# An Interpolated Accent Map for Scotland

Examination Number 3787347

MSc Speech and Language Processing

The University of Edinburgh

2010

# Contents

# List of Figures

# List of Tables

A proposal has been made to create an application which displays a map of Scotland and, when clicked, synthesizes speech approximating the dialect of the location clicked on, regardless of whether or not data has been collected for that precise area. This is to be accomplished by creating HMM voices for a number of locations across Scotland and interpolating appropriately between these voices at locations specified by the user. This report serves as a proof of concept for this project. It addresses two concerns: one, demonstrating that the Voices of Young Scots corpus, which collects recordings of youth across Scotland, is a suitable corpus for training these models; and two, building an extensible framework for the application and testing it using a set of voices with known geographical information. Both of these concerns are demonstrated to be surmountable, and some discussion is made of future work.

# Declaration

I have read and understood the University of Edinburgh guidelines on Plagiarism and declare that this written dissertation is all my own work except where I indicate otherwise by proper use of quotes and references.

# Part I.

# Introduction

# 1. Eventual goal

The English spoken in the UK is anything but uniform. While there may be only a handful of varieties normally heard on TV, variations in dialect can be detected from town to town. These variations generally occur on a continuum, with locations close to each other sounding more alike than those which are far apart. Of particular interest is the situation in Scotland, where a separate literary and linguistic tradition flourished for many years and acquired its own distinct dialects, and where until relatively recently two unrelated languages prevailed in different regions. For entities concerned with promoting and educating about Scottish culture, such as the national museums, the ability to display this diversity would be a great asset.

One way to present the diversity of Scottish English would be to provide a way for the interested user to hear every local dialect. Unfortunately it is infeasible to get recordings from every location in a country, and even were it feasible a huge body of recordings would be required from each area to demonstrate every relevant feature. Speech synthesis can solve this problem. We propose that, having trained high-quality parametric models representative of several dialects, the remaining regional variation can be estimated by weighted interpolation. These models and their interpolated intermediates can be implemented into a map-based application so that a user can choose any location in a region and hear an approximately correct voice say anything they like, within reason.

The eventual goal is to create such an application to cover the whole of Scotland and possibly the rest of Great Britain for contrast.

# 2. Proof of concept

Due to time and resource constraints, it is impossible for a single developer to build a high-quality dialect map of Scotland over the course of one summer. The best-quality synthesized voices take a great investment of time and expertise and a large number of voices are required to make a useful map. This project should be seen as a proof of concept for the final application. This proof of concept takes two parts:

- First, demonstrating that the Voices of Young Scots corpus can be used to create hidden Markov model-based voices of adequate quality even with relatively minimal resources. It should then be possible to create good quality voices from this corpus when more advanced methods are applied. This is an important point to demonstrate, as the VOYS corpus is unique in two regards: it uses adolescents instead of trained adults, with the associated decrease in professionalism, and collection of recordings is decentralized, with some effects on quality control. [3] It is worthwhile to explore whether such corpora can still produce quality voices, as considering the relative ease of data collection they are more attractive to the prospective corpus builder, and what particular pitfalls might be encountered. It is also important to demonstrate that the regional variations in pronunciation are reflected accurately in the synthesized voice. Other aspects of dialect apart from pronunciation will not be addressed at this time.

- Second, demonstrating that a map with embedded voice interpolation can be used as an intuitive and efficient way of organizing and displaying dialect continuums. A successful speechmap should, when a particular location not represented by a constructed voice is specified, be able to produce intelligible and

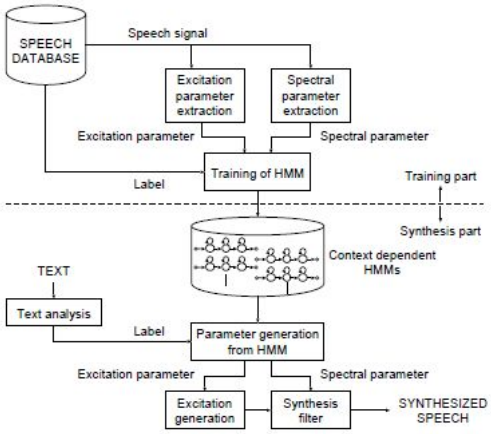
reasonably accurate speech from trained models for nearby locations. For this project, influence is modelled as flowing from the two nearest speech innovation centers only, proportional to their geographical closeness, but more complicated algorithms can easily be integrated. Due to the low number of Scottish voices that were able to be completed in the time available, a subset of WSJCAM0 voices from England will be used instead; conclusions about the functioning of the WSJ-CAM0 speechmap should however generalize to a VOYS speechmap or indeed any other corpus.

# Part II.

# HTS and HMM Synthesis

# 3. Theory

Figure 3.1.: Flowchart giving an overview of the training and synthesis procedure for an HMM voice.




Flowchart courtesy [11]

It should be noted that there are some points of variation in HMM-based speech synthesis. Where there are different approaches, I have described only the one used in building voices for this project, particularly the approach used by early versions of HTS, described more fully in chapter 4.

## 3.1. Structure of an HMM voice

In broadest terms, as a parametric approach to speech synthesis, an HMM voice must consist of statistics for sounds and a method of determining which statistics are associated with which sounds. These statistics can be trained from data and then used

to estimate the appropriate characteristics for a desired utterance, which can then be synthesized.

So what are these statistics? Duration, $F_0$, and spectral characteristics are modeled separately. (Aperiodicity is also modelled in more recent models, but was not used in building the VOYS voices and will not be considered in depth here.) This represents a source-filter model, where speech is considered to be a combination of source energy, or $F_0$, and the vocal tract acting as filter, represented by spectral correlates.

- Pitch or $F_0$ is the fundamental frequency of a speech signal, corresponding to the frequency at which the vocal folds vibrate during a voiced utterance. Speech can also be unvoiced, with excitation coming from white noise instead of a periodic vibration. Unvoiced speech can be hard to model with an HMM, as it does not have a defined frequency and cannot be produced. Therefore $F_0$ is modeled not by a simple HMM with a Gaussian and probabilities, but by a multi-space distribution or MSD-HMM which can produce results of varying dimension. These allow for the production of a zero-dimensional result for unvoiced F0 without losing the desirable properties of HMMs. [11]

- Spectrum: While we refer to spectral characteristics of the vocal tract, it is more accurate to refer to the cepstrum when talking about the actual statistics used by an HMM voice. Mel-generalized cepstral coefficients are the result of a series of transforms on the speech signal to reduce correlation and separate the shape of the envelope from the periodic peaks due to $F_0$. These are skewed according to the Mel scale to emphasize the particular frequencies that the human mind is most attuned to when listening to speech. The key point is that the MGCCs of a particular speech window are a series of numbers which indicate the vocal tract shape, essentially the phone being uttered aside from voicing. [12]

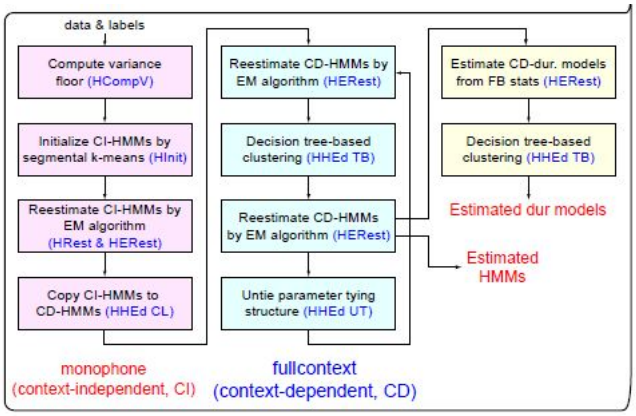
- Duration, Traditionally when modelling with hidden Markov models, the duration spent in any particular state is modelled via the self-transition probability on that state. However, this is not appropriate for modelling speech, as it produces

> a geometric decay in likelihood. Actual phone durations, however, are likely to cluster around a mean observation with some variance to either side, much like the other characteristics of speech. What we have referred to as HMMs to model the other characteristics are therefore more rightly called HSMMs: semi-Markov, as transition probabilities are governed by a Gaussian trained on observed statistics rather than by the probability of a particular number of self-transitions. [14]

For the dialectal modelling that is our goal, we expect spectral characteristics to be the most relevant as they correspond to pronunciation. Average overall pitch is fairly irrelevant, as it varies by individual, but pitch curves across entire sentences may reflect dialect; witness the "sing-song" description given to the intonation of some Welsh accents, for example. Duration's importance is unclear; speakers of some dialects are anecdotally considered to talk more slowly or quickly than others, and accurately representing duration may be part of modelling accents convincingly.

We have referred several times to "phones", but a good acoustic model does not only consider the current phoneme. This is especially true of HMM-based speech synthesis, where context can include two phonemes in either direction, the position of the current phoneme in its syllable, the position in the word, the position in the sentence, the stress on the current syllable, the number of syllables in the current word, part of speech, and any number of other context features which may be relevant to the realization of the current phoneme. [12] For the sake of brevity, we will call this full-context structure a "phone" for the remainder of this paper. Just about any feature which can be determined or predicted automatically given an utterance text can be used as part of the context for an acoustic model. This is far more context than is often used for tasks like ASR. The goal is to aid in making synthesized speech more natural; it is not clear what features will or will not be influential, and it is better to consider any possibility as unimportant contexts will simply not be used. While ASR only has to consider enough context to produce a model which is close enough to be recognizable, speech synthesis seeks to go beyond and reflect the very minute distinctions which can move a voice out of the Uncanny Valley into naturalness.

Figure 3.2.: Flowchart giving more detail on the training process.




Flowchart courtesy [12]

Of course not every possible phone will be observed in training, as there are millions of possibilities. Even those which are observed will often be so rare that accurate statistics cannot be drawn from them. Clustering is necessary to create groupings of similar-sounding phones which are frequent enough to use for training, and a decision tree is built in the clustering process to allow navigation. This decision tree can later be used during synthesis to determine statistics for other unseen phones, and therefore the decision tree is as a necessary part of any HMM voice as the corresponding cluster statistics.

## 3.2. Training an HMM voice

The first step in training a voice is to extract the relevant statistics from the recordings to be trained on. Excitation statistics and spectral statistics are extracted separately, the former via an automated pitch tracking algorithm and the latter via a series of transforms and warpings on windowed sections of speech. After this extraction, the only data being trained on is numbers, a set describing estimated pitch and spectral characteristics for each window length. The goal of training is to correctly align each window to the appropriate phone label and determine from those statistics how to model that phone.

Given a transcription of each utterance, labels are generated which predict the surface realization of that transcription, usually with Festival or a similar text analysis system. Consistent estimation is key to success. For the first iteration of alignment and training, only monophones are considered; there is unlikely to be enough data for any one full-context phone to be able to train it correctly from scratch.

An initial alignment for each utterance is produced using $k$-means, with the observed signal divided into the expected number of phones and similar contiguous signals aligned alike. From this initial alignment it is possible to create initial statistics for $F_0$ and spectral coefficients for each estimated phone. Although the alignment may be well off in some cases, as long as there is not too much silence in this recording it is likely that a significant portion of the alignments will correspond to some of the correct parts of the recording. Mean and variance can then be computed from the aligned statistics and used to find a new most likely alignment in an application of the EM algorithm. It is worth noting, however, that a universally bad alignment is very difficult to recover from, as no models useful for improving alignment can be produced.

After this process has performed a few loops, a reasonable alignment should have arisen and an acceptable monophone model created for each context-free phone. (Unsurprisingly this procedure will not work if there are monophones in the phoneset not attested in the training data; HTS will shut down if it detects such a situation.) This monophone model can then be copied for every possible context, so that a model exists of each full-context phone but is simply the same model as its context-free associate. These full-context models can be retrained using full-context labels, which should differentiate observed versions of the former monophone that have different surface realizations.

When the observed full-context models have been trained enough that differences arise, it is possible to build a decision tree to cluster phones so that there is enough data to reasonably train. This tree, with questions at each node about the context of the current phone and "yes" and "no" branches below, is built by repeatedly choosing the question from a preset list which divides the current phone models into the most self-

similar groups. This process stops for any particular leaf when the group falls below a particular threshold. The decision tree will be different for every voice and reflects the particular phonological mergers used by that speaker. Therefore the structure of the decision tree is also key to modelling a particular dialect.

