Characterization of Speakers for Improved Automatic Speech Recognition

Mike Lincoln

A thesis submitted for the Degree of
Doctor of Philosophy
in the
School of Information Systems,
University of East Anglia, Norwich

December 13, 1999

Abstract

Automatic speech recognition technology is becoming increasingly widespread in many applications. For dictation tasks, where a single talker is to use the system for long periods of time, the high recognition accuracies obtained are in part due to the user performing a lengthy enrolment procedure to 'tune' the parameters of the recogniser to their particular voice characteristics and speaking style. Interactive speech systems, where the speaker is using the system for only a short period of time (for example to obtain information) do not have the luxury of long enrolments and have to adapt rapidly to new speakers and speaking styles.

This thesis discusses the variations between speakers and speaking styles which result in decreased recognition performance when there is a mismatch between the talker and the systems models. An unsupervised method to rapidly identify and normalise differences in vocal tract length is presented and shown to give improvements in recognition accuracy for little computational overhead.

Two unsupervised methods of identifying speakers with similar speaking styles are also presented. The first, a data-driven technique, is shown to accurately classify British and American accented speech, and is also used to improve recognition accuracy by clustering groups of similar talkers. The second uses the phonotactic information available within pronunciation dictionaries to model British and American accented speech. This model is then used to rapidly and accurately classify speakers.

Contents

List of Figures

List of Tables

Publications

These are the publications which have been produced by the author during his Ph.D candidacy.

1. M. Lincoln, S. Cox and S. Ringland. A fast method of speaker normalisation using formant estimation. In Proceedings of Eurospeech'97, Volume 4, PP 2095-2098

2. M.Lincoln, S.Cox and S. Ringland. A comparison of two unsupervised approaches to accent identification. In Proceedings of The 5th International Conference on Spoken Language Processing, 1998, PP109-112

Acknowledgements

Thanks to my supervisor Dr. Stephen Cox for the guidance, support and advice given over the course of the work, and to Mr S. Ringland for the interesting discussions and ideas.

Thanks to Mr. Shaun McCullagh and the rest of the SYS support team for keeping everything up and running, even with Bill around.

Finally, thanks to all those I've known who've passed through S2.27, S2.28, and The Wolfson Lab, for making it a cool place to work.

This work was funded by British Telecom Research Laboritories.

Chapter 1

Introduction

1.1 ASR and Speaker Variation

The most natural means of communication for humans is that of spoken language, augmented as necessary by other means such as gesture, written language, diagrams, maps etc. It has long been a goal to create a machine which can automatically recognise and understand spoken language input. Recognition systems, such as those shown in science-fiction films (e.g. HAL in 2001, and R2D2 in Star Wars) would remove the need for the contrived, machine driven, input techniques such as keyboards and mice that we use at present and which many people find difficult to use. Such devices would be replaced with an entirely natural means of communicating ideas and requests to the system upon which the machine could react — that of spoken language. There are several advantages to the use of speech input: The integration of computer systems in situations where the use of a keyboard is impossible (by doctors during surgery for instance) would become a practical possibility. Disabled users who have difficulty using standard input devices would be able to use machines to the same effect as their able bodied colleagues and the use of computer technology where no keyboard exists (eg at the end of a phone line) would also become a possibility.

Much of the reason why the problem of automatic speech recognition has *not* been solved to date is that speech is widely variable. We have few problems understanding the speech of, for example, an American child, despite the fact that his/her speech will be acoustically very different from our own. The same is not true of automatic recognition systems which generally try to model the acoustic patterns of speech and as such are highly sensitive to variations in speaking styles. That is not to say that current systems are of no use — in recent years, companies such as Dragon, Microsoft, IBM etc, have released automatic dictation systems with usable recognition levels implemented on desktop personal computers. However, such systems gain a large proportion of their accuracy by having long enrolment procedures in which the system is trained to recognise the particular speaking style of each user. This time consuming procedure is necessary if anything approaching reasonable accuracy is to be achieved, and users are generally willing to perform the enrolment for the additional benefits of having speech recognition available as an alternative to keyboard control.

