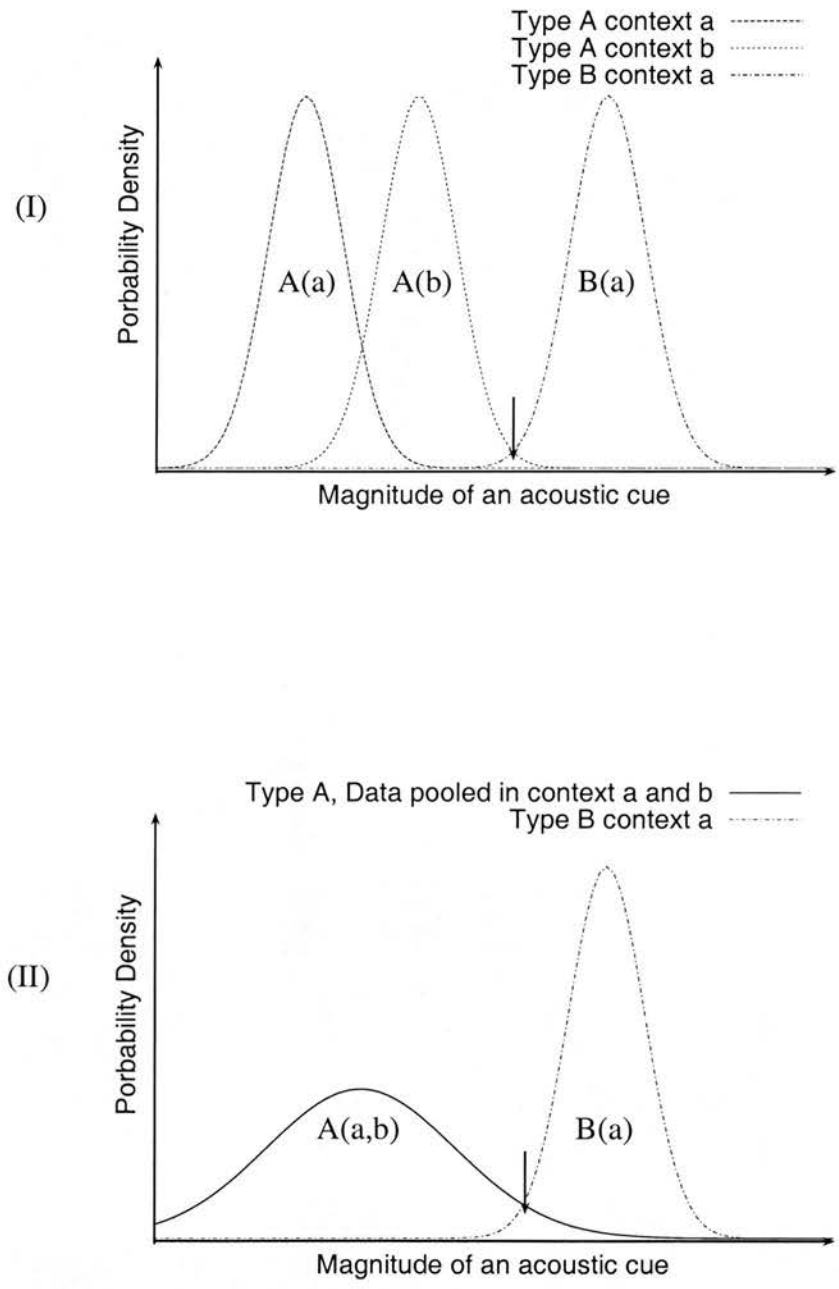



Figure 2.6: Schematic representation of increased overlapping area caused by data pooling. Left two distributions in (I) represent the same type of phones in different contexts. In (II), pooling the data in context a and b brings about greater standard deviation. ↓ points at the overlapping area between Type A and Type B.

We have also seen that the recognition of stops in an automatic phone recogniser was worse than the overall recognition rate and that there was much functional load of stops in various datasets for Korean. The results support the idea that improving stop recognition may lead to overall improvement of an ASR system for Korean.

In order to use the acoustic cues for stops in ASR, we have examined various factors that may affect the characteristics of acoustic cues. The magnitude of each acoustic cue can vary depending on previous phone environments, prosodic positions, speech rates and so on. However, we have seen that the variation does not change the order of overall magnitude of each acoustic cue for each type of stop.

We have found that each acoustic cue has a limitation on its availability depending on prosodic positions. For example, closure duration is unavailable in the Utterance initial position, but even in this position, other acoustic cues such as VOT and vowel F0 are still available. Thus, considering multiple acoustic cues will be more helpful to identify the stop in any prosodic position than a fixed single cue.

In ASR, multiple acoustic cues have not been used for improvement of stop recognition. Only the segmental F0 effect has been exploited for this purpose by Jang (2000b). Using multiple acoustic cues for stop recognition in ASR looks more appropriate at least in phonetic studies. So, on the basis of our investigation so far, we are going to see multiple acoustic cues are actually helpful for stop recognition in a Korean ASR system. But first, in the next chapter, the three acoustic cues are statistically tested to confirm their validity of differentiating the three types of stops in speech with contextual variability, as well as controlled speech.

# CHAPTER 3

# Statistical Analyses of Korean Stops

## 3.1 Introduction

In this chapter, characteristics of Korean stops we have examined in the previous chapter will be statistically tested using two databases. Even though these characteristics have been previously observed in the literature, previous results are incomplete in four respects. First, all acoustic parameters have never been examined in a single database. Each database in the previous studies was created to describe characteristics of one or two acoustic cues. Even if two cues were examined, their relative strengths and interactions have not been explored. Second, statistical tests have not been done for all stops and places of articulation. Third, multiple acoustic cues have never been treated as a multivariate way. So, only the ANOVA test, rather than MANOVA, was adopted for statistical tests. Fourth, a continuous speech database with contextual variability has never been used either for description of acoustic characteristics or for statistical tests.

To confirm whether all reported acoustic characteristics can be found in a single dataset, we are going to use a speech database recorded with carrier sentences. An advantage of using a database recorded with carrier sentences is that other linguistic factors than the ones under investigation can be controlled, so that we can have clearer view of acoustic

characteristics of all stops in all places of articulation. For confirmation of the same characteristics in continuous speech as in the database with carrier sentences, the KAIST speech database will be used. Both databases will be described in the next section.

Compared to the work done by Lisker & Abramson (1964), Kim (1965), and Han & Weitzman (1965, 1970), ours is focused more on statistical tests, which, we believe, will give more an accurate description of stop characteristics. In Lisker & Abramson (1964), fewer than 35 tokens for each type in each place of articulation were used for the measurements of VOT. In Kim (1965) and Han & Weitzman (1965, 1970), on the other hand, many more tokens were used. Although the objective of these works was to show relative comparisons of VOT for each type, statistical tests seem to have been ignored.

Statistical tests were done in Han (1996) but the study was limited to lax and tense stops, and velar stops were excluded because of their frequent change to homorganic fricatives in intervocalic position. Closure duration was the only acoustic cue she used for the statistical test. Each group of stops consisted of fewer than 13 tokens.

In Silva (1992), various statistical tests were done using closure duration and "vowel lag", which is a duration from the burst of a stop to the onset of the second formant for the following vowel. However, in most cases, the target stops of his ANOVA tests were restricted to bilabials.

Using the first database that we mentioned above, Jang (2000b) did statistical tests to see whether three types of stops can be statistically differentiated by vowel F0 after stops. Other acoustic cues such as VOT and closure duration were not considered in his statistical tests.

The goal of our statistical tests is the same as in Jang (2000b), but we will use three acoustic cues all together at the same time whenever they can be measured. And this is the major distinction from other statistical works on the same subject. Even though it has been known that multiple acoustic cues exist for differentiating the three types of stops in

Korean, a statistical approach using multivariate analysis has not previously been tried. So, whenever a statistical test on three or more groups of samples is needed, MANOVA tests will be adopted.

As we mentioned in the Introduction, this statistical test is done to support the idea that the acoustic cues can be used in ASR. Since the second database is designed for ASR, contextual variability is not avoidable. So, it should be independently verified whether the acoustic cues still play a role under such circumstances.

This chapter is organised as follows. Firstly, the two databases we used will be described. Secondly, the criteria for labelling of each acoustic cue will be described. Thirdly, methods and procedures for statistical tests will be explained. Next, the statistical results will be presented. Finally, on the basis of the statistical tests, the acoustic characteristics of Korean stops will be revisited.

## 3.2 Data

The first database, which we will call the *stop-rich database* and will use for general information on the acoustic characteristics for Korean stops, was created by me and my former colleague, Tae-Yeoub Jang, who used this database for his segmental F0 study (Jang 2000b)[1].

Two hundred and sixteen two-syllable isolated words were prepared. Each word was chosen from a Korean dictionary, which means the words were not imaginary nonsense words but real ones. At least one of the two syllables had a stop or an affricate in onset position. To prevent or minimise prosodic effects from affecting the acoustic characteristics of targeted phones, each word was read in a carrier sentence as follows (Jang 2000b:78):

[1] Since Jang has already used the data, we refer to his description of it (Jang 2000b:77-78).

(3.1) *ikəs-in* _____ *$c^h$ələm* *po-i-nta*
this-TOP like see-PAS-END
'This looks like _____'

TOP: Topic marker
PAS: Passive marker
END: Verbal ending

We tried to balance the numbers of each stop in the onsets of first and second syllables, but it was hard to find enough two-syllable words with stops in both positions. And we also tried to balance the number of phones in the coda of the first syllable. As stated earlier on page 42, seven phones can be realised as coda consonants. They are the lax stops [p, t, k], three nasals [m, n, ŋ], and a liquid, [l]. By balancing the number of these phone groups, we can observe the effect of the previous phone on the acoustic characteristics of onset stops in the second syllable of the word. Since the words were from a dictionary, we could not completely control the number of phones in the various contexts, so the numbers of tokens in the statistical tests are by and large unbalanced.

Four male speakers (JSH, KHK, TSS and WHY), including myself (WHY), took part in recording. All speakers spoke Seoul Korean, which is regarded as standard Korean. We all lived in Korea for over 30 years and currently reside here in the UK for studying as postgraduate students.

Two hundred and sixteen words were iterated 5 times for each speaker. A total of 4320 (216 words x 4 speaker x 5 iterations) sentences were recorded. Among the 216 words, there are 70 words that include one of the three affricates either in the onset of the first syllable or in that of the second syllable. Even though affricates were recorded, they are not used in the statistical tests, simply because they have not been reported to have the same durational pattern as stops do. Sentences were recorded in random order in a sound proof booth. 16 bits and a 16 kHz sampling rate were applied for analog to digital conversion.

The second database, which is called the KAIST speech database (Park *et al.* 1995), is, as we have already mentioned above, designed for developing an ASR system. The database consists of five data sets, one of which is continuously read speech and is used here for the statistical tests and for the continuous speech recognition experiments in Chapter 4 and 5. The database was provided by the *Communications Research Laboratory* of the *Korea Advanced Institute of Science and Technology* (KAIST)[2]

The scripts for the continuous speech data were composed from sentences occurring in a conversation textbook for trade negotiation. Words denoting time, date and geographical names were added. Some honorific endings were also added. These changes and additions of utterances were done in order to make natural utterance tokens.

Before we go into the number of words in the database, we have to discuss the definition of a word in Korean. While a word in English text can be crudely defined as a lexical item between two white spaces, a Korean word cannot be defined in the same way. The corresponding unit in Korean, surrounded by two white spaces, is called '*eo-jeol*'. *Eo-jeol* is a spacing unit in Korean orthography and is composed of either a single morpheme or a set of morphemes. When an *eo-jeol* is composed of a set of morphemes, it usually represents a lexical item followed by case markers or postpositions.

Using an *eo-jeol* as a word makes the dictionary grow rapidly when all noun adjective and verb forms are considered. Looking ahead to a large vocabulary speech recognition system, a morpheme dictionary will be much smaller than an *eo-jeol* dictionary, due to agglutination. However, complete separation of morphemes can cause a problem in ASR because some morphemes are very short. These short morphemes are generally verb endings, postpositions or suffixes. So, in ASR, the short morphemes are concatenated to their stems and these concatenate morphemes are called *pseudomorphemes*. Using the

[2]Continuous speech data is one of the five sets in the KAIST database. Since the continuous speech data is the only one that I use in the KAIST database, the KAIST database refers only to the continuous speech data in this thesis.

| word item | structure | glossary |
|---|---|---|
| na | STEM | '1st person singular' |
| na-neun | I + TOP | 'I (subject) ...' |
| na-eke | I + DAT | 'to me' |
| na-man | I + RES | 'only I' or 'only to me' |
| na-eke-neun | I + DAT + TOP | 'I (possess) ...' |
| na-eke-man | I + DAT + RES | 'only to me' |
| na-man-eun | I + RES + TOP | 'only I ... (negative)' |
| na-eke-man-eun | I + DAT + RES + TOP | 'at least only to me' |

Table 3.1: Example of Korean morpheme agglutination. The grammatical markers used are: TOP(ical), DAT(ive), RES(trictive). (Reproduction from Jang (2000b:55))

*pseudomorpheme* as a word unit, improvement was achieved in several studies (Kwon *et al.* 1999, Kwon 2000, Park & Chung 2001).

Jang (2000b) separated nouns and pronouns from their grammatical markers and other forms were used as an *eo-jeol* unit, which is similar to the *pseudomorpheme*. An example of separation is reproduced here in Table 3.1. After the separation of these grammatical morphemes, the vocabulary size was 2920. Without this separation, there would be 3200 unique *eo-jeol*s found in the utterances in the database, according to Jang's estimation. The average number of words in a sentence was 8.4.

The database will be also used in Chapter 4 and 5 for developing an ASR system using acoustic cues for Korean stops. More explanation on how the data is divided for training and testing will be presented in that chapter. For the use of statistical tests, 433 utterances were selected and hand-labelled[3]. Each of 150 speakers read 90 to 100 sentences and a total of 14746 sentences were recorded. In 433 hand-labelled data for statistical tests, 4 or 5 sentences of 85 speakers were used. 33 were female and 52 were male. Speakers were in their twenties or thirties and had received higher education. Recording was carried

[3]Formerly in Yun & Jang (1999), 628 utterances were hand labelled but some of the utterances from a specific dialect, which is quite different from standard Korean, were excluded here. This dialect has different tonal structure and is spoken in Kyeong-sang province, which is in the south east part of South Korea.

out in a relatively quiet room. Because the purpose of the database was for developing an ASR system, a normal speech environment was chosen when the data was created. Speech data were digitally processed using 16 bits and a 16 kHz sampling rate.

## 3.3 Hand-labelling Criteria and F0 Extraction

Hand-labelling was done by Jang and me using *xwaves* and *xlabel* of *Entropic* (Entropic 1998). The hand-labelling procedure was explained in Jang (2000b:52), and we summarise it here again. However, he did not explain what criteria were employed for labelling stops. Mostly, labelling criteria were in line with Silva (1992:92-96), which will also be summarised below. The same labelling criteria were applied to both databases.

Phones were assigned phoneme, rather than allophone, labels. In particular, allophonic variants, [b, d, g] were marked the same as their unvoiced counterparts, [p, t, k]. Because in voiced stops the vocal folds vibrate while the occlusion is made at the stop's place of articulation, the voiced stops will have non-zero acoustic energy during the closure. Since the closure duration is regarded as an acoustic silence, the voiced stops will not have closure durations. So, after completing the labelling, voiced allophones can be picked up if necessary by a simple programming script because any lax stop that is not preceded by the closure label is a voiced one. A 50 ms arbitrary closure duration was given to sentence initial stops and stops after a pause, which is also a silence and is not caused by an articulatory movement for stops[4]. These arbitrary closure durations were not counted for statistical tests.

Because a stop for our purposes is divided into two parts, a closure and a VOT, we need at least three demarcations;(1) the beginning of a closure, which is the same as the end of a preceding phone, (2) the beginning of a VOT, which is, in other words, a point where a stop burst starts, (3) the end of the VOT, which also means the starting of a following vowel.

[4]It would be more appropriate to say that the closure duration is included in the pause but since they are all silence acoustically, there is no visible cue for separating the two in the waveform.

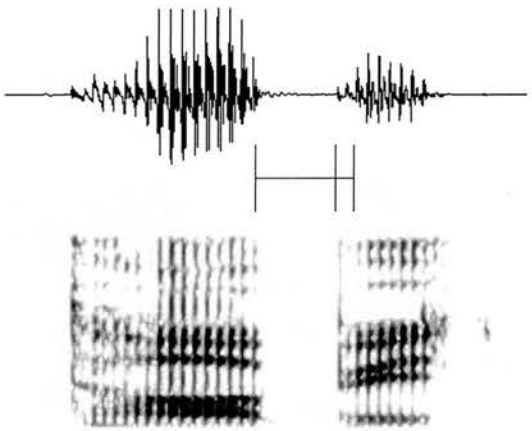
Silva (1992) used more detailed demarcations for stops and other related labels. He added a mark when vocal cord vibration of a preceding phone ended. For a stop burst, which he called 'stop release', he used two marks, one for the beginning and the other for the end of the burst. For the end of a VOT, he also used two separate marks. The first one is for the end of a VOT, which starts a periodic energy in the waveform or a voicing bar in the spectrogram. The second one, which he called 'F2 Onset', lies on the waveform when the spectrogram starts to show both the first and the second formant.

We used the same method as Silva for the beginning of a stop closure. Two cases were regarded as cues for determining the start of a stop closure. The first case was when complex periodic waves in the waveform were reduced. The second one was when the formant structure, particularly the second formant, began to disappear in the spectrogram.

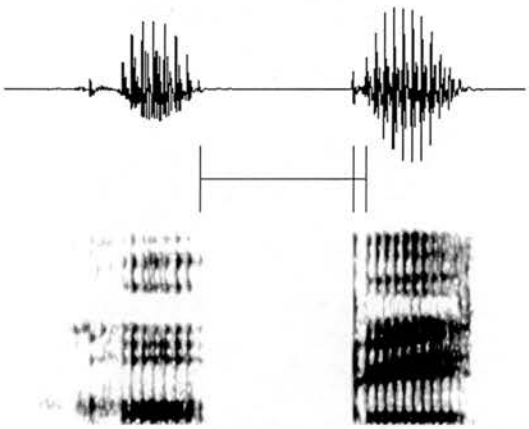
We did not record the duration between the start and the end of a burst. Nor did we record the duration between the end of what Silva called 'voice onset' and the start of what he called 'F2 onset', which represents the beginning of the first and the second formants for a following vowel in the spectrogram. The reason we ignored these durations was that they did not show up clearly in the waveform. As Jang (2000b) illustrated in an example of three types of hand-labelled stops, it is very hard to separate these durations. Even though these durations exist, they are short enough to be ignored. we reproduce Jang's figure here in Figure 3.1. Instead of making two marks for a stop burst, we considered the beginning of a burst as a beginning of a VOT. Likewise, for the end of a VOT, we used the F2 onset as the end of a VOT.

The demarcation was based on the speech waveforms. To have more detailed information for formants and acoustic energy spread throughout the frequency domain, wide-band spectrograms were used, but only supplementarily.

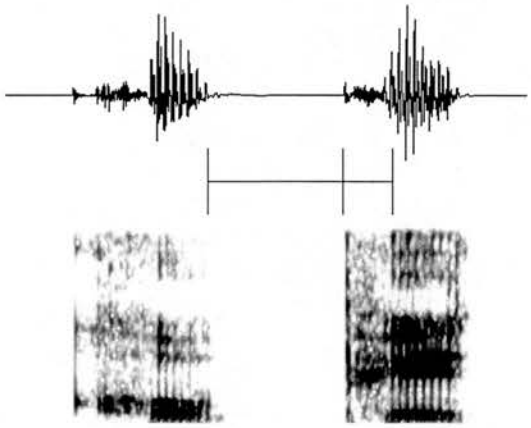
Errors that were made during hand labelling were automatically and manually corrected. For example, if a label for closure was shown before a non-stop phone label, it was



(a) CV p i



(b) CV p' i



(c) CV ph i

Figure 3.1: Demarcation of a stop closure and VOT for three types of stops especially in intervocalic positions. Words are (a) *teo-pi* 'race', (b) *ko-ppi* 'reins', and (c) *kho-phi* 'nosebleed'. (Reproduction from Jang (2000b:43))

automatically picked up and corrected manually. After the correction, all labels were examined again by looking at the waves and labels at the same time.

In order to get vowel F0 after stops, the F0 extraction program called *get_f0* provided by Entropic (1998) was used[5]. As Jang (2000b:81) explained, vowel F0 after a stop, as an acoustic parameter for statistical analyses, is regarded as the overall mean of F0 values of all the non-zero frames of the vowel period. *get_f0* sometimes assigns non-zero F0 values to frames which have been labelled as being inside the stop. Such F0 values are included in the calculation of the mean F0.

## 3.4 Additional Labelling and Sample Grouping

After hand-labelling, preceding phonetic contents for stops were grouped according to their natural classes. The possible contextual environments for the slot preceding a stop are silence, vowel, coda stop, and sonorants such as nasals and liquids. Silence as a contextual environment means that the closure duration for a stop cannot be measured and thus, does not participate in statistical analysis.

Since the two syllable words were inserted in a specific carrier sentence in the stop-rich database, the onset stops of the first syllable are not actually sentence initially positioned stops. They are AP initial stops and can be potentially IP initial[6]. If they are IP initial, they are preceded by a pause. To avoid unwanted mixing of two different sources of silence, closure durations for initial stops in the target words were excluded from the statistical analysis.

Since the KAIST database does not have prosodic labels and there are too many sentences for hand-labelling, we do not have prosodic boundaries for all utterances. We decided

[5]This program is based on normalised cross correlation and dynamic programming proposed in Talkin (1995).

[6]There was no single case found where a lax stop was realised as a voiced stop. Since an AP boundary blocks the phonological rule of *Lenis Stop Voicing*, it is clear that onset stops in the first syllable in the stop-rich database should be regarded as at least AP initial stops.

not to attempt automatic prosodic labelling. As Jang (2000b:134) already pointed out the segmental F0 effect might mask real AP boundaries, which makes automatic detection of APs unreliable. Nevertheless, word initial stops can be labelled as PW initial stops in an automatic way because a PW is syntactically defined[7]. A simple manipulation of the utterance script can divide stops into three groups:(1) PW initial, (2) PW medial, (3) resyllabified onset stops.

The first grouping of the samples was made according to their places of articulation. Then, the samples from the same place of articulation were divided according to their type, lax, tense or aspirated. Further groupings were made depending on natural classes of preceding phone and on prosodic position of the stop. Having four criteria for grouping of samples, there are four factors. The summarisation of factors and factor levels is as follows:

(3.2) Factor 1: Phonation Types

factor level: lax, tense, aspirated

Factor 2: Place of Articulation

factor level: bilabial, alveolar, velar

Factor 3: Preceding Phone Context

factor level: silence, sonorant, vowel, stop

Factor 4: Prosodic Positions

Stop-rich data

factor level: AP (or IP) initial, PW medial[8]

KAIST database

factor level: PW initial, PW medial, Resyllabified Syllable initial

Although we have 4 factors, they do not have an equal status. Since what we are trying to test is whether different types of stops can be statistically differentiated when they are

[7]Refer to the definition of PW in page 24.

[8]There are only two prosodic positions in the stop-rich database and they happen to correspond to word initial and medial positions. So, I use these terms interchangeably when using this stop-rich database.

grouped according to the other factors, the effect of each factor, except for type, will be tested as a function of the three-way distinction of phonation types. For example, for the factor place of articulation, what we are interested in is not whether three places of articulation are differentiated by the stops' acoustic parameters, no matter what the type is, but whether three types of stops are differentiated in each place of articulation.

Some of the groups have fewer tokens than other groups. For example, since intervocalic lax stops in PW medial position are supposed to be realised as voiced stops, the number of these stops is small, compared to the lax stops in PW initial position. In the stop-rich data, for example, there are 1060 bilabial lax stops. 493 tokens occur in PW initial position whereas only 43 are PW medial. 524 lax stops were actually realised as voiced stops. Token imbalance is more problematic in the KAIST database. Some groups have fewer than one or two samples. These groups were ignored and the samples were pooled with other groups if possible.

For each sample, three acoustic parameters were extracted. These samples were used as inputs to a statistical analysis package SPSS Version.10.0 (Statistical Package for Social Science) (SPSS 1999).

## 3.5 Some Notes on Statistical Procedure

The main objective for using MANOVA is to see the differences among the groups formed by the factors mentioned above. Before we present the MANOVA results, we will show means and standard deviations of each group.

MANOVA tests provide two merits over ANOVA. First, the probability of a Type I error rate can be controlled experiment-wide. Second, it can reveal an overall group difference introduced by correlations of dependent variables[9]

[9]Discussions on general statistical background information concerning MANOVA test in this section and throughout are based on Hair Jr *et al.* (1998).

Data for a MANOVA test must conform to at least three assumptions. First, dependent variables should follow multivariate normal distributions. Second, there should be equal variance and covariances across the groups. And finally, observations of a variable should be independent of one another, which means observations should not be related to the earlier or later observations.

To be statistically correct and accurate, one should test these assumptions. However, the MANOVA test is robust in the sense that violation of the assumptions does not affect the result of the test very much. Even though multivariate normality is violated, it is known that the violation can be tolerable if there is a large number of samples. Homogeneity of variance and covariance matrices can be also violated if the size of the largest group is no more than 1.5 times bigger than that of the smallest group. And even if there is a serious violation, the MANOVA test still can be done when a conservative level of significance is used. Independence among sample observations causes the most serious problem in MANOVA tests when violated. The reason for randomisation of sentences in recording is mainly to provide independence among observations. Since there is no complete test for independence, if no extraneous interference with independence is detected in statistical design, or data collection, independence among observations is assumed in normal practice.

SPSS provides various significance test for MANOVA tests. Specifically, Wilks' lambda, which is most commonly used when there are more than two groups for comparison, is used to test the null hypothesis. When there is a significant difference found, a post-hoc test follows up. SPSS also provides several ways of doing the post-hoc test. Among those, the Games-Howell pairwise comparison test is used because it does not have to assume equal variances and equal sample sizes.

| | | LAX | TNS | ASP |
|---|---|---|---|---|
| CLS | Mean | 36.2 | 90.2 | 70.7 |
| | St. Dv. | 13.6 | 33.7 | 29.5 |
| | n | 168 | 1157 | 899 |
| VOT | Mean | 43.7 | 20.2 | 50.7 |
| | St. Dv. | 17.0 | 9.3 | 19.5 |
| | n | 1724 | 1675 | 1558 |
| F0 | Mean | 114.1 | 137.7 | 144.8 |
| | St. Dv. | 24.3 | 25.8 | 32.9 |
| | n | 1724 | 1675 | 1558 |

Table 3.2: Means, standard deviations and token numbers of acoustic parameters for different types of stops regardless of the place of articulation. CLS is for closure duration, F0 is for vowel F0 after stops. The scale for closure duration and VOT is millisecond (ms) and for F0, Herz (Hz). PW initial and medial stops are pooled for VOT and vowel F0 after the stops.

## 3.6 Statistical Analysis of the Stop-Rich Data

### *3.6.1 Effect of phonation type*

A general description of the acoustic parameters of different types of stops is presented in Table 3.2. Stops of the same type are grouped together without considering their place of articulation or any other factor in order to see whether the reported characteristics still hold regardless of other factors. The acoustic characteristics we have seen so far are based on the comparison of different types of stops that are produced at the same place of articulation. If there is a durational difference depending on the place of articulation, pooling the samples across places of articulation might result in weakening the saliency of acoustic characteristics for each type of stops. However, the characteristics can still be found irrespective of the place of articulation.