Each cluster, denoting phones grouped together in a leaf of the decision tree, is grouped together for the next rounds of training. They share the same model and will be realized the same way, barring future reclustering. Having a good alignment before initial clustering is quite important, therefore, as the decision of which phones sound sufficiently alike is dependent on the models pre-clustering. After a few rounds of re-estimation, phones can be untied, retrained, and a new decision tree generated, which may exhibit different clustering if the alignment has changed. This gives some capacity for recovery if the initial clustering is imperfect.

After repeated retrainings - the more the merrier - a final alignment is produced for each utterance. These alignments are used to fit a Gaussian for the duration of each phone cluster to its observed lengths in the training data. At this point all three statistics have been estimated for every observed phone and a decision tree built from the observed mergers in this data set which can later be traversed to find the best predicted cluster for an unseen phone. [12]

The choice of recordings for training has a great impact on the type of voice produced. Voices can be speaker-dependent, trained entirely on the recordings of one speaker in an attempt to emulate that speaker's voice, or speaker-independent, where training data comes from many different speakers and the resulting voice is a weighted average of the training speakers. Speaker-dependent voices are the easiest way to emulate a very particular speaking style, but require a lot of data from a single speaker. Speaker-independent voices can be harder to make specific, but the ability to use multiple speakers opens up the possibility of using far more data.

A popular method of overcoming sparse data while maintaining similarity is to use speaker adaptation. In speaker adaptation, an average voice is built with a large amount of data from different speakers, possibly using normalization of some sort to better re-

duce variability between individuals. Adaptive techniques inspired by ASR are then used to build a transformation from the average voice to a target speaker; this adaptive transformation can be trained on relatively little data while the final voice still has the high quality of the average voice. [14] While speaker adaptation was not used for the voices I built, it was used to build the WSJCAM0 voices which I used for my proof of concept speechmap and is proposed for future work with the VOYS corpus.

## 3.3. Synthesis with HMM voices

The first step in synthesizing an utterance from a prepared, trained HMM voice is to generate labels for the desired utterance text. These labels must be developed using the same phoneset and context characteristics as the labels used to train the data on. This is frequently done using Festival, as unlike HTS Festival already contains all the appropriate framework for determining context, and this project will do the same. Note that the label sequence is not affected by any of the trained models; variation between voices arises from how those models are realized given their fixed contexts.

Having determined the desired label sequence, HMMs can be retrieved from the trained voice. Given the vast number of possible labels, most labels will not have been seen during training; appropriate statistics are determined by navigating the trained decision tree until a leaf is reached. Once the appropriate HMMs are determined for each phone, these are concatenated to represent the desired utterance. The duration models are then used to determine the most likely sequence of phone durations for the full HMM.

If synthesis was done at this point, using these estimated characteristics, speech would be very choppy in both pitch and pronunciation. Choosing the most likely HMMs in sequence tends to output a sequence of the mean values of each HMM for the mean duration, as that is their most likely value. As the mean values for each statistic differ for each phone (or so we hope!), values for each statistic over time will be discontinuous. Thankfully the model contains delta and delta-delta statistics for both cepstral

coefficients and $F_0$, representing the derivative at that point and the derivative's derivative. This information allows for calculation of the most likely smooth curve, hitting the most likely value but connected between those peaks and valleys in a way that is both smooth and probable. [12]

Having determined the sequence of $F_0$ and spectral characteristics, as well as their respective durations, speech can now be synthesized. As the excitation and spectral models have been deliberately separated throughout this process, it is possible to simulate them separately and combine them. There are several methods for producing these two elements, depending on the vocoder and type of excitation desired, but the underlying theme is that a vocoder resynthesizes speech by producing an excitation signal corresponding to the predicted $F_0$curves and filtering it through the calculated cepstral coefficients, much like the human vocal tract does.[11]

# 4. HTS

The HMM-Based Speech Synthesis System, abbreviated as H Triple S or HTS, is created and maintained by the HTS Working Group, a group of volunteer developers centered around the Nagoya Institute of Technology and the Tokyo Institute of Technology. It functions as an patch to the HTK Hidden Markov Model toolkit which adapts and extends HTK's existing speech recognition tools to be more useful for speech synthesis. HTS also modifies the clustering mechanism of HTK and implements some of the structural innovations described in the previous section, such as separate state duration modelling and MSD-HMMs. The voices built in Part III use HTS version 2.1. [11]

A standalone synthesis engine called hts_engine is also included, which can be run without installing HTK. This is useful for making portable applications, and a modified and updated form of hts_engine is used for the speech interpolation map described in Part IV. This updated version is not publicly released and was graciously provided by Junichi Yamagishi.

# 5. Interpolation

The final key theoretical concept for this project is interpolation. Broadly speaking, interpolation is a combination of the features of two or more voices to produce a synthesized utterance which is a weighted average of the input voices. Interpolation can be performed separately on each characteristic, as their decision trees and cluster statistics are separate, but for simplicity unless otherwise stated all characteristics will be interpolated alike.

While it would theoretically be possible to create a new voice whole by interpolating two or more voices, this is not practical unless the voices use the same decision trees. This can be done - perhaps by simultaneously building the best tree that applies to both when the voices are trained, if interpolation is expected, or by untying and retying - but it is inefficient, and in a case where arbitrarily many weighted combinations of voices are possible it is not feasible. Having matching trees would also conceal the particular dialectal differences we wish to draw out. It is better to use the respective decision trees of each voice to be interpolated, get the appropriate statistics for each phone in the utterance to be synthesized, and average between the prediction from each model to produce the final set of statistics to be used. This minimizes the amount of time and processing required. [10, 17] An arbitrarily large number of voices can be used, as the final step would simply involve finding more statistics and averaging them appropriately, but we will use two here for theoretical simplicity and efficiency.

Interpolation is generally used whenever data does not exist for some target speaker, but voices do exist for other speakers that can be combined in some way to represent the target. Simulating a dialect continuum as we are doing in this project is an obvious

example, but it is also possible to use interpolation as a more general form of speaker adaptation, where voices trained on a large amount of data can be interpolated appropriately to create a new voice to match a target speaker. [17] It can also be used to build intermediate voices between two extremes, for entertainment or other purposes.

# Part III.

# Building Voices from Amateur Data

# 6. Voices of Young Scots corpus

## 6.1. Speakers and Speech Recording

The Voices of Young Scots corpus is innovative in two ways, both of which affect its use for speech synthesis: it focuses on adolescent speakers, with most of the participants between the ages of 13 and 18, and speech collection was done remotely via the Internet using the WikiSpeech content management system. One representative school was chosen for each of ten locations from seven distinct dialect regions, with a preference for schools with a mixture of working and middle class students approaching the local average. Each school was sent an identical set of recording equipment but all organization and recording was done by the schools. Prompts were displayed to the speaker by a web application; while a teacher was on site to organize the recording session, most interaction with the speaker appears to have been automated, although the existence of repeated recordings of a single prompt implies that manual override is possible. The recorded data was then automatically uploaded to central storage.[3]

The version of the corpus used for this project was unfinished, containing only five of the planned ten cities, and due to time constraints only three were used: Ayr, representing the West Mid accent; Jedburgh, representing a Southern accent; and Inverness, representing a Highland accent. These were chosen for two reasons. First, the three represent very different accents and histories; Inverness is a historically Gaelic-speaking area, unlike the Scots influence of the other two, and the Borders area where Jedburgh is located has historically had a very distinct accent to the Central Belt influence of Ayr. [3, 6] Second, as can be seen in figure 6.1, these three cities cover between them the largest area of Scotland of any three cities in the survey. This allows more space for mi-

Figure 6.1.: VOYS recording locations.


Ayr is location 10, Jedburgh location 8, and Inverness location 1.

nor gradations in the interpolation, and makes for a potentially more appealing project, as a larger map covering more of the country can be used.

Participants, being drawn from middle schools, range in age from 11 to 18 with most falling between 13 and 18. This was a deliberate decision on the part of researchers, who wanted to counteract the lack of speech recordings of younger speakers along with the paucity of resources for Scottish accents. Recordings of a wide variety of adolescent speakers will allow evaluation of physical and sociolinguistic effects on their voices and manners of speech. [2] However, for the purposes of this project, the ages are primarily interesting as a note that voices may have different qualities to adult speech (for example, higher $F_0$), as well as a possible explanation for some of the challenges noted in section 7.3. It is also worthwhile to use children's voices because a good portion of visitors to a hypothetical museum exhibit will be young and may be pleased to hear someone more like themselves.

Figure 6.2.: Prompt statistics for VOYS corpus.

| Code | item description | Count | |
|---|---|---|---|
| | | total | per session |
| X | test item | 5 | 5 |
| A | command word | 63 | 10 |
| B | mobile keys | 2000 | 2 |
| C | telephone and credit card numbers, PIN codes | 1257 | 3 |
| D | dates | 802 | 3 |
| E | command phrase | 145 | 1 |
| I | digit | 10 | 2 |
| L | spelling | 1112 | 3 |
| M | amount of money | 1308 | 2 |
| N | natural numbers, company names, geographical locations | 778 | 1 |
| O | name of a person | 1812 | 8 |
| Q | yes/no question | 2 | 2 |
| S | phonetically rich sentence | 2528 | 9 |
| T | time expression | 517 | 2 |
| W | phonetically rich word | 1200 | 4 |
| 001-110 | dialect word | 110 | 22 |
| 111- 129 | dialect sentence | 19 | 19 |
| Y | narrative | 4 | 4 |
| total | | | 102 |

The speakers chosen for synthesis range more or less evenly across the 13-18 range within each voice, apart from the Inverness cohort which, due in part to a smaller number of participants to choose from, consists entirely of 17 and 18-year-olds.

## 6.2. Prompts

'An Overview of the VOYS project in Scotland' ([3]) gives the table in figure 6.2 to describe the prompts contained in the corpus:

The prompts contain a mixture of long and short selections; many of the shorter prompts consist of a single word in isolation, such as a command a smartphone user might need. Longer selections range from sentences to a full story. Some prompts are unique to the speaker, but many are repeated between speakers; this is good for

sociological or dialectological studies, since direct comparison can be done, but less useful for acoustic modelling where a greater variety of phones and contexts is desirable. Certain prompts, both of the individual words and the longer sentences, are used to indicate dialect by containing particular phonetic combinations that would be realized differently in different dialects. The selections do not appear to be phonetically balanced. [3]

Some prompts are questions; two are meant to elicit a yes or no response, and several provoke an arbitrarily long response. A series of pictures tell a story which the speaker is asked to describe. The speaker is also asked to describe their way to school and their favorite book, movie, or videogame. These recordings frequently sound quite different to the others, as the speaker is not reading words off a screen but has to speak spontaneously. There is more pausing and false starts for this reason. The character of the voice is also often affected, although how depends on the speaker; some seem to get quite nervous and speak very little with frequent pauses, while a few seem to get very excited at the chance to talk about themselves and give lengthy and exuberant responses. This has rather interesting implications from a synthesis perspective: most synthesis or ASR databases are entirely prompted speech, rather than spontaneous, and while this gives a slightly artificial basis for acoustic modelling it is at least consistent. Having both prompted and spontaneous recordings from the same speaker, which often sound very different but are treated the same, may complicate model building as the same predicted phone can be realized differently depending on the speaker's attitude.

The total amount of speech for each speaker is about six minutes. This is a very small amount, especially for speech synthesis. The voices built in this experiment therefore use several speakers each, although the number is still kept low to explore the effect of sparse data on synthesis quality.

## 6.3. Sociolinguistic observations

It is perhaps not surprising that there is a great variety in the level of dialectal characteristics exhibited by the youth recorded for this corpus. The dialects of Scots exist on a continuum of sorts with standard Englishes, although we will not take a position on the breadth or implications of that continuum. Even speakers of the broadest dialects can shift into Scottish Standard English, a Scots-flavored variety of standard English, under the right circumstances, although their pronunciation may hew closely to their preferred register. The shifts are often related to the social class of the speaker; for this reason, the VOYS corpus deliberately uses a balance of speakers from middle- and working-class families. All VOYS recordings which I listened to fell on the SSE side of the spectrum, which is to be expected in the more urban regions represented; it is however reasonable to assume that, since local pronunciation in SSE is affected by the variety of Scots spoken there, studying the local Scots dialects will help predict appropriate pronunciation for SSE voices. These characteristics are addressed more directly at the end of the next chapter. [6]

These observations raise an interesting question about how the speakers considered their context, if possibly subconsciously. It is known that many speakers from Scots regions change register depending on who they are speaking with, using more mainstream SSE with outgroup members and broader Scots with ingroup members. [6] While younger speakers in more urban areas may not have as nuanced a grasp of this usage as others, it is likely that they modify pronunciation somewhat depending on their context. This raises the question of how each speaker would have evaluated the context of speaking to a computer while alone in a room. Some may have thought about the strangers who would be listening to their recordings and instinctively used their best SSE; others may have considered that the strangers wanted to hear a broad accent and hammed it up appropriately. Some speakers, however, may have thought of it as talking to themselves and used whatever level of dialect they were most comfortable with.