Around the same time as the introduction of these automatic dictation systems, BT's 'Callminder', an 'in the network' answering machine with voice control and other voice activated telephony services became available. Such systems do not have the luxury of large amounts of enrolment data from each speaker (imagine the popularity of a system which required reading a twenty minute passage every time you wished to check your voice messages!) and as such rely on explicitly modelling *all* the acoustic variations in the speech sounds. Because of the range of variability between speakers and despite the fact that these usually have more restricted vocabularies and grammar their recognition rate is usually much lower than that of systems with an enrolment phase, sometimes unacceptably so.

The purpose of this investigation is to identify methods of rapidly classifying speakers based on their particular speech characteristics. Given this knowledge of the speaking style, we then aim to show how this information may be used to improve the accuracy of recognition systems. We will concentrate on classification based on parameters related to vocal tract length and accent since normalising for these factors has been shown to give improvements in recognition accuracy. It is to the class of systems general termed ’interactive speech systems’, rather than 'dictation systems', that the procedures described in this thesis are directed, and this imposes several restrictions on the methods. They must have very low computational overhead since they may be being used in situations where many speakers may be using the system at a given time (for instance in the case of ’Callminder’) and therefore increasing the computation for a single speaker results in an unacceptably large increase in overall system processing requirements. They must also work in an unsupervised manner since labelled enrolment data will not generally be available and it would be unacceptable to expect the user to provide it. Since the systems are only likely to be used for a short time by each user, it is also a requirement that the methods achieve an improvement in accuracy after only a very small amount of adaptation data has been received from the speaker.

To summarise, we are seeking to identify and exploit characteristics of a speaker's speaking style to improve recognition accuracy given the constraints that the methods must have low computational overhead, work in an unsupervised manner and require very small amounts of adaptation data.

1.2 Structure of the Thesis

In Chapter 2 the human speech production system is described and used to classify the various sounds used in speech. A simplified model of the vocal apparatus, the source-filter model, is also presented. Chapter 3 describes the methods used to extract information relevant to speech recognition from the acoustic signal, and describes the most popular method used to perform automatic speech recognition - Hidden Markov Modelling. Chapter 4 describes the ways in which speakers may vary and the effect of such variation on both the speech signal and recognition task. It also describes the methods currently used to overcome such variation. Chapter 5 develops a new method for rapid, unsupervised speaker normalisation based on formant modeling. Chapter 6 describes a new technique for unsupervised clustering of talkers with similar speaking styles and uses the technique both for accent identification and to improve performance an automatic speech recogniser. Chapter 7 describes a second technique to perform accent identification. Results are summarised in chapter 8 and conclusions and further work identified.

Chapter 2

Speech Production system and Modelling

In this chapter, the mechanism for speech production is introduced. This leads to a classification of speech sounds in terms of their production process. A model is then introduced which can be used to approximate the characteristics of a speaker's vocal apparatus.

2.1 Speech Production System

Figure 2.1 shows a cross section of the vocal apparatus, consisting essentially of the lungs, trachea, larynx and the oral and nasal tracts. The manner in which these are used in producing speech sounds will be briefly described — more detailed descriptions are given in most introductory phonetics books such as [39, 55].

The lungs act as the energy source for speech generation. They are filled with air by the expansion of the rib cage and the lowering of the diaphragm. As the rib cage contracts, air is forced out of the lungs along the trachea. The velocity at which air exits the lungs is used to control the volume of the produced speech.

The first section of the vocal apparatus which the air encounters is the larynx which controls the voicing of the subsequent sound. The larynx consists of two folds of skin called the vocal cords, with the space between them known as the glottis. The vocal cords may be in one of three states — closed, open, or vibrating. In the closed state, air

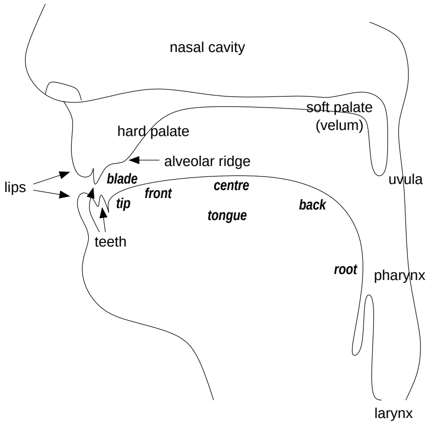


Figure 2.1: Primary features of the human vocal apparatus. After [49].