As reported in various phonetic studies, tense stops have the longest closure durations and lax stops have shortest ones. For VOT, aspirated stops are the longest and tense stops

| | Group I | | Group II | |
|---|---|---|---|---|
| | F value | P value | F value | P value |
| Wilks' Lambda | 1072.819 | .001 | 311.521 | .001 |

Table 3.3: F values of MANOVA tests for different types are illustrated. Group I is the stops in all positions and Group II is the stops in word medial position.

are the shortest. For vowel F0 after stops, aspirated stops have the highest F0s and lax stops have the lowest ones.

Samples of the same type are divided into two groups. The first group is for all stops in all positions. The second one is for the stops in word medial position. Since closure duration in word initial position is not used in this statistical test, it is left out of consideration for Group I. Because the stops in the second group are in word medial position, they will be dealt with at greater length where the effect of prosodic position is discussed in Section 3.6.3 on page 71.

One-way MANOVA tests were done for both groups. The results of the tests are shown in Table 3.3. The test for Group I, which ignores closure duration, reveals that the null hypothesis that there is no difference in mean vectors among three groups is rejected. Subsequent post-hoc tests in Table 3.4 show that mean difference can be found in all pairs, lax-tense, tense-aspirated and lax-aspirated stops at the significance level of $\alpha = 0.01$.

MANOVA tests confirm that the three types of stops are statistically different from each other. In the following sections different groupings of samples for other factors will show how the stops are differentiated.

### *3.6.2 Effect of place of articulation*

It has frequently been observed that place of articulation has an influence on the duration of stops. In English, closure durations are found to be the longest in bilabial stops,

| | | Group I | | Group II | |
|---|---|---|---|---|---|
| | | Significance Level | | Significance Level | |
| | | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| Closure Duration | ASP-LAX | na. | na. | √ | √ |
| | ASP-TNS | na. | na. | √ | √ |
| | LAX-TNS | na. | na. | √ | √ |
| VOT | ASP-LAX | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | | |
| F0 | ASP-LAX | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | √ | |

Table 3.4: Games-Howell post-hoc tests for different types of stops. Empty slots represent the acceptance of the null hypothesis at the corresponding significant level. Group I is the stops in all positions and Group II is the stops in word medial position. na. represents 'not available'.

and they are longer in velar stops than in alveolar stops (Port 1977, Luce & Charles-Luce 1985, Crystal & House 1988). On the other hand, VOTs tend to be longer in velar stops than bilabial or alveolar stops (Lisker & Abramson 1967). It is rare to find a study of vowel F0 which is affected by neighbouring onset stops' place of articulation. Pillau (1997) investigated vowel F0 after stops in Sosokhui[10]. She reported that place of articulation for stops did not seem to bring about any noticeable variation of vowel F0. More studies of vowel F0 have been done in relation to the place of articulation for vowel itself, which was dealt with in Section 2.4.3. We have seen from Figure 2.4 that the intrinsic vowel F0 does not preempt stops' segmental F0s.

Lee (1998b) studied effects of the place of articulation on closure durations and VOTs for different types of Korean stops. In order to check the interaction effect between type and place of stop articulation, a two-way ANOVA was done in Lee (1998b). She found that there was not only a main effect of each factor, but also an interaction effect. A one-way

[10]The language is spoken in the Republic of Guinea, West Africa.

ANOVA test was also made to see differences between groups classified by stops' type and place of articulation at the same time. She found a tendency for bilabial stops to have longer closure durations, but no statistically significant difference between bilabial and velar lax stops. She also reported that VOTs are longer in velar stops regardless of type and that bilabials' VOTs are shorter than those of alveolars in tense and aspirated stops but no statistical difference was found between bilabial and alveolar lax stops.

A general description of each acoustic parameter is presented in Table 3.5. It shows that the acoustic characteristics of different types of stops are maintained in every place of articulation.

Instead of using ANOVA, two-way MANOVA tests were done for Group I and II to verify the interaction effect. Group I, as in the previous section, uses only two parameters, VOT and F0. The F-value of Wilks' Lambda for the interaction effect is 19.217 and p-value is less than 0.001, which rejects the null hypothesis that there is no interaction effect. In Group II, the null hypothesis is also rejected (F-value 3.617, $P < 0.001$).

To check whether the 9 groups (3 types and 3 places of articulation) are still differentiated when interaction effects are not taken into account, a one-way MANOVA was performed. Table 3.6 shows the result of the MANOVA test, in which the null hypothesis is once again rejected.

The subsequent post-hoc test compares all possible pairs. Since there are 9 groups, there will be 36 pairs to compare. But what we are interested in is the pairs of groups that have their place of articulation in common. Put differently, whether or not the three types of stops at the same place of articulation are statistically differentiated is more important than whether a type at one place of articulation is different from a type in another place of articulation. Table 3.7, 3.8 and 3.9 respectively represent post-hoc tests for the three types of stops in the same places of articulation. For Group I, the three types are statistically verified to be different. Group II will be discussed in the next section in connection with prosodic positional effects.

| | | LAX | TNS | ASP |
|---|---|---|---|---|
| LABIAL | CLS | 48.3 | 95.2 | 84.5 |
| | | 12.3 | 32.3 | 27.1 |
| | | 43 | 359 | 399 |
| | VOT | 38.7 | 15.3 | 47.0 |
| | | 14.1 | 5.4 | 17.7 |
| | | 535 | 518 | 633 |
| | F0 | 111.7 | 137.4 | 145.1 |
| | | 25.1 | 25.5 | 32.9 |
| | | 535 | 518 | 633 |
| ALVEOLAR | CLS | 32.9 | 93.3 | 74.9 |
| | | 10.6 | 34.2 | 31.2 |
| | | 92 | 318 | 255 |
| | VOT | 31.0 | 14.6 | 49.6 |
| | | 14.0 | 4.6 | 20.1 |
| | | 403 | 517 | 545 |
| | F0 | 117.4 | 135.1 | 141.7 |
| | | 23.0 | 24.1 | 34.2 |
| | | 403 | 518 | 545 |
| VELAR | CLS | 30.7 | 84.4 | 68.7 |
| | | 13.7 | 33.5 | 28.5 |
| | | 33 | 480 | 245 |
| | VOT | 53.5 | 28.8 | 58.4 |
| | | 14.1 | 8.3 | 19.2 |
| | | 786 | 640 | 380 |
| | F0 | 114.0 | 140.0 | 148.7 |
| | | 24.3 | 27.2 | 30.4 |
| | | 786 | 640 | 380 |

Table 3.5: Means, standard deviations and token numbers of acoustic parameters for different types of stops at different places of articulation. The acoustic characteristics of different types of stops is the same as we discussed in Chapter 2.

| | Group I | | Group II | |
|---|---|---|---|---|
| | F value | P value | F value | P value |
| Wilks' Lambda | 382.859 | .001 | 118.414 | .001 |

Table 3.6: F values of MANOVA tests for stops of different types and places of articulation are illustrated.

| | | Group I | | Group II | |
|---|---|---|---|---|---|
| Bilabial | | Significance Level | | Significance Level | |
| | | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| CLS | ASP-LAX | na. | na. | √ | √ |
| | ASP-TNS | na. | na. | √ | √ |
| | LAX-TNS | na. | na. | √ | √ |
| VOT | ASP-LAX | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | | |
| F0 | ASP-LAX | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | |
| | LAX-TNS | √ | √ | √ | |

Table 3.7: Games-Howell post-hoc tests for different types of bilabial stops.

| | | Group I | | Group II | |
|---|---|---|---|---|---|
| Alveolar | | Significance Level | | Significance Level | |
| | | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| CLS | ASP-LAX | na. | na. | √ | √ |
| | ASP-TNS | na. | na. | √ | √ |
| | LAX-TNS | na. | na. | √ | √ |
| VOT | ASP-LAX | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | | |
| F0 | ASP-LAX | √ | √ | | |
| | ASP-TNS | √ | √ | | |
| | LAX-TNS | √ | √ | | |

Table 3.8: Games-Howell post-hoc tests for different types of alveolar stops.

| | | Group I | | Group II | |
|---|---|---|---|---|---|
| Velar | | Significance Level | | Significance Level | |
| | | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| CLS | ASP-LAX | na. | na. | √ | √ |
| | ASP-TNS | na. | na. | √ | √ |
| | LAX-TNS | na. | na. | √ | √ |
| VOT | ASP-LAX | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | | |
| F0 | ASP-LAX | √ | √ | | |
| | ASP-TNS | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | | |

Table 3.9: Games-Howell post-hoc tests for different types of velar stops.

### *3.6.3 Effect of prosodic position*

Having seen the effect of place of articulation on the differentiation of types of stops, we are going to consider the effect of prosodic position as a variable that can affect the three-way distinction of phonation types. Two issues are involved in the effect of prosodic position. The first issue is whether different stops in the same prosodic position can be differentiated by durational and F0 cues. The second one is whether stops in one prosodic position are different from those in another prosodic position.

In relation to the first issue, what we have not tested is the stops in AP (or IP) initial position, so we are focusing on them here. First of all, to see the overall magnitude of the difference in the acoustic parameters in the two prosodic positions, parameters from each position, which is either AP (or IP) initial, or PW medial, are grouped together. Since closure durations were not measured in the initial position, we cannot take them into account for this purpose. The behaviour of VOT and F0 is shown in Table 3.10.

Table 3.10 shows that VOT of lax stops is longer than that of tense stops whereas for word medial position the order is reversed. When the stops are further grouped according to

| | | LAX | TNS | ASP |
|---|---|---|---|---|
| AP (IP) init. | VOT | 46.3 | 18.5 | 64.8 |
| | | 15.6 | 8.7 | 16.4 |
| | | 1556 | 518 | 659 |
| | F0 | 112.6 | 148.1 | 154.8 |
| | | 23.8 | 23.4 | 32.7 |
| | | 1556 | 518 | 659 |
| Word med. | VOT | 19.6 | 21.0 | 40.4 |
| | | 8.3 | 9.5 | 14.4 |
| | | 168 | 1157 | 899 |
| | F0 | 128.0 | 133.0 | 137.4 |
| | | 24.9 | 25.5 | 31.1 |
| | | 168 | 1157 | 899 |

Table 3.10: Means, standard deviations and token numbers of acoustic parameters for different types of stops in different prosodic positions.

their place of articulation, the order reversal is confined to velar stops. Table 3.11, and 3.12 show the durational and F0 statistics between the two prosodic positions.

A two-way MANOVA was done for type and prosodic position. Excluding closure duration, we can find main effects and an interaction effect. P-values for all effects are less than 0.001. On the basis of the two-way MANOVA test, we compared each type of stop in one prosodic position to the same type of stop in the other prosodic position. A one-way MANOVA was performed and we found that each type in AP (or IP) initial position is different from the same type of stop in PW medial position. Another one-way MANOVA test was done for three types of AP initial stops. Post-hoc tests for AP (or IP) initial stops in Table 3.13 show that both acoustic parameters of all types are differentiated.

For PW medial stops, Group II in Table 3.4 on page 67 shows that lax and tense stops are not differentiated by VOT. This is in line with previous observations that VOT of lax stops is reduced in PW medial position.

| | | | LAX | TNS | ASP |
|---|---|---|---|---|---|
| AP (or IP) init. | LABIAL | CLS | na. | na. | na. |
| | | VOT | 40.5 | 14.0 | 63.4 |
| | | | 13.2 | 4.4 | 13.1 |
| | | F0 | 111.0 | 147.1 | 154.9 |
| | | | 25.3 | 23.4 | 37.4 |
| | | **n** | **492** | **159** | **234** |
| | ALVEOLAR | CLS | na. | na. | na. |
| | | VOT | 35.3 | 13.4 | 61.8 |
| | | | 12.9 | 3.3 | 18.7 |
| | | F0 | 113.3 | 145.0 | 154.8 |
| | | | 20.8 | 22.3 | 28.2 |
| | | **n** | **311** | **199** | **290** |
| | VELAR | CLS | na. | na. | na. |
| | | VOT | 54.7 | 29.5 | 73.9 |
| | | | 13.3 | 6.3 | 12.7 |
| | | F0 | 113.3 | 153.0 | 154.8 |
| | | | 24.0 | 24.1 | 33.0 |
| | | **n** | **753** | **160** | **135** |

Table 3.11: Means, standard deviations and token numbers of acoustic parameters for different types of stops in different places of articulation. All stops are positioned AP (or IP) initially.

To assess the role of place of articulation, samples were regrouped according to their type, place of articulation and prosodic position. A one-way MANOVA was performed to see whether three types of AP (or IP) initial stops in each place of articulation can still differentiated. PW medial stops were already shown in Group II, Table 3.7, 3.8 and 3.9.

The fact that the F-value for Wilks' Lambda is 374.626 and p-value is less than 0.001 rejects the null hypothesis. Post-hoc tests compare each pair of groups in AP initial position. The pairs of interest are illustrated in Table 3.14.

| | | | LAX | TNS | ASP |
|---|---|---|---|---|---|
| PW med. | LABIAL | CLS | 48.3 | 95.2 | 84.5 |
| | | | 12.3 | 32.3 | 27.1 |
| | | VOT | 19.3 | 15.9 | 37.5 |
| | | | 8.3 | 5.7 | 12.2 |
| | | F0 | 119.5 | 133.1 | 139.4 |
| | | | 21.8 | 25.2 | 28.4 |
| | | **n** | **43** | **359** | **399** |
| | ALVEOLAR | CLS | 32.9 | 93.3 | 74.9 |
| | | | 10.6 | 34.2 | 31.2 |
| | | VOT | 16.2 | 15.5 | 35.8 |
| | | | 4.9 | 5.1 | 10.5 |
| | | F0 | 131.2 | 128.9 | 126.8 |
| | | | 24.6 | 23.2 | 34.5 |
| | | **n** | **92** | **318** | **255** |
| | VELAR | CLS | 30.7 | 84.4 | 68.7 |
| | | | 13.7 | 33.5 | 28.5 |
| | | VOT | 28.6 | 28.7 | 49.9 |
| | | | 9.8 | 8.9 | 16.8 |
| | | F0 | 130.1 | 135.7 | 145.3 |
| | | | 27.2 | 26.8 | 28.4 |
| | | **n** | **33** | **480** | **245** |

Table 3.12: Means, standard deviations and token numbers of acoustic parameters for different types of stops in different places of articulation. All stops are positioned PW medially.

| | | Significance Level | |
|---|---|---|---|
| | | $\alpha = 0.05$ | $\alpha = 0.01$ |
| VOT | ASP-LAX | √ | √ |
| | ASP-TNS | √ | √ |
| | LAX-TNS | √ | √ |
| F0 | ASP-LAX | √ | √ |
| | ASP-TNS | √ | √ |
| | LAX-TNS | √ | √ |

Table 3.13: Games-Howell post-hoc tests for different types of AP (or IP) initial stops. Because of no availability of closure duration, two parameters were used.

| | | Labial | | Alveolar | | Velar | |
|---|---|---|---|---|---|---|---|
| | | Significance Level | | Significance Level | | Significance Level | |
| | | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| VOT | ASP-LAX | √ | √ | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | √ | √ | √ | √ |
| F0 | ASP-LAX | √ | √ | √ | √ | √ | √ |
| | ASP-TNS | | | √ | √ | | |
| | LAX-TNS | √ | √ | √ | √ | √ | √ |

Table 3.14: Games-Howell post-hoc tests for different types and different place of stop articulation in AP (or IP) initial position.

To confirm whether there is a *domain-initial strengthening effect*[11] in the stop-rich database, we compared each type of stop at each place of articulation in one prosodic position to that in the other prosodic position. Since we do not have closure durations in AP (or IP) initial position, VOT is the only available acoustic parameter for the investigation. A one-way ANOVA and post-hoc tests confirm that except for the velar tense stops, there is a VOT difference between each type of stop in a higher prosodic position and the same type of stop in a lower prosodic position. Even though bilabial and alveolar tense stops showed statistical differences in the two prosodic positions, we could say they are not much affected by the effect. As summarised in (2.8) on page 30, tense stops cannot be expected to have a systematic difference, and indeed none is found in means of VOT for tense stops shown in Table 3.11 and 3.12. While the means of VOT for bilabial and alveolar tense stops are slightly increased, that for velar stop is decreased. So, our result is in accordance with Cho & Keating (2001).

For AP (or IP) initial stops, VOT seems to be the major acoustic cue, since aspirated and tense pairs in bilabial and velar positions do not show a statistical difference at the 0.05 significance level. If we were to use a single parameter to differentiate the three types

[11] Refer to the discussion on page 24.

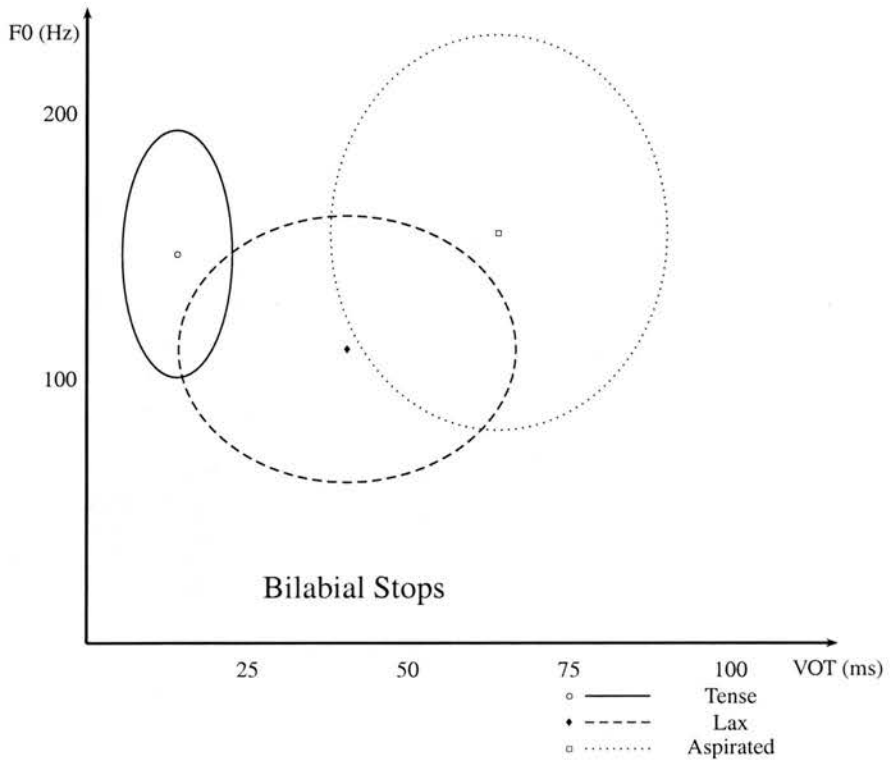



Figure 3.2: Bivariate distribution of bilabial stops in AP initial position. 95.4 % of samples in each type is included in each ellipse.

of stops in Korean, using vowel F0 only would not be much help. However, as seen in Figure 3.2, 3.3 and 3.4, stops can be differentiated when multiple acoustic cues are used. In PW medial position, VOT and vowel F0 are weaker acoustic cues than closure duration as seen in Table 3.9. However, using all acoustic cues together will help to differentiate three types of stops.

### *3.6.4 Effect of preceding phone context*

The effect of preceding phones was investigated in Silva (1992) but only for bilabial stops. We are trying to examine all stops in all places of articulation. We consider the cases where preceding phones are silence, vowel, stop, and sonorant consonants such as nasals and liquids. Stops with silence as a preceding phone context are the same as the

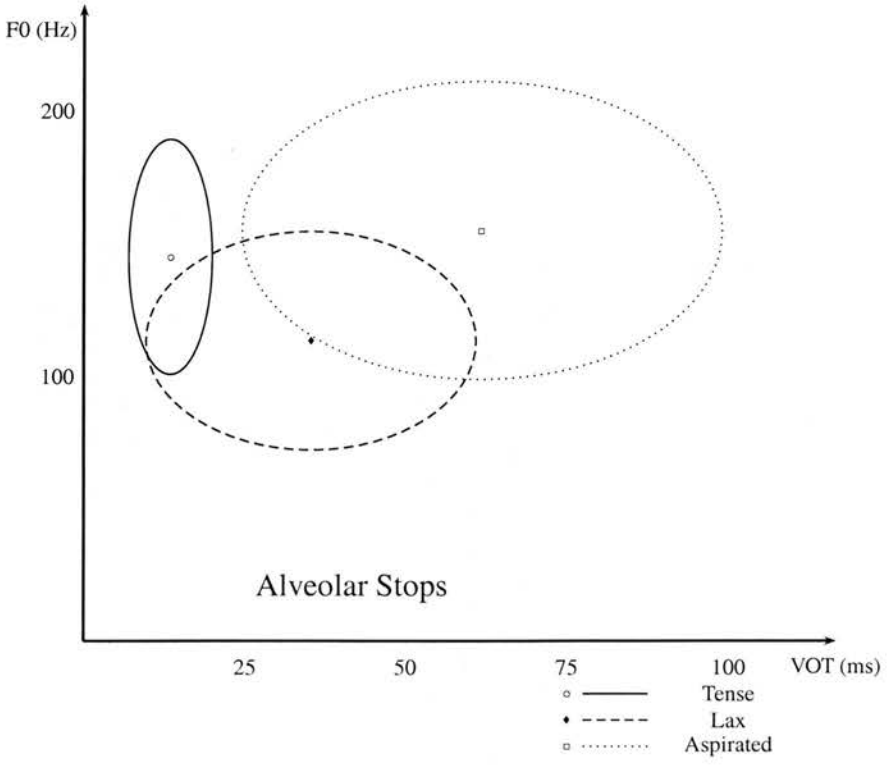



Figure 3.3: Bivariate distribution of alveolar stops in AP initial position. 95.4 % of samples in each type is included in each ellipse.

stops in AP (or IP) initial position. So, the statistical result for the stops preceded by silence will be the same as the one in the previous section.

Since the phonological rule *Post-obstruent Tensing* applies to lax stops preceded by a stop, it is only tense and aspirated stops which can have a stop as a preceding phone context. The general description of acoustic parameters in each preceding phone context is summarised in Table 3.15[12].

As in Silva (1992), stops with sonorant consonants as a preceding context have shorter closure duration than the ones after vowels. A stop context lengthens closure duration because it is realised as a silence in coda position. Compared to aspirated stops, tense

[12]Place of articulation is taken into account when stops are grouped.

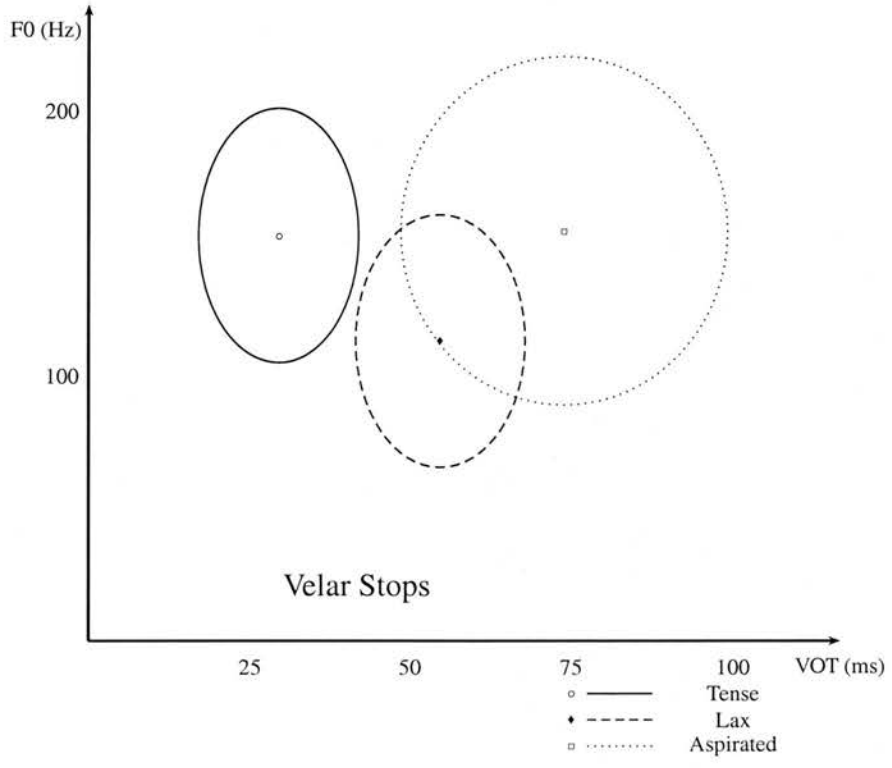



Figure 3.4: Bivariate distribution of velar stops in AP initial position. 95.4 % of samples in each type is included in each ellipse.

stops have longer closure duration when they have a preceding stop. Stops do not seem to vary much in VOT according to the type of the previous phone. It looks obvious that there is an influence of preceding phone contexts on acoustic parameters. So, MANOVA tests to see whether stops of the same type can be differentiated according to their preceding phone contexts are not needed. More interesting is to compare one type of stop in a particular preceding phone context to other types of stops in the same preceding phone context. For example, bilabial lax stops with a vowel as a preceding phone context are different from the other two types in the same preceding phone context. A one-way MANOVA was done and followed by post-hoc tests. The result is summarised in Table 3.16.

| | | LAX | | | TNS | | | ASP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Stop | Vowel | Son | Stop | Vowel | Son | Stop | Vowel | Son |
| LAB | CLS | na. | 48.2 | nc. | 121.5 | 89.7 | 66.9 | 118.3 | 87.8 | 56.2 |
| | | na. | 12.3 | nc | 28.9 | 24.1 | 19.9 | 26.9 | 17.8 | 14.7 |
| | VOT | na. | 19.2 | nc. | 14.3 | 17.2 | 15.7 | 34.5 | 38.3 | 37.1 |
| | | na. | 8.3 | nc | 3.4 | 7.1 | 4.8 | 12.2 | 11.4 | 13.7 |
| | F0 | na. | 119.5 | nc. | 128.2 | 138.6 | 129.6 | 132.2 | 139.0 | 144.8 |
| | | na. | 21.8 | nc | 24.7 | 25.8 | 22.7 | 22.6 | 30.2 | 26.4 |
| | **n** | **na.** | **43** | **nc.** | **120** | **159** | **80** | **61** | **238** | **100** |
| ALV | CLS | na. | 34.5 | 20.5 | 116.4 | 88.6 | 64.0 | 106.5 | 82.0 | 49.7 |
| | | na. | 10.0 | 5.9 | 30.8 | 22.2 | 20.1 | 30.8 | 23.9 | 16.2 |
| | VOT | na. | 16.5 | 17.6 | 14.1 | 16.7 | 16.3 | 34.9 | 36.7 | 35.4 |
| | | na. | 4.8 | 6.1 | 4.6 | 4.8 | 5.7 | 10.3 | 10.6 | 10.6 |
| | F0 | na. | 132.8 | 119.4 | 126.9 | 132.4 | 128.8 | 134.5 | 120.1 | 128.7 |
| | | na. | 24.0 | 27.0 | 22.0 | 25.7 | 22.4 | 35.0 | 38.3 | 29.2 |
| | **n** | **na.** | **81** | **11** | **140** | **80** | **98** | **59** | **95** | **101** |
| VEL | CLS | na. | 34.4 | 16.9 | 117.0 | 79.6 | 56.3 | 98.2 | 70.6 | 49.1 |
| | | na. | 13.0 | 4.3 | 38.0 | 20.4 | 19.6 | 31.9 | 18.6 | 17.1 |
| | VOT | na. | 29.0 | 27.2 | 27.1 | 29.8 | 27.4 | 54.5 | 48.7 | 48.3 |
| | | na. | 10.8 | 4.0 | 8.6 | 9.7 | 6.2 | 22.5 | 14.7 | 14.2 |
| | F0 | na. | 128.3 | 136.5 | 130.6 | 139.5 | 132.3 | 152.7 | 147.7 | 138.5 |
| | | na. | 26.8 | 29.9 | 27.5 | 26.9 | 24.5 | 24.5 | 24.2 | 32.8 |
| | **n** | **na.** | **26** | **7** | **123** | **257** | **100** | **57** | **93** | **95** |

Table 3.15: Means, standard deviations and token numbers of acoustic parameters for stops with preceding phones classified by natural classes. na. and nc. represent 'not available' and 'no case found', respectively.