# 7. Voice Building

## 7.1. Method

Voices were built speaker-independently for two reasons: most practically, since there was only around six minutes of data for each speaker, it was necessary to combine speakers in order to have enough data to successfully train a voice. More theoretically, the goal of the project was not to emulate a particular individual, as is often the case with speech synthesis, but to capture the character and common elements of a local accent. The more speakers who are used to build one voice, the more "average" relative to that cohort it becomes; therefore whatever unusual (compared to voices from outside the region) characteristics persist after that leveling can be confidently assumed to be particular to the group, not to the whims of an individual. These unusual characteristics are expected to be primarily spectral, representing differences in pronunciation, but may include duration or tendencies in pitch contour as well, reflecting the varying rates of speech or tonal qualities which may differ in certain accents. Five speakers were used for the Ayr voice, four for Jedburgh, and three for Inverness.

Some pre-processing had to be performed on the speakers even before listening to their recordings. Speakers were separated by gender, in part so that synthesized voices sound more natural and in part so that possible gender effects on accent strength can be observed. The Ayr voice is trained on female speakers, while the Inverness and Jedburgh speakers are trained on males. The VOYS corpus contains a description of its speakers, including their first language and the dialect they consider themselves to speak; second-language English speakers were excluded from this experiment, and speakers were sorted according to their dialect, not their location, although the two

were usually synonymous. Speakers of non-Scottish dialects were also excluded. Since the voices were trained speaker-independently, it was important that the desired features (i.e., the particular accent) be present in as much of the training data as possible so that they would persist to the final result. The goal of the project is not to present an average adolescent from these locations, but to present an adolescent example of the expected accent from these locations; while both aims are laudable, they may not always be the same. What is attainable will fall somewhere in the middle.

Voice building was done by executing a modified version of the scripts used for the official HTS demo[1]. This is a fairly simple recipe with embedded re-estimation done five rounds at a time, and models untied, re-estimated, and re-tied once after the initial clustering and retraining. Several versions of the voice were produced, including one which modeled variance as a one-component simple Gaussian and one which used a two-component mixture model. Generally the dialectal features were the same, but the samples given in 15 are taken from the two-mixture model.

## 7.2. Lexicon

A General American lexicon was used to create the phone labels used for training and later on for synthesis. The same lexicon must be used for all voices in order to enable interpolation; even if varying pronunciation dictionaries could be used for varying dialects, no specialized lexicon for each Scottish dialect appears to exist at the present time. This is expected to lead to mismatches between the predicted pronunciation and the observed realization, especially in vowels. However, there is mounting evidence that choice of lexicon may not have much effect on speech synthesis quality. A model trained from data should learn the correct surface realization of a particular phone in a particular setting, regardless of what the symbol attached to that phone during training is. [8] Some Scottish models, for example, should learn to weaken or drop /l/at the end of a word in certain contexts and to realize it as dark otherwise. Considering the very

[1]Located at http://hts.sp.nitech.ac.jp/?Release%20Archive.

high level of context used for phones in this system, enough information should be available to the learning algorithm to smooth out most discontinuities, and more data will improve this further.

That said, there was a deliberate choice in using an American lexicon over RP. Anecdotally, pairing a rhotic accent with a rhotic accent seems to work better than choosing a non-rhotic lexicon. This is sensible; a Scottish voice should be at least as able to realize/r/as a Midwestern American voice, and making sure these rhotic endings can be trained separately and specifically is wise. Intrusive /r/ would probably be modelled away even if predicted, especially since full context should distinguish this from other /r/ that should occur between vowels in a Scottish accent, but it is still preferable to avoid the situation. Rhoticity aside, there remains a possibility for false mergers in either case if vowels are merged in the American or RP case but not in the Scottish; however this has not been observed and if present is not sufficiently damaging to be a major concern.

## 7.3. Challenges

As the VOYS corpus is rather different to the purpose-built corpora usually used for speech synthesis, or even to most ASR corpora which are increasingly being adapted for speech synthesis purposes, there are several challenges specific to this corpus that must be addressed. None is impossible to overcome, and I have made progress on each, as I will describe below. I have also identified areas of possible improvement.

### 7.3.1. Pauses and duration

As recording appears to have begun automatically as soon as the prompt was displayed to the speaker, there are often long silences at the beginning of an utterance; even longer silences appear at the end of the utterance, since stopping the recording of a particular sound was automated as well. [2] These long silences can cause trouble for automatic alignment, especially since they are not usually pure silence; there are noises from the

rooms outside, the sound of the speaker shuffling in their chair or opening their mouths to prepare to speak, and so on. While background noise is good for ASR purposes, it can be quite problematic when using those recordings for synthesis. Training may begin by aligning labels to the noise at the beginning of the recording, rather than to the actual speech, and be unable to recover since these models are consistent, if undesirable. Building a voice on data which has not been cleaned at all results in a voice which produces only simulated microphone noise and static, with the occasional throat-clearing or a rare vowel.

This problem can be overcome by trimming the silence at the beginning and end of the recording; this step could likely be automated if enough care was taken distinguishing silence and static from speech. Alignment was also improved by cutting long, multi-sentence utterances into shorter single-sentence utterances, as the fewer alignments to be performed the less room for error. Large files can also cause errors as HTS will not load files above a certain size. This step could likely also be automated by splitting up utterances at long pauses, although there is then the risk of splitting mid-sentence if the speaker pauses at the wrong time.

Long beginning or trailing silences, even if they are noiseless, also have an undesirable effect on pause duration. This is even more egregious when utterances have pauses mid-sentence: when a speaker stops to read the next clause of a long sentence to themselves or has to stop to think about the structure of the spontaneous sentence they are forming. If these pauses occur between clauses or other places where Festival would predict a pause, the duration model of all inter-clause pauses is affected, since these will hopefully be clustered together (especially with a relatively small amount of data). Pauses in the middle of clauses may also cause trouble, as the pause may be mistaken for part of a sound and then reproduced during synthesis as an inappropriately long duration or sudden silence mid-clause. Removing boundary silence but not reducing inner silences leads to a voice that makes very long, awkward pauses in the middle of sentences.

The obvious solution to this problem, and the one used in this experiment, was to

trim out excessive silences; however, this does have some implications for speaker similarity. It is unclear how much trimming is appropriate; the nature of prompted speech from non-professional voice actors is such that there is no way of knowing even as a human listener what that individual's pause duration during natural speech would be. Even among the recorded speakers some paused frequently and at length, and some spoke very quickly. As this particular project's goal was not to accurately reproduce a particular individual but rather to capture characteristics of regional speech, silences were generally trimmed to a length that seemed adequately natural but was otherwise rather arbitrary; however, this may be an area of caution if a particular speaker must be reproduced accurately. This problem is one easily solved adequately, but difficult to solve optimally.

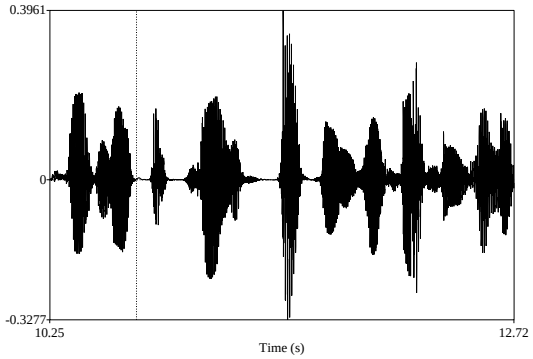
### 7.3.2. Sound quality and volume

Due to the decentralized nature of the recording process, sound was uploaded automatically regardless of whether or not the quality was acceptable. Thus utterances may contain pops or fuzz from a relatively untrained speaker getting too close to the microphone, background sounds from the noisy school environment of the recordings, or abrupt cutoffs if the speaker began the utterance too long after that file began recording. Some speakers had consistent problems; one speaker had to be discarded from his synthesis cohort because nearly every utterance contained microphone feedback noises. Unfortunately, since these sound problems distort the spectral envelope in a way that can't easily be counteracted, these problematic utterances had to be discarded in order to avoid recreating those noises in synthesis; with such little material to work with and so many potential contexts, a bad instance of a particular phone could easily be without any good examples to counter it. More data may increase robustness against this kind of problem.

Inconsistent speech amplitude is also a problem with this corpus; perhaps due to the lack of external oversight, some speakers tend to have quieter recordings than others. Thankfully the first prompt is used to help the speaker calibrate the microphone

so they do not usually speak loudly enough to cause clipping and ruin their recordings, but amplitude still varies widely between and occasionally even within speakers. One particular speaker had the tendency to start each sentence speaking loudly and decrease amplitude over the course of the utterance. Amplitude trouble seems to be a greater problem at some locations than others, indicating it may be related to local setup and could perhaps be overcome for future corpora with better training of local coordinators. As an experiment, I built a voice from the particularly inconsistent Jedburgh recordings without modifying amplitude beforehand; as might be expected, the resulting voice tended to change amplitude from syllable to syllable, as demonstrated in figures 7.1 and 7.2. This is perhaps due to certain context-dependent phones only occurring for particular speakers, giving amplitude for those phones no chance to average out. Modifying amplitude to be uniform before training should prevent this, although it may be difficult to equalize speakers with constantly changing amplitude.

Figure 7.1.: An utterance fragment from the Jedburgh voice.

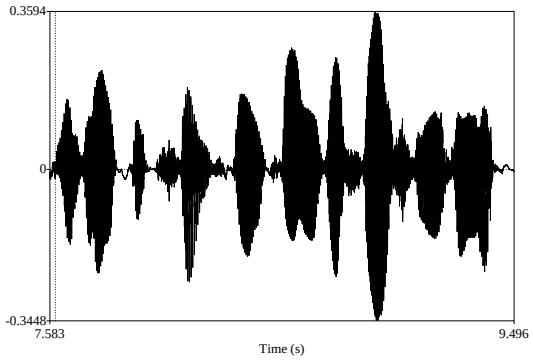



An utterance fragment ("had no pictures or conversations in it") from the synthesized Jedburgh speaker, showing inconsistent volume/amplitude and frequent quiet sections.

### 7.3.3. Pitch tracking errors

The $F_0$ statistics used to train models for the synthesized voice are derived automatically from the sound files. As can be expected from an automatic process applied to complex data, the pitch tracking algorithm is sometimes wrong. Incorrect pitch esti-

Figure 7.2.: An utterance fragment from a control voice.




The same utterance synthesized with a different, speaker-dependent voice without volume inconsistencies, showing more consistent amplitude.

mation can lead to inconsistent or unnatural pitch in the synthesized voice due to bad training. Unfortunately pitch ranges can vary widely, even within a fixed region and gender. This is made even more true by the fact that while all of the student speakers used in the VOYS corpus occupy the same age range, that age range also contains puberty, and especially for male speakers two individuals may have very different pitch ranges. The algorithm used to determine $F_0$ requires that a pitch range be set; the tighter this range can be made, assuming it is accurate, the better results from the algorithm can be. However, setting the range to cover all observed $F_0$ for the entire set of speakers does not work for our young males, and this leads to too many inaccurate estimations for speakers that tend to either end of that range.

The best solution found was to run the algorithm on each speaker separately, using the particular range for each. While this range can still be wide for individual speakers, results are much better for the baritone with a very narrow, low range who no longer runs the risk of accidentally being marked much higher-pitched than would be accurate. There are still some problems with pitch contours in the final voice, although some of that may simply be due to inconsistent pitch in the source material (see 7.4.2). More advanced pitch tracking algorithms would improve performance in this area.