builds up to high pressure behind the vocal cords and can then be released by the vocal cords parting. This is known as a glottal stop and may be heard in many accents such as Cockney, Glaswegian and Birmingham in words such as ‘water’ and ‘butter’. With the vocal cords open, air passes unimpeded through the glottis. Such sounds are known as *voiceless*, for example /t/ as in ten. If the vocal cords are held close together, but not tightly closed, the air builds up behind them until it reaches sufficient pressure to force them apart. The pressure then drops, the cords close again and the pressure begins to build once more — the vocal cords effectively act like a mechanical oscillator. Air is let through the glottis in short bursts, though the bursts are in such rapid succession (from 70 to 1000 per second) that they are perceived as a constant vibration. Sounds such as /u/ as in ‘boon’ which are produced in this way are referred to as *voiced*.

Having passed through the glottis the air is then directed into either just the mouth, or the nose and mouth simultaneously, depending upon the position of the velum. Sounds made with the velum open, that is with air passing into both cavities, are referred to as *nasal*, while those produced with the velum closed, and air passing into the mouth only, are referred to as *oral*. The sound is then modulated by the various articulators within the oral cavity and the resulting sounds may be classified based on their manner and place of articulation. The *contoid* sounds are produced by forming different closures within the vocal apparatus which interfere with the air stream. The *manner* of articulation describes the degree of closure produced and may be one of the following :

- *Stop:* In which the air flow is completely blocked by the articulators, e.g. the first and last sounds in ‘top’.
- *Fricative:* In which the articulators are brought close enough together to cause a turbulent airflow e.g. ‘zoo’.
- *Approximant:* In which the articulators are close, but not enough to cause a fricative, e.g. ‘we’.
- *Nasal* In which air flow is blocked in the oral cavity, but the velum is open, allowing air to pass through the nasal cavity, e.g. ‘my’.
- *Affricate* In which a stop is immediately followed by a fricative, e.g. the first sound in ‘cheap’.
- *Lateral* In which the air stream is obstructed at a point along the centre of the oral tract, with incomplete closure at the sides of the tongue, e.g. the first sound in ‘lie’.

The *place* of articulation describes which of the articulators cause the interference and may be one of the following :

- *Bilabial:* The sound is produced by the action of both the lips working together, e.g. ‘pop’.
- *Labiodental:* The lower lip and the upper teeth are brought together, e.g. ‘fudge’.
- *Dental:* The tongue and the upper teeth are used to form a constriction, e.g. ‘thigh’.
- *Alveolar:* Between the tongue tip or blade and the alveolar ridge, e.g. ‘die’.
- *Retroflex:* Between the tip of the tongue and the hard palate. This is not used in English.
- *Palato-Alveolar:* Between the blade of the tounge and the back of the alveolar ridge. e.g. ‘shy’.
- *Palatal:* Between the front of the tongue and the hard palate e.g. ‘huge’.

- *velar:* Between the back of the tongue and the soft palate, e.g. ‘gang’.

The *vocoids* are produced if there is no contact between the articulators. The sounds are then classified based upon the position of the tongue with respect to the cardinal vowel space, and are described as front, centre or back and low, middle or high and are also differentiated by the degree of lip rounding present. The cardinal vowels are shown in Figure 2.2 and Figure 2.3 shows the vowels used in RP English.

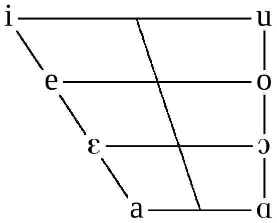


Figure 2.2: The cardinal vowels. Front is to the left

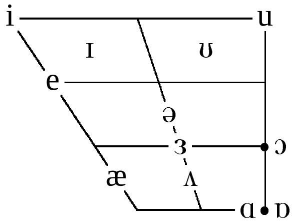


Figure 2.3: Vowels used in received pronunciation (RP) English. Front is to the left.