As seen in Table 3.16, vowel F0 on its own does not differentiate the three types of stops. Since segmental F0 is not salient in PW medial position, this result is somewhat expected. However the problem is not confined to vowel F0. No single acoustic parameter can by itself differentiate all types of stops at the significance level of 0.05. This establishes the need to use multiple cues for differentiation of the three types of stops in any position.

In the next section we are going to simulate the effect of the speech rate on VOT.

| | | LAX-TNS | | | TNS-ASP | | | ASP-LAX | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Stop | Vowel | Son | Stop | Vowel | Son | Stop | Vowel | Son |
| LAB | CLS | na. | √√ | na. | | | √ | na. | √√ | na. |
| | VOT | na. | | na. | √√ | √√ | √√ | na. | √√ | na. |
| | F0 | na. | √√ | na. | | | √√ | na. | √√ | na. |
| ALV | CLS | na. | √√ | √√ | | | √√ | na. | √√ | √√ |
| | VOT | na. | | | √√ | √√ | √√ | na. | √√ | √√ |
| | F0 | na. | | | | | | na. | | |
| VEL | CLS | na. | √√ | √√ | | √ | | na. | √√ | √√ |
| | VOT | na. | | | √√ | √√ | √√ | na. | √√ | √√ |
| | F0 | na. | | | √√ | | | na. | | |

Table 3.16: Games-Howell post-hoc tests for different types of stops with preceding phone contexts. Two significance levels 0.05 and 0.01 were used in each pair. Empty slots represent the acceptance of the null hypothesis at both significance levels and one check represents the rejection at the lowest significance level, which is 0.05.

### *3.6.5 Simulation of speech rate*

To see whether the similar relationship holds up between speech rate and VOT that we have discussed in Section 2.4.3 on page 33, we are going to examine the correlation between syllable durations and VOTs. The stop-rich database was not designed with consideration of speech rate in mind. The speech rate of the database does not vary much across the speakers. Even though the database does not have much variation of speech rate, if we can find a similar pattern to that which Kessinger & Blumstein (1998) found for syllable duration and VOT, it will be reasonable to suppose the relationship between VOT and speech rate is maintained in Korean stops[13].

As revealed in Kessinger & Blumstein (1998), if speech rate slows, syllable duration increases and vowel duration is also increased. So, the syllable and vowel durations are positively correlated. Likewise, VOT increases when syllable duration is increased. Table 3.17 shows Pearson's correlation coefficients between vowel and syllable durations,

[13]We assume that speech rate is a major influence on syllable duration.

| | LAX | TNS | ASP |
|---|---|---|---|
| | syllable | syllable | syllable |
| VOWEL | .739 | .919 | .670 |
| VOT | .366 | .212 | .477 |

Table 3.17: Correlation coefficients between VOT and syllable duration, and between vowel and syllable duration for three types of stops.

| | | LAX | TNS | ASP |
|---|---|---|---|---|
| | | syllable | syllable | syllable |
| AP init. | VOWEL | .782 | .932 | .736 |
| | VOT | .325 | .070 | .463 |
| PW med. | VOWEL | .898 | .915 | .815 |
| | VOT | .230 | .288 | .404 |

Table 3.18: Correlation coefficients between VOT and syllable duration, and between vowel and syllable duration for three types of stops in each prosodic position.

and VOTs and syllable durations for the three types of stops. There is a high correlation between vowel and syllable duration in each type of stop. And also we can find a correlation between VOTs and syllable durations.

We can also see how much VOT and syllable duration are correlated and whether correlation coefficients are different depending on types of stops. In Kessinger & Blumstein (1997), a stop category that has a short VOT is not affected much by speech rate whereas a category that has longer VOT is much more affected by it. In other words, longer VOTs increase or decrease in the same way as the syllable duration does due to the speech rate. A similar relationship can be found in Korean as shown in Table 3.17. This also applies to stops in AP initial position. Stops in PW medial show a slightly different profile because VOTs of lax and aspirated stops are reduced, which is an effect of prosodic position, not of speech rate.

| | | | LAX | TNS | ASP |
|---|---|---|---|---|---|
| | | | syllable | syllable | syllable |
| AP ini. | LAB | VOWEL | .825 | .974 | .763 |
| | | VOT | .271 | -.129 | .398 |
| | ALV | VOWEL | .892 | .993 | .718 |
| | | VOT | .127 | -.096 | .495 |
| | VEL | VOWEL | .815 | .940 | .781 |
| | | VOT | .421 | .004 | .525 |
| PW med. | LAB | VOWEL | .942 | .972 | .857 |
| | | VOT | .477 | .149 | .369 |
| | ALV | VOWEL | .955 | .973 | .912 |
| | | VOT | -.110 | .081 | .066 |
| | VEL | VOWEL | .779 | .925 | .771 |
| | | VOT | .635 | .487 | .682 |

Table 3.19: Correlation coefficients between VOT and syllable duration, and between vowel and syllable duration for stops in different prosodic positions and in different places of articulation.

Further grouping of stops depending on their place of articulation shows in Table 3.19 that stops in all positions follow the same pattern as AP initial stops, except for alveolar aspirated stops. Correlation coefficients between VOT and syllable duration of alveolar stops are very small. This is not limited to tense stops but to all types of alveolar stops. It is probably because the tip of the tongue, which is the articulator of alveolar stops, moves faster than other articulators in PW medial position.

Speech rate simulation with syllable duration shows behaviour similar to what Kessinger & Blumstein (1997) found for Thai. A difference between this simulation and their work is that there was little overlapping between the short VOT category and other two in Thai, whereas we found overlapping between lax stops and other two types[14]. It is hard to say that these overlapping areas are introduced directly by speech rate but it seems to be obvious that acoustic parameters of Korean stops have greater variation than those of

[14]The overlapping areas for Korean stops can be seen in the three figures in Section 3.6.3.

Thai stops. However, as we have seen, the overlapping in Korean stops can be eventually reduced by considering other acoustic parameters.

## 3.7 Statistical Analysis of the KAIST Data

The purpose of doing statistical analysis with the KAIST speech database is to see whether the same acoustic characteristics as we have seen in the stop-rich database can still be found in the presence of contextual variation. If the acoustic characteristics are still maintained in the KAIST data, the characteristics are very robust and it can be regarded as invariant characteristics.

As we described on page 57, among 150 speakers, 433 utterances of 85 speakers were used for labelling and statistical tests of stops. Because of the large number of speakers, vowel F0 after stops varies more widely than in the stop-rich database recorded by 4 speakers. Unlike the stop-rich database, the KAIST database contains utterances by female speakers, so there are inter-gender pitch range differences. For these reasons, vowel F0 after stops in the KAIST speech data were used for statistical tests after two kinds of normalisation, which had been already done by Jang (2000b). In the next section, we will briefly summarise how the two types of normalisation were performed[15]

In 433 sentences, a total of 1589 stop sounds were found, among which 797 stops were from female speakers and 792 stops from male speakers. We considered using three prosodic positions in the analysis, PW (or AP) initial and PW medial and resyllabified syllable initial positions, but there were only 29 resyllabified stops in the data, so only two prosodic positions, PW initial, and PW medial were used as prosodic positional factor levels.

[15]The summarisation of the normalisation refers to Jang (2000b:120).

### 3.7.1 Normalisation

*Declination normalisation*

It is known that declination of F0 during the course of an utterance can be observed cross-linguistically (Pierrehumbert 1979). For Korean, Koo (1986) conducted two experiments concerning declination phenomena. In the first experiment, he examined the interaction between declination and prominence and found that the F0 of a later emphasised phrase never exceeded that of an earlier one. He concluded that this tendency could be attributed to declination. He also investigated pitch range and the slope of declination in relation to length of a sentence. He found that pitch range remained constant without being affected by sentence length. For the slope of declination, if the sentence is long, the slope stayed flat except in the first and the last part of the sentence. In other words, if the sentence is short, the slope would be steep. Koo's experiments confirm that declination can be found in Korean.

To minimise the effect of declination, the F0 contour of an utterance should be normalised when considering F0 as a cue for differentiating the three types of stops. The method Jang (2000b) used for declination normalisation was a simple linear regression. First, he eliminated abrupt bumps and reduced the size of microprosodic perturbation using three-frame median smoothing. Second, using linear regression analysis, he estimated the slope of declination of each utterance token and found the intercept of the regression line at the beginning of the sentence. And then, for each frame, he calculated a distance between the regression line and the horizontal line at the level of the initial intercept. Finally, the distance was added to the original F0 value of each frame to get a normalised value for declination. A schematic representation of F0 normalisation is illustrated in Figure 3.5 and as a real speech example, we reproduced Jang's graphical illustration of declination normalisation in Figure 3.6.

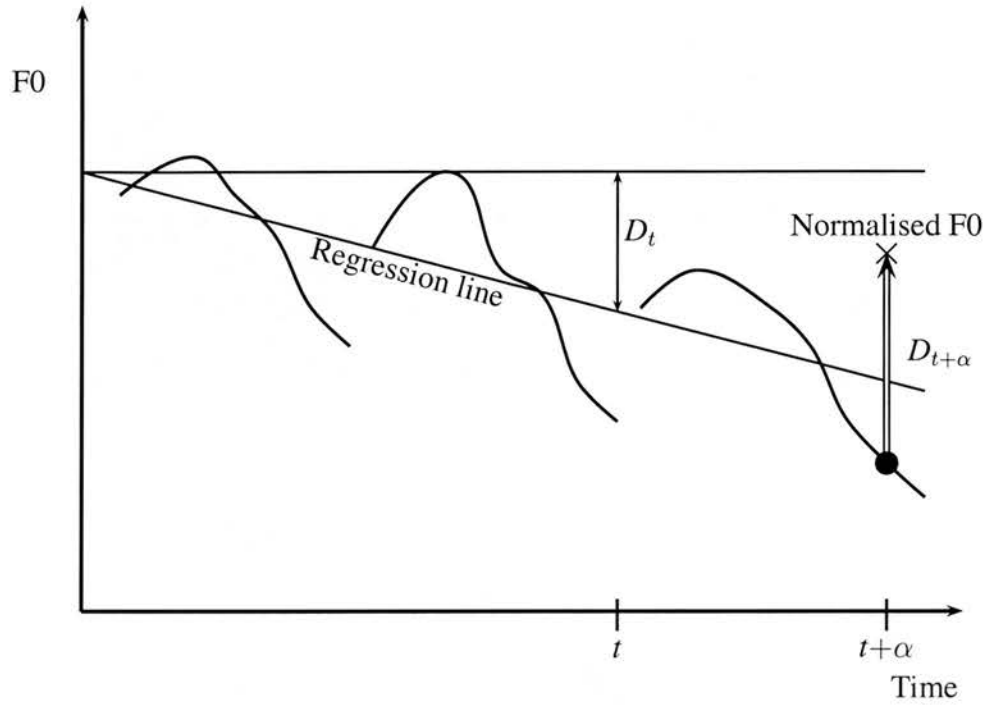



Figure 3.5: Schematic representation of declination normalisation. $D_t$ is a distance between the two lines at time, $t$. The distance at time, $t+\alpha$ is added to the original F0 value to cancel out the declination effect.

*Speaker normalisation*

To reduce the magnitude of inter-speaker and inter-gender F0 variation, Jang normalised F0 for each utterance with respect to a global pitch range. The global range was obtained by calculating the mean and standard deviation of all the voiced vowel tokens spoken by the male speakers in the database. The mean and standard deviation for the global pitch were approximately 130 Hz and 25 Hz[16]. Jang chose 2 standard deviations for the range of normalisation and adjusted F0 values to fall between 80 Hz and 180 Hz by using the algorithm which I quote below (Jang 2000b:124).

[16]The exact values are 132 Hz and 24.52 Hz, respectively.

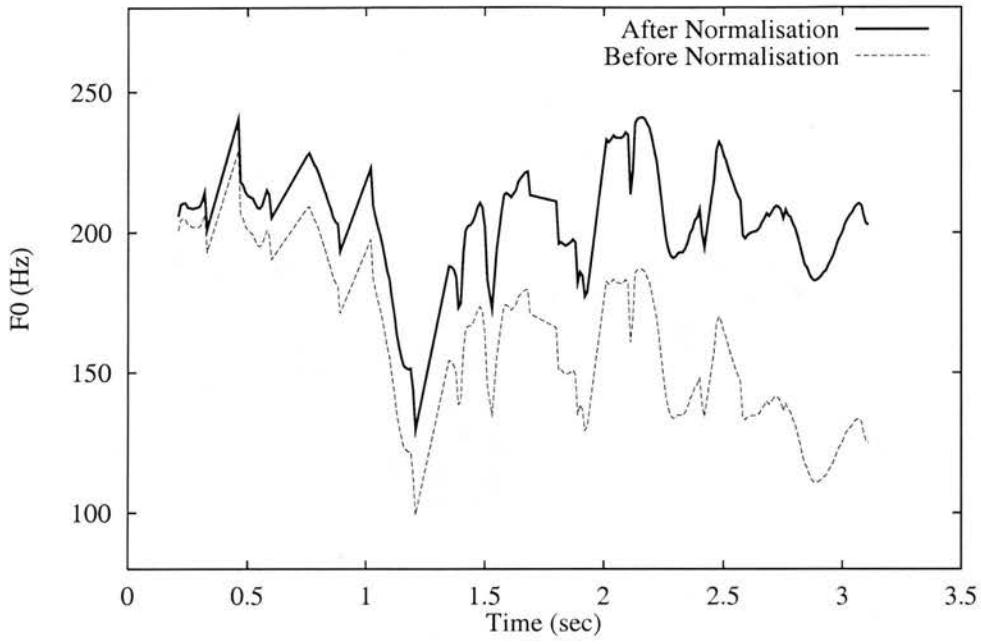



Figure 3.6: F0 declination normalisation of an utterance in our speech data. (Reproduction from Jang (2000b:124))

$$F0[i]_{normalised} = \left( \left( \frac{F0[i] - \mu_s}{4\sigma_s} + 0.5 \right) (High - Low) \right) + Low$$

where $\mu_s$ and $\sigma_s$ stand for the *mean* and *standard deviation* of F0 value for the utterance being normalised and $Low$ and $High$ specify the designated pitch range, i.e. they are equal to 80 Hz and 180 Hz respectively for this case.

After the two types of normalisation, F0, together with other two acoustic parameters is used for statistical description and tests of different types of stops.

### *3.7.2 Effect of phonation type*

As in Section 3.6.1, each type of stop is grouped regardless of place of articulation. The mean and standard deviation of each type is presented in Table 3.20. Compared to the

| | | LAX | TNS | ASP |
|---|---|---|---|---|
| CLS | Mean | 29.5 | 80.1 | 58.3 |
| | St. Dv. | 23.9 | 29.8 | 28.6 |
| VOT | Mean | 31.6 | 16.5 | 37.5 |
| | St. Dv. | 19.7 | 8.4 | 17.3 |
| F0 | Mean | 118.8 | 137.7 | 150.3 |
| | St. Dv. | 22.4 | 20.9 | 20.8 |
| n | | 769 | 442 | 324 |

Table 3.20: Means, standard deviations and token numbers of acoustic parameters for different types of stops in the KAIST database.

durations and vowel F0 in Table 3.2, the magnitude of each acoustic parameter in the KAIST speech data is smaller. However, the acoustic characteristics we have seen are still maintained under contextual variation.

In the stop-rich database, the number of tokens was different for closure duration because we ignored AP (or IP) initial stops' closure durations. However, in the KAIST speech database, the initial stops are basically PW initial and possibly AP initial. The closure durations for stops in these prosodic position are not mixed up with a silence introduced by a pause which signals a start of an IP boundary.

A MANOVA test confirms that there is a difference among the three types of stops at the significance level of 0.001. Subsequent post-hoc tests also confirms that differences are found in all pairs in all acoustic parameters at both significance levels, $\alpha = 0.05$ and 0.01.

### *3.7.3 Effect of place of articulation*

Each type of stop was further divided into three groups, according to the stops' places of articulation. We can see that the acoustic characteristics found in the stop-rich database are still maintained in the KAIST speech database, as shown in Table 3.21.

| | | LAX | TNS | ASP |
|---|---|---|---|---|
| LABIAL | CLS | 44.5 | 76.1 | 55.9 |
| | | 28.2 | 36.2 | 28.6 |
| | VOT | 32.0 | 11.7 | 37.9 |
| | | 16.6 | 4.8 | 18.0 |
| | F0 | 125.7 | 142.9 | 151.7 |
| | | 17.2 | 21.4 | 21.2 |
| | **n** | **104** | **163** | **147** |
| ALVEOLAR | CLS | 30.9 | 84.3 | 57.5 |
| | | 18.3 | 27.2 | 35.4 |
| | VOT | 21.0 | 13.1 | 33.5 |
| | | 12.5 | 5.5 | 14.8 |
| | F0 | 119.6 | 141.9 | 145.3 |
| | | 18.9 | 19.3 | 23.4 |
| | **n** | **324** | **61** | **83** |
| VELAR | CLS | 23.5 | 81.9 | 62.9 |
| | | 24.9 | 24.5 | 20.5 |
| | VOT | 41.5 | 21.1 | 40.6 |
| | | 20.9 | 8.7 | 17.8 |
| | F0 | 116.0 | 132.7 | 152.6 |
| | | 26.1 | 19.8 | 17.0 |
| | **n** | **341** | **218** | **94** |

Table 3.21: Means, standard deviations and token numbers of acoustic parameters for different types of stops at different places of articulation in the KAIST speech database.

A two-way MANOVA test was done to check if there is an interaction effect between type and place of articulation. The F-value of Wilks' Lambda for the interaction effect is 10.004. The null hypothesis is rejected at the 0.01 significance level, which means there is an interaction. A one-way MANOVA test was done to see whether groups of stops can be statistically differentiated. The F-value is 92.785 and the null hypothesis is also rejected. The result of post-hoc tests is shown in Table 3.22.

In bilabial and velar positions, aspirated and lax stops are not statistically differentiated by means of VOT. Aspirated and tense stops at the alveolar position are not differentiated

| | | Labial | | Alveolar | | Velar | |
|---|---|---|---|---|---|---|---|
| | | Significance Level | | Significance Level | | Significance Level | |
| | | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| CLS | ASP-LAX | √ | | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | √ | √ | √ | √ |
| VOT | ASP-LAX | | | √ | √ | | |
| | ASP-TNS | √ | √ | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | √ | √ | √ | √ |
| F0 | ASP-LAX | √ | √ | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | | | √ | √ |
| | LAX-TNS | √ | √ | √ | √ | √ | √ |

Table 3.22: Games-Howell post-hoc tests for different types of stops at different places of articulation. An empty slot means that the null hypothesis can not be rejected at the corresponding significance level.

by means of vowel F0, either. However, as we observed in the stop-rich database, other acoustic cues can differentiate the three types of stops.

### *3.7.4 Effect of prosodic position*

Stops in two prosodic positions, PW (or AP) initial and PW medial were investigated. As we mentioned above, even though we marked resyllabified syllable initial stops, they did not participate in the statistics because the tokens were too few to use a sample mean and standard deviation of each type of stop.

Table 3.23 shows a statistical description of acoustic parameters of each type of stop in the two prosodic positions. The characteristics observed for the stop-rich database are maintained for stops in PW (or AP) initial position. Further grouping of the stops according to their place of articulation shows two minor differences. First, as shown in Table 3.24, the mean closure duration of aspirated labial stops is almost the same as that

| | | LAX | TNS | ASP |
|---|---|---|---|---|
| PW (AP) initial | CLS | 28.1 | 70.5 | 47.1 |
| | | 27.1 | 35.3 | 36.4 |
| | VOT | 38.7 | 13.8 | 43.2 |
| | | 19.2 | 7.2 | 19.8 |
| | F0 | 114.7 | 140.4 | 156.0 |
| | | 21.2 | 22.3 | 21.9 |
| | **n** | **501** | **188** | **97** |
| PW medial | CLS | 32.0 | 87.2 | 63.2 |
| | | 15.9 | 22.5 | 23.0 |
| | VOT | 16.3 | 18.6 | 35.1 |
| | | 8.2 | 8.6 | 15.6 |
| | F0 | 126.7 | 135.8 | 147.9 |
| | | 22.4 | 19.6 | 19.9 |
| | **n** | **268** | **254** | **227** |

Table 3.23: Means, standard deviations and token numbers of acoustic parameters for different types of stops in different prosodic positions.

of lax stops. Second, as seen in Table 3.26, PW medial tense stops in velar position have longer VOT than lax stops in the same place of articulation.

A two-way MANOVA was performed for type and prosodic positions. Main effects and an interaction effect are found. P-values for all effects are less than 0.001. As for the stop-rich database, we compared the three types of stops in PW (or AP initial) position to those in PW medial position in order to check whether the *domain-initial strengthening effect* can still be found in the KAIST speech database. A one-way MANOVA was done and post-hoc tests were subsequently done. Except for closure duration of the lax stops, other acoustic parameters of other types of stops are different in PW (or AP) initial and PW medial prosodic position.

Post-hoc tests indicate that different types of stops in PW (or AP) initial position can be differentiated in all acoustic parameters except for VOT of lax and aspirated stops. Table 3.25 shows the results of post-hoc tests on a more detailed regrouping of stops with

| | | | LAX | TNS | ASP |
|---|---|---|---|---|---|
| PW (or AP) init. | LABIAL | CLS | 45.9 | 71.5 | 45.3 |
| | | | 30.5 | 37.3 | 34.1 |
| | | VOT | 35.9 | 10.8 | 42.3 |
| | | | 14.6 | 3.9 | 20.4 |
| | | F0 | 121.5 | 142.0 | 157.3 |
| | | | 13.4 | 22.6 | 20.4 |
| | | **n** | **83** | **124** | **56** |
| | ALVEOLAR | CLS | 30.7 | 59.1 | 43.6 |
| | | | 23.6 | 32.5 | 46.9 |
| | | VOT | 27.8 | 15.3 | 39.5 |
| | | | 15.2 | 8.7 | 15.9 |
| | | F0 | 119.3 | 142.8 | 148.7 |
| | | | 19.0 | 22.2 | 25.8 |
| | | **n** | **141** | **14** | **26** |
| | VELAR | CLS | 21.5 | 71.4 | 59.7 |
| | | | 25.1 | 30.8 | 19.1 |
| | | VOT | 46.9 | 20.8 | 53.0 |
| | | | 18.9 | 8.0 | 21.4 |
| | | F0 | 110.2 | 135.7 | 163.8 |
| | | | 23.2 | 21.3 | 17.6 |
| | | **n** | **277** | **50** | **15** |

Table 3.24: Means, standard deviations and token numbers of acoustic parameters for different types of stops in different places of articulation. All stops are positioned PW (or AP) initially.

place of articulation. Judged from the post-hoc tests, closure duration seems to be a weak acoustic cue specially for all types of alveolar stops while vowel F0 and VOT look quite strong because they differentiate stops in more places and more types of stops.

Other post-hoc tests for stops in PW medial position reveal that the three types of stops can be differentiated in all acoustic parameters at the 0.01 significance level. Table 3.26 shows a general statistical description of acoustic parameters of PW medial stops that were grouped according to their types and place of articulation. The result of post-hoc tests is presented in Table 3.27. Compared to the post-hoc tests for PW (or AP) initial

| | | Labial | | Alveolar | | Velar | |
|---|---|---|---|---|---|---|---|
| | | Significance Level | | Significance Level | | Significance Level | |
| | | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| CLS | ASP-LAX | | | | | √ | √ |
| | ASP-TNS | √ | √ | | | | |
| | LAX-TNS | √ | √ | | | √ | √ |
| VOT | ASP-LAX | | | √ | | | |
| | ASP-TNS | √ | √ | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | √ | √ | √ | √ |
| F0 | ASP-LAX | √ | √ | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | | | √ | √ |
| | LAX-TNS | √ | √ | √ | | √ | √ |

Table 3.25: Games-Howell post-hoc tests for different types of PW (or AP) initial stops in the KAIST speech database.

stops above, they show a strong indication that closure duration is better than other two acoustic cues for differentiating the three types of stops in all places of articulation. In particular, vowel F0 is not a good acoustic cue in this position. These two post-hoc tests show that no single acoustic cue is enough to differentiate the three types of stops.

Finally, we tested the *domain-initial strengthening effect* for stops grouped by type and place of articulation. Each type of PW (or AP) initial stop at each place of articulation was compared to the same stop in PW medial position. We did not find any difference for closure duration. For VOT, all lax stops in three places of articulation show statistical differences but the other two types of stops do not show any differences between the two prosodic positions. For vowel F0, bilabial lax stops in PW (or AP) initial position differ from those in PW medial position only at the 0.05 significance level. F0 for velar lax stops rejects the null hypothesis at the 0.01 significance level[17]. Other pairs of stops do

[17]Vowel F0 is not related to the *domain-initial strengthening effect*, but we proceeded to examine the parameter because we did not find systematic difference in the other two parameters.

| | | | LAX | TNS | ASP |
|---|---|---|---|---|---|
| PW med. | LABIAL | CLS | 39.1 | 90.8 | 62.5 |
| | | | 16.3 | 28.1 | 22.5 |
| | | VOT | 16.5 | 14.5 | 35.2 |
| | | | 15.2 | 6.2 | 15.9 |
| | | F0 | 142.3 | 145.8 | 148.3 |
| | | | 20.8 | 17.0 | 21.0 |
| | | **n** | **21** | **39** | **91** |
| | ALVEOLAR | CLS | 31.0 | 91.8 | 63.9 |
| | | | 12.8 | 20.4 | 26.8 |
| | | VOT | 15.7 | 12.5 | 30.7 |
| | | | 6.0 | 3.9 | 13.5 |
| | | F0 | 119.9 | 141.6 | 143.8 |
| | | | 18.8 | 18.6 | 22.4 |
| | | **n** | **183** | **47** | **57** |
| | VELAR | CLS | 32.6 | 85.1 | 63.5 |
| | | | 22.1 | 21.5 | 20.8 |
| | | VOT | 18.0 | 21.2 | 38.2 |
| | | | 10.2 | 8.9 | 16.1 |
| | | F0 | 140.9 | 131.9 | 150.4 |
| | | | 23.2 | 19.3 | 16.1 |
| | | **n** | **64** | **168** | **79** |

Table 3.26: Means, standard deviations and token numbers of acoustic parameters for different types of stops in different places of articulation. All stops are positioned PW medially.

| | | Labial | | Alveolar | | Velar | |
|---|---|---|---|---|---|---|---|
| | | Significance Level | | Significance Level | | Significance Level | |
| | | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ | $\alpha = 0.05$ | $\alpha = 0.01$ |
| CLS | ASP-LAX | √ | √ | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | √ | √ | √ |
| | LAX-TNS | √ | √ | √ | √ | √ | √ |
| VOT | ASP-LAX | √ | √ | √ | √ | √ | √ |
| | ASP-TNS | √ | √ | √ | √ | √ | √ |
| | LAX-TNS | | | √ | √ | | |
| F0 | ASP-LAX | | | √ | √ | | |
| | ASP-TNS | | | | | √ | √ |
| | LAX-TNS | | | √ | | | |

Table 3.27: Games-Howell post-hoc tests for different types of PW medial stops.

not show any statistical difference. These results suggest that the *domain-initial strengthening effect* can be found in the controlled speech data, but not in the data with contextual variability.

### 3.7.5 *Effect of preceding phone context*

Stops are regrouped according to their preceding phone contexts. Since not much difference was found in two prosodic positions, prosodic position is excluded in the grouping. Due to the data imbalance, if a group has fewer than 7 tokens, this group is excluded in statistics[18]. Stops in Utterance or IP initial position have neither a previous phone context nor closure duration distinguishable from a pause. Descriptive statistics of stops with a preceding phone context are presented in Table 3.28 and those of Utterance or IP initial stops are in Table 3.29.