### 7.3.4. Mistaken speech

The VOYS corpus is accompanied by a list of all of the prompts for each speaker, but not a transcript of the responses, a key distinction. Most prompts consist only of a word or a sentence which the speaker is to read, so it might be an acceptable risk to use those prompts as transcripts; however, several are questions to be answered, in a word or a few sentences, and the answers are not provided. I have transcribed these responses for the speakers I used to build my voices. Unfortunately it also proved necessary to check the provided transcription for the rest of the speech; since responses are recorded whether or not they correspond to the prompt, there are some cases of speech cutoffs, misreadings, or even the speaker responding with "I can't pronounce that word." While less than 3% of recordings are truncated, mistakes are more common. [3] Checking is also required for ambiguous prompts, such as the number series; 0 can be pronounced reasonably as "oh" or "zero" or part of a larger number, even by the same speaker, which cannot be predicted by Festival. A larger corpus might be more robust against mistranscription, but with only around 6 minutes of speech per speaker, a few bad utterances can throw off alignment for even good utterances. This can be solved by carefully listening to every utterance and adjusting the transcript appropriately. This requires a time commitment but is not difficult.

Mispronunciations also occur, as well as many words that are not in Festival's lexicon, most notably place names. This could lead to bad prediction of phones, which would then worsen models. Having made it my rule to not throw out an utterance unless absolutely necessary, as it was in the case of background noises, I attempted to overcome this problem by changing transcriptions where necessary to attempt to force Festival to make consistent predictions. This was usually done by spelling phonetically, or by using words that sound like the mispronunciation. 'Consistent' is a key word; as noted previously, the Generalized American lexicon and phoneset were being used, not a Scottish lexicon, so predicted pronunciations did not always match with actual pronunciations even when the speaker did not make a mistake. However, as demonstrated in 7.2, because the differences are reasonably consistent, results tend to be good any-

way. For this reason, when attempting to force a prediction, I chose spellings and words that would produce a pronunciation with the corresponding American vowels, rather than completely matching the observed Scottish pronunciation. For example, for the prompt "Newtownabbey", I adjusted the transcription to "new town abbey"; while the observed pronunciation naturally did not match the American pronunciation of "new town abbey", it did match the expected pronunciation for that speaker for "new town abbey". I also Americanized spellings to make sure words in the transcription occurred in the lexicon whenever possible.

There is a slightly more subtle issue at play here as well. Responses to question prompts were generated spontaneously and frequently suffer from grammatical errors and false starts. In conversation between humans, this is fine; spontaneous speech is often more flexible grammatically, and errors are less noticeable and more easily forgiven than in written text or prepared speech. However, Festival runs a part-of-speech tagger on input as part of its label-generation process, and the part-of-speech tagger can reasonably be assumed to be trained on grammatically stringent, written text. It may, therefore, give unusual results when applied to a transcription of ungrammatical spontaneous speech, and since this affects the predicted pronunciation and features used by the full-context models, there may be some effect on the resulting acoustic models. However, it seems impossible to avoid this, unless all ungrammatical speech is removed from the training set; this has the unpleasant side effect of removing most of the spontaneous speech, a good portion of the speech which is unique to each speaker and improves phonetic coverage. It seems reasonable to assume that this problem is fairly low-priority, as it would be difficult to overcome and likely has little effect.

### 7.3.5. Contextual issues

There is a mismatch between the type of speech recorded for this corpus and the type of speech we would wish to produce. This is a problem to some degree with many corpora of speech recordings; speech read from a prompt will always be somewhat more flat and less natural than true conversational or spontaneous speech would be,

and synthesized speech should sound somewhat spontaneous even if it is in actuality quite literally scripted. However, most adults are self-aware enough to be able to fake naturalness even when reading a prompt from a sheet. The young VOYS speakers frequently do not succeed at this; while some are able to sound natural, many either stumble over the words or even take up an exaggerated swagger, as if they wish to prove that they are too cool for the prompts. Question prompts can be even worse, as the speakers stumble over their words while thinking. None of these paradigms is one we really want to reproduce, yet the effects of these situations on both speech duration and spectral characteristics make it difficult to completely remove them. I have attempted to solve the problem of awkward, halting speech by removing pauses wherever appropriate, as discussed in 7.3.1. However, the flatness in the read prompts seems difficult to modify away. Performance could be improved by choosing only the most natural-sounding speakers and recordings, although this becomes less plausible for the low levels of data available from VOYS.

### 7.3.6. Lack of data

As is often said of any task requiring training of a model from examples, there is no data like more data. Voices in speech synthesis are frequently trained on recordings measured in terms of hours, not minutes. In stark contrast, the VOYS recordings only contain around 6 minutes of actual speech per speaker. With many locations there are enough speakers that an hour or two of data could be acquired, given enough time to perform all the required data cleanup, but for some groups, such as Inverness males, there are only fifteen or twenty minutes of usable data. Considering this hurdle, it is rather amazing that the speaker-independent voices came out as well as they have at all. Unfortunately, considering the scarcity of corpora like VOYS, it is unlikely that more data can be found to patch up this problem in a straightforward fashion.

It is also unclear whether having a good portion of the recordings for any given speaker be words in isolation, rather than sentences, has an adverse effect on synthesis above and beyond the effect of decreased data. Certainly words in isolation have dif-

ferent qualities and provide fewer instances of long strings of phones without a pause, which would be more important for the synthesis of full sentences. Some researchers have chosen to avoid single words when building voices, although that was not an option here. [15] With more cleaned data, however, it might be possible to build using only sentences and compare the results.

Faced with all this, speaker-adaptive methods are likely to prove more effective. High-quality average voice models can be trained on other, larger data sets; the amount of data which poses problems for speaker-independent voice training is plenty to adapt an existing model. Indeed, this is the method used to create the high-quality WSJCAM0 voices described in 9, also built from a small amount of data - around 80 sentences per speaker, similar to the amount available for the VOYS speakers. [15] This triumph over a lack of data may come with a price, however; it is unclear how the average voice model would interact with the desired goal of accent modelling. An out-of-dialect average voice would presumably cluster in a very different way, and there may be unexpected leveling of the very features that the project sets out to capture. It may be possible to overcome this by training the average voice on the VOYS corpus as a whole, but care must still be taken not to erase some of the particular distinctions within Scottish speech that we seek to preserve. Work on dialects of German indicates that while adaptive methods can be effective on speech of different dialects, what method performs best depends greatly on the distribution of data available. With a corpus such as VOYS, where each speaker has approximately the same amount of data, these results seem to indicate that best performance is likely to be achieved with a speaker-independent average voice, trained on data normalized for dialect differences, coupled with dialect-specific transformations. [10]

## 7.4. Results

### 7.4.1. Ayr

Ayr falls in the West Mid dialect region, dominated by Glasgow but containing several subtypes within which the Ayrshire dialect of Scots is reasonably distinct. It was this dialect that primarily influenced Robert Burns. As the only example of a Mid-Scots speaking area in the three voices created, the Ayr voice is expected to be the most stereotypically "Scottish". Due to its closeness to the Central Belt, it is expected that many speakers will be "monolingual" in Scottish Standard English rather than switching between dialects, although as previously noted it is unclear what the sociolinguistic implications of talking to a computer might be. Either way, word choice should be fairly standard and stereotypical Scots-influenced pronunciation is to be expected, with a small vowel system and influence and innovations flowing primarily from Glasgow. [6]

With five female speakers chosen from Ayr, it is possible to attempt to catalogue some particular dialectal features which were common among all of them and which we would therefore hope to find in the synthesized Ayr speaker. For example, all observed speakers tended to have BIT drift towards /e/, although not quite reaching it. The BITE dipthong was very pronounced, with both vowels somewhat elongated. Interestingly the duration of most fricatives, especially coronals, tended to be greatly lengthened compared to other areas. BOOT is realized as /y/, giving the voice a distinctly Scottish sound. Most speakers tended to speak rather slowly overall.

The Ayr voice was the first built, and the only voice built using female speakers. It was chosen because there were no Glasgow recordings and being able to model the west section of the Central Belt, if somewhat indirectly, seemed important. Originally, six speakers were chosen arbitrarily and their recordings were cleaned of bad utterances and long silences in accordance with the troubleshooting laid out in 7.3. Five of these speakers exhibited a reasonably uniform accent, although they ranged widely in age and some were certainly more confident than others. However, one of the others

displayed a markedly different and more traditionally Southern English accent, presumably for sociolinguistic reasons as her hometown was Ayr. Including this speaker noticeably dampened the appearance most of the above features in the synthesized voice, even though they were present in the vast majority of speakers. The conclusion drawn is that training data must be chosen carefully in order to properly cultivate a representative model.

After training, the synthesized voice exhibited many of the same characteristics as the training speakers. Fricatives definitely had extended durations. BIT was somewhat fronted, although not as dramatically as some of the speakers. The BITE prominence did not come through as much, however. /y/ definitely appears for BOOT, even though the General American lexicon would have predicted /u/ for these words; this demonstrates that even a small amount of training can overcome a mismatch between the lexicon and the target speaker's pronunciation.

### 7.4.2. Inverness

Inverness is distinct from the other two chosen locations in that it has never had a Scots-speaking majority; while few of its current inhabitants speak Gaelic, the historical influences on the English currently spoken there are far more likely to be Celtic than Scots. It is worth comparing to Scots-influenced voices for that reason alone. While there is generally a discontinuity between dialect in the Highlands, of which Inverness is the only representative in the VOYS corpus, and in the Scots-speaking Lowlands, there are regions where Highland English and Scottish Standard English come into conflict and even intermingle; Inverness is close to the border with areas speaking Northern Scots and has been observed to affect their dialect. [6] Interpolation between these types of voices is not unreasonable, then, as some mutual influence exists.

The recorded speakers for this area are a mixed bag, with very different levels of realization of their accent and none exhibiting a stereotypical "pure" Invernesian accent. There are, however, some consistent features. /r/ is slightly trilled for all speakers and there is some fronting of /θ/. BOOT is clearly /y/, in common with the other regions.

Interestingly, one speaker realizes BAIT as /i/.

I chose my second voice to stand in contrast to the Ayr voice in case I did not have time to build more. I wanted to use male speakers because my first voice had been female. I also wanted a region which was likely to have a very different accent, and the Gaelic influence of the Highlands seemed wise to investigate. Unfortunately, having made these choices I discovered that there were only five male speakers recorded from Inverness, the only Highlands city represented, and two of those were unusable due to quality issues. The remaining three speakers were definitely varied; one seemingly shy with a halting voice and more Scots-influenced accent and two with consistent but not notably Invernesian accents, one of whom spoke more slowly and one of whom spoke very quickly and often sarcastically. I chose to go ahead with building the voice to study the result of averaging three very different speakers. It would also be useful to compare a voice built with a mixture of accents to the relatively consistent Ayr voice.

Without a large number of consistent features in this dataset, it is not surprising that the voice, while intelligible, is difficult to place regionally. /y/ is definitely part of the vowelset, which does give it a Scottish feel, but it is difficult to tell from the level of vocoding quality whether /r/ is trilled. It does, however, change /ŋ/ to /n/ word-finally far more than the other voices, presumably due to the influence of the overly-casual speaker. This speaker gave far longer answers to the question prompts than his compatriots and therefore is slightly overrepresented in the final voice. $F_0$ curves are inconsistent and may cause the voice to sound like different people interrupting each other, due in part to the very different speakers who may have covered different phones. The overacting of the casual speaker may also affect the occasionally wild swings in intonation. If data needs were not so pressing, removing this speaker from the synthesis set would be wise, although he certainly has value for sociolinguistic studies.

### 7.4.3. Jedburgh

Jedburgh represents Southern Scots and the SSE varieties influenced by it. Southern Scots is distinguished by a tendency towards dipthongization, notably of MEET and OUT, and a slightly different vowelset to many Scots regions. The Jedburgh region does however contain similar vowels to the Edinburgh area and is somewhat influenced by it. Younger speakers are particularly prone to glottalization both finally and intervocalically. The boundary between Southern Scots dialects and Northumbrian dialects is historically well-defined. Traditionally Southern speakers have perceived their dialect as a purely local occurrence, rather than as part of a national language which they identify with the Central Belt, and nearly all older natives speak a form of Scots even if they can switch into SSE. However, immigration from the Central Belt threatens its distinctness, and it is unclear whether younger speakers would be likely to maintain this pride and distinctness. [6]

The young speakers from Jedburgh whose recordings I used tended to display very similar pronunciation features, apart from one speaker whose accent was notably more Central Belt-influenced. All had a tendency to glottalize /t/ both word-finally and intervocalically, which was not nearly so common in the other locations. Another feature shared among all speakers apart from the Central Belt representative but not seen in the other areas was a strong tendency to realize CAT as /ɑ/. BIT is shifted towards /ə/, in contrast to Ayr's shift to /e/, and the KITE dipthong is again elongated. /r/tends to be trilled or even pronounced rather like /l/; rhoticity is often lost word-finally, but less often mid-word. Two of the four speakers studied had a tendency to front /θ/. As with the others, however, BOOT is always /y/.

The Jedburgh recording site seemed to have some particular problems with speech volume, as some speakers were uniformly louder than others. Another speaker tended to start out loud and become quiet over the course of an utterance, for reasons which were unclear; this speaker was excluded from the data, as there did not seem to be any simple form of normalization which would be able to fix it. I chose to proceed with training a voice on this data without normalizing the volume in order to see what effect

this inconsistency would have.

As may have been expected, inconsistent volume in the training data does leave the final synthesized voice with some inconsistency of its own, although it is less prominent than one might fear. The observed $F_0$ curves have a somewhat "sing-song" quality, resulting in a distinctly melodic intonation unlike the other two voices. Word-final rhoticity is disappearing, and both the CAT-ɑ and BIT-ə realizations are present. Unfortunately with the buzziness of the vocoder, it is impossible to tell whether the /r/ is trilled.

### 7.4.4. Intelligibility

It is difficult, although not impossible, to do a more formal evaluation than the above of these voices' accuracy in reproducing the appropriate accents; this is discussed in more detail in chapter 14. However, a simple evaluation of their intelligibility can be and was performed. Five semantically unpredictable sentences were synthesized for each of the three voices, as well as a control General American voice built on the author's ARCTIC recordings. Thirteen raters were then recruited to listen to the sentences once and attempt to transcribe them. Word accuracy rates were computed from these transcriptions and scored against each other. This functions as a simple intelligibility test; while a larger sample would be more statistically significant, some conclusions can still be drawn from these results.

As seen in figure 7.3, while there is a statistically significant difference in intelligibility, the full range of scores does overlap with the control for some of the VOYS voices. It is to be expected that the control voice would be more intelligible: although the training method is equally simple, the control is trained using four to eight times more data than the VOYS voices, and the data was purposely recorded for speech synthesis. What is promising is that the VOYS voices can still approach the intelligibility of the control, especially when only native speaker data is used (as in figure 7.4). While this dataset will be more of a challenge than some to use, it is demonstrably possible to build a voice which is somewhat intelligible, and there already exists a large body of literature

Figure 7.3.: Relative word accuracy rate for the four tested voices

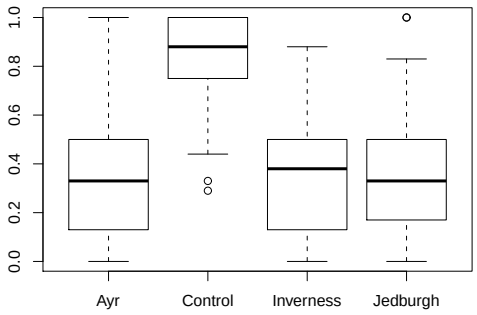
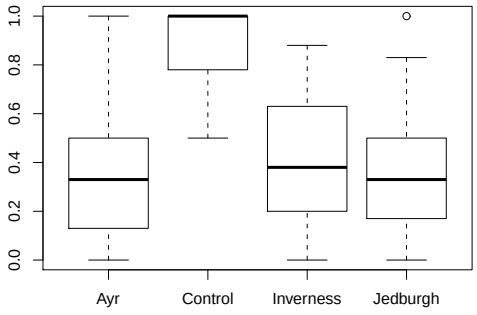



Figure 7.4.: Relative word accuracy rate for the four tested voices, only native English speakers considered.




on improving the intelligibility of voices.

Noting the types of errors listeners made is also interesting as a corollary to my own phonetic observations. None of the listeners recruited were Scottish and some had never been to Scotland, so errors consistent across users point to a marked difference between the Scottish speech and the English they are more used to (in most cases General American English).

- Many listeners mistook initial “how” for either “hi” or “hey” in all three VOYS voices, but not the control. This points towards a vowel shift in OUT.

- Several listeners consistently heard initial “how” as “rarely”, “highly”, or even “hairy”, but only for the Ayr voice. That particular vowel is realized differently for Ayr.

- Listeners often misheard “bail” as “bagel” in an utterance from the Jedburgh voice; this confusion indicates that this word sounded like two syllables to them This region is known to exhibit exaggeration of both vowels in the BITE dipthong almost to the point of separating into syllables.

- As noted above, the Jedburgh speakers are losing rhoticity and the synthesized voice is expected to emulate this. Many listeners mistook “houses” for “horses” for the Jedburgh voice, but did not make similar rhoticity errors for the other voices.

- Many of the errors are around confusion of stops - “bat” for “bad”, “steel” for “school”. This may be due to the buzziness associated with the vocoder used on these voices; a higher-quality vocoder may greatly improve accuracy simply by reducing these errors.

# Part IV.

# The Speech Interpolation Map

# 8. Dialect Continuums

The map portion of this proof of concept will demonstrate that a simple map-based interface can intuitively demonstrate dialectal variation over a geographical area. As this is only a proof of concept, several simplifying assumptions regarding dialectal variation have been made, with flexibility built in so that more complex ideas of influence can be used to achieve more accurate results. Note that for this report we are only looking at pronunciation; word choice or morphology must be handled in part at a higher level than interpolation, and while an interesting problem is beyond our current scope.

## 8.1. Cultural center assumption

First is the assumption that distinct accents or dialects can be identified as primarily occurring in or characteristic of a particular location, presumably a larger city or cultural center. These cities are, ideally, the ones we collect data and build voices for. Predicted accents for other areas come from a function of representative speech patterns from nearby cultural centers. We will refer to this function as "influence" and use it to determine which prebuilt voices to interpolate and what weights to use with that interpolation.

For the sake of simplicity and easy demonstration of effects, the influence function discussed here will be a simple distance function; the two influential cities are the two closest cities to the click and they influence the chosen location in proportion to their distance. That is, a location equidistant between cities A and B will sound like a perfect average between the two, whereas a location 9 times closer to A as to B will sound 90% like A and 10% like B. This function can occasionally give bizarre results when

moving outside of the convex hull of the representative cities - moving out beyond a city into a realm without any new influences will paradoxically increase the influence of the second closest city. For this reason, this simplification works best when many influential cities occur on map borders.

When adapting this to the problem of interpolating the VOYS voices, this means assuming that the cities chosen by the VOYS project are the best, or at least sufficiently accurate, representations of their dialect areas. The project description does not directly address this issue, but does state that the cities chosen are major urban areas in their dialect regions, and as the cities are reasonably well-distributed across dialect regions, it does not seem too large a leap to assume them transmitters of influence. Indeed, since part of the purpose of VOYS is sociolinguistic and dialectological data collection, presumably this assumption informs the decisions made by the group. A few of the proposed cities do occupy the same dialect region - for example, Dundee, Perth, and Stirling are all North Mid - and it might be prudent to consider whether all three cities should be used as a better indication of the shape of the region or whether just one central city should be used, so as not to disproportionately favor that dialect. VOYS does have one notable weakness regarding coverage of urban centers: there is no Glaswegian data, nor any plans for such, since many other recordings of Glasgow exist. Yet Glasgow has an undeniable influence on the Scottish accent landscape. It may be necessary to attempt to use these other recordings to build a compatible Glaswegian voice, possibly using adaptation on an average voice from the VOYS recordings to minimize other differences. [3]

If we are to use interpolation as a tool for estimating local accent, it is necessary to use some form of this assumption, otherwise our entire premise is flawed. This particular implementation is, of course, a simplification of dialect geography, and some suggestions for weakening this assumption and improving influence modelling may be found in 13.3. However, using this simple distance function is not without merit. It is common in dialectology to find a chain of dialects which are mutually intelligible with their neighbors, yet each quite distinct from those dialects far away from it. One of the

most famous examples of these is the West Romance dialect continuum, which stretches from Portugal through Belgium including forms normally referred to as French, Italian, Catalan, Spanish, and Portuguese. While the standard varieties of each of these are not normally considered to be remarkably mutually intelligible, speakers of dialects from bordering regions have little trouble understanding each other. [1] There is thus evidence that dialects often occur as a fairly smooth shift over distance. In addition, literature on Scottish dialects refers to local dialects as occurring in either the core or periphery of the influence zone of particular cities, with transitional areas between. [6] This supports the assumption that influence radiates from a central area.

## 8.2. Phoneset assumption

In order to interpolate between two voices, it is necessary that both use the same phoneset. While different decision trees are not and cannot be an obstacle to building an interpolated voice, at least for a particular utterance, this process requires matching phones. Looking for statistics for a phone which only exists in one of the sets of statistics can only lead to disaster. For this reason, it is necessary to assume that any two voices that will be interpolated in normal use of the map use the same phoneset. [17]

This is a less restrictive requirement than it may seem. Should premade voices be used in a map, it is likely that any particular set will all use the same phoneset simply to make it easier for the builders. Any one group of voices will therefore likely allow interpolation between members of that group. The most likely source of conflict with this requirement, therefore, is a situation where multiple voice sets are being combined to create a larger map. It is reasonable to assume that multiple voice sets won't be used for the same region; differences in those voices stemming from different recording conditions or purposes for their training corpora would be rather distracting, and if a given set does not have enough voices to cover that area by itself, why is it being used? Different sets should then cover different regions, and probably rather distinct ones. Initial plans for this project, for example, included using the VOYS voices for Scotland

and the WSJCAM0 voices for England. WSJCAM0 uses an RP phoneset, eminently reasonably, and would not be compatible with the American phoneset used for VOYS; however, we would not want it to be! There is a rigid boundary between Scottish and English accents, and it would not be accurate to interpolate between the two. In this situation it would be appropriate to use two different influence functions: one, using Scottish cities, when a Scottish location is chosen and one, using English cities, when an English location is chosen.

However, if we are building new voices for a speechmap, we can easily overcome this by simply building all voices on the same phoneset. This has already been accomplished for the existing VOYS voices and there is no reason to change for future voices. Indeed, as 7.2 implies, it may well be just as reasonable to train all voices on the same phoneset regardless of where they are from; necessary boundaries can be enforced by changing the influence function instead.

## 8.3. Uniformity assumption

The third assumption is that influence from cultural centers affects all qualities of a voice equally. In other words, if a location is equidistant between major cities A and B, every utterance that results for that location will be uniformly 50% A and 50% B, rather than, say, having the vowels be more strongly influenced by A and a particular subset of consonants be much like B's. This is not an entirely vacuous assumption to make: different phones may have very different influence patterns, and some are subject to gradual change while others shift abruptly as a marker of dialect regions. This has been observed between Austrian and Viennese German. [10]

Requiring that all phones follow the same gradual interpolation is not necessary, however, should there be more precise knowledge of the shifts to be modelled. In the interpolation step, HMMs are made for the sentence to be synthesized under both voices and then combined according to rules. It is possible to adjust the rules for this combination so that some types of target phone are taken wholesale from one model or

the other, depending on the intended location's position on one side of a dialect border or the other. Indeed, it is even possible to elide phones in this fashion. [10] These more complex rules were not implemented for the WSJCAM0 speechmap used as demonstration for this project, as they would have obscured some of the basic functionality, so this assumption remains in force for now.

While implementing such a feature is not theoretically complex on the synthesis side, it depends on very precise dialectological knowledge. The ability to interpolate some phones and abruptly shift others is somewhat irrelevant if there is little to no knowledge of where the appropriate boundaries lie and what shifts should be made. Unfortunately there is relatively little research on regional variation in Scottish pronunciation, and not enough to implement such complex distinctions at this time. [9] Indeed, this is part of why dialectal voices must be trained from data. Under these circumstances, it is appropriate to fall back on this assumption to make it possible to model the dialect geography of Scotland at all, with hopes that more research can be used to improve the model later.