As in the stop-rich database, closure duration is reduced in stops that have a sonorant as a preceding phone context. However, closure duration of tense stops is longer than for the other two types of stops in the same preceding phone context. For VOT, stops in various phone contexts do not seem to have any consistent tendency. For example, the mean VOT for intervocalic aspirated bilabial stops is not longer than those for lax stops with a vowel as a preceding phone context. The observation applies to velar stops in a sonorant phone context. Nevertheless, F0 is in accordance with the general description that aspirated stops have higher F0 than lax stops. The Post-hoc tests in Table 3.30 recapitulate the usefulness of each acoustic cue.

Post-hoc tests for Utterance or IP initial stops were done only for pairs of bilabial stops and lax and aspirated pairs of alveolar stops because there was not enough data for other groups. The results are that an aspirated-tense pair of bilabial stops can be differentiated by VOT at the 0.01 significant level and a lax-tense pair of bilabial stops is also differentiated at the 0.05 significance level. But no difference can be found for vowel F0 in any

[18]Since the smallest token number of a group in the stop-rich data was 7, we use that number as a criterion.

| | | LAX | | | TNS | | | ASP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Stop | Vowel | Son | Stop | Vowel | Son | Stop | Vowel | Son |
| LAB | CLS | | 49.5 | 32.8 | 112.6 | 85.8 | 71.4 | | 69.2 | 44.0 |
| | | | 23.5 | 17.5 | 32.1 | 25.3 | 26.5 | | 19.6 | 19.5 |
| | VOT | | 31.5 | 28.0 | 13.2 | 21.1 | 10.4 | | 32.6 | 37.4 |
| | | | 17.6 | 13.9 | 6.7 | 5.2 | 3.6 | | 15.9 | 21.0 |
| | F0 | | 126.8 | 127.2 | 155.2 | 146.3 | 141.0 | | 149.0 | 160.2 |
| | | | 18.0 | 15.1 | 17.9 | 13.8 | 23.9 | | 20.0 | 21.3 |
| | **n** | | **66** | **24** | **15** | **79** | **55** | | **89** | **38** |
| ALV | CLS | | 33.1 | 27.1 | 106.2 | 85.2 | | 108.1 | 77.2 | 44.2 |
| | | | 13.3 | 16.5 | 14.2 | 22.1 | | 21.6 | 20.4 | 15.9 |
| | VOT | | 19.2 | 21.4 | 11.6 | 13.4 | | 35.0 | 34.2 | 31.3 |
| | | | 11.4 | 12.7 | 2.2 | 6.2 | | 7.3 | 18.0 | 13.1 |
| | F0 | | 121.4 | 121.0 | 154.8 | 140.7 | | 132.6 | 145.9 | 149.2 |
| | | | 18.6 | 14.4 | 7.5 | 19.6 | | 15.8 | 25.0 | 17.6 |
| | **n** | | **251** | **44** | **9** | **45** | | **8** | **34** | **29** |
| VEL | CLS | | 37.1 | 33.8 | 110.3 | 85.8 | 62.6 | | 64.5 | 46.4 |
| | | | 20.2 | 20.2 | 21.4 | 22.5 | 15.9 | | 19.6 | 22.0 |
| | VOT | | 33.8 | 44.2 | 19.3 | 21.6 | 20.3 | | 40.4 | 41.7 |
| | | | 21.7 | 23.2 | 8.6 | 9.1 | 7.6 | | 18.4 | 11.3 |
| | F0 | | 125.5 | 119.4 | 146.2 | 134.5 | 123.9 | | 152.3 | 154.8 |
| | | | 25.2 | 16.9 | 18.6 | 18.6 | 20.0 | | 17.4 | 12.9 |
| | **n** | | **139** | **75** | **17** | **146** | **55** | | **85** | **9** |

Table 3.28: Means, standard deviations and token numbers of acoustic parameters for stops with preceding phones classified by natural classes. Groups of fewer than 7 tokens remain empty.

pair of stops. A lax-aspirated pair of alveolar stops has no statistical difference for both VOT and vowel F0.

## 3.8 Acoustic Characteristics of Korean Stops Revisited: Summary

So far, we have done statistical tests for differentiation of stops in two databases. In the stop-rich database, the reported acoustic characteristics are well exhibited in the data, namely that tense stops have longer closure durations and shorter VOT whereas aspirated

| | | LAX | TNS | ASP |
|---|---|---|---|---|
| LABIAL | VOT | 39.4 | 11.7 | 50.0 |
| | | 11.8 | 3.6 | 14.3 |
| | F0 | 115.7 | 117.2 | 141.6 |
| | | 17.9 | 29.6 | 21.6 |
| | **n** | **8** | **14** | **15** |
| ALVEOLAR | VOT | 38.3 | | 35.3 |
| | | 11.3 | | 12.7 |
| | F0 | 97.9 | | 142.9 |
| | | 16.4 | | 33.1 |
| | **n** | **24** | | **12** |
| VELAR | VOT | 48.5 | | |
| | | 14.9 | | |
| | F0 | 103.2 | | |
| | | 26.8 | | |
| | **n** | **124** | | |

Table 3.29: Means, standard deviations and token numbers of acoustic parameters for Utterance or IP initial stops in the KAIST speech database. Empty slots are the groups that are either no token or few than 7 tokens found.

| | | LAX-TNS | | | TNS-ASP | | | ASP-LAX | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Stop | Vowel | Son | Stop | Vowel | Son | Stop | Vowel | Son |
| LAB | CLS | na. | √√ | √√ | na. | √√ | √√ | na. | √√ | |
| | VOT | na. | √√ | √√ | na. | √√ | √√ | na. | | |
| | F0 | na. | √√ | | na. | | √ | na. | √√ | √√ |
| ALV | CLS | na. | √√ | na. | | | | na. | √√ | √ |
| | VOT | na. | √√ | na. | √√ | √√ | √√ | na. | √ | |
| | F0 | na. | √√ | na. | | | | na. | √√ | √√ |
| VEL | CLS | na. | √√ | √√ | na. | √√ | | na. | √√ | |
| | VOT | na. | √√ | √√ | na. | √√ | √ | na. | | |
| | F0 | na. | | | na. | √√ | √√ | na. | √√ | √√ |

Table 3.30: Games-Howell post-hoc tests for different types of stops with preceding phone contexts in the KAIST speech database. Two significance levels 0.05 and 0.01 were used in each pair. Empty slots represent the acceptance of the null hypothesis at both significance levels and one check represents the rejection at the lowest significance level, which is 0.05.

stops have longer VOT and higher vowel F0. As we saw in the figures of bivariate distributions for each type of stop in each place of articulation, lax stops seem to be positioned in the middle of each acoustic parameter's measure.

Even when more factors are introduced and stops are further grouped into small number of tokens, the acoustic characteristics can still be found. MANOVA tests confirm that the three types of stops can be differentiated by these acoustic parameters. However, as we have seen in post-hoc tests, when we look at the individual parameters, we cannot differentiate some of the pairs. In other words, although each acoustic cue was previously studied independently and was statistically proved to be meaningful for differentiation of the three types of stops, its validity in all situations cannot be guaranteed. However, using multiple acoustic cues can resolve this problem. Even if each acoustic cue may not be enough to differentiate the three types of stops by its own, other acoustic cues can help to differentiate them. In sum, acoustic cues work as a whole to achieve their goal to differentiate the three types of stops in Korean.

A difference between our findings and previous ones is that the *domain-initial strengthening effect* does not stand out in the KAIST speech database. We do not find in general that stops in higher prosodic position have longer closure durations and VOTs than the ones in the lower prosodic position. The effect seems to be limited to two cases. One is that when two or more stops in one utterance are observed in different prosodic positions, the ones in higher prosodic position may be stronger than the one in lower position. The second case is that when two utterances have a similar prosodic structure, a stop in higher prosodic position in one utterance may be stronger than the one in lower prosodic position of the other utterance. These facts would be hard to use in ASR because they refer to relative properties of pairs of phones rather than to individual characteristics of the phones.

It is clear that for the purpose of differentiation of the three types of stops, the three acoustic cues should all be used together. Since the factors, place of articulation and prosodic

position, differentiate the three types, it will be easier to predict its type by looking at the three acoustic cues if we know the place of articulation and prosodic position.

In Chapter 4 and 5, we will develop a system that can use these acoustic cues to improve performance of the system.

# CHAPTER 4

# Baseline ASR Model

## 4.1 Introduction

In this chapter, we are going to explain how we built a baseline ASR system for the purpose of evaluation and comparison with our final system, which uses extra acoustic cues to differentiate stop consonants and which will be described in Chapter 5.

Jang (2000b) has already built a similar baseline system for comparison with one using F0 for better recognition of Korean obstruents. In his baseline system, HMMs were trained with the same database as we use, which is the KAIST speech database. Instead of using his baseline system, we attempt to modify three parts to see if these modifications could improve the performance of the baseline system.

First, we introduce a tailored lexicon that reflects cross-word phonological change for a given word sequence to get better label files for HMM training.

Second, a short pause model is created to model silence between two words. In doing so, some minor changes compared to the conventional way of creating the short pause model and adjusting the silence model are considered. Conventionally, the silence model is given a backward transitional probability and the short pause model has a transition that

skips an emitting state. We experiment with the system, including and excluding these transition probabilities, to see if these changes affect the performance of the system.

Third, we try to modify the language model, so that we can force a word sequence to have a legitimate sequence of pronunciation variants. When we use multiple pronunciation variants, some of the variants cannot be connected to a certain pronunciation of the previous word due to cross-word phonological rules. Whether eliminating these ill-formed pronunciation sequences by way of giving low probability to an undesirable word sequence can improve performance of the system is examined.

Because of the similarity of our baseline system to the one by Jang (2000b), the first thing is to explain how he built his baseline system. Minor differences in our methods are mentioned in the course of explaining his system. Then we discuss how we get the label file using a tailored lexicon, how we create a short pause model and how we modify the language model. Finally, results of various experiments with new labels, the short pause model and the language model are presented.

## 4.2 Baseline Model

### *4.2.1 General description of the system*

Jang's (2000b) baseline system has been already introduced briefly in Section 2.5.1 on page 40. Differences between the system in the previous stop recognition experiment and the one described here are mostly the addition of a lexicon and a language model. In the previous experiment, the lexicon was composed of phones, not words. Accordingly, "language" models modelled sequences of phones, not words. However, initialisation and training of the models are the same in these two systems.

As mentioned earlier in Section 2.5.1, the system was built with HTK (Young *et al.* 1996) and the speech data was the KAIST speech database. Jang used 8790 utterance tokens for training and 2073 utterances tokens for testing. He excluded utterances spoken in a

dialect that has a different tonal structure from standard Korean because characteristics of vowel F0 after stops can be different from those revealed in standard Korean. However, in our baseline model, the unused speech data is also exploited. So, 11781 utterances are used for training and 2965 for testing.

Each model in the system has 3 states, each of which produces an output probability. We use left to right continuous density HMMs without skipping states. Each model was initialised with hand-labelled data and retrained with all of the training data.

Jang used a set of 37 phones, including a silence model[1]. We use a short pause model which will be dealt with in detail in Section 4.3.2. The phone set is given below in International Phonetic Alphabet (IPA) transcription.

(4.1) Obstruents:

p p' $p^h$ t t' $t^h$ k k' $k^h$ c c' $c^h$ s s' h

Monophthongs:

a e i o u ɨ ə

Diphthongs:

wa we wi wə ɨi ya ye yo yu yə

Nasals and a liquid:

n m ŋ l

Silence: sil

The phones above are phoneme-like units. In the case of diphthongs, semivowels like w and y are not treated as independent phonemes. As Jang (2000b:51) pointed out, whether or not diphthongs are treated as a single unit does not make much difference in system performance because using context-dependent triphone models has a similar effect to treating a semivowel and nucleus as one unit. Combining semivowels and following vowels into a single unit makes hand-labelling and data processing easier. It is quite hard

[1] In a personal communication, he said that a short pause model was created but not used in testing the HMMs because he found it worsened the result.

to determine the boundary between a semivowel and a following vowel in the speech signal because there is no abrupt change in formant structure. Also, if we treat semivowels and their following vowels as a single unit, Korean syllable structure can be described as at most CVC structure, as we have already seen on page 10.

The same parameterisation of speech waveforms into sequences of feature vectors and the same window size (25 ms window, 10 ms frame shift) were used as in the phone recognition experiment. 12 Mel Frequency Cepstral Coefficients (MFCCs) plus energy, and their first and second derivatives were the parameters we used as in the phone recognition experiment.

### 4.2.2 *Pronunciation variations*

Strik & Cucchiarini (1998) describe knowledge-based and data-driven methods for deriving pronunciation variants for ASR. The knowledge based method is to generate pronunciation variants using abstract linguistic information (Top-down method). The data-driven one is to obtain pronunciation variations using the speech signal by hand or automatic labelling (Bottom-up method). In Lee *et al.* (1995), Jeon *et al.* (1998), Kwon *et al.* (1999) and Kwon (2000), the top-down method was employed and in Yun *et al.* (1997) and recently in Lee & Chung (2002), the data-driven approach was attempted for Korean pronunciation variations. The lexicon we are going to use is the one Jang (2000b) built. The lexicon was based mainly on the top-down method and more pronunciation variations were added with the data-driven method.

Constructing a Korean lexicon for ASR is relatively straightforward because letters and phonemes in Korean are in one to one correspondence. Jang (2000b) created what he called his *base lexicon* in four steps. First, Korean orthography was romanised. Second, each letter was mapped into a corresponding phoneme, and third, word internal phonological rules were applied. The final step was a manual adjustment of irregular pronunciations.

Irregular pronunciations seem to be produced in two ways. First, there are some cases where the general letter to phoneme relationship does not hold. In other words, there is a discrepancy between writing and pronunciation. Second, there could be overgeneration of pronunciations by phonological rules. For example, even when a phonological rule environment is met, sometimes there are cases when the rule should not be applied. Considering these two patterns of irregularities of pronunciations, the words with ad-hoc pronunciations were treated manually[2]. After these manipulations, each lexical item in the preliminary lexicon has one canonical pronunciation.

Jang (2000b) improved the system by modifying the base lexicon. He added pronunciation variants to the lexicon in two ways. The first one was to apply more phonological rules. Phonological rules were selected from Huh (1985) and Lee (1996)[3]. While phonological rules already used in the preliminary lexicon summarised in Table 4.1, are obligatory rules, these rules are optional. Cross-word pronunciation variations were also included in the lexicon.

The second way was to use real pronunciations obtained from hand-labelling. During the hand-labelling, we found some lexical items were pronounced differently from the way they were "supposed" to be. As Jang (2000b:69) illustrated, for example, a lexical item, *kwi-sa* meaning, your company, is believed to be pronounced as [kwisa], which is a pronunciation form in the base lexicon. However, we found a different pronunciation form, [kisa] in hand-labelling the data. The frequency of this pronunciation is 5 among 32 appearances of this lexical item in the data. Since these pronunciations were spoken by various speakers, it would not be appropriate to regard them as idiosyncratic. Jang categorised these pronunciations into several groups to capture phonological regularities. Phonological rules that Jang used for the modified lexicon are reproduced in Table 4.2.

[2]A failure of rule application can be seen as being regular, if we have morphological information in each lexical item. Rule application can be different depending on whether the word is native Korean or sino-Korean, which is morphological information. This should be taken into account when building a large vocabulary ASR system. Since our data has fewer than 3000 words, this issue will be set aside.

[3]*Lenis Stop Voicing*, for example, changes a lax stop to a voiced counterpart, which is not used as a recognition unit in our system. These rules are not needed in the modified lexicon.

| | Phonological Rule | Conversion Example | |
|---|---|---|---|
| a. | Cluster simplification | k a p s → k a p | 'price' |
| b. | Coda neutralisation | i $p^h$ → i p | 'leaf' |
| c. | Consonant nasalisation | s i p yu k → s i m n yu k | 'sixteen' |
| | | k u k m u l → k u ng m u l | 'soup' |
| d. | Tensing | s u k c e → s u k c' e | 'homework' |
| e. | Aspiration | s i l h t a → s i l $t^h$ a | 'hate+END' |
| f. | Palatalisation | k u t i → k u c i | 'obstinately' |
| g. | Laterallisation | $k^h$ a l n a l → $k^h$ a l l a l | 'blade' |

Table 4.1: Obligatory phonological rules for pronunciation of the preliminary lexicon. (Reproduction from Jang (2000b:57))

| | Phonological Rule | Conversion Example | |
|---|---|---|---|
| a. | Vowel deletion | o s i p → o s p | 'fifty |
| b. | Glide deletion | s a m wv l → s a m v l | 'March' |
| c. | h-deletion | s e s i m h a n → s e s i m a n | 'careful' |
| d. | Place assimilation | i m n i t a → i m m i t a | 'Be+RESPECT+EN D' |
| e. | Cluster simplification | $p^h$ u m m o k → $p^h$ u m o k | ' item' |
| f. | Aspiration | s i l t a → s i l $t^h$ a | 'hate+END' |
| g. | Tensification | h yo k wa → h yo k' wa | 'effect' |

Table 4.2: Phonological rules for generating pronunciation variants for the modified lexicon. (Reproduction from Jang (2000b:69))

After adding pronunciation variants to the base lexicon, Jang deleted some of the pronunciation variants to avoid creating an unnecessarily large network for Viterbi decoding. A total of 87093 variants was reduced to 12059 pronunciation variations. To reduce the number of variants, he first created automatic labels for all training data using the forced alignment provided by HTK and then deleted variants that were infrequent with respect to the automatic labels. The canonical pronunciation was kept even if it did not appear in the automatic labels. Jang's three steps for selection of variants are quoted below (Jang 2000b:72).

1. The base form is always kept.
2. For each of the remaining variants $V_i$, the probability $P(V_i|W)$ is calcu-

lated as

$$P(V_i|W) = \frac{C(V_i)}{C(W)}$$

where $C()$ is a counting function.

3. Discard $V_i$ if $P(V_i|W) < T$ where $T$ is a cutoff threshold

In our experiment, the same modified lexicon is used. Due to the introduction of the short pause model, all pronunciation variants begin with a short pause.

### *4.2.3 Language model*

The eventual language model we use in our experiment is the same language model as Jang constructed, although we experimented with changes as described in Section 4.3.3. Adopting a statistical method, he built a bigram language model[4]. A bigram language model expresses the probability of having two words, '$w_1$ $w_2$' consecutively. The estimation of the probability is calculated as follows:

(4.2)

$$\hat{P}(w_1, w_2) = \frac{C(w_1, w_2)}{C(w_1)}$$

It is highly probable that the frequency of a word or a word sequence can be biased towards the training data used for the language model and also that two consecutive words in the test data have never been seen together in the training data. To reduce the influence caused by these problems, two methods, *discounting* and *backing off* were adopted in Jang's bigram language model.

*Discounting* is a way of reducing the bias towards the training data. It is possible that a word sequence that is not observed in the training data does occur in the test data. So,

[4]Jang found that a bigram language model was more suitable for our data than a trigram language model (Jang 2000b:59).

the bigram probability for an observed sequence should be reduced by multiplying by a *discounting coefficient* and the remaining probability mass is given to word sequences that are not observed in the training data (Clarkson & Rosenfeld 1997). Jang adopted Ney *et al.* (1994)'s method called *absolute discounting*, in which the *discounting coefficient* $D$ is calculated as:

(4.3)

$$D = \frac{C(w_1, w_2) - b}{C(w_1, w_2)}$$

and the constant $b$ is calculated by

$$b = \frac{n_1}{n_1 + 2n_2}$$

where $n_1$ is the number of bigrams which occur once and $n_2$ is the number of bigrams which occur twice.

*Backing off* is a way of distributing the probability mass gained through *discounting* among the word sequences that have never occurred or have been observed fewer times than a certain threshold in the training data. In the case of a bigram language model, the unigram probability is used instead of a bigram probability which does not reach to the threshold. Considering the discounting coefficient, $D$ and the unigram probability of the word, $w_2$, the formula for the final bigram language model Jang used is as follows (Jang 2000b:61).

(4.4)

$$P(w_1, w_2) = \begin{cases} \frac{C(w_1, w_2) \times D}{C(w_1)} & \text{if } C(w_1, w_2) > k, \\ \alpha P(w_2) & \text{otherwise} \end{cases}$$

where the back-off weight $\alpha$ is calculated to ensure that the sum of $P(w_1, w_2)$ for all the vocabulary words $w_2$ equals to 1.

Jang calculated the *perplexity* of the bigram language model, which was 25.74. Perplexity $B$ is defined as (Rabiner & Juang 1993:450):

(4.5)

$$B = 2^{H_p} = \hat{P}(w_1, w_2, ..., w_Q)^{-\frac{1}{Q}}$$

where the $H_p$ is calculated as:

(4.6)

$$H_p = -\frac{1}{n}[log\hat{P}(w_1, w_2, ..., w_n)]$$

What the value of *perplexity* represents is, according to Rabiner & Juang (1993:450), "when the recogniser uses this language model [estimated by $\hat{P}(w_1, w_2, ..., w_n)$] for the task, the difficulty it faces is equivalent to that of recognising a text generated by a source that chooses words from a vocabulary size of $B$ independently of each other and with equal probability." They also state that $H_p$ is the average difficulty or uncertainty in each word based on the language model (Rabiner & Juang 1993:450).

In Section 4.3.3, we will explain our attempt to modify this language model for better modelling of cross-word pronunciation variation. However, in other experiments, the same language model as described here is used.

## 4.3 Modifications of the Baseline Model

Starting from Jang's baseline system, we make three main modifications. Firstly, a tailored lexicon is introduced to obtain better labels. Secondly, a short pause model is created to model a short silence between words. Thirdly, the language model is modified to better reflect the phonological change between two words. We start this section with phone label generation with a tailored lexicon.

### 4.3.1 Phone labels for retraining of HMMs

Since we have a small amount of hand-labelled data, these hand labels are used for initialisation of the HMMs. HTK provides a tool called HInit, which is an implementation of a *segmental k-means* algorithm (Rabiner & Juang 1993). Parameters are then re-estimated by the *Baum-Welch* algorithm which can be done with the HRest command. After the initialisation of the HMMs, the models are retrained with all of the training data. For this stage, the time information in the label files is not needed, but accurate phone sequences should be provided.

Because we have multiple pronunciation variants, a pronunciation has to be chosen for each training word. Jang (2000b) created labels using forced alignment. In this process, the selection of the pronunciation for each word was decided by the recogniser on the basis of pronunciation candidates having the highest probability. However, there is a problem concerning phonological rules that apply across word boundaries. For example, in Table 4.1, the rule *Aspiration* can occur between two words and the pronunciations that are possible in this word sequence can be limited to the ones that start with a changed phone. Suppose we have a word sequence as in (4.7). The pronunciation after *Aspiration* is shown below as well.

(4.7) ... *pap* *ha-neun* ...
... rice(noun) do-REL ...
... [p a (p) $p^h$ a n i̱ n]
... 'to cook rice' ...

REL: Relativiser

The rule *Aspiration* is triggered by lax stops /p, t, k/, and the target phone /h/ in *ha-neun* is changed to the aspirated form of the triggering lax stops. As a result, the lexicon includes at least four pronunciation variations for the word, in which [h a n i̱ n] is a baseform. The two words with pronunciation variations are described in the lexicon as follows[5]:

[5]The number at the end of each word is added for the convenience of explanation.

(4.8)

| | | |
|---|---|---|
| pap | p a p | 1 |
| pap | p a | 2 |
| ha-neun | h a n ɨ n | 3 |
| ha-neun | $p^h$ a n ɨ n | 4 |
| ha-neun | $t^h$ a n ɨ n | 5 |
| ha-neun | $k^h$ a n ɨ n | 6 |

As shown in (4.7), the proper pronunciation forms for the word sequence is either 1-4 or 2-4 in (4.8)[6]. It is also possible to have 1-3 as a pronunciation sequence when the rule is not applied but given the word sequence, pronunciation sequences that include a pronunciation form 5, or 6 cannot be realised in any case.

If the lexicon is not restricted, there is a chance that an undesirable pronunciation sequence would be selected by the automatic labeller which constructs word networks using all the possible pronunciation variants present in the lexicon. To get proper pronunciation sequences, we tailor the lexicon to have correct cross-word pronunciation variations in a given utterance. Since each utterance is an input to the automatic labeller, we can produce a tailored lexicon on-the-fly and replace the original lexicon with it. Figure 4.1 represents how the automatic labeller is interfaced with the tailored lexicon.

After the preparation of the labels generated with the tailored lexicon, the labels are fed to the process of *embedded training*, which updates all of the HMMs at the same time. The process is invoked by HERest command in HTK.

### *4.3.2 Short pause model*

Silence and pauses can be found in an utterance. The term "silence" is used for an acoustically silent and non-verbal part before or after an utterance. It can include background and impulse noises before or after an utterance. Silence in ASR plays a role of signalling

[6]The pronunciation sequence 2-4 is regarded as a coalescence.

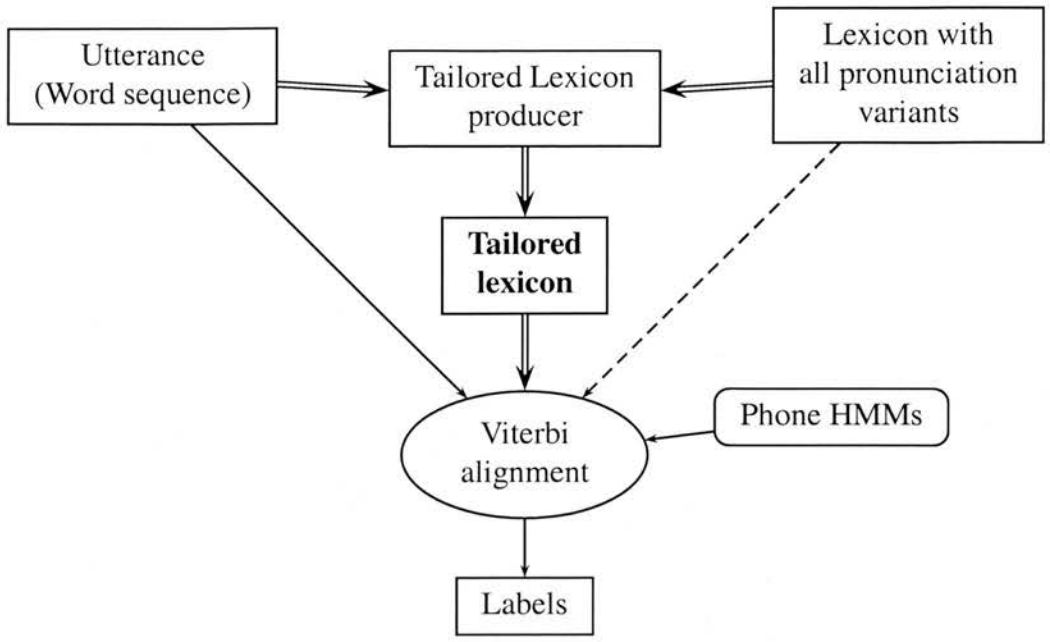



Figure 4.1: Automatic labeller with the forced alignment using a tailored lexicon. The tailored lexicon producer picks up the valid pronunciations given the utterance from the original lexicon with all pronunciation variants. For the normal procedure of the automatic labelling, instead of using tailored lexicon, the lexicon with all variants is put directly to the labeller, which is represented with a dashed arrow.

the start and the end of an ASR process, so that the first word of an utterance is always preceded by silence and the final word is always followed by silence. On the other hand, pauses are silent parts between words. Continuous speech can contain pauses in the course of an utterance but these pauses are not associated with a phoneme unless an independent recognition unit is trained to capture them.