# 9. WSJCAM0 voices

## 9.1. Voice selection

The voices used for the speechmap were pre-existing, provided by Junichi Yamagishi of CSTR. These voices are high-quality, with a Mean Opinion Score of nearly 3 for naturalness and only around 20% WER. [15] I chose to use these voices for the speechmap demonstration instead of my VOYS-trained voices for several reasons. First, a large number of voices already existed in this set, meaning that the intricacy of the map I could produce was not limited by how many voices I could build in a short period of time. The eventual final product will presumably have a sizable set of voices, even though those voices do not currently exist, so it would be more accurate and interesting to demonstrate with more voices. Furthermore, the quality of these voices is beyond what I could produce with the resources of an MSc student; these voices are built as adaptations of a well-trained average voice which I did not have access to, and the experience and computing power of CSTR and their associates outstrips mine. Higher-quality voices better show off the potential of a speechmap by not distracting the user with buzziness or unnaturalness and better reflect the projected end product.

The WSJCAM0 voice set was chosen specifically because not only was it high-quality and sizable, it contained geographic information about the speakers, which was not common for other sets of voices. Furthermore, the majority of the speakers were British, although mainly English; this seemed closer to the spirit of the original goal, since a large body of high-quality Scottish voices were not available. A subset of the voices were chosen so that effects were clearer; these speakers are all female, from northern England and southern Scotland. The screenshots in chapter 16 show their locations.

## 9.2. Training corpus description

The WSJCAM0 corpus was created as a counterpart to the American English WSJ0 recordings. Prompts are drawn from the WSJ text corpus, specifically from the sub-corpus chosen for training purposes by the WSJ0 coordinators. It was designed for training speech recognizers, making it a good candidate for HMM synthesis. Each of 140 speakers recorded between 80 and 90 sentences, with little repetition of sentences between speakers. Unlike VOYS, prompts are normalized to decrease pronunciation variation; for example, numbers in texts are spelled out, rather than left to the individual to determine.[5]

All participants were native English speakers recruited from the Cambridge University population. As a consequence most are between 18 and 28 years of age, although there are a few who are older; none are younger. Because recording occurred on campus, all speakers were living or working in Cambridge at the time. Speakers were given the chance to review the sentences before reading them and were specifically instructed to use their normal voices and accents while recording. [5]

## 9.3. Differences between WSJCAM0 and VOYS

There are several key differences between the WSJCAM0 corpus and the VOYS corpus. The most obvious is age: WSJCAM0, like most corpora, consists of recordings of adults - most between the age of 18 and 28, but still significantly older than the majority of the 13-18-year-olds that make up the VOYS. This has one very significant effect: older speakers are more likely to understand why they are doing the recording and have some investment in it, while some of the VOYS speakers did not take the task seriously.

WSJCAM0 also contains none of the spontaneous speech that VOYS does, being entirely readings of Wall Street Journal text. This may simplify training, since all the speech is of the same type. The fact that WSJCAM0 was designed for training speech recognizers favors the builder of HMM voices as well. [5] VOYS, on the other hand, seems aimed as much at sociolinguists, for whom pauses and mumbles may be as

meaningful as formants, as at speech technology researchers. It is certainly still useful for speech technology purposes, as demonstrated in Part III, but a pure ASR project like WSJCAM0 does have an advantage.

Finally, if most anecdotally, there is far less variation in accent within the WSJCAM0 corpus than in the VOYS corpus. The majority of the Cambridge speakers appear to have acquired the same accent, regardless of how they may have spoken at home. In contrast, VOYS speakers, even of the same gender and attending the same school, may display different levels of their local accent in the recordings, and features common in one location are often rare at another. For this reason, WSJCAM0 as a whole would not be much good for sociolinguistic or dialectal research, and interpolating between these voices shows more of the spectrum of individual variation than accent variation. This is certainly a drawback of this set of voices, considering the goal of the project is to display accents. However, I believe that the benefits outweigh this disadvantage. Each speaker still has a unique voice with characteristics that can be blended during interpolation. The unique accent characteristics identified in the VOYS voices can be inferred to behave the same way. Furthermore, the subset of voices chosen for the demonstration contains one woman with a rather more Scottish accent than the others, and a gradient of accent can be seen between her and the more English voices, although perhaps not as a reflection of reality[1].

[1] A precomputed demonstration of map-based interpolation between this speaker and an English speaker can be found at the resources in chapter 15

# 10. Interface

## 10.1. Implementation

The decision was made to implement the interface as a webpage, although the map is designed to run offline for possible installation in a museum exhibit. This should enhance portability, as it only requires that the computer have a web browser installed. Not all browsers are the same, of course, especially in terms of how they handle audio; the map is guaranteed to work in the freely-available and common Firefox 3, however, so should no installed browser a compatible one can easily be acquired. Implementing as a webpage has several other advantages. It is reasonably easy to produce a clean and appealing visual design with HTML, and the necessary data cleanup and script execution can be done with Javascript and CGI as well as a program built from scratch would. Best of all, most people are familiar with webpages and the visual shorthand thereof, and using some of the same motifs, even something as simple as forms, can help make the program less intimidating to a wider audience. Some technical ability may be necessary to set up the program, although there is an attempt to avoid that; it should not be required to interact with it.

To start up the speechmap, a small shell script is included which will launch an Apache server to allow the web browser to use CGI. This also launches a Festival server to handle requests and, finally, a Firefox window which will display the page. This only needs to be done at startup of the display; in a museum the speechmap could continue to run all day after this script is run, barring any accidents.

## 10.2. User's perspective

Screenshots showing the progression of the user through the following steps can be found in chapter 16.

The visual presented to the user is simple; there is a map of the region, a section of northern England and southern Scotland in this case, and space for a description of the project. The map is simply a static image file taken from http://www.openstreetmap.org/, an open-source map repository. Markers laid over the map indicate the locations of the canonical voices. When the user clicks on the map, a Javascript click handler processes its location relative to the image and determines the two closest voices to use and the ratio of their distances. The two voices chosen are displayed to the user by a change in the color of their displayed markers. As noted earlier, using the simple distance function may give odd results when the click moves outside the convex hull of the voice locations. As long as this function is in use, it is best to choose a map that does not have much space beyond the edges of where voices occur.

Once the voices and ratio are determined, a textbox appears below the map to request text from the user to be uttered. A popup is not used for this task to allow the user a moment to think about their choice and fix any mistakes they have made. This textbox is part of a for, which sends the text, along with the names of the two selected voices and the ratio between their influences, to a CGI script, further described in chapter 11.

The user does not see any of the calculations behind the scene; from their perspective, a blank popup appears and, after a few seconds, displays audio controls. The WAV file produced behind the scenes autoplays, but can be replayed at the user's desire. A small popup is employed so that the user can clearly see that the map still exists, ready to be clicked again. However this does mean that the installer should be careful that popup blockers are configured to allow this.

## 10.3. Extendability

Adding new voices to the map or changing to an entirely different map and voice set is relatively trivial. Once a new map image is chosen, the relative coordinates associated with each voice must be determined, but this can be accomplished quickly. An array holds the location relative to the image, as well as the name of the voice in Festival, and can be updated as necessary. Markers are automatically placed at startup based on the array. The distance formula can also be modified as appropriate.

# 11. Inner workings

The CGI script which organizes the hidden creation of the requested sound file is written in Python, primarily due to personal preference. It checks the text for length and, so long as it is under 75 characters long (to avoid very long waiting times), continues on to perform the rest of the procedure. Calls are performed as Python subprocesses, with results written to a log rather than displayed.

## 11.1. Production of labels

Before either voice can be loaded it is necessary to convert the input text into full-context labels for synthesis. The speechmap accomplishes this by making a subprocess call to a Festival client, stored locally and included in the distribution so that the correct voices can be included. The Festival server should already be running after startup. Before creating an utterance structure, Festival must be set to use a voice of the same lexicon and phoneset that the voices to be interpolated use; for the WSJCAM0 voices that would be the RP Unisyn lexicon. An utterance structure is produced quite quickly and the predicted speech characteristics drawn from that. All this is accomplished by using the Python script to write a short Festival script to a temporary file, then making the Festival call on that script. The program dumpfeats is then called to convert these labels into HTS-compatible labels.

## 11.2. Interpolation and synthesis

The final subprocess call is to another shell script, which calls the appropriate version of HTS on the two voices which were chosen as the closest, the labels developed in the previous step, and the appropriate ratio. As most versions of HTS support interpolation, this should be adaptable to voices built for any version, although some changes may require investigation of the proper syntax as it often changes greatly between versions. HTS then carries out the interpolation procedure described in 5 and saves a WAV file to disk which is then loaded into the browser.

The interpolation step tends to be the most time-consuming part of the process by far, and unfortunately requires the user to wait while voices are loaded and the appropriate calculations are done. This is a good argument against using more than two voices to interpolate, unless the computer being used is very fast; interpolation requires estimating statistics for the utterance for each voice before averaging, and doubling the number of voices would double the amount of time required for this section of interpolation.

Synthesis of the WSJCAM0 voices is done with a higher-quality vocoder and analysis system called STRAIGHT. This vocoder uses aperiodicity information, representing fluctuations in pitch which is otherwise assumed to be perfectly periodic. While this complicates analysis of a speech signal, it does allow for better synthesis as less information about the signal is lost. [7] STRAIGHT also makes better use of $F_0$ information in analysis to create smooth spectral envelopes. [14] Very high quality and natural speech can therefore be produced using STRAIGHT. For intellectual property reasons I was not able to use this higher-quality vocoder for the voices I built; being able to use it would definitely improve naturalness.

## 11.3. Efficiency

11.1 shows timings for the three phases of production after the user chooses a location and submits text; unfortunately it was not straightforwardly possible to break down the steps of the interpolation process. Clearly interpolation is the biggest contributor to

Table 11.1.: Time taken for steps of synthesis.

| Sentence synthesized | Festival call | | Conversion to HTS labels | | Interpolation | |
|---|---|---|---|---|---|---|
| | real | user+sys | real | user+sys | real | user+sys |
| "Hello, world." | 0m0.56s | 0m0.004s | 0m0.13s | 0m0.11s | 0m16.47s | 0m16.44s |
| "A longer test sentence." | 0m0.76s | 0m0.003s | 0m0.14s | 0m0.12s | 0m16.50s | 0m16.40s |

inefficiency in this process. Both real and CPU times are included, as some of the delay in real timing may be relevant as well; Firefox will still be running in the background for a museum installation of the speechmap and may take up some of the same resources. While the absolute times are system-dependent, the relevant times are important.

Interestingly, increasing the length of a sentence - the second is approximately twice the length of the first - does not seem to significantly affect performance except perhaps in conversion of labels. There is not enough evidence to draw a solid conclusion, however. Interpolation is clearly the most time-intensive step, but the entire process is still not too long on a reasonably modern computer.

It must be noted that for this particular implementation, due to the structure of the University of Edinburgh's Informatics department computers, all files needed to be stored on a network rather than locally. All required files were on the same groupspace, which may have helped, but this still greatly increases the amount of time required as access to multiple files is required for nearly every task in this project. Interpolation especially is slow, as there are several files which must be loaded for each voice. It is presumed that an installed version would be run with the files stored on the same computer. The above numbers can thus be considered to be an upper bound on the actual time, assuming a reasonably modern computer.

# Part V.

# Conclusions and Future Work

# 12. Conclusions

It has been demonstrated that given appropriate preprocessing, the VOYS recordings of youth in Scotland can be used to train HMM voices for synthesis which capture many features of the regional variation of speech in Scotland. While the quality of the voices which have been produced is somewhat lower than desirable, that can be overcome using a better recipe for voice training, judicious use of speaker adaptation, better vocoders, and more of the data; low intelligibility in HMM voices is a known problem with known solutions. Unfortunately, appropriate preprocessing is quite time-consuming, which prevented many voices from being built at this time. The corpus has many interesting features, some of which, such as the youthful target and the often characterful speakers, may aid the project and some, such as the inconsistent recording quality and low levels of data for each speaker, may complicate it. Overall, though, this corpus offers unique information about the speech of Scotland and is well worth the time that must be put into it.

It has also been demonstrated that interpolation and a clickable map interface can be combined to produce an intuitive exploration of speech continuums. A proof of concept speechmap has been produced and will be left to the speech synthesis experts at the University for demonstration and further work. While the WSJCAM0 voices are not ideal for this project, they suffice to show how interpolation would work in such an application and how variation can be displayed. The basic framework for future maps of many areas has been laid down and can be expanded with reasonably little effort to handle any desired set of voices and model of speech center influence. Each project, whether Scottish or elsewhere, will have its own challenges and may require particular

features to be implemented for an ideal result, but the framework presented here can serve as a starting point and as an easy way to produce an acceptable first version.

# 13. Points of future improvement

While the voices and map application produced for this project are a suitable first step, they are only intended to be a first step in preparation for a larger project or projects. There are therefore many possible directions for improvement or future work, some of which will be explored in this section.

## 13.1. Improve voices

Far better recipes exist for building HMM voices from data than the rather simplistic and relatively quick one that I used, described in 7.1. Any subset of training could be repeated more times, and the voices that were built were designed for an old version of the hts_engine API rather than more modern methods which use better vocoders and more complex statistics such as aperiodicity modelling. For the purpose of this proof of concept, it was enough to discuss the strengths and weaknesses of the VOYS corpus and demonstrate that intelligible voices can be trained on it that capture some of the particular regional pronunciation variation found in its speakers. Since these adequate voices were built in a rather basic fashion, it follows that applying more expertise and resources could potentially lead to voices that are nearly as good as those trained on purpose-built corpora.

## 13.2. Improve speechmap interface

The most obvious improvement to accomplish the final goal of the project is to implement a speechmap which truly covers Scotland, or indeed all of Great Britain should

suitable English voices be found. I have striven to make the speechmap easily adaptable to new voice sets, so implementing the map itself should not be a major challenge for future workers. The greatest challenge in time and resources will be building an appropriate number of high-quality voices, as discussed above.

As the goal of the project is to provide a map deployable to museums, it had to be built to function offline. It would be quite embarrassing for a museum to lose an exhibit because their internet connection was down! However, it might be useful to have a version of this project which can run online as well, making it available and advertisable to a wider audience. This could perhaps be a smaller demonstration version only, to maintain the appeal of the full product to museums and cultural centers.

Furthermore, if the map is planned to be deployed in museums it would be wise to consider filtering the input, as by the nature of the project anything the user chooses to synthesize will be perceivable by others in the vicinity. Children will not always appreciate the opportunity to explore the variety of expression in their country as much as they appreciate the opportunity to make a robotic voice say something rude to a roomful of strangers.

Currently a large factor of what makes the speechmap rather slow is that it must constantly retrieve network resources, and we expect a purely local version to run more quickly given approximately the same level of computing power. However, there are still improvements to be made. Although out of the scope of this project, it may be possible to modify the version of hts_engine so that the appropriate voices are preloaded in some fashion, as it can be known before the interpolation step which voices will be used. The remainder of the interpolation process is probably about as efficient as it is likely to be, barring new breakthroughs.

## 13.3. Improve theoretical underpinnings

Weighting one city or regional representative as being more influential could be used to simulate places where the shift in dialect does not occur at a constant rate, but re-

mains continuous. For example, a larger city or capital logically may have a stronger influence, so that even a location geographically halfway between that capital and a smaller source of training data would be more influenced by the capital. This can be accomplished by a slight modification of the influence function to consider weights as well as distances; as the influence function is clearly defined in the Javascript for the interface page, modifying it should also be simple.

It may also be necessary to enforce some stricter boundaries. As previously noted, there appears to be a very firm dialect boundary between Scotland and England without the gradual change exhibited on either side of the border. This could easily be modeled by using different influence functions, which each only consider cities of the appropriate country, when clicking on either side of the border. A line is also known to exist between the formerly Gaelic-speaking Highlands and the Scots-speaking areas south and east of them. A similar approach, or at least one that strongly favored one side of the line or another, could also be used there. [6] As more research is done on these areas, possibly even using the VOYS recordings, a more complex or piecewise function can be introduced accordingly.

More complex interpolation methods could be used if there is research to support them. Interpolation can be done between more than two voices, should there be evidence that more distant cities or dialect regions have an influence; in the Scottish case it may be appropriate to always include a little bit of influence from a Central Belt city, for example. It is also possible to use separate pronunciation dictionaries for two different regions as long as they have the same phoneset so that overall, target phones can be matched. This may help if there are inconsistent pronunciation differences between regions, or perhaps even to do some simple tailoring to local dialect words so long as these words are used wholesale, not interpolated. [10] There is a reasonably good, if slightly out of date, literature on Scots dialect words; a project to model these traditional differences, while somewhat tangential to the goal of this project, might be quite interesting.

# 14. Formal evaluation

In 7.4 I gave an informal evaluation of the voices built from the VOYS corpus, demonstrating with phonetic features that the synthesized voices had learned the appropriate pronunciations and doing a preliminary survey of their intelligibility. A more formal evaluation would be the next step once a final set of voices has been produced. Native English speakers - preferably Scottish, but a sufficient number of Scots may be difficult to find - could be trained to distinguish between the accents in question by listening to recordings of native speakers of the different dialects, preferably the recordings used to build the voices so that age issues do not complicate matters. The synthesized voices could be evaluated for accuracy by playing sentences to these trained listeners and having them guess the location data they were trained on. Listeners could also be used to evaluate the similarity of predictions to observed speech by training on samples from locations not represented by a precomputed voice, perhaps from cities held back from the original map (such as the multiple North Mid cities). They could then be asked to choose the best match from a number of interpolated utterances, one of which is the prediction for that area. These are very important to conclusively demonstrate the accuracy of both the trained voices and the interpolation process, but require a great investment of time and resources which is not appropriate for a proof of concept like this.

It would also be wise to perform more usability tests for the map interface itself once a more final version has been produced. While I and those I have asked to try it feel that the interface is straightforward and easy to use, more opinions should be sought outside of the speech technology community. It is important that a new user be able to

tell at a glance what the purpose of the application is, how they are expected to interact with it, and what they can expect it to do. This is difficult to predict without a survey, as every user brings a different perspective.

# Part VI.

# Appendices

# 15. Samples

Samples of both successful and unsuccessful synthesis, along with a canned version of the demo, can be found at http://design-by-darwinism.com/speechmap/ , graciously hosted by a friend of mine.

# 16. Interface

Figure 16.1.: The speechmap before the user's click.

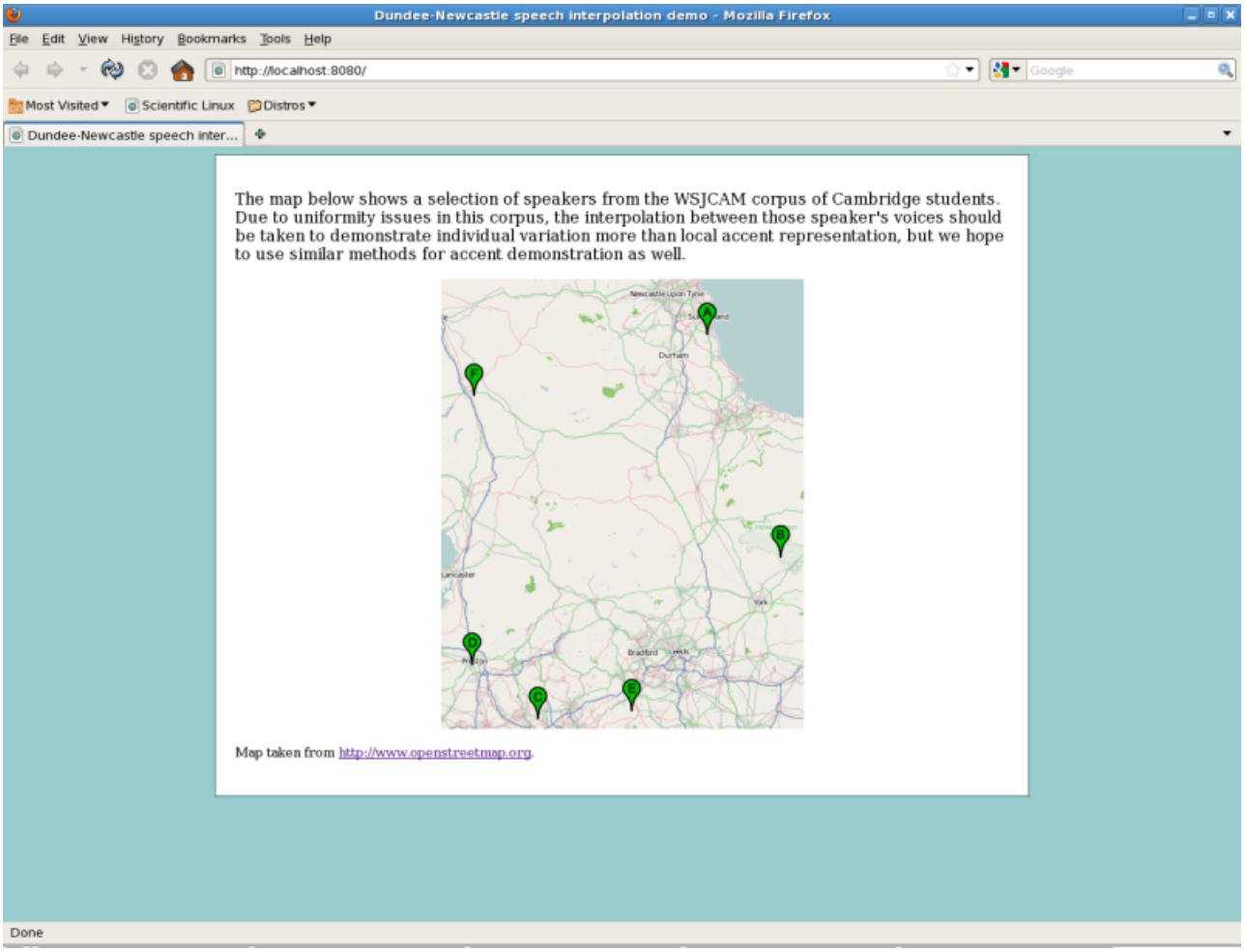


Figure 16.2.: The speechmap after the user clicks on a location.

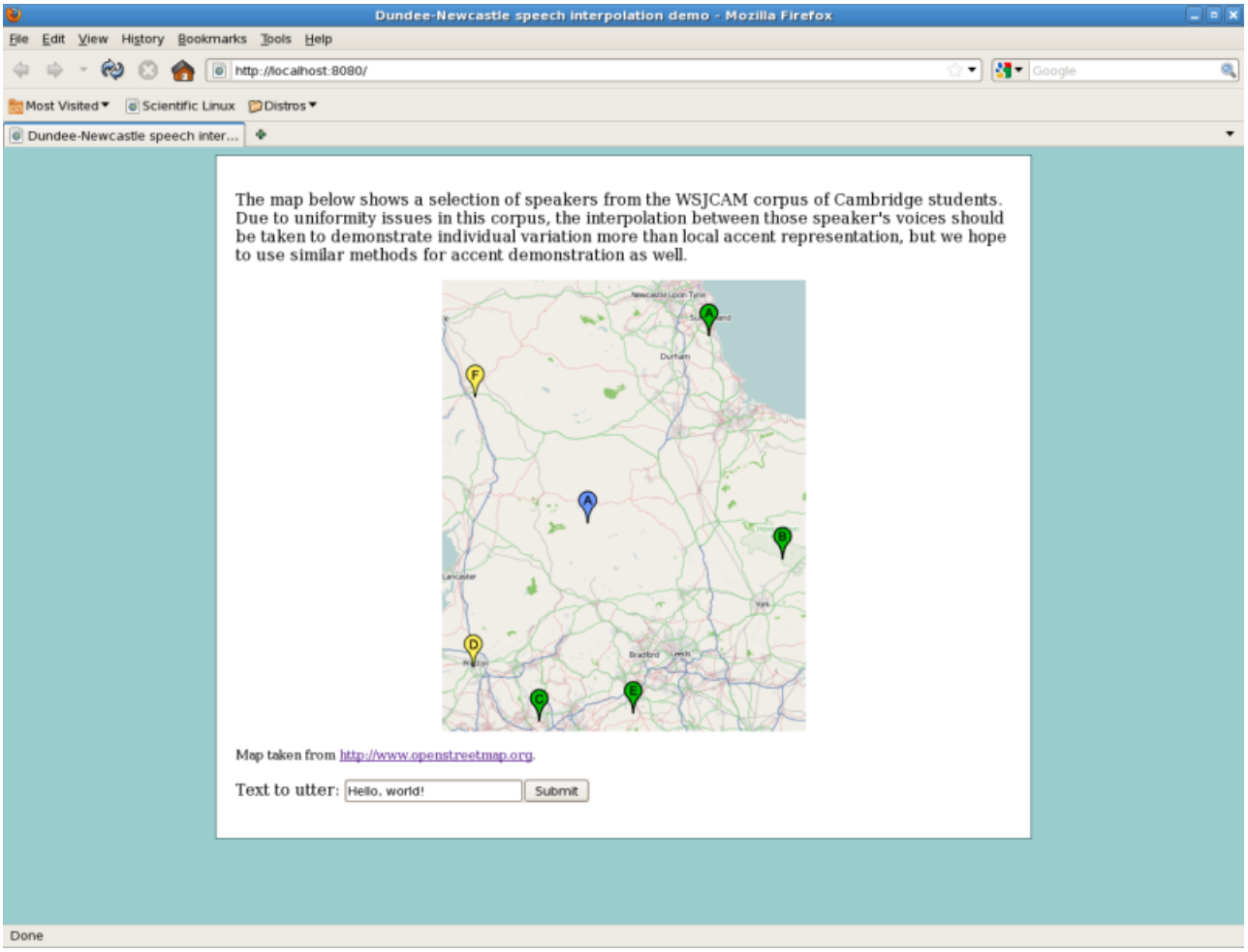


Figure 16.3.: The speechmap after the Submit button is clicked.