A problem for training pauses with an independent recognition unit is that pauses in speech data should be hand-labelled for proper initialisation. However, in continuously read speech such as the KAIST database, there are not many tokens found. Instead of initialising the short pause model independently, we can employ the idea that the pause

shares its acoustic nature with silence and it can be created by copying the centre state of the silence model.

The HMM for the short pause consists of three states. The second state has an output probability whereas the first and third, are non-emitting states. These non-emitting states are used as the entry and the exit of the model. Since the acoustic characteristic of a short pause is that it is short and uniform, it is legitimate to have one emitting state, and having one emitting state is a standard way of setting up a short pause model.

The ordinary procedure for preparing the short pause model as given in the HTK manual is that the centre state of the silence model is copied to the second state of the short pause model and these states are tied together so that whenever the silence model is updated by retraining, the short pause model is also updated. After copying and tying the two states, one extra transition should be made in order to model the cases where there is no short pause between two words. The transition from the first to the last state enables the short pause model to be skipped and to be connected to a following phone model (Young *et al.* 1996).

When the short pause model is created, the silence model is also altered, the HTK manual gives instructions to add forward and backward transitions for the silence model. Since a total number of states, including the first and the last non-emitting states, for the silence model is five, the forward transition is from the second to the fourth state and the backward transition is from the fourth to the second state. The additional transitions are needed to absorb various noises and to accommodate a long silence so as not to pass the silence over to the next phone model.

Silence and short pause models are related to an articulatory closure for a stop sound in that they are all acoustically silent. So, when a silence or short pause is followed by a stop, it is possible that the silence or short pause model can be affected by the closure introduced by production of the stop, and conversely, it is also possible that the stop model can be affected by the silence or short pause model. After creating the short pause

model, new labels for retraining are generated by forced alignment. When updating the silence, short pause and stop models with these labels, the closure part, which is "supposed" to be modelled in the first emitting state of the stop HMM, could happen to be modelled in the silence or short pause model. If there are not many cases where there is a silence or short pause model and a stop is in succession, it may not be a problem because a small number of extreme cases can be tolerable in statistics. However, when the functional load of stop sounds is considered in Korean as seen in Table 2.6 on page 45, the cases where the silence or short pause model followed by a word starting with a stop is too large to be ignored. In fact, in our training data the silence or short pause model appears 77267 times and 30084 cases are found where the silence or short pause is followed by a stop. So, it is possible that the silence or short pause model could disrupt the integrity of the following stop HMM.

Another problem involved with the short pause model is that it could also absorb a word final coda stop. Stops in coda position are realised as unreleased stops. So, a word final stop can be trained and recognised as a short pause, which does not seem to make the system deteriorate seriously but it could affect the calculation of overall probability of hypotheses in the process of decoding, which can lead to a change of priority of hypotheses.

In order to find a better model for silence and short pause, we did an experiment in which an HMM for silence was trained with and without additional forward and backward transitions. What we have in mind is that in the case of utterances that start with one of the stops, VOT could be regarded as noise, which may result in selecting a wrong hypothesis at the end of decoding. In a second experiment, we did tests with and without a transition to permit skipping the emitting state of the short pause model. When new labels were generated by forced alignment, the decision to include or exclude the short pause between two words was left up to the automatic labeller. By doing so, we assumed that the short pause model would not take up the closure part of the following stop. In a third experiment, labels were manipulated to have no coda stops at the end of a word on the

assumption that the short pause model could replace it. A possible additional advantage would be that HMMs for stops would model only onset stops. It means that HMMs for stops do not have to consider coda stops, which have a different acoustic manifestation.

Our result does not show much difference in the three experiments. An HMM without the forward and backward transitions is a little better than the one with the transition. The result also shows that having coda stops or not does not make much difference. Nevertheless, the short pause model is important because the system is improved with the short pause model compared to the one without the model. More detailed results will be discussed in Section 4.3.4.

### *4.3.3 Modification of the language model*

The bigram language model that Jang (2000b) built is modified to reflect cross-word phonological change. Cross-word pronunciation variants are already included in the lexicon but the lexicon needs to be tailored not to overgenerate ill-formed pronunciation variants. Even though we successfully created a tailored lexicon for a given utterance to get better phone labels with forced alignment, it cannot be used during recognition because in recognition, no specific utterance is given.

HTK creates a word network using all pronunciation variations. For example, if a word A is followed by a word B, then all pronunciation variants of A are connected to the pronunciation variants of B to construct the network. This causes a similar problem to the one we saw in Section 4.3.1. It is not a major issue to see which pronunciation for a word is used when all words in utterance are correct. However, if the selection of pronunciation variants is left entirely to the trained HMMs, a wrong word sequence could be chosen. To see how it could select a wrong sequence, we introduce a word sequence in relation to the one in (4.7).

(4.9) ... *pap* *tha-neun* ...
... rice(noun) burn-REL ...

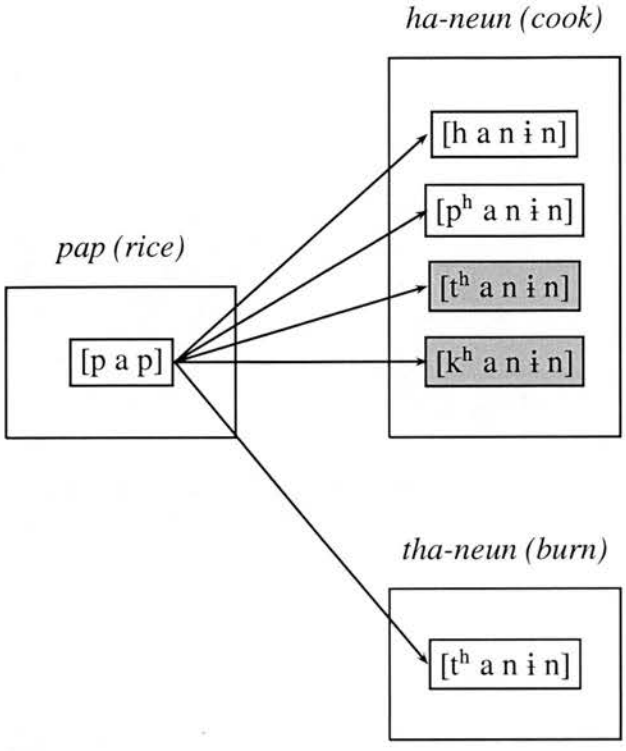



Figure 4.2: Word connection and cross-word phonological change in recognition. Sequences of *pap* and shaded pronunciation variants are phonologically ill-formed.

... [p a (p) tʰ a n i n]
... 'to burn rice' ...

REL: Relativiser

The lexicon we saw in (4.8) should include the word, *tha-neun* as [tʰ a n i n]. The word connection during the recognition is represented in Figure 4.2.

Suppose the language model probability of the sequence, *pap ha-neun* is higher than the sequence of *pap tha-neun* and also, suppose the acoustic probability is the highest when the phone sequence is [p a p tʰ a n i n], which two word sequences share. Even though the target sequence that a recogniser should pick up is *pap tha-neun*, the word sequence, *pap ha-neun* will be selected due to its higher language model probability.

If we can manipulate the network so that the undesirable sequences in the network, which are shaded in Figure 4.2, are deleted or have enough of a penalty to cancel out the benefit that the wrong sequence would get from the language model probability, the correct word sequence *pap tha-neun* will be chosen because of its higher acoustic probability. Yu *et al.* (2000) controlled the network between the end of a previous word and the start of a current word and succeeded in reducing word error rate from 32.9 % to 15.4 %. However, as Kwon & Waibel (2002) pointed out, controlling the connection of the end and the start of words in the network is very complicated.

A different way of having a similar effect to the previous method on the network is to manipulate the language model probability that will be imposed on top of the acoustic probability. A similar approach was taken in Cremelie & Martens (1995). As seen in 4.2.3, each sequence '$w_1$ $w_2$' has either a bigram or a unigram probability. However, the probability is applied to a word item regardless of its pronunciation variants. That is why the undesirable pronunciation given a previous word is as probable as the correct one. If a different pronunciation variant can be assigned a different probability given a previous word, so that the undesirable pronunciation variant has a low probability from the language model, the chance that the wrong pronunciation sequence is selected is very low.

In order to manipulate the language model probability, each pronunciation variant should be registered in the lexical item. For example, in (4.8) the word, *ha-neun* should have different lexical forms for pronunciation variants arising from phonological rules as follows:

(4.10)

| | |
|---|---|
| pap | p a p |
| pap | p a |
| ha-neun | h a n ɨ n |
| p_ha-neun | $p^h$ a n ɨ n |
| t_ha-neun | $t^h$ a n ɨ n |

k_ha-neun k$^{h}$ a n i̯ n

If the (log) probability of having the sequence (*pap*, *ha-neun*) is -3.05, the new probability for modified bigram model would be as follows[7]:

(4.11)

$$P(pap, ha\text{-}neun) = -3.05$$

$$P(pap, p\_ha\text{-}neun) = -3.05$$

$$P(pap, t\_ha\text{-}neun) = -20.01$$

$$P(pap, k\_ha\text{-}neun) = -20.01$$

By manipulating the bigram probability, the sequence (*pap*, *t_ha-neun*) or (*pap*, *k_ha-neun*) will become disadvantageous, but the correct pronunciation sequence has the same probability from the language model as is originally allocated to the word sequence. The determination of the probability that is given to the undesirable sequence is arbitrary. It is less than the smallest probability in the language model.

This method has also a drawback. Since each pronunciation variant is registered in the lexicon, the lexicon will be very large, which creates a large network. The problem caused by the large lexicon and network is that it needs to have a large memory and great amount of computation. The problem is alleviated to a small extent during decoding because hypotheses with undesirable sequences get a low probability and will be deleted in the network by the beam search. In this experiment to check if cross-word pronunciation variation can be modelled by modifying the language model, the computational problems will be ignored. The result of this experiment, together with others, will be shown in the following section.

[7]The probability represented here is the log probability. The reason of using the log probability is that the log form is convenient for handling small numbers

| Transition in the silence model | Skipping of emitting state in sp. | Word Accuracy (%) |
|---|---|---|
| no | no | 88.89 |
| no | yes | 89.01 |
| yes | yes | 88.99 |

Table 4.3: Results of experiments with and without the forward and backward transition in the silence model and skipping of emitting state in the short pause model. The results do not show great differences between systems, but the system with the best result has no forward and backward transition.

### 4.3.4 *Final result of baseline system*

The performance of an ASR system is often measured by *word error rate* (WER), which is calculated by subtracting *word accuracy* from 100 (%). And the accuracy is obtained by calculating:

(4.12)

$$Accuracy(\%) = \frac{N - (Substitution + Insertion + Deletion)}{N} \times 100$$

Tests were done to see if skipping of the emitting state of the short pause model has an effect on the recognition accuracy. The first two results seen in Table 4.3, show that the skipping does not make a significant difference. Next, forward and backward transitions in the silence model were set up. Since the system with skipping shows a slightly better result, the system that is used to see if the transition in the silence model affects performance includes skipping in the short pause model. The result shows that presence or absence of forward and backward transitions in the silence model does not affect the system performance much.

To check how much can be gained by retraining labels with a tailored lexicon, other phone model systems were prepared. These systems include skipping in the short pause model but exclude forward and backward transitions, based on the results shown in Table

| Selection of the short pause by the automatic labeller | Word Accuracy (%) |
|---|---|
| no | 88.33 |
| yes | 88.40 |

Table 4.4: Results of experiments with two types of tailored lexicon. The first one has the short pause in every pronunciation and the second one has at least two pronunications one with the short pause and the other without it.

4.3. When retraining labels were generated by forced alignment, the tailored lexicon was used to model cross-word pronunciation variation. In doing so, two types of lexicon were provided. The first one is composed such that every lexical item has at least two variants, one with the short pause and one without it. This lexicon enables the automatic labeller to decide whether the pronunciation with the short pause was selected or not. The second one includes a short pause in every pronunciation variant. Table 4.4 shows that using the tailored lexicon with two pronunciations performs slightly better. However, compared to the results shown in Table 4.3, using the tailored lexicon somehow makes the results worse.

The tailored lexicon does not seem to yield any system improvement over labels obtained by the standard procedure. However, we checked the effectiveness of the tailored lexicon through N-best recognition. First, the 10 best recognition hypotheses of each test sentence were stored and then each hypothesis was rescored with forced alignment. Since there is a given utterance, a tailored lexicon can be produced. The forced aligner picked the best fitting pronunciation variants within the tailored lexicon and also produced a new acoustic probability for each word. Using the new acoustic probabilities gives a new probability for each utterance. Then, these 10 new utterance probabilities were compared and the utterances were reordered according to the new probabilities. From this rescoring, we achieved about 0.3% improvement in word accuracy.

| | Num. of correct utt. | Word accuracy (%) |
|---|---|---|
| Original hypotheses | 4 | 70.52 |
| Reordered hypotheses | 31 | 80.02 |

Table 4.5: Effect of a tailored lexicon before and after rescoring the acoustic probabilities in 96 affected utterances by the tailored lexicon.

Reordering of hypotheses affects 96 utterances among 2965. The 2869 other utterances are not affected by the tailored lexicon and preserve the order of the N-best hypotheses. The first best hypotheses of these utterances are either correct or incorrect and cross-word pronunciation could be involved in either cases. If there is a cross-word pronunciation problem and the utterance is correct, the total probability of the hypothesis is high enough to survive to the end of decoding. If the utterance is wrong, we may conclude that the phone model must have been badly trained.

The best hypotheses of the affected 96 utterances can be also either correct or incorrect. That means rescoring could have a negative effect on the recognition result. To confirm the effect, the original first best hypotheses and the newly ordered first best hypotheses in the 96 utterances were gathered and their word accuracy calculated. As seen in Table 4.5, in the original first best hypotheses, only 4 utterances are correct whereas the new hypotheses get 31 utterances correct. Even though the labels for retraining generated with the tailored lexicon did not make an improvement, the importance of cross-word pronunciation is demonstrated.

An experiment related to a sequence of a word final coda stop and a short pause was done. Since the Korean coda stops are unreleased, a word final stop followed by a short pause does not have a clear boundary. This may cause a problem in retraining the coda stops and short pause. A possible way of getting around this problem is to let the coda stops be recognised as a short pause. As we mentioned above in Section 4.3.2, this could make modelling stops more robust because in other systems, onset, coda, and voiced stops have

| Coda modelling | Short pause | Word Accuracy (%) |
|---|---|---|
| yes | no | 79.74 |
| no | no | 85.17 |
| no | yes | 88.30 |

Table 4.6. Results of experiments with and without coda stops and short pause modelling.

to be modelled in one HMM whereas in this experiment, coda stops do not have to be modelled in the HMM.

For this experiment, we changed the hand-labels so that the coda stops would not appear, and initialised the HMMs. Recognition experiments were done with and without the short pause model on the basis of the new HMMs. The result in Table 4.6 shows that the system without coda stops does not help the short pause to be modelled better than the one in the other systems. Comparing two systems without a short pause model, the one that does not model coda stops is actually better. There seems to be compensatory relationship between coda stops and the short pause model. Without the short pause model the system is better than the others but once the short pause model is introduced, the advantage disappears.

In the experiment with the language model manipulation for modelling cross-word pronunciations the word accuracy of the system is 88.83 %, which is very similar to the results of various systems we made. That means modelling cross-word pronunciation variation does not make much difference in this system. This could result from the database. As we have seen in the experiment of rescoring with the tailored lexicon, there are not many utterances that are affected by cross-word pronunciation. Or, it could be the case that although there were many utterances affected by it, the affected phone sequences could be biased toward a specific phone sequence, so that the other phone sequences would not be properly trained and modelled.

We investigated our data to see how many utterances include the sequence of a lax stop followed by 'h'. We found that majority of the cases are the 'k-h' sequence, which was found in the 1227 utterances at least once. The number of 'p-h' sequences is 172 and of 't-h' sequences only 3. This means that the data is heavily biased towards 'k-h' sequences which may result in a low language model probability for the words ending with 'p' or 't'. So, the attempt to model cross-word variation may not have been fairly tested.

Our various results with the silence and short pause model modification, labels with a tailored lexicon, and language model modification did not show much difference, but including the short pause model makes a great difference. We next created a context-dependent system with and without the short pause model to see how much effect the short pause model can bring in. Our results of context-dependent models are compared with other systems and are summarised in Table 4.7, showing that the context-dependent system with a short pause model gives the best results of any system so far reported for Korean ASR.

The test datasets in Table 4.7 are not the same but the database that all the systems were developed on is the same. So, small differences in evaluation of the best word accuracy cannot be depended on but we can compare the improvement to each system due to short pause models and context dependent phone models.

In our experiments, the effect of the short pause model seems to be larger in the monophone models than in the context-dependent triphone models. However, since the triphone model without a short pause showed quite high word accuracy, there is less room for improvement. The increase of word accuracy in the context-dependent triphone model with a short pause is 6.57% but it reduced the WER from 11.89% to 4.44%, which is a 62.66% of reduction in error rate[8]. So, the effect of the short pause in the context-dependent triphone models is as large as in the monophone models. Figure 4.7 also shows the improvement of word accuracy made by our systems with or without a short pause is much larger than the systems of Choi *et al.* (1995).

[8]Error reduction rate of the monophone model between those with and without a short pause is 45.76%.

| | Choi *et al* (1995) | Yun *et al* (1997) | Jang (2000b) | Our model |
|---|---|---|---|---|
| No sp | 65.3 | | 78.47 | 79.74 |
| With sp | 69.4 | 68.00 | | 89.01 |
| C.D. no sp | 89.5 | | 87.93 | 88.11 |
| C.D. with sp | 90.7 | 91.11 | | **95.56** |
| Demi. no sp | | | 91.03 | |
| Demi. with sp | | | | 91.98 |

Table 4.7: Comparison of word accuracy results of 4 systems. The systems with the short pause model show the improvement and the improvement is greater in our systems with the short pause than in the others. C.D., Demi. and sp represent "context-dependent", "Demisyllable" and "short pause", respectively. The demisyllable system was trained with two more feature vectors, F0 and its first derivative.

Jang's (2000b) demisyllable system using the segmental F0 effect performed with 91.03% word accuracy. The system included F0 and its first derivative together with other 39 parameters. We created a demisyllable based system similar to Jang's and a short pause model was included. We trained and tested the system with the same speech files used in building Jang's demisyllable system. After increasing mixtures and data-driven clustering, we got 91.98% word accuracy. Our demisyllable model with a short pause performed a little better than Jang's demisyllable model. But it did not show the performance we had in our monophone and context-dependent models. Even though two more speech parameters for the demisyllable model are used, compared to our context-dependent triphone model, the demisyllable model with a short pause does not look attractive. So, our final baseline system that will be compared to a system implemented with the three acoustic cues we have seen in the previous chapters is a context-dependent triphone system with a short pause.

## 4.4 Summary

In this chapter, we summarised how Jang created his baseline model. We also described his preliminary and revised lexicon. Based on his system, we modified the lexicon to

have better labels in an attempt to get better trained HMMs. Pronunciation variants that are affected by cross-word phonological rules in a given utterance are picked out from the original lexicon to form a tailored lexicon, which is input to the automatic labeller with forced alignment.

Short pause and silence models are created and manipulated. The silence model without forward and backward transitions is very slightly better than the one with them. Labels for retraining generated with a tailored lexicon do not make much contribution on producing a better system, but N-best recognition and rescoring with the tailored lexicon confirms that the cross-word pronunciation modelling makes a difference in performance. Modification of the language model to incorporate the cross-word pronunciations into the network is not successful but we found that our data did not reflect the cross-word variation fairly. The results of systems with the tailored lexicon and the modified language model mean that our hypothesis that a system with a better cross-word pronunciation modelling will eventually decrease the WER is valid.

Compared to the results of other systems, our baseline monophone and triphone systems with the short pause model are improved significantly. Even though our attempts to improve the performance our baseline models by way of making automatic labels with a tailored lexicon and of modifying the language model did not help, creating the short pause model itself makes a large difference. In particular, the context-dependent triphone model with a short pause is better than the similar system built by Jang (2000b), which has two more speech parameters. In other words, we may not need to have additional parameters to model stop sounds.

In the next chapter, we will describe how we develop a system using the three acoustic cues analysed in Chapter 3. We will also compare our final system with the baseline model described here.

# CHAPTER 5

# ASR System Using Multiple Acoustic Cues

## 5.1 Introduction

We saw that confusion of different types of stops brought about the low rate of stop recognition in Section 2.5.1, and also that there was a large functional load of stops in Korean words in Section 2.5.2. To properly differentiate the three types of stops, we saw that three acoustic cues namely, closure duration, VOT, and vowel F0 after stops should be used all together. This was confirmed by statistical tests in Chapter 3. If we can extract these acoustic cues and use them in ASR, we may expect overall improvement of an ASR system.

In Jang (2000b), an ASR system based on a demisyllable unit was successfully developed for implementation of F0. In addition to F0, our final system will employ two more acoustic cues, closure duration and VOT. Problems involving implementation of such a system are: first, how to get accurate durations from the speech signal, second, how we model these parameters to get an appropriate probability given the durations and F0 and finally, how the stop probability is integrated with the probabilities of other phones.

One of the obstacles to implementing a system that can use durations is that durational information cannot be parameterised in the same way as Jang (2000b) did with vowel F0.

The F0 of a 10 ms frame of speech can be treated like any other parameter in an HMM. However, duration is not a property of individual frames and has to be treated differently. In standard HMMs it is effectively modelled via transition probabilities, although this is known to give poor probability distributions for durations (Rabiner & Juang 1993). Since we will not use the HMM for the durations, external modelling is needed, and we must find a way to integrate the durational probabilities with the rest of the system.

For vowel F0, what we have seen to be significant in relation to characteristics of stops is the mean F0 of the vowel after a stop, not a frame by frame F0 value. So, instead of parameterising F0 in the speech feature vectors, we can externally model it as we do with durations. In doing so, we can use the three parameters externally all together. However, we also include experiments of implementing systems with F0 parameterisation and external modelling of two durations only. The result will be reported in Section 5.6.

In order to obtain accurate durations, we propose an automatic segmentation of closure and VOT for stops by clustering and picking up a large difference of log power in two consecutive windows. For modelling of the parameters we are using, two approaches with parametric pdfs, one for univariate, and the other for multivariate are attempted. A post-processing method is employed to integrate the stop probability into the system. To validate the method we are employing and to see if phonation type confusion of stops can be cleared up using the external modelling of three acoustic cues, we experiment phone-level recognition and then, we build a word-level recogniser.

This chapter is organised as follows. First, problems of getting durations and modelling them with a conventional HMM will be investigated. Second, an automatic way of extracting durations from the speech signal will be explained. Third, the way in which we create duration models from the training data using automatic segmentation will be also explained. Fourth, an overview of an ASR system using a post-processing method will be described. Fifth, phone-level recognition using the post-processing method will be also described. Finally, results of the word-level recognition system and analyses will follow up.

## 5.2 Problems of Duration Modelling with an HMM

The inherent duration probability density $P_i(d)$ associated with state $i$, with self-transition coefficient $a_{ii}$ is calculated as (Rabiner & Juang 1993:358):

(5.1)

$$P_i(d) = (a_{ii})^{d-1}(1 - a_{ii})$$

Probability of staying in $i$ state for time $d$ exponentially decreases with time. Levinson states, "this exponential distribution of state durations is inappropriate if the states of the HMM are to represent linguistically meaningful components of the speech code" (Levinson 1986:31). To overcome this durational probability problem in the conventional HMM, revised models have been proposed. A common aspect of these revised models is that the self-transition coefficient, $a_{ii}$ is set to zero and a new $P_i(d)$ is specified explicitly and added to the HMM. In Ferguson (1980), non-parametric modelling of $P_i(d)$ was attempted. However, since a non-parametric approach needs a lot of training data and computation, an alternative, namely parametric probability density function (pdf) for modelling duration was introduced. In Russell & Moore (1985), duration was modelled with the Poisson distribution and in Levinson (1986) and Hochberg & Silverman (1993), a gamma and a normal pdf were used for parametric modelling of duration, respectively.

As Wang (1997) pointed out, the explicit durational pdf was brought in to properly model duration of staying in state $i$, not of an entire phone, which is generally composed of several states not just one state. Modelling the whole phone duration seems to be more important than modelling state duration, when considering durational difference of a phone depending on its position in an utterance (Klatt 1976). However, with respect to Korean stops, state duration modelling appears to be more important if the first and the second states are believed to model closure duration, and VOT, respectively, with a three-state HMM[1]. There are still two problems remaining. First, we have to be sure that the HMM

[1]The third state is considered to model a transition from VOT to the following vowel.

states really do correspond to closure duration and VOT, and second, the standard 10 ms frame shift is too large to mark closure and VOT boundaries accurately. In fact, the second problem is more serious than the first because the VOT of a tense stop can be less than 10 ms, which is less than one frame shift. One may parameterise the speech with a shortened frame shift. Shortening frame shift increases the number of frames, which causes a lot more unnecessary computation for training of HMMs of not only stops, but also other phones. So, we are in need of a new method for automatic segmentation that makes it possible to get an accurate closure duration and VOT. In the next section, the new method for automatic segmentation for stops is described.

## 5.3 Segmentation Algorithm for Closure Duration and VOT

The number of demarcation points needed for closure duration and VOT is at least three, except for stops in utterance initial position. The first boundary is the starting point of closure, the second one is the end of the closure, which is again the starting point of VOT. The third one is the end of VOT, which is in turn, the start of a following segment. Since closure duration is not available in utterance initial position, the starting boundary for closure duration is not needed.

Stop segmentation was attempted with a knowledge-based approach in Ali *et al.* (1999) without any help of stochastic modelling. But detailed segmentation for the internal events of a stop was not performed. In Suh *et al.* (1998), voiced, unvoiced and silence were extracted using a recurrent neural network. In Lee & Rhee (1999), speech and non-speech were segmented based on wavelet transform. A closure was regarded as a part of non-speech elements. These works might provide boundary information for a stop but the window size used for parameterisation of features in their systems was the conventional one.

Out strategy for getting an accurate boundary for the internal segmentation of stops is firstly, labelling, secondly, sorting the labels and thirdly, fine adjustment. Labelling is

done using *clustering* (Duda *et al.* 2001). Through clustering, closure and VOT parts are vaguely identified. Second, using the log power difference between two consecutive windows we can isolate the accurate transition moment from previous phone to closure, from closure to VOT (McKenna 2000). For a following vowel demarcation, we depended on voicing probability. By selecting the onset of voicing from the voicing probability, a better demarcation for the vowel was obtained. The parameters used for this automatic segmentation procedure are described in the following section. Then, the clustering procedure is explained in detail. Finally, how to select a boundary using the power difference and voicing probability is described.

### *5.3.1 Speech parameters*

The speech parameters that we used for clustering are some of the output from *get_f0*, which is an implementation of F0 estimation algorithm (Entropic 1998). *get_f0* produces four fields for its output. They are F0, Root Mean Squared power, 'ac_peak' and voicing probability. 'ac_peak' is the peak normalised cross-correlation value that is found to determine the output F0. Among these, ac_peak and voicing probability are multiplied together and the result is used as a parameter. Since the value for voicing probability is either zero or one, ac_peak will be all zero in unvoiced regions. The reason for the multiplication of the two fields is that the value from the ac_peak is reliable within a voiced area, such as a vowel, but it does not show a systematic difference in its values within an unvoiced area. For example, the ac_peak at the end of a vowel is high due to glottal pulses, which should not be regarded as a part of the vowel. So the multiplication removes the unreliable ac_peak value caused by glottal pulses at the end of a vowel,

Other parameters used for clustering are log power and zero crossing rate (ZCR). Log power is expected to help in determining the boundary between VOT and the following vowel. ZCR is normally used for fricative sounds. Since the VOT is the part of frication, ZCR is useful for the boundary between closure and VOT, and possibly, between VOT and the following vowel.

All speech signals were high pass filtered with a cut-off frequency of 35 Hz to eliminate low frequency drift, which gives a misleading ZCR. Then, log power and ZCR were calculated from a 20 ms window shifted in 3 ms steps. The ac_peak and voicing probability were obtained just by running the *get_f0* command.

### 5.3.2 *Clustering*

For the segmentation of closure and VOT, we need to distinguish at least three types of frames. They are closure, VOT and the rest. The rest includes vowels and nasals and other consonants, such as fricatives. Considering syllable structures in Korean, phones that could occur right before or after a stop are vowels and nasals. Given the parameters described above, VOT and vowel have a big difference in terms of power. The difference is much larger than that between closure and VOT, which causes weak vowels and nasals to be clustered as VOT. To avoid this, it is better to have 4 rather than 3 types of frames. We call the additional frame type "nasal" because the power of nasal sounds falls in between that of VOT and vowels. By setting up a nasal frame type, the power differences between closure and VOT, between VOT and nasal, and between nasal and vowel are similar. This will reduce the number of frames that are clustered to an incorrect frame type.

The clustering technique we used is similar to *k-means clustering*. *K-means clustering* is clustering data points into $k$ groups by means of comparing distances between each data point's coordinates and the each group's initial mean coordinates. The group that has the shortest distance to the data point is the one that the data point belongs to. After the first iteration, new means for the reference group are calculated and the same process is repeated until there is no change in current means and the newly calculated ones.

We defined the first initial coordinates of each group intuitively based on the expected behaviour given the parameters. For example, a closure has the smallest log power whereas a vowel has greatest. A VOT seems to fall in between the two. As we described above, the nasal would be positioned between the VOT and the vowel.

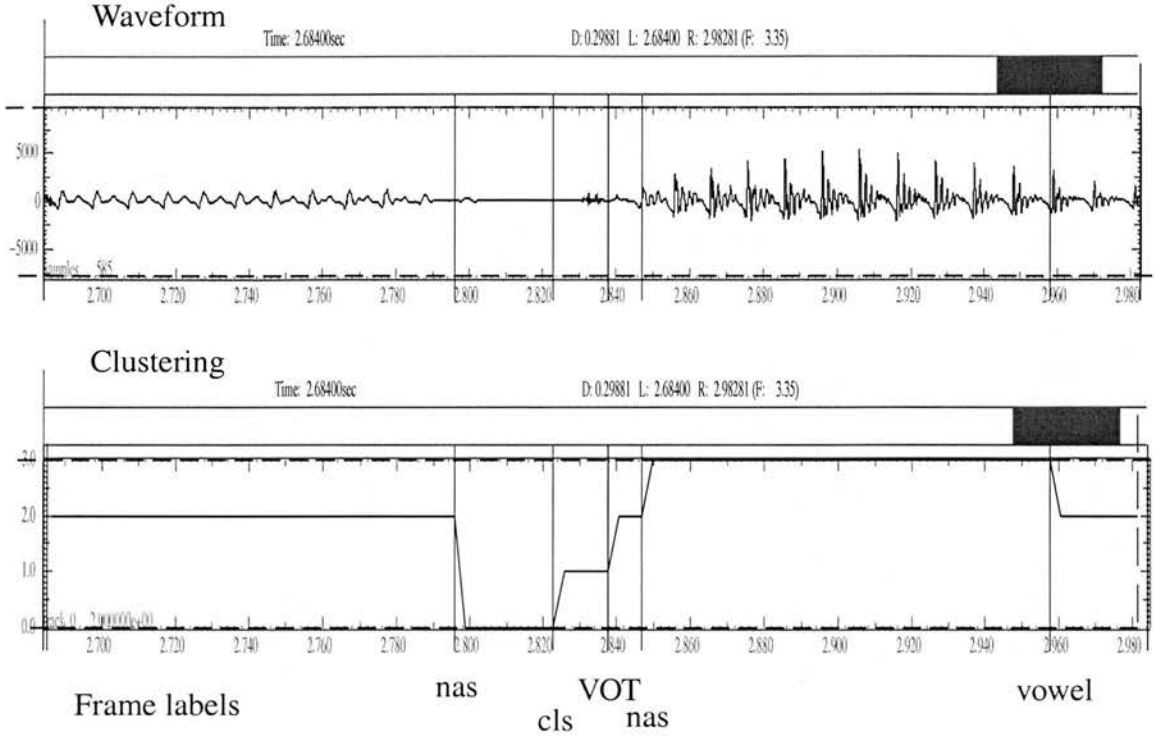



Figure 5.1: Frame labels after clustering. *nas*, and *cls* represent nasal and closure, respectively. The numbers 0, 1, 2, 3 are assigned to frame types for graphic display. From the left in the second window, 2 is assigned to nasal, 0 to closure, 1 to VOT, 2 to nasal and 3 to vowel frames. Labels are placed at the righthand end of the segments to which they apply.

For purposes of graphic display, we assigned the numbers, 0, 1, 2, 3 to closure, VOT, nasal, and vowel, respectively. An example of clustering outcome is illustrated in Figure 5.1.

### *5.3.3 Power difference between two consecutive windows*

Clustering gives better boundary information than the labels from forced alignment because it uses a smaller window shift. Yet, it is not accurate enough to demarcate the

boundary we need. To pinpoint the boundary, we also used power difference between two consecutive windows, as described in Section 5.3.4 below. What we mean by consecutive windows is that if the first window for the log power is from 0 to 20 ms, the second window is from 20 ms to 40 ms. We subtracted the power of the first window from that of the second. The difference was marked as a value at time 20 ms.

We used a finer window size (1 ms) to calculate the log power than the size used in clustering. After each power calculation, the two windows are shifted one speech sample to the right, which is 0.0625 ms at a 16 kHz sampling rate. The bottom window in Figure 5.4 plots the power difference calculated at each sample point.

### *5.3.4 Integration of clustering and power difference*

After finishing clustering, the outcome of clustering is smoothed in such a way that a type of a frame that is not in accordance with the previous and the next frame type is changed to the same frame type as the neighbouring ones. More specifically, the label of each frame is compared to that of the previous one. If the labels are different, the label of the frame we are looking at is compared to the label of the next frame. If they are also different, then the previous label is compared to the next label to see whether they are the same. If they are, the label we are looking at is considered as *a stranded label.* This label is changed to the same as the previous one. After that, the frame labels are reformatted in *xlabel* style, in which a sequence of the same type of label is marked as one label. Demarcations are made only the beginning and the end of the same type of label sequence, so that the labels look like phone labels as in Figure 5.1. From now on, the frame labels we mention in this section are reformatted labels.

For the internal structure of a stop, maximally 3 demarcations are needed. First, the end of the previous phone, which is the beginning of the stop closure. Second, the end of the closure, and third, the end of VOT. We have a three-stage procedure for a fine adjustment of labels we need. First, forced aligned labels marking the beginning and end of the stop are replaced with two frame labels approximating the beginning and end of the stop

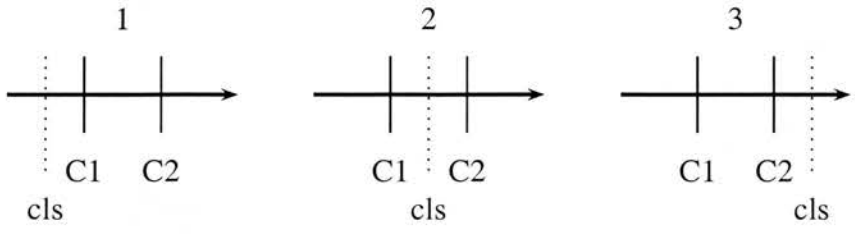



Figure 5.2: Three possible cases of closure type frame label with respect to forced aligned labels of a stop. C1 and C2 represent the start of stop and the end of stop, respectively. *cls* is closure type frame label.

closure, respectively. These two frame labels localise the stop area. Second, we search for the onset of voicing using voicing probability and consider the onset of voicing as the end of VOT. Third, the beginning and end of the closure are adjusted using power difference. We will explain each stage in detail.

There are 2 steps in replacing forced aligned labels with two frame labels. Call the beginning and end labels C1 and C2, respectively. First, we identify the position of the closure type frame label which is nearest to C2. The position will be one of the three cases in Figure 5.2. In case of 1 in Figure 5.2, we regard the stop as voiced. This stop is processed no further and skipped. In the other two cases, we next find a nearest frame label to C1. It can be positioned either left of right side of C1 as in Figure 5.3. We could select either $S_l$ or $S_r$. If $S_l$ is far away from C1, the localisation of stop area is quite large. So, heuristically, if the duration is shorter than 9 ms, the start of the localisation is $S_l$, otherwise, $S_r$.

It is possible that there could be other frame labels between $S_l$ (or $S_r$) and *cls*. If there is another closure type frame label between the two, we have to check whether the label *cls* which is the closest to C2 is a short spurious closure that would sometimes be found

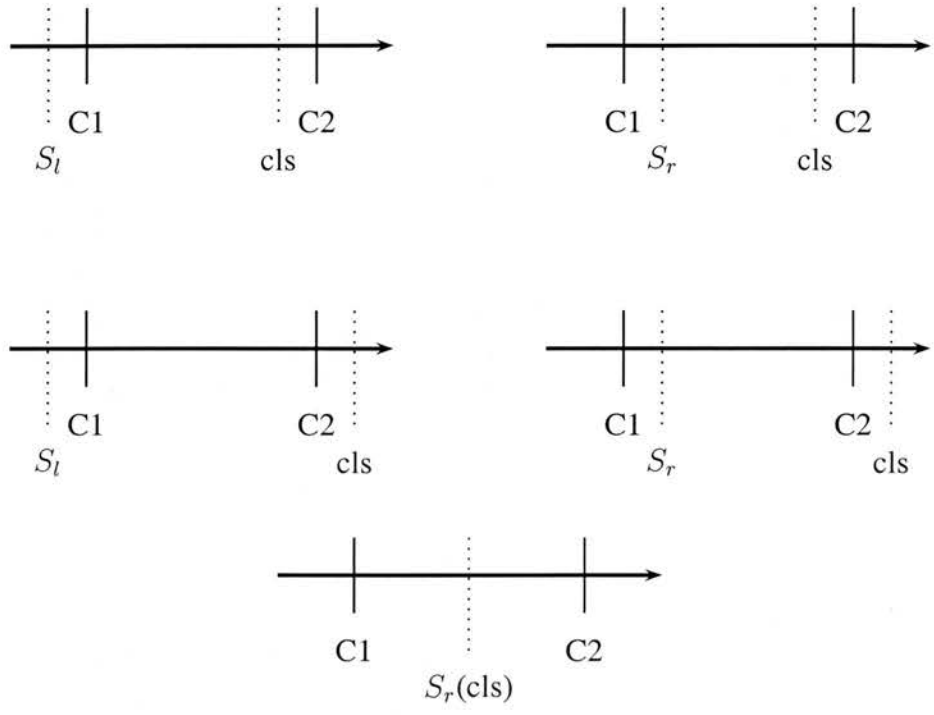



Figure 5.3: Possible positions of the start of stop area. $S_l$ ($S_r$) represents the frame label which is left (right) side of C1.

between two VOT type frame labels. If the duration of *cls* is 10 ms or more, we adopt it as a genuine closure. Otherwise, we measure the duration of each closure label between $S_l$ (or $S_r$) and *cls* and the end of stop area is moved to the end of the longest closure.

The second stage of the segmentation procedure is searching for the onset of voicing using voicing probability. The onset will be the demarcation of the end of VOT, which is the beginning of the following vowel. The reason to do the vowel (or end of VOT) demarcation first is that we can find the beginning of the VOT between *cls* and the vowel demarcation. The demarcation is made by selecting the first frame of voicing according to voicing probability. If C2 is positioned in a voiceless part, we search rightward to find the onset of voicing. If C2 is positioned in a voiced part, we search for the onset of

voicing going leftward. Once it is found, we mark the beginning time line of the onset voicing as a demarcation of the following vowel.

The third stage of the segmentation procedure is making a fine adjustment of $S_l$ (or $S_r$) and *cls* using power difference. The search area for *cls* is from *cls* to the vowel demarcation just determined. We pick the greatest power difference in the search area. If the search area is too long, there is a chance of picking up a spurious power difference which is an accident of VOT frication. So, the area where the greatest power difference is searched for should be minimised. As a rule of thumb, if the area is longer than 30 ms, it is halved and the peak of power difference is searched for from the *cls* to the end of the first half of the area. If it is less than 30 ms, the whole area will be the search area. The method for the adjustment of $S_l$ (or $S_r$) is the same as the one we described for *cls*, but the search area is different. First, the duration between $S_l$ (or $S_r$) and *cls* is measured. If it is longer than 30 ms, it is halved and the first half is used for searching the greatest power difference. If it is shorter than 30 ms, the search area is from $S_l$ (or $S_r$) to *cls*.

Finally, if $S_l$ (or $S_r$) is itself a closure type frame label and is longer than the previously chosen *cls*, *cls* is moved to $S_l$ (or $S_r$). When this happens, there will be no search area for the start of the closure. In this case, we search from the 6 ms left and to 3 ms right of C1.

In Figure 5.4, the labels are compared to those obtained from forced alignment. As it shows, the end of the previous vowel and start of VOT are segmented correctly. The end of the VOT is better than the label from forced alignment.

After getting the labels for closure duration and VOT, the labels were used to get the mean F0 of a vowel after stop. An overview of the automatic segmentation procedure is presented in Figure 5.5.

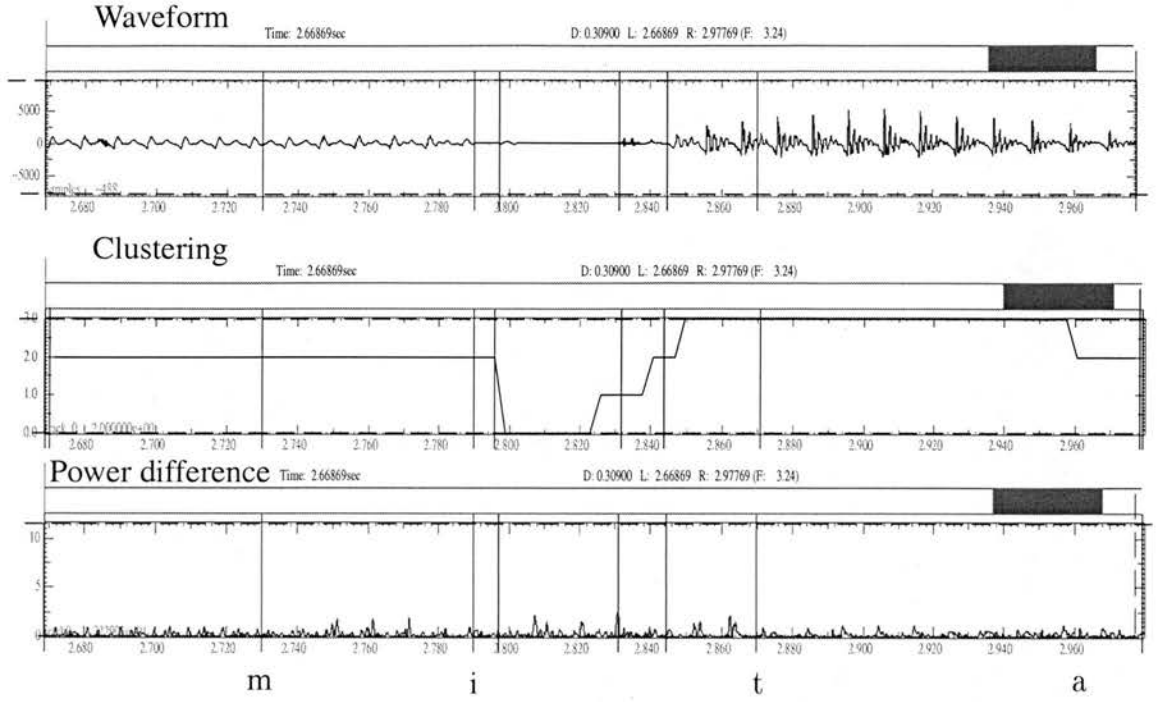



Figure 5.4: Automatic segmentation of closure and VOT. Compared to the label from the forced alignment (upper label), the automatically segmented label (lower label) is more accurate. *cls* represents a closure for the stop [t] and *t* in the lower label represents VOT of the stop.

## 5.4 Post-processing ASR System

A system in which durations are directly modelled in the HMM as in Levinson (1986) produces a probability reflecting acoustic and durational aspects all together. In such a system, the training and testing procedure would be as simple as in a standard HMM-based recognition system, even though the system is computationally complex and costly (Mitchell *et al.* 1995).

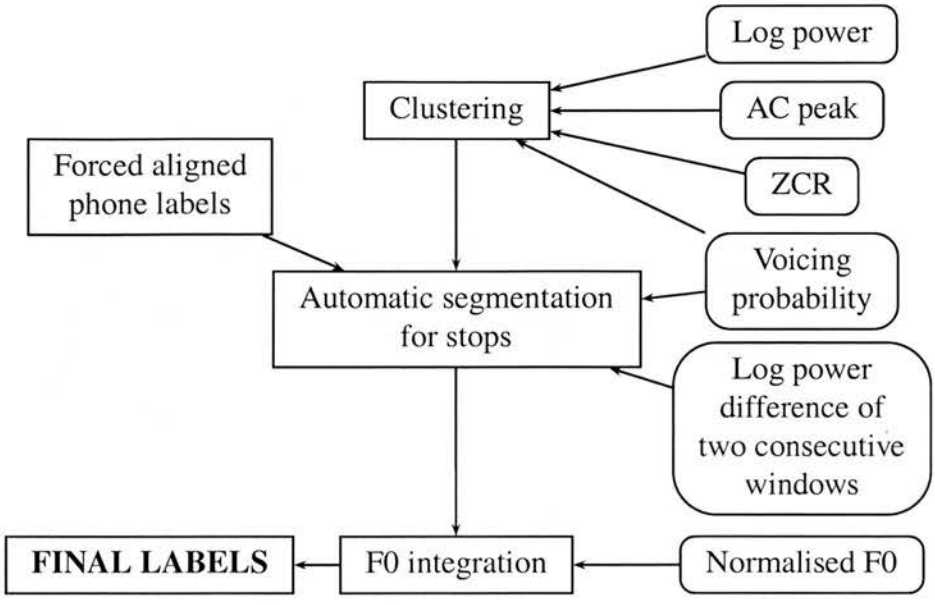



Figure 5.5: Overview of automatic segmentation for stops. Round boxes represent speech parameters that are used in different processes.

A system that models phone durations independently from the HMM must incorporate the durational probability with the acoustic probability. Systems developed in Wang *et al.* (1996) and Molloy & Isard (1998) used a post-processing technique to utilise the durational probability. In these works, the N-best recognition candidates were first obtained and the durational probability was added to the output probability of each N-best hypothesis[2].

In our system, the same post-processing technique is adopted. After having N-best hypotheses, a stop durational probability of each hypothesis is added to that of the hypothesis and then the hypotheses are reordered with the new probability. More detailed calculation of stop probability will be described in Section 5.4.4.

[2]In order to avoid complexity caused by multiplication, HTK calculates probabilities in logarithmic domain. So, the probability is a log probability.

Another issue in developing the ASR system is how to model the three parameters. The three parameters are collected from the training data and modelled with parametric pdfs in two ways. The first one is to use a univariate pdf, such as a normal or a gamma pdf. The second one is to use a multivariate normal pdf. We will look into the modelling procedure in Section 5.4.2 in greater depth.

The system we are describing in this section is a word-level recogniser. A phone-level recogniser will be described in Section 5.5, but the procedure for the post-process and calculation of stop probability using external modelling remain the same in the two recognisers. The system we are referring to is a word-level recogniser, unless otherwise specified.

### *5.4.1 System overview*

The system can be described in two parts, testing and training. Testing mode is about how the probability of each N-best hypothesis is incorporated with the probability from the models. Training mode is about how the closure duration, VOT and F0 from the training data are modelled for generation of stop probabilities.

For testing, the first thing is to do N-best recognition. Since our context dependent HMM system with a short pause model performed with 95.56 % word accuracy, we use this system for N-best recognition. Each of the N hypotheses is labelled with forced alignment so that the boundaries of each phone are demarcated. Stops at this stage have external boundaries but the internal segmentation for closure and VOT are not made yet. Through clustering and automatic segmentation, the beginning and the end of the closure and VOT are demarcated. Using this detailed labelling, vowel F0 after the stop can be calculated.

Figure 5.6 shows an overview of the system in testing mode. For a complete view of the system, it can be put together with Figure 5.5. The boxes labelled as *Automatic segmentation*, *F0 integration* and *FINAL LABELS* correspond to the same labels in Figure

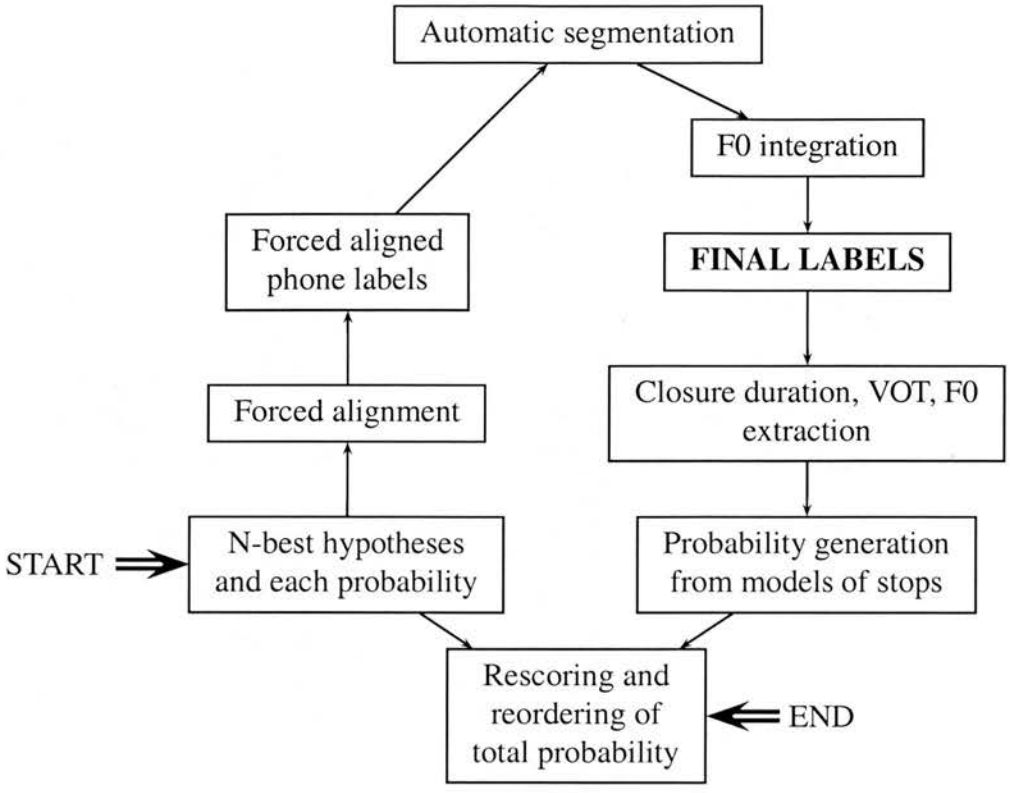



Figure 5.6: Overview of an ASR system

5.5. *Forced aligned phone labels* here are from each N-best hypothesis but in training mode, the phone labels are from each training utterance.

For N-best recognition, we set the N as 10, so that 10 best hypotheses were obtained. Table 5.1 shows the results of N-best recognition from the context dependent HMM system with a short pause model. Word accuracy goes up by about 3 % from the first to the fifth, whereas from the fifth to the last hypothesis, it is less than 0.5 %. Although the word accuracy seems to be saturated at the fifth hypothesis and above, we used all ten hypotheses for rescoring and reordering with stop probabilities.

| n-best | Word accuracy (%) | Sentence correctness (%) |
|---|---|---|
| 1st | 95.56 | 75.92 |
| 2nd | 97.49 | 86.58 |
| 3rd | 98.11 | 89.85 |
| 4th | 98.38 | 91.43 |
| 5th | 98.54 | 92.21 |
| ⋮ | ⋮ | ⋮ |
| 9th | 98.97 | 94.40 |
| 10th | 99.01 | 94.54 |

Table 5.1: Results of n-best recognition. After first five hypotheses word accuracy is saturated.

### *5.4.2 Modelling three acoustic parameters*

Using the automatic segmentation technique described above, closure durations, VOTs and F0s were extracted from the labels of training data. Among 11781 training data, 11280 files had at least one stop sound in the utterance. During the automatic segmentation, 84 files failed to have proper demarcations. The automatic labels of these files have an incorrect phone sequence in a stop area. When a stop demarcation is followed by a closure demarcation, it is regarded as an automatic segmentation failure. Our automatic segmentation technique for closure duration and VOT recorded a 0.74 % error rate.

When modelling the three parameters, contextual variabilities were taken into consideration. Since a stop is conditioned by various factors and factor levels, which were previously described in (3.2) on page 62, each model is specified with these factors and factor levels. Some of the factor levels could be more specific than ones in page 62. For example, instead of using four broad phonological classifications such as sonorant, or vowel for the factor *preceding phone context*, the actual preceding phone could be a factor level by itself, so that a model could be specifically designed for stops preceded by, say, "a". However, there is a trade-off. Using more specific factor levels may produce a model that represents the stop in question better but it may bring about a problem of data

sparsity so that a parametric representation of the model may have an inaccurate mean and variance.

Since the stops were trained with their identities in terms of place and phonation types in HMM, all factor levels for these two factors were specified in the models. Preceding phone contexts and prosodic positions were also used for specification of the models but factor levels for prosodic position were slightly different from what we described in (3.2) on page 62. As we found only a small number of stops in resyllabified syllable initial position in Section 3.7, there were not many tokens for this position in the training data, either. So, as we did in the statistical tests for the hand-labelled KAIST data, we pooled the stops in this position with those in the PW medial position for the models.

We modelled the three parameters with parametric family of continuous pdfs. In doing so, two different approach were attempted. First, each parameter was modelled independently with a gamma or a normal pdf. A total probability of closure duration, VOT and F0 was calculated by multiplying these probabilities. The determination of which univariate pdf should be used was made by calculating Root Mean Squared Error (RMSE) as follows:

(5.2)
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(M_i - x_i)^2}{n}}$$

where $M_i$ is a value predicted from the model, which is either a gamma or a normal pdf.

A normal pdf is defined as:

(5.3)
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

And a gamma pdf is defined as follows:

(5.4)

$$f(x) = \begin{cases} \frac{\beta^{\alpha} x^{\alpha-1} e^{-x\beta}}{\Gamma(\alpha)} & if \quad x > 0 \\ 0 & otherwise \end{cases}$$

where the Gamma function, $\Gamma(z)$, is

$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx \qquad z > 0$$

and $\alpha$, $\beta$ are,

$$\alpha = \mu^2/\sigma^2, \qquad \beta = \mu/\sigma^2$$

Before we calculated the RMSE, we excluded outliers to get robust statistical parameters. For the univariate models, the Euclidean distance between each data point and the mean of each parameter was normalised and data points that had their normalised distance larger than $+3$, or smaller than $-3$ were excluded[3].

Results of the RMSE shows that the gamma pdf is better for all parameters. In other words, the real data of closure duration, VOT and F0 are skewed and that is why a gamma distribution fits better than the normal distribution. As an illustration, a histogram of closure durations of $p^h$ preceded by a vowel and positioned in PW initial is drawn against a gamma pdf in Figure 5.7[4]. On the basis of the result from the RMSE, gamma pdfs were used for all models.

The second way of our modelling the three parameter was to use the multivariate normal pdf[5]. Since we investigated the durational and F0 behaviour in a multivariate way, a

[3]Each area represents 0.3 % of the distribution. Since there are two tails for the distribution of normalised distance, the normalised distance which are in 0.6 % are excluded.