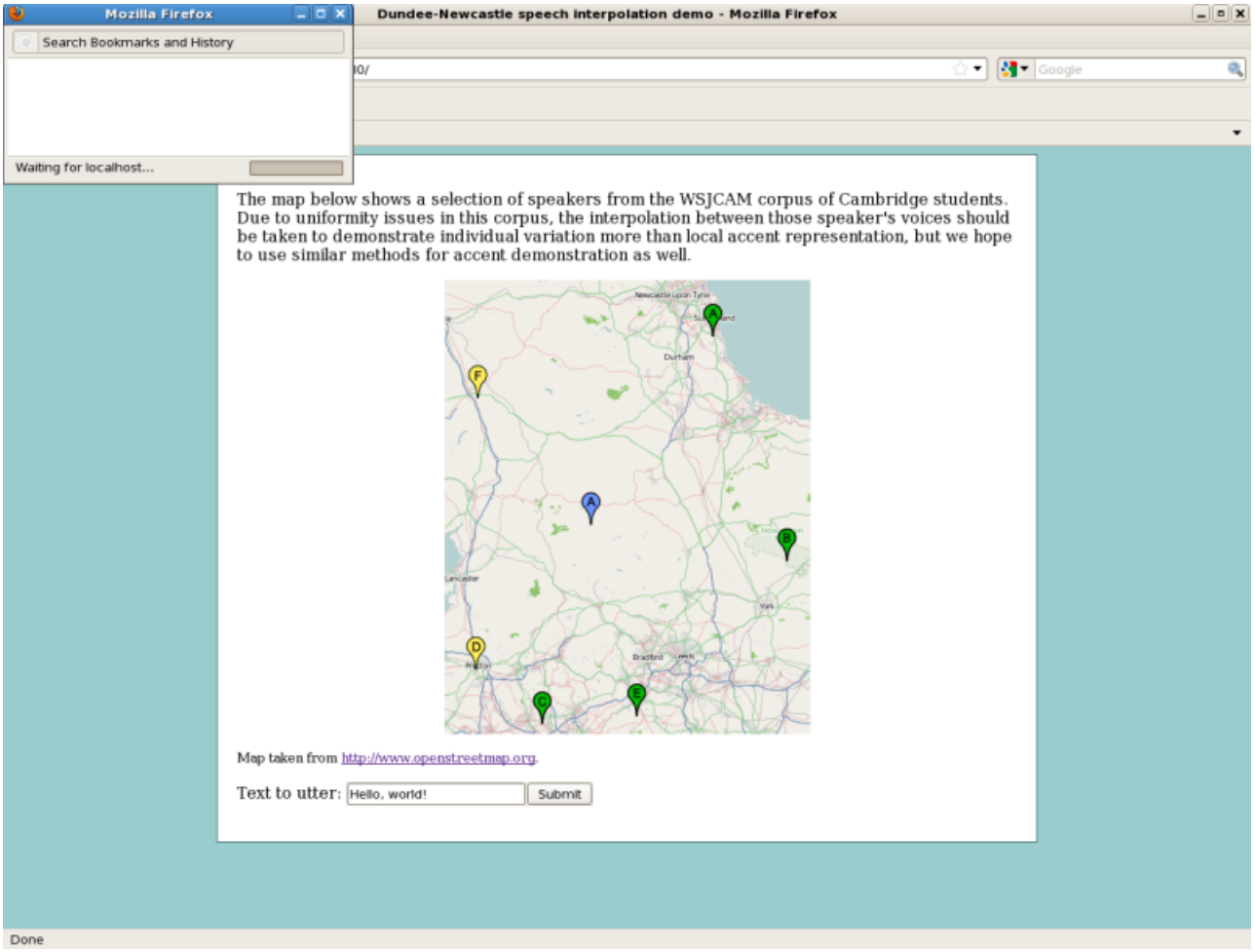
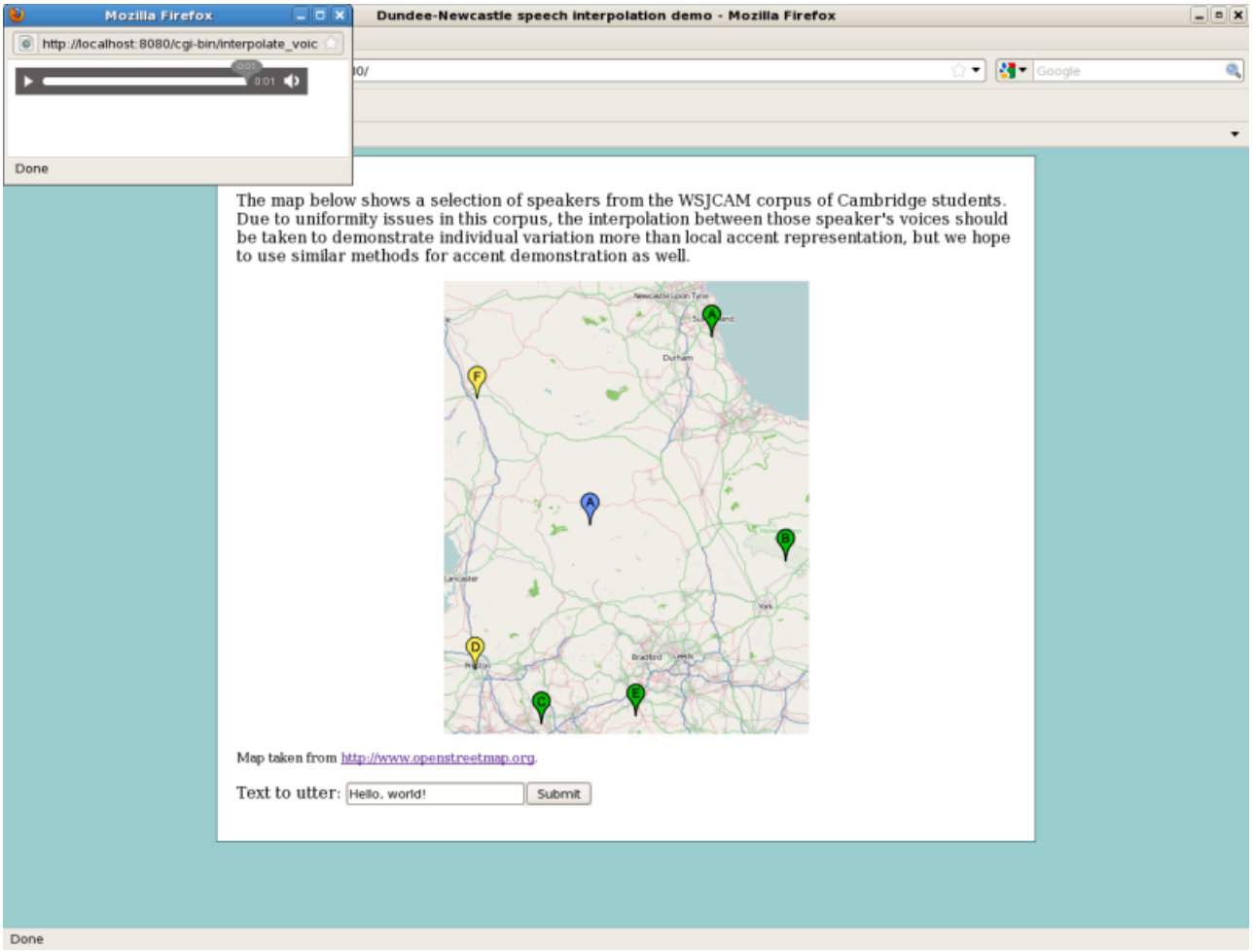


Figure 16.4.: The speechmap after the synthesized speech clip is loaded and played.



# Bibliography


[1] Chambers, J.K., & Trudgill, P(1980). Dialectology. Cambridge : Cambridge University Press..

[2] Dickie, C, Draxler, C, Schaeffler, F, & Jänsch, K (2010). 'Updating the Scottish accent map: preliminary formant data from the VOYS corpus.' British Association of Academic Phoneticians colloquium (BAAP), London, March 2010

[3] Dickie, C., Schaeffler, F., Draxler, Chr., Jänsch, K. (2009) 'Speech Recordings via the Internet: An Overview of the VOYS project in Scotland', Proc. of Interspeech, Brighton, UK. Retrieved from http://sites.google.com/site/cathlinguistics/ .

[4] Draxler, Chr., Jänsch, K. (2008) 'WikiSpeech - A Content Managment System for Speech Databases', Proc. of Interspeech, pp. 1646-1649, Brisbane, Australia

[5] Fransen, J., Pye, D., Robinson, T., Woodland, P., Young, S. 'WSJCAM0 Corpus and Recording Description.' Cambridge University Engineering Department (CUED) Speech Group. http://www.ldc.upenn.edu/Catalog/readme_files/wsjcam0/wsjcam0.html

[6] Johnston, P. (1997). 'Regional Variation.' In Jones, C. (ed), /The Edinburgh history of the Scots language/ (pp. 433-513). Edinburgh: Edinburgh University Press

[7] Kawahara, H., Estill, J., Fujimura, O. (2001). 'Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT', 2nd MAVEBA.

[8] Miller, C. (1998). 'Pronunciation modelling in Speech Synthesis.' Ph.D. thesis, University of Pennsylvania.

[9] Miller, J. 'Syntax and Discourse in Modern Scots.' The Edinburgh companion to Scots. Edited by Corbett, J., McCClure, J., and Stuart-Smith, J. Edinburgh University Press, c2003.

[10] Pucher, M., Schabus, D., Yamagishi, J., Neubarth, F. (2010). 'Modelling and Interpolation of Austrian German and Viennese Dialect in HMM-based Speech Synthesis,' Speech Communication, Volume 52, Issue 2, Pages 164-179.

[11] Tokuda, K., Zen, H., Black, A.W. (2002) 'An HMM-based speech synthesis system applied to English', Proc. of 2002 IEEE SSW, Sept.

[12] Tokuda, K., Zen, H. (2009) 'Fundamentals and recent advances in HMM-based speech synthesis', Interspeech 2009, Tutorial, Brighton, U.K., September 6, 2009.

[13] Tokuda, K., Zen, H., Yamagishi, J., Black, A.W., Masuko, T., Sako, S., Toda, T., Nose, T., Oura, K. The HMM-based Speech Synthesis System (HTS) version 2.1 Readme. http://hts.sp.nitech.ac.jp/archives/2.1/README_HTS.txt

[14] Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K. King, S., Renals, S. (2009) 'A Robust Speaker-Adaptive HMM-based Text-to-Speech Synthesis,' IEEE Audio, Speech, & Language Processing, vol.17, no.6, pp.1208-1230.

[15] Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Hu, R., Guan, Y., Oura, K., Tokuda, K., Karhila, R., Kurimo, M. (2010). 'Thousands of Voices for HMM-based Speech Synthesis – Analysis and Application of TTS Systems Built on Various ASR Corpora', IEEE Audio, Speech, & Language Processing.

[16] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T. (1999) 'Simultaneous modelling of Spectrum, Pitch and Duration in HMMBased Speech Synthesis,' Proc. of EUROSPEECH, vol.5, pp.2347– 2350.

[17] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., (2000). 'Speaker interpolation for HMM-based speech synthesis system.' Acoust. Sci. Technol. 21 (4), 199–206.