[4]In order to have a better fit to the gamma pdf, the statistical parameters for the gamma pdf was obtained from the data that a minimum closure duration is subtracted from. Whether or not subtracting the minimum, the gamma pdf has smaller RMSE than the normal pdf.

[5]There is no analytic form for a multivariate gamma pdf.

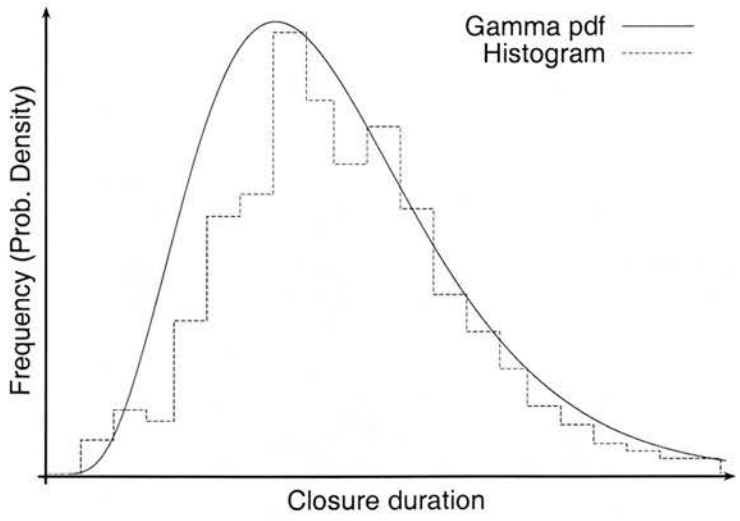



Figure 5.7: Histogram of closure durations and its parametric representation with a gamma pdf. The phone is $p^h$ in intervocalic and PW initial position.

model using the multivariate normal pdf is worth testing. The multivariate normal pdf is defined as follows:

(5.5)

$$f(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$$

where $\mu$ is a mean vector and $\Sigma$ is a covariance matrix, and $n$ is the dimensionality of observation $o$.

Since there are three parameters, closure, VOT and F0, the dimensionality of the observation is three. However, models which represent stops positioned utterance initially have only two dimensions due to the lack of closure duration in that position.

When modelling parameters with the multivariate normal pdf, we also excluded outliers in the same way as we did with the univariate pdf. That is, we excluded the data points

that had their normalised distance outside of $\pm 3$ standard deviations. However, the calculation of distance from each data points to the centre of the distribution is different. Because of the covariance, Euclidean distance is not suitable for the distance measure. So, we used the Mahalanobis distance measure. The distance $r$ from $x$ to $\mu$ is defined as:

(5.6)

$$r = [(x - \mu)^t \Sigma^{-1} (x - \mu)]^{\frac{1}{2}}$$

After having the data trimmed, means and covariance matrix were obtained from the data for the multivariate normal pdf.

### *5.4.3 Performance of automatic segmentation: hand label vs automatic label*

We compared closure durations, VOT and vowel F0 obtained in hand labelled data with those in the same dataset but labelled by automatic segmentation. For the purpose of the comparison of the two groups, a multivariate test was adopted. The multivariate test used for comparing two groups is Hotelling's $T^2$ test, which is a multivariate version of univariate independent t-test. Hotelling's $T^2$ is known to have the same F-value, degree of freedom and significance level as Hotelling's Trace. It is also known that statistical tests based on Hotelling $T^2$ and Wilks Lambda are the same. These tests compare the mean vectors of the two groups and see whether there is no difference in two mean vectors at the significance level of $\alpha$.

Among 433 hand-labelled speech files, boundaries for closure duration and VOT in 6 files failed to be demarcated by the automatic labelling. So, the comparison was made excluding the samples in these files. As shown in Table 5.2, the multivariate test does not reject the null hypothesis, which means that there is no statistical difference between the mean vectors of the two groups. A test of between-subject effects, also shown in Table 5.2, shows that there is no mean difference in the two groups of each parameters. On the basis of these statistical tests, we assume that durations and F0 from the automatic segmentation are not substantially different from those from the manual demarcation.

| Multivariate test | F value | P value |
|---|---|---|
| Wilk's Lambda | 2.566 | .053 |

| Between-Subject Effects | F value | P value |
|---|---|---|
| Closure duration | .539 | .463 |
| VOT | 3.019 | .082 |
| F0 | 5.497 | .019 |

Table 5.2: Multivariate test for comparing two mean vectors of the two groups, one from hand-labels and the other is from automatic labels. Between-subject effects test compares two means of each dependent variable.

### *5.4.4 Probability of each acoustic parameter*

Each probability from the N-best recognition is composed of two parts. One is the acoustic probabilities of phones and the other is the probability from the language model. The way we put our stop probability on top of the probability from the N-best recognition is similar to the method used in Molloy & Isard (1998). So, the total probability of each hypothesis is calculated as follows:

(5.7)

$$P_{total} = P_{N-best} + \alpha P_{stop}$$

All terms are log probabilities and the $\alpha$ is a scaling factor.

If there is more than one stop in the hypothesis, $P_{stop}$ is a mean probability of stops in the hypothesis. However, if there is no stop in the hypothesis, a default probability should be given to the calculation replacing $P_{stop}$ in order not to be prioritised over hypotheses that happen to have one or more stop sounds. Eventually, $P_{stop}$ can be obtained by:

(5.8)

$$P_{stop} = \frac{\sum_{i=1}^{num} P_{stop_i}}{num}$$

$num$ is the number of stops in the hypothesis. If $num$ is zero, then $P_{stop}$ is replaced by $P_{default}$ in which the parameters of stops are from the training data, so $P_{default}$ is an average stop probability.

When the probability is calculated with a univariate pdf, the actual $P_{stop}$ is a sum of three probabilities, one from each parameter. When the parameters are modelled with a multivariate normal pdf, the summation is not needed. That is:

(5.9) Univariate pdf $P_{stop_i} = P_{closure} + P_{VOT} + P_{F0}$

Multivariate pdf $P_{stop_i} = P_{(closure,VOT,F0)}$

As described in 5.4.2, a gamma and a multivariate normal pdf were used for the univariate and the multivariate pdf. After getting total probabilities of all the hypotheses, they were reordered by their probabilities. The final results of the two models in word-level recognition will be presented in Section 5.6. But first, phone-level recognition using univariate modelling with gamma pdfs by the post-processing will be explained in the following section.

## 5.5 Phone-level Recognition with the Acoustic Cues

We have seen in Section 2.5.1 that stop recognition needs to be improved. In order to see whether our external modelling of three acoustic parameters by way of the post-processing method, which we explained in the above section, leads to an improvement, we replicated the experiment done by Jang (2000b) and then reordered the 10 best hypotheses by adding in the stop probabilities just described.

Unlike the Jang's (2000b) phone recogniser, our phone recogniser used context-dependent models. Since the context-dependent models already achieved great improvement compared to the monophone models, there may be fewer stops that are not properly differentiated in type, but if we got improvement with these stops, it would show the value of the external modelling for Korean ASR.

| n-best | Word accuracy (%) |
|---|---|
| 1st | 74.47 |
| 2nd | 75.66 |
| 3rd | 76.47 |
| ⋮ | ⋮ |
| 9th | 78.29 |
| 10th | 78.46 |

Table 5.3: Results of n-best phone-level recognition. About 4% improvement was made.

Since this is phone-level recognition, modelling the acoustic parameters is slightly different from the modelling for word-level recognition. At the phone-level, we cannot predict prosodic position of a stop, so we pooled the samples of word initial and medial position. As shown in Figure 2.6, pooling samples may increase the overlapping area between two acoustically neighbouring types. However, as seen in Section 3.7, the three types of stops are statistically differentiated by the multiple acoustic cues working together. Excluding the prosodic position, we modelled the parameters with three factors, which are phonation type, place of articulation and preceding phone context. We used univariate pdfs for modelling in this experiment.

### *5.5.1 Results and analyses*

Training and test data sets are the same as we used in our baseline system and reference phone sequences were made with forced alignment. Results of N-best recognition are presented in Table 5.3.

Taking the 10 best hypotheses gives only a 4% improvement over the single best hypothesis, which limits the best possible gain from rescoring to 4%. After rescoring total phone sequence probabilities with the stop probability of each sentence, 180 sentences were reordered. The phone accuracy after reordering was 74.48%. No significant improvement

| | Phone-level recognition |
|---|---|
| Post-processing results | Phone acc. (%) |
| All 1st best | 74.48 |
| Num. | 2965 |
| Reordered dataset only | Phone acc. (%) |
| Before | 73.04 |
| After | 73.23 |
| Num. | 180 |

Table 5.4: Post-processing results of stop probability using univariate pdfs.

was made. We compared the 180 reordered hypotheses to the corresponding original first-best hypotheses to see how many stops were corrected by the post-processing method. The result is shown in Table 5.4.

The improvement after reordering is too small to confirm that our method is useful. However, what the post-processing result really shows is there are not many correct phone sequences in any of the 2nd through 10th hypotheses, and there is not much room for improvement in the first place. Nevertheless, if we can find any improvement in recognition of stop sounds, it will confirm that the post-processing is useful for stop recognition.

Table 5.5 shows the stop accuracy in the 180 reordered sentences before and after reordering. Small improvement for bilabial and alveolar stops can be found, but for velar stops, performance deteriorates slightly. The overall impact of introducing acoustic parameters looks small. We further investigated confusions of phonation type and place of articulation of stops. In Table 5.6, confusion of phonation type is generally greater than confusion of place of articulation. This is in line with the result of previous phone recognition in Table 2.5 on page 43. Deterioration of velar stop recognition comes from the confusion of place of articulation.

| | Stop recognition accuracy (%) | | |
|---|---|---|---|
| | Bilabial | Alveolar | Velar |
| Original | 60.60 | 68.14 | 58.63 |
| Reordered | 61.90 | 72.58 | 56.98 |
| Num. of stops | 231 | 383 | 544 |

Table 5.5: Results of stop recognition in each place of articulation.

| | Phonation type confusion | | |
|---|---|---|---|
| | Bilabial | Alveolar | Velar |
| Original | 28 | 29 | 69 |
| Reordered | 27 | 27 | 67 |
| | Place of articulation confusion | | |
| Original | 6 | 22 | 27 |
| Reordered | 7 | 23 | 40 |

Table 5.6: Numbers of stop confusions in phonation type and place of articulation. Each number in phonation type confusion represents stops recognised as different types of stops at the same place of articulation and the number in place confusion represents stops recognised as different stops in other place of articulation.

Since our main concern is phonation type differentiation, we focus on the phonation type confusion. We examined all 126 instances of phonation type confusion in the original hypotheses. We found for the majority of confusions (98) there were no hypotheses that included the correct type of stop in the top 10 hypotheses. We also found 15 stops were corrected in the reordered hypotheses,

We looked at hypotheses containing those 15 corrected stops and found that 13 of them were the only incorrect stop in the original hypothesis. The other two hypotheses had more than two stops, and no alternative on offer where all were correct. When there are more than two stops involved, the possibility of selecting a hypothesis that has all correct stops seems to be slim. However, when there is only one stop, the post-processing works well.

We also investigated cases where stops were correctly recognised in the original first-best hypotheses before reordering and later on, they were recognised with the same place of articulation but with different phonation type. There were 14 stops that were mis-recognised with wrong types of stops. Among these, 2 cases were caused by automatic segmentation failure. 8 cases were found where the values of the acoustic parameters are at the edge of the distribution, which can be considered a failure of the statistical modelling. The remaining 4 cases should be considered as correct. The reason they were picked was because of the stops in the reference phone sequences. As explained in the above section, the reference phone sequence was generated by forced alignment but cross-word phonological rules were not reflected based on our experiments in Section 4.3. The 4 cases were stops that underwent a cross-word phonological rule.

Since the success of the rescoring technique depends on the quality of the hypotheses available to it, it is possible that it will work better in word-level recognition, where the top hypotheses have been further refined by word-level constraints. We can also use probability distributions based on prosodic context at word level. In the next section, we will present our results for word-level recognition.

## 5.6 Results and Analyses of the Word-level Recognition System

Word accuracy was calculated with the first best hypotheses. Few hypotheses were re-ordered. It seems that stop probability does not affect the total probability of each hypothesis enough to change the rank from the N-best recognition. To maximise the influence of stop probability, the scaling factor was increased by 2 and only a few more hypotheses were reordered but the results remained almost the same as shown in Table 5.7. Word accuracy is increased by 0.03 % compared to the result of the original first-best hypotheses from the N-best recognition.

| | Gamma pdf | | Multivariate normal pdf | |
|---|---|---|---|---|
| Post-processing results | Word acc. (%) | Sentence (%) | Word acc. (%) | Sentence (%) |
| All 1st best | 95.59 | 75.99 | 95.56 | 75.95 |
| Num. | 2965 | | 2965 | |
| Reordered dataset only | Word acc. (%) | Sentence (%) | Word acc. (%) | Sentence (%) |
| Before | 65.85 | 0.00 | 83.75 | 33.33 |
| After | 82.93 | 40.00 | 82.50 | 33.33 |
| Num. | 5 | | 9 | |

Table 5.7: Post-processing results of stop probability and comparison of two models. Word accuracy of the dataset that are reordered is also calculated to see the impact of the stop probability in two models.

Since our post-processing experiments with two approaches in modelling the three acoustic parameters for stops do not show improved results, we have to investigate the effectiveness of stop probability in the post-processing system, as we did in phone-level recognition. For the analysis of the system, we explored each reordered hypothesis, whether it was correct one or not, to see if the stop probability was produced correctly.

For the modelling with univariate pdfs, 5 second-best hypotheses were reranked to best. Among these, two hypotheses were correct. The original first-best hypotheses from the N-best recognition had insertion, deletion, and substitution errors in each hypothesis. The three other reordered hypotheses are not correct, but the original first-best hypotheses are all incorrect as well, so incorporating the stop probability has not done any harm. On the other hand, for the multivariate modelling, a few more hypotheses were reordered, compared to the models of single pdfs. Three reordered hypotheses were correct but three original first best hypotheses that were correct were pushed out by incorrect ones. So, the effect cancels out.

We have too little outcome from the addition of the stop probability to claim that the stop probability with the post-processing technique works. However, there is an interesting

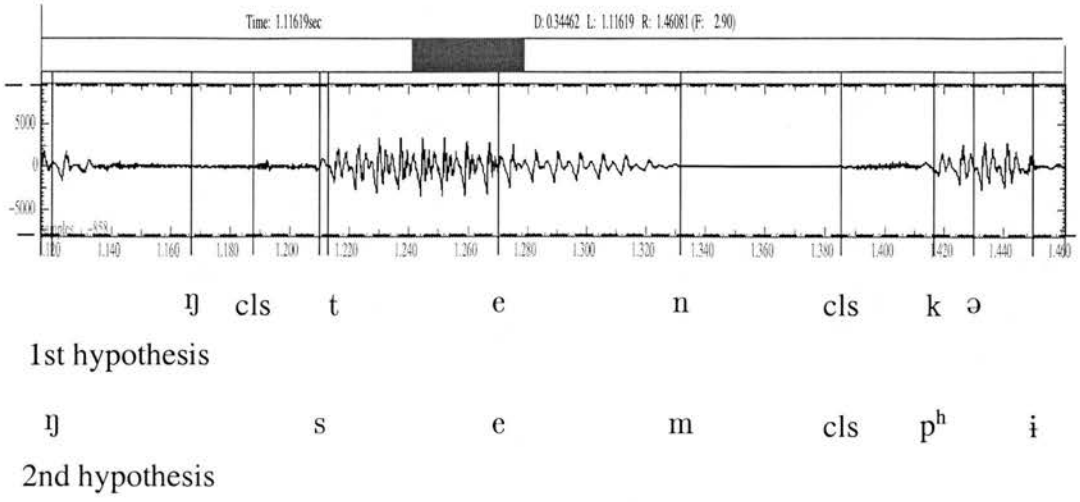



Figure 5.8: A speech part that shows differences in labels of two hypotheses. Labels in the other part of the speech are the same in both hypotheses. *cls* represents a closure for the stops.

thing to draw our attention to. The two reordered and correct hypotheses from the model with a single pdf are the same hypotheses that were picked up by the model with a multivariate normal pdf which has three correct reordered hypotheses.

The first test sentence, for which the models with two different approaches commonly reranked the top two hypotheses, actually has three stops. We illustrate the part of the speech in Figure 5.8, in which we can see two stops in the first hypothesis and one stop in the second one. In other parts of the speech, there are two more stops in both hypotheses and their labels are all the same in both hypotheses. So, the only different part is in Figure 5.8 and we can see the stop [t] in the first hypothesis is replaced by [s] in the second one.

Apart from the wrongly inserted stop [t] in the first hypothesis, its second stop [k] is also different from the corresponding stop in the second hypothesis. They are different in place of articulation and phonation type. Our main concern is the differentiation of

| | Gamma pdf | | | Multivariate normal pdf |
|---|---|---|---|---|
| | Closure | VOT | F0 | All in one |
| 1st ([k]) | 0.0547 | 0.2243 | 1.057e-5 | 0.0127 |
| 2nd ([$p^h$]) | 0.1012 | 0.2599 | 8.984e-5 | 0.0840 |

Table 5.8: Stop probability of each parameter in the gamma pdf model and the multivariate normal pdf one. A word medial [$p^h$] with a nasal for the previous phone context in the second hypothesis shows higher probability in all parameters and all models than the word initial [k].

three types of stops in the same place of articulation, so we did not check whether every possible pair is statistically different regardless of place of articulation in statistical tests in Chapter 3. However, generally speaking, there is a difference in the parameters of different places of articulation and phonation types, which was confirmed by a MANOVA test on page 88 for continuous speech. In this particular case, our models made the correct distinctions because they produced an appropriate probability for each stop shown in Table 5.8.

The second test sentence in which stop probability works nicely shows a more persuasive reason why the stop probability is needed somehow. As seen in the crucial part of the speech illustrated in Figure 5.9, the first and the second hypotheses have very similar phone sequences. In fact, only two phones are different, which are the preceding nasals and the stops. The acoustic probability of the first hypothesis is higher than the second one whereas the probability from the language model is lower in the first one than in the second. When we look at the each phone level acoustic probability just before the nasal, it has the same probability. And the difference of acoustic probability between [n] in the first hypothesis and [ŋ] in the second one is less than $-0.8$ in log probability. But the difference between [k] and [k'] in the first and the second hypotheses is about $-157.9$ log probability. Considering the difference of the total acoustic probability between the two hypotheses ($-30.14$), the difference coming from the stops is the most influential on the

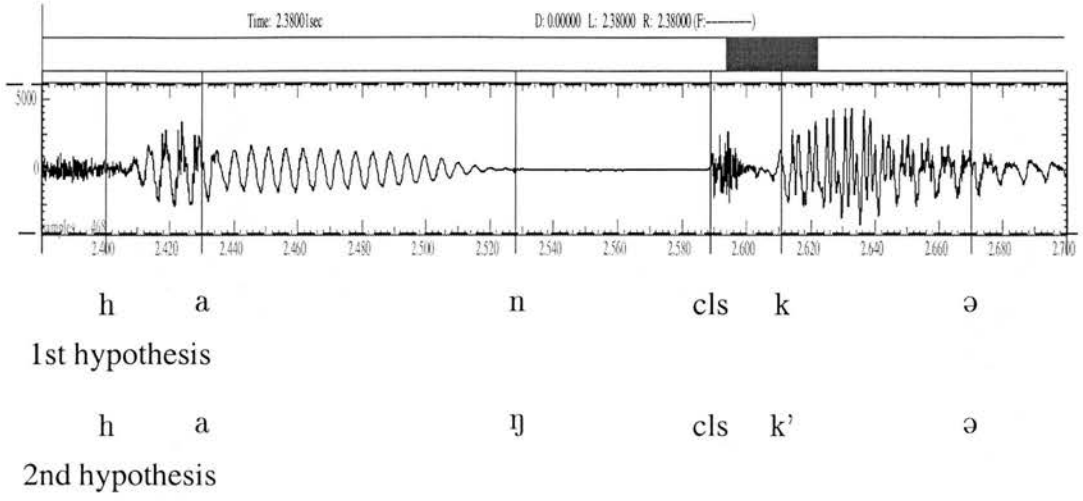



Figure 5.9: A speech part that shows differences in labels of two hypotheses. All other labels are the same in both hypotheses.

| | Gamma pdf | | | Multivariate normal pdf |
|---|---|---|---|---|
| | Closure | VOT | F0 | All in one |
| 1st ([k]) | 0.0478 | 0.1278 | 9.311e-7 | 0.0020 |
| 2nd ([k']) | 0.0500 | 0.2123 | 3.869e-5 | 0.1883 |

Table 5.9: Stop probability of each parameter in the gamma pdf model and the multivariate normal pdf one. A word medial [k'] with a nasal for the previous phone context in the second hypothesis shows higher probability in all parameters and all models than the word initial [k].

difference of the total probabilities. In other words, the second hypothesis is penalised due to the performance of the HMM for [k'].

Since the previous phone context is nasal and the stop's place of articulation is alveolar in both hypotheses, the critical decision on the type of the stop is brought down to the three parameters. And our models predict the type correctly with a higher probability for the tense stop shown in Table 5.9. So, the stop probability can be a supplementary method to contribute to the calculation of proper acoustic probability of a stop.

This second example strongly suggests that stop probability is needed and it would work if we have two competing hypotheses where the word sequences for both hypotheses are the same except for one word pair that is a minimal pair for a stop sound, similarly to what we found in the phone-level recognition experiment. However, in order to confirm this we would need to have a substantial number of minimal pairs in the speech database.

Three points can be made in relation to our non-significant results. First, considering the way stop probability works in the above examples, our speech database does not have enough minimal pairs. We investigated the number of minimal pairs in our database used in the ASR experiment and found that there were only eight minimal pairs. And most of their lexical categories are different, so the bigram language model could produce quite different language model probability for the word sequence that contains one word of the pair.

There is a possibility that the errors made by the N-best recognition may involve stop sounds. Some of the substitutions involving stops are shown in (5.10).

(5.10)

| Correct word | Substituted word |
|---|---|
| **pyən-toŋ** | **pyən-kyənŋ** |
| *variation* | *change* |
| **pe-tal** | **te-tap** |
| *delivery* | *answer* |
| **toŋ-kyəŋ** | **koŋ-caŋ** |
| *Tokyo* | *factory* |
| **to-wa** | **po-yə** |
| *help* | *to be seen* |
| **pul-ə** | **p$^{h}$al-wəl** |
| *French language* | *August* |

In the first-best hypotheses, 36 sentences have at least one substituted word involving stops. Among those, 14 sentences have phonation type confusions, 6 have both phonation type and place confusions, and 16 have place confusions. Since our primary concern was differentiation of phonation types, we took a closer look at the 14 sentences that had phonation type confusions. First, we listened to the speech to see whether the stop was properly spoken. We found that the stops in three sentences could be perceived as different types, so we excluded them for further investigation. Among 11 sentences, 2 have no corresponding stop words in the second through ninth hypotheses. In other words, even though the first-best hypothesis has an erroneous word involving a phonation type of a stop compared to the correct reference sentence, there is no chance for a word with a correct corresponding stop, whether or not the word is correct, to be selected because the rest of the hypotheses have no corresponding word with a correct stop. These 2 sentences were also excluded. We calculated stop probability of 9 sentences. Instead of summing all the stop probabilities in each sentence, we calculated only the stops in the erroneous words. We also calculated probability of the corresponding correct type of stop. 4 in 9 sentences are correctly predicted by the stop probability. Note that there are some cases where the word is not correct, even though the stop used in that word is correct, so the overall performance is not changed. Through this investigation, we found that there are too few examples of stops in our database and too few examples that are involved in phonation type confusions.

In order to check if having a small number of minimal pairs is a general tendency in Korean, we investigated the number of minimal pairs in the POSTECH dictionary which Jang (2000b) used to see the functional load of stops[6]. Table 5.10 shows the number of minimal pairs for each stop pair. The total number of pairs is 13352, which is 23.58% of the total number of vocabularies in the dictionary. The number of minimal pairs looks quite large in this dictionary because some morphemes occur in more than one entry. The dictionary based on *pseudomorphemes* which have been employed in current Korean

[6]Refer to Table 2.6 on page 45. The second dataset in the table is the POSTECH dictionary which has 113194 items.

| | p' | $p^h$ | t | t' | $t^h$ | k | k' | $k^h$ |
|---|---|---|---|---|---|---|---|---|
| p | 991 | 623 | 782 | 153 | 297 | 1337 | 296 | 285 |
| p' | | 188 | 104 | 69 | 63 | 114 | 82 | 47 |
| $p^h$ | | | 397 | 110 | 259 | 583 | 222 | 175 |
| t | | | | 643 | 584 | 959 | 192 | 110 |
| t' | | | | | 418 | 200 | 151 | 81 |
| $t^h$ | | | | | | 475 | 124 | 101 |
| k | | | | | | | 1218 | 560 |
| k' | | | | | | | | 359 |

Table 5.10: Number of minimal pairs of stops in 113194 dictionary words.

| | Number of triplets |
|---|---|
| Bilabials | 165 |
| Alveolars | 120 |
| Velars | 254 |

Table 5.11: Number of triplets of stops in 113194 dictionary words.

ASR systems, as mentioned in page 56, is similar to the POSTECH dictionary in that the same pseudomorpheme can occur in multiple entries. As a result, a similar number of minimal pairs would be expected in a large dictionary constructed for the use of an ASR system.

One of the characteristics of minimal pairs of stops in Korean is that many of the minimal pairs have a synonymous relationship between the two words in the pair[7]. A subtle difference between the two can be found. Sometimes, a word with a stop is interchangeably used with the other type of stop without any obvious reason. Especially, in spontaneous speech, many words are pronounced with an alternative stop in order to give a slight variation in linguistic sense. Some words are found in a dialect or child language. Some of the words are exemplified below.

(5.11)

[7]Stops in this type of minimal pair maintains their places of articulation.

| | | |
|---|---|---|
| **ko-ki** | **k'o-ki** | *meat* |
| standard | child language | |
| **kəm-ta** | **k'əm-ta** | *black* |
| normal | strong | |
| **tol-a-i** | **t'ol-a-i** | *crazy man* (slang) |
| no fun | more fun | |
| **piŋ-kɨl** | **p$^{h}$iŋ-kɨl** | *spin-spin* |
| normal | strong | |

The fact that many of the minimal pairs are in a synonymous relationship, and that the one word can fit to the place where the other word of the pair can be used can cause a problem in training a language model. Even though these words are not much used in written text, it is syntactically acceptable to interchange one word with the other. Language models trained on text will have the wrong probabilities for the alternative pronunciations in spontaneous speech. Thus, it is important to make the distinction at the acoustic level, in which case, stop probability is the most critical criterion for the distinction.

We also counted the number of triplets as summarised in Table 5.11. Some of real examples have already been illustrated in (2.1) on page 9. Triplets are the maximum confusion that can be caused by stop sounds. The more minimal pairs and triplets are found in a dictionary used for ASR, the more importantly the stop probability would be needed. But as seen our experiment, our database does not contain minimal pairs and triplets enough to prove our claim.

The second point is that the non-significant results might be attributed to our manipulation of the three parameters. In particular, vowel F0 was externally modelled unlike in Jang (2000b). It may be possible if we parameterise the F0 in the feature vectors and use only two parameters for external modelling, the system could perform better than the previous one. In order to confirm this, we replicated Jang's experiment in which F0

| | Demisyllable | | Triphone | |
|---|---|---|---|---|
| N-best | Word acc. (%) | Sentence (%) | Word acc. (%) | Sentence (%) |
| 1st | 91.98 | 63.05 | 94.21 | 70.32 |
| 2nd | 94.70 | 76.27 | 96.52 | 82.63 |
| 3rd | 95.57 | 79.84 | 97.20 | 85.87 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 9th | 97.29 | 86.93 | 98.42 | 91.57 |
| 10th | 97.40 | 87.46 | 98.49 | 91.91 |

Table 5.12: Results of N-best recognition of two systems. F0 and its first derivative were added to the feature vectors. 41 features were used in total.

was parameterised in feature vectors along with other features and the HMM was effectively used. In doing so, we set up two systems using different recognition units. First, as Jang did, we used a demisyllable model as a recognition unit because a demisyllable model is good for taking advantage of characteristics of vowel F0 after obstruents. Second, we also used a context-dependent triphone model because the context-dependent model performed better than the demisyllable model as seen in Table 4.7. The advantage that is taken by using a demisyllable could be compensated by better performance of the triphone model.

The new systems used two more feature vectors, which are F0 and its first derivative. Because the F0 was already used in HMM training, two remaining parameters were externally modelled in the same way as we did for our stop probability modelling. Tests were done with 10 best hypotheses. The results of the N-best recognition are in Table 5.12.

Compared to the result from a system that did not use the F0 in feature vectors as presented in Table 5.1, both systems perform less effectively. However, since we will rescore and reorder the probabilities of 10 hypotheses, the maximum word accuracy we can get is the same as the tenth result in Table 5.12. So, the result of the first best hypotheses

| | Demisyllable | | Triphone | |
|---|---|---|---|---|
| Post-processing results | Word acc. (%) | Sentence (%) | Word acc. (%) | Sentence (%) |
| All 1st best | 92.00 | 63.19 | 94.24 | 70.49 |
| Num. | 2073 | | 2965 | |
| Reordered dataset only | Word acc. (%) | Sentence (%) | Word acc. (%) | Sentence (%) |
| Before | 66.96 | 7.14 | 82.89 | 11.11 |
| After | 70.54 | 28.57 | 92.11 | 66.67 |
| Num. | 14 | | 9 | |

Table 5.13: Post-processing results of stop probability using univariate pdfs. N-best recognition were performed by HMMs trained on F0 and its first derivative feature vectors along with other 39 features. A small effect was made by external modelling of closure duration and VOT in the context-dependent model.

does not restrict the potential performance that might be obtained after post-processing procedure.

In these experiments, we modelled the closure duration and VOT only with univariate pdfs. After increasing the scaling factor by two, 14 and 9 hypotheses were reordered as presented in Table 5.13. Judging from the word accuracy obtained from these two systems and the previous one in which F0 was externally modelled, the effect of stop probabilities of these two systems is not as great as that of the previous one whose result is summarised in Table 5.7, even if a very small number of hypotheses were reordered. As the results of all the first best hypotheses from both methods show, the previous system which does not have F0 parameterisation performs better. These experiments reveal that F0, excluded from the external modelling and parameterised in feature vectors, does not make much difference.

The last point would be that the post-processing technique might not be an appropriate method to incorporated our models in the system. In both the phone-level and word-level recognition experiments, we saw that the post-processing method seemed to work

in minimal pairs. In order to exploit the probability from modelling the three acoustic parameters in other environment as well as minimal pairs, an alternative way should be considered. One possible way of using the probability would be to incorporate the stop probability with the acoustic phone probability within the decoder. Modification of the decoder to reflect the stop probability will be left for our future work.

In relation to modelling with a univariate pdf and a multivariate normal pdf, what we learned from our experiments is that independent modelling of the three acoustic parameters is more effective than using the multivariate normal pdf. Even though we have a very small number of reordered hypotheses, the model with a gamma pdf seems to work better than the multivariate model, which may imply that modelling skewed data with a proper parametric pdf is more important than using the covariance. This should be also tested with a proper speech database that includes a lot of minimal pairs of stop sounds.

In sum, stop probability does not make significant effect on the performance of the system by way of the post-processing technique, but it should be tested with proper database that contains more minimal pairs.

## 5.7 Summary

An ASR system was developed using stop probability and tested on a 2965 test dataset which did not participate in training. Stop probability was calculated by the three acoustic parameters, which are closure duration, VOT and vowel F0 after a stop. To get an accurate closure duration and VOT, accurate segmentation must be performed beforehand. However, a general way of getting labels is by forced alignment, which is not satisfactorily accurate for our purpose in two reasons. Firstly, a durational probability density from an HMM is not adequate for modelling duration, and secondly, the conventional frame shift is so large that the HMM can only create crude segmentation.

To get more accurate closure duration and VOT, We implemented an automatic segmentation for stops' internal structure. The automatic segmentation demarcated boundaries

for closure duration and VOT by way of clustering and picking up a peak of difference in log power between the two consecutive windows. Comparison of hand-labels with labels by automatic segmentation revealed that they were not statistically different.

Closure duration, VOT and F0 were modelled in two ways, with univariate and multivariate pdfs. For univariate pdfs, normal and gamma pdfs were compared. A gamma pdf showed a better fit. For a multivariate pdf, a multivariate normal pdf was used. When each stop was modelled, preceding phone contexts and positions in a word were considered.

A post-processing technique was employed to see whether stop probability helped to improve the performance of the system. A log stop probability was added to the log probability of each hypothesis from N-best recognition. If there were more than two stops or no stop at all in the processed hypothesis, a mean stop probability or default probability was given to the calculation of total probability. After adding the probability, the rank of the hypotheses were reordered.

Our result did not show a significant improvement on the performance of the ASR system. The stop probability also did not reorder many hypotheses. However, as shown in the phone-level recognition and two reordered hypotheses from the word-level recognition experiment, modelling of the acoustic parameters is needed in sorting out a minimal pair of stops, if the pair is a source of the confusion of two competing hypotheses. External modelling of acoustic parameters without using F0 showed a similar result to the result from our previous experiment. We found that the HMM system trained with F0 parameters in feature vectors did not outperform our baseline model which conventionally used 39 feature vectors such as MFCC, energy and their first and second derivatives. We also found that the independent univariate modelling with a gamma pdf was better than the modelling with a multivariate normal pdf.

Stop probability using multiple acoustic cues should be tested on a database that contains a lot of minimal pairs to see whether the stop probability obtained from modelling mul-

tiple acoustic parameters actually improves performance of an ASR system. Rather than post-processing to use stop probability, modification of decoder to incorporate the stop probability with the acoustic probability can be a better approach. We will pursue the modification in the future.

# CHAPTER 6

# Conclusion

There are three main conclusions to be drown from this research. First, phonetic cues found in experiments based on controlled speech can also be found in speech with contextual variability. Second, a well trained HMM system with a short pause outperforms the other systems developed on the same speech data. Third, even though we cannot confirm whether our external modelling of stop durations and vowel F0 really helps in differentiating three types of stops in our ASR system, a possibility still remains that the external modelling can be useful in ASR when there are more minimal pairs of stops in lexicon and when there are more stop related substitutions from the N-best recognition. In this chapter, these three things are explained more in detail as our conclusion of this research.

## 6.1 Statistical Tests

Since the purpose of this research is to find out whether specific phonetic information can help the improvement of an ASR system, prerequisites for this purpose are firstly to establish what characteristics the stop durations and vowel F0 after a stop have, secondly, to prove whether these characteristics still play a role in the speech having contextual variability because in ASR speech cannot be restricted to a certain type of sentence.

Korean stops show three-way rather than two-way contrast as shown in English stops. To differentiate three types of stops in phonation, closure duration, VOT and vowel F0 after a stop is used as acoustic cues. The characteristics of these acoustic cues are that tense stops have the longest closure duration, aspirated stops have the longest VOT, and lax stops have the lowest vowel F0 after the stops. These cues are affected by various factors. In particular, the magnitude of acoustic parameters varies depending on prosodic position of stops. However, we confirmed with a stop-rich database that the differentiation of the three types of stops is still possible in spite of these factors.

Unlike earlier phonetic studies, our tests of acoustic parameters on speech databases used multivariate analyses. These three parameters can be viewed as dependent variables for each type of stop, and a MANOVA test is more appropriate as a statistical test than univariate analyses, such as ANOVA.

What we have found through the statistical tests can be summarised in two conclusions. First, different types of stops in Korean can be differentiated by these three acoustic parameters not only in controlled speech but also in speech with contextual variability. Second, differentiation of stops in any prosodic position can only be possible when these three acoustic cues are used altogether. In other words, differentiation can not be done by using only one acoustic parameter among the three.

Statistical significance of the tests on both database implies that the multiple acoustic cues are good source of information that can be used in ASR, if we have a proper way of extracting them and implementing them in a system.

## 6.2 Upgraded Baseline Model

For comparison purposes, we built a baseline ASR system using various modifications. For getting better label files used for further training, we developed a tailored lexicon reflecting cross-word pronunciation variants. Retraining the HMMs with the tailored lexicon made little difference to the word error rate, but rescoring the top hypotheses by

forced alignment recognition with the tailored lexicon showed that acoustic scores were actually better as shown in Table 4.4 on page 118. We also modified the language model that constrains undesirable word sequences by giving low language model probability to those sequences. However, it was not successful because of biased word sequences.

Silence models were also tried with and without forward and backward transitions. And at the same time, a short pause model was created. Systems with a short pause and a silence model without forward and backward transition produced a great improvement compared to the other systems built similarly to our system. In particular, the performance of the context-dependent triphone system was much better than any other systems developed on the same speech database we used. This system even outperformed Jang's (2000b) demisyllable model, which were trained with two more feature vectors, F0 and its derivative.

## 6.3 Implementation of Multiple Acoustic Cues in ASR

Since duration modelling with HMM has theoretical problems, we developed an automatic method of segmentation of closure and VOT. Classifying each 3 ms speech frame as one of 4 groups and picking up the largest power difference between two consecutive windows were employed to adjust the preceding phone label obtained from forced alignment and to demarcate the beginning of VOT. For the adjustment of the end of VOT, we selected the onset of voicing on the basis of voicing probability.

We tested our automatic segmentation technique for accuracy and confirmed that there was no statistical difference between the durations got from the labels using this method and those from hand-labels.

The multiple acoustic cues were implemented in a phone-level and a word-level ASR systems by post-processing. The three parameters were externally modelled with univariate pdfs in phone-level recognition, and with both univariate and multivariate normal pdfs in word-level recognition. Stop probability was added to each N-best hypothesis

and reordered according to the total acoustic probability. Statistically significant system improvement was not achieved. However, analyses of the results and some examples showed a room for possible improvement by using external modelling of closure durations, VOT and vowel F0 after a stop. How the possible improvement can be made will be discussed in the next section.

## 6.4 Further Work

Three kinds of work can be suggested. As we have already mentioned, our non-significant result may be attributed to the speech data our system was developed on. Even though we saw the functional load of stops in Korean to be heavy, adding stop probability did not make much difference. If our data had included a lot of minimal pairs, the result could have been different. In fact, the number of minimal pairs in our dictionary used in the system is very small compared to a general large vocabulary dictionary that can be used in an ASR system. Thus, construction of a new speech database including abundant minimal pairs of stops will be the first work to be done.

Secondly, the way we incorporate stop probability can be reviewed. Instead of using post-processing technique, each stop probability can be incorporated with the stop's acoustic probability. By doing so, we can take advantage of the viterbi algorithm, so that the higher probability of a stop can still remain as one of the hypotheses and possibly be selected at the end of decoding process. This will require intense modification of the code of viterbi implementation.

Thirdly, the automatic segmentation method can be improved. When we localised the area where three demarcations needed for closure duration and VOT were located, we used some thresholds which were chosen on the basis of observations of a limited number of cases. These thresholds could be trained automatically from large numbers of examples.

# Bibliography

ALI, AHMED M. ABDELATTY, JAN VAN DER SPIEGEL, & PAUL MEULLER. 1999. Automatic detection and classification of stop consonants using an acoustic-phonetic feature-based system. In *International Conference on Phonetic Science 99*, 1709–1712, San Francisco.

AMERMAN, JAMES D., & MARTHA M. PARNELL. 1981. Influence of context and rate of speech on stop-consonant recognition. *Journal of Phonetics* 9.323–332.

CATFORD, JOHN CUNNISON. 1988. *A Practical Introduction to Phonetics*. New York: Oxford University Press.

CHO, TAEHONG, & PATRICIA A. KEATING. 2001. Articulatory and acoustic studies of domain-initial strengthening in korean. *Journal of Phonetics* 29.155–190.

CHO, YOUNG-MEE YU. 1987. Phrasal phonology of korean. In *Harvard Studies in Korean Linguistics II*, ed. by S. Kuno etal, Cambridge, MA. Harvard University Press.

CHOI, I. J., O. W. KWON, J. R. PARK, Y. K. PARK, D. Y. KIM, H. Y. JEONG, & C. K. UN. 1995. On the development of a large-vocabulary continuous speech recognition system for the Korean language. *The Journal of the Acoustical Society of Korea* 14(5).44–50. (in Korean).

CLARKSON, PHILIP, & RONALD ROSENFELD. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *European Conference on Speech Communication and Technology 97*, volume 5, 2707–2710.

CREMELIE, NICK, & JEAN-PIERRE MARTENS. 1995. On the use of pronunciation rules for improved word recognition. In *European Conference on Speech Communication and Technology 95*, 1747–1750, Madrid, Spain.

CRYSTAL, THOMAS H., & ARTHUR S. HOUSE. 1988. The duration of American-English stop consonants: an overview. *Journal of Phonetics* 16.285–294.

DUDA, RICHARD O., PETER E. HART, & DAVID G. STORK. 2001. *Pattern Classification*. New York: John Wiley & Sons, 2nd edition.

ENTROPIC. 1998. *ESPS/Waves+ with EnSig 5.3*. Version 5.3.

FERGUSON, J. D. 1980. Variable duration models for speech. In *Proceedings of the Symposium on the Application of Hidden Markov Models to Text and Speech*, ed. by J. D. Ferguson, 143–179, Princeton, New Jersey.

FOUGERON, CÉCILE, & PATRICIA KEATING. 1997. Articulatory strengthening at edges of prosodic domains. *The Journal of the Acoustical Society of America* 101(6).3728–3740.

HAIR JR, JOSEPH, ROLPH ANDERSON, RONALD TATHAM, & WILLIAM BLACK. 1998. *Multivariate Data Analysis*. New Jersey: Prentice Hall Inc., 5th edition.

HAN, J. I., 1996. *The Phonetics and Phonology of Tense and Plain Consonants in Korean*. Cornell University dissertation.

HAN, MIEKO S., & R. S. WEITZMAN. 1965. Studies in the phonology of asian languages iii: Acoustic characteristics of Korean stop consonants. Technical report, University of Southern California.

HAN, MIEKO S., & R. S. WEITZMAN. 1967. Studies in the phonology of asian languages v: Acoustic features in the manner-differentiation of Korean stop consonants. Technical report, University of Southern California.

HAN, MIEKO S., & R. S. WEITZMAN. 1970. Acoustic features of Korean /P, T, K/, /p, t, k/ and /$p^h$, $t^h$, $k^h$/. *Phonetica* 22.112–128.

HARDCASTLE, W. J. 1973. Some observations on the tense-lax distinction in initial stops in Korean. *Journal of Phonetics* 1.263–272.

HOCHBERG, M. M., & H. F. SILVERMAN. 1993. Constraining the duration variance in hmm-based connected-speech recognition. In *European Conference on Speech Communication and Technology 93*, 323–326, Berlin, Germany.

HOLMES, WENDY, & MARK HUCKVALE. 1994. Why have HMMs been so successful for automatic speech recognition and how might they be improved? *Speech, Hearing and Language:work in Progress* 8.207–219.

HUH, W. 1985. *Korean Phonology*. Seoul, Korea: Sam Munhwasa. (in Korean).

JANG, TAE-YEOUB. 2000a. Fundamental frequency in manner differentiation of Korean stops and affricates. *Speech Sciences* 7(1).217–232. The Korean Association of Speech Sciences.

JANG, TAE-YEOUB, 2000b. *Phonetics of Segmental F0 and Machine Recognition of Korean Speech*. University of Edinburgh dissertation.

JEON, JEHUN, SUNHWA CHA, MINHWA CHUNG, JUN PARK, & KYUWOONG HWANG. 1998. Automatic generation of korean pronunciation variants by multistage applications of phonological rules. In *International Conference on Spoken Language Processing 98*, 675–678, Sydney.

JIN, NAM-TAEK. 1992. On the functional load of the Korean phonemes: A quantitative linguistic approach. *Eon-Eo-Yeon-Ku* 5.79–101. (in Korean).

JUN, SUN-AH, 1993. *The phonetics and phonology of Korean prosody*. The Ohio State University dissertation.

JUN, SUN-AH. 1996. Influence of micorprosody on macroprosody: a case of phrase initial strengthening. *UCLA Working Papers in Phonetics* 92.97–116.

JUN, SUN-AH. 1998. The accentual phrase in the Korean prosodic hierarchy. *Phonology* 15(2).189–226.

KAGAYA, RYOHEI. 1974. A fiberscopic and acoustic study of the Korean stops, affricates and fricatives. *Journal of Phonetics* 2.161–180.

KANG, ONGMI, 1992. *Korean Prosodic Phonology*. University of Washington dissertation.

KENSTOWICZ, MICHAEL. 1994. *Phonology in Generative Grammar*. Blackwell.

KESSINGER, RACHEL H., & SHEILA E. BLUMSTEIN. 1997. Effects of speaking rate on voice-onset time in Tai, French, and English. *Journal of Phonetics* 25.143–168.

KESSINGER, RACHEL H., & SHEILA E. BLUMSTEIN. 1998. Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. *Journal of Phonetics* 26.117–128.

KIM, CHIN-WU. 1965. On the autonomy of the tensity feature in stop classification (with special reference to Korean stops). *Word* 21(3).339–359.

KIM, CHIN-WU. 1970. A theory of aspiration. *Phonetica* 21.107–116.

KLATT, DENNIS H. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America* 59(5).1208–1220.

KOO, HEE SAN, 1986. *An Experimental Acoustic Study of the Phonetics of Intonation in Standard Korean.* University of Texas at Austin dissertation.

KOREMAN, JACQUES, WILLIAM J. BARRY, & BISTRA ANDREEVA. 1997. Relational phonetic features for consonant identification in a hybrid ASR system. *Phonus* 3.83–109.

KWON, OH-WOOK. 2000. Performance of LVCSR with morpheme-based and syllable-based recognition units. In *International Conference on Acoustics, Speech, and Signal Processing 2000*, 1567–1570, Istanbul, Turkey.

KWON, OH-WOOK, KYUWOONG HWANG, & JUN PARK. 1999. Korean large vocabulary continuous speech recognition using pseudomorpheme units. In *European Conference on Speech Communication and Technology 99*, 483–486, Budapest.

KWON, OH-WOOK, & ALEX WAIBEL. 2002. Korean broadcast news transcription using morpheme-based recognition units. *The Journal of the Acoustical Society of Korea* 21(1E).3–11.

LADEFOGED, PETER. 1982. *A Course in Phonetics.* Harcourt Brace Jovanovich Inc., 2nd edition.

LAVER, JOHN. 1994. *Principles of Phonetics.* Cambridge University Press.

LEE, H., M. ZHI, & Y. KIM. 1993. The acoustic effects of major Korean coarticulatory processes. *HAN-GEUL* 220.5–27. (in Korean).

LEE, H. B. 1973. A phonetic study of the accent in Korean. *Mullidaehakpo (Seoul National University)* 19.1–16. (in Korean).

LEE, HO-YOUNG. 1996. *Korean Phonetics*. Seoul, Korea: Taehaksa. (in Korean).

LEE, HYUCK-JOON. 1998a. Non-adjacent segmental effects in tonal realization of accentual phrase in Seoul Korean. In *International Conference on Spoken Language Processing 98*, volume 3, 623–626.

LEE, JUN-HWAN, & SANG-BURM RHEE. 1999. A study on consonants/vowels phonetic segmentation of Korean isolated words based on a rule-based system for the phenomenon of Korean vocalization. In *IEEE TENCON*, 1351–1354.

LEE, KIYOUNG, & MINSUCK SONG. 2000. Automatic detection of intonational and accentual phrases in Korean standard continuous speech. *Speech Sciences* 7(2).209–224. (In Korean).

LEE, KYONG-NIM, & MIN-HWA CHUNG. 2002. Discriminative allophonic rules for optimizing pronunciation dictionary in Korean LVCSR. *The Journal of the Acoustical Society of Korea* 21(7). (in Korean).

LEE, SOOK-HYANG. 1998b. On the Effects of Places of Articulation of Stops on Their Closure Duration in Korean. *JASK* 17(6).8–13. (in Korean).

LEE, WONIL, GEUNBAE LEE, & JONG-HYEOK LEE. 1995. Phonological modeling for continuous speech recognition in Korean. In *Proceedings of the natural language pacific rim symposium*, Seoul, Korea.

LEVINSON, S. E. 1986. Continuously variable duration hidden markov models for automatic speech recognition. *Computer Speech and Language* 1.29–45.

LINDBLOM, BJORN. 1983. Economy of speech gestures. In *The Production of Speech*, ed. by Peter MacNeilage, 217–245. New York: Springer-Verlag.

LISKER, L., & A. ABRAMSON. 1964. A cross-language study of voicing in initial stops: Acoustic measurements. *Word* 20.384–422.

LISKER, L., & A. ABRAMSON. 1967. Some effects of context on voice onset time in english stops. *Language and speech* 10.1–28.

LISKER, LEIGH. 1957. Closure duration and intervocalic voiced-voiceless distinction in English. *Language* 33.42–49.

LUCE, P. A., & J. CHARLES-LUCE. 1985. Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America* 78(6).1949–1957.

MCKENNA, JOHN, 2000. *Kalman Filtering Towards Automatic Speaker Characterisation*. Edinburgh: University of Edinburgh dissertation.

MILLER, J. L., K. P. GREEN, & A. REEVES. 1986. Speaking rate and segments:a look at the relation between speech production and speech perception for the voicing contrast. *Phonetica* 43.106–115.

MITCHELL, CARL, MARY HARPER, & LEAH JAMIESON. 1995. On the complexity of explicit duration HMM's. *IEEE Transactions on Speech and Audio Processing* 3(3).213–217.

MOLLOY, LAURENCE, & STEPHEN ISARD. 1998. Suprasegmental duration modelling with elastic constraints in automatic speech recognition. In *International Conference on Spoken Language Processing 98*, 2975–2978.

NESPOR, M., & I. VOGEL. 1986. *Prosodic Phonology*. Dordrecht: Foris Publication.

NEY, H., U. ESSEN, & R. KNESER. 1994. On structuring probablistic dependencies in stochastic language modelling. *Computer Speech and Language* 8(1).1–38.

OH, MIRA. 1999. Korean prosodic structure and focus. In *International Conference on Phonetic Science 99*, 1517–1520, San Francisco.

OHALA, JOHN J. 1978. Production of tone. In *Tone: A linguistic survey*, ed. by Victoria Fromkin, 5–39. Academic Press.

PAE, J., J. SHIN, & D. KO. 1999. Some acoustical aspects of Korean stops in various utterance positions :focusing on their temporal characteristics. *Korean Journal of Speech Sciences* 5(2).139–159. (in Korean).

PARK, HEA SUK, HAJIME HIROSE, HIROHIDE YOSHIOKA, & MASAYUKI SAWASHIMA. 1982. An electromyographic study of laryngeal adjustments for the korean stops. In *Linguistics in the Morning Calm*, ed. by The Linguistic Society of Korea, 659–671, Seoul, Korea. Hanshin publishing company.

PARK, JONG RYEAL, OH WOOK KWON, DO YEONG KIM, IN JEONG CHOI, HO YOUNG JEONG, & CHONG KWAN UN. 1995. Speech data collection for Korean speech recognition. *The Journal of the Acoustic Society of Korea* 14(4).74–81. (in Korean).

PARK, YOUNG-HEE, & MINHWA CHUNG. 2001. Automatic generation of concatenate morpheme based language models for Korean LVCSR. In *International Conference on Speech Processing 2001*, Taejon, Korea.

PIERREHUMBERT, JANET. 1979. The perception of fundamental frequency declination. *The Journal of the Acoustical Society of America* 66(2).363–369.

PILLAU, FLORIAN. 1997. An experimental investigation of sosokhui tone perturbation. *Forschungsberichte des Instituts Für Phonetik und Sprachliche Kommunikation der Universität München* 35.149–164.

PORT, ROBERT F., 1977. *The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words.* Bloomington: Indiana University dissertation.

RABINER, L. R., & B. H. JUANG. 1993. *Fundamentals of Speech Recognition.* New Jersey: Prentice Hall.

REPP, BRUNO H. 1984. Closure duration and release burst amplitude cues to stop consonant manner and place of articulation. *Language and Speech* 27.245–254.

RUSSELL, M. J., & R. K. MOORE. 1985. Explicit modelling of state occupancy in hidden markov models for automatic speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing 94*, volume 1, 5–8, Tampa, Florida.

SELKIRK, E. 1984. *Phonology and Syntax: The Relation between Sound and Structure.* Cambridge, Massachusetts: MIT Press.

SHATTUCK-HUFNAGEL, STEFANIE, & ALICE E. TURK. 1996. A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* 25(2).193–247.

SILVA, DAVID J. 1992. *The Phonetics and Phonology of Stop Lenition in Korean*. Ithaca: Cornell University.

SPSS, 1999. *Statistical Package for Social Science*. SPSS Inc. Version 10.0.

STRIK, H., & C. CUCCHIARINI. 1998. Modeling pronunciation variation for asr: overview and comparison of methods. In *Proceedings of the ESCA workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, ed. by H. Strik, J. Kessens, & M. Wester, 137–144, Rolduc, The Netherlands. University of Nijmegen.

SUH, YOUNGJOO, KYUWOONG HWANG, OH-WOOK KWON, & JUN PARK. 1998. Improving speech recognizer by broader acoustic-phonetic group classification. In *International Conference on Spoken Language Processing 98*, 1107–1110, Sydney.

TALKIN, D. 1995. A robust algorithm for pitch tracking (RAPT). In *Speech Coding and Synthesis*, ed. by K. K. Paliwal. Elsevier.

VOLATIS, LYDIA E., & JOANNE L. MILLER. 1992. Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America* 92(2).723–735.

WANG, XUE, 1997. *Incorporating Knowledge on Segmental Duration in HMM-Based Continuous Speech Recognition*. University of Amsterdam dissertation.

WANG, XUE, LOUIS F. M. TEN BOSCH, & LOUIS C. W. POLS. 1996. Integration of context-dependent durational knowledge into HMM-based speech recognition. In *International Conference on Spoken Language Processing 96*, 1073–1076, Philadelphia.

WHALEN, D. H., & A. G. LEVITT. 1995. The universality of intrinsic f0 of vowels. *Journal of Phonetics* 23.349–366.

YOUNG, S., J. JANSEN, J. OLLASON, & P. WOODLAND. 1996. *HTK Book*. Entropic.

YU, HA-JIN, HOON KIM, JOON-MO HONG, MIN-SEONG KIM, & JONG-SEOK LEE. 2000. Large vocabulary korean continuous speech recognition using a one-pass algorithm. In *International Conference on Spoken Language Processing 2000*, Beijing, China.

YUN, SEONG JIN, HWAN JIN CHOI, & YUNG HWAN OH. 1997. Stochastic pronunciation lexicon modeling for large vocabulary continuous speech recognition. *The Journal of the Acoustical Society of Korea* 16(2).49–57. (in Korean).

YUN, WEONHEE, & TAE-YEOUB JANG. 1999. Statistical analysis of Korean stop durations and its application for speech recognition. In *International Conference on Speech Processing 99*, volume 1, 305–310, Seoul, Korea.

ZHI, M. 1993. The duration of sound. *Sae-kuk-eo-saeng-hwal* 3(1).39–57. (in Korean).

ZHI, M., Y. J. LEE, & H. B. LEE. 1990. Temporal structure of Korean plosives in /VCV/. In *Proceedings of The Seoul International Conference on Natural Language Processing 90*, 369–374, Seoul, Korea. Language Research Institute, Seoul National University.