If the articulators remain in a steady state during the vocoid, they are referred to as *monophthongs*. Those in which they move during articulation are known as *diphthongs*.

Each of the sounds which may be made by combinations of phonation, manner and place of articulation are referred to as phonemes. Since in general there are more sounds than letters in most alphabets, the International Phonetic Alphabet (IPA) is used to describe each of the sounds [66]. This alphabet is common across all languages, and defines a set of sounds and associated symbols which may be used to unambiguously transcribe any utterance. A list of the phonemes used in the RP production of English is given in Table 3.1.

2.2 Source-Filter Model of the Vocal Tract

The vocal apparatus, when producing vocoid sounds may be modelled as a simple tube, open at one end (the lips) and with a sound source at the other (the larynx) as shown in Figure 2.4.

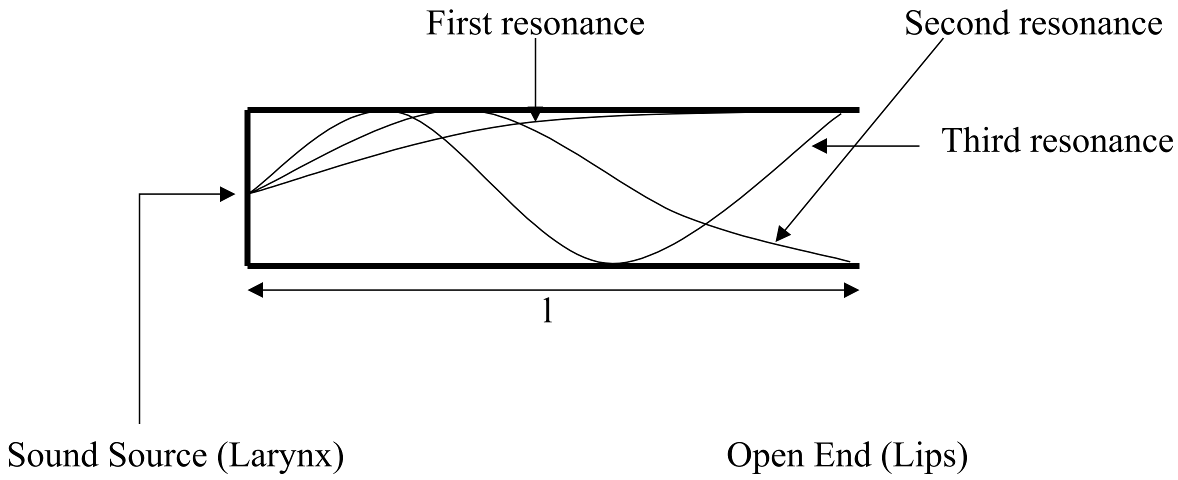


Figure 2.4: Uniform tube model of vocal tract.

Such a system has resonances at odd harmonic frequencies shown by the curves in the tube in Figure 2.4. They denote the standing wave vibrations which air in the tube will form and are determined by the way in which the cross sectional area varies along the length of the tube. In the case of the central, mid vowel sound, /e/, the resonances occur at approximately $f_0, 3f_0, 5f_0$ etc where $f_0 = c/4l$, c being the speed of sound in air and l the length of the tube. Taking l to be 17cm (the average length for a male talker) and c to be 340 m/s gives $f_0 = 500$Hz and third and fifth harmonics at 1000Hz and 1500Hz respectively. This model is a large over simplification however, since it does not take into account the separate resonances of the oral and nasal tracts, the effects of the tongue, or constrictions along the vocal tract.

The resonances of the vocal tract are referred to as the *formants* and the position of the first three formants is highly correlated to the perceived quality of the vowel sound being produced [51, 62]. Although the vocal tract has an infinite number of such resonances, because the glottal excitation source rolls off at -12dB/octave, it is only necessary to consider the first 3 or 4. Variations between speakers' vocal tracts will manifest themselves in changes in the values of the formant frequencies. Speakers with longer vocal tracts will have lower frequency formants, while those with shorter vocal tracts will in general have higher formant frequencies. This model is a large oversimplification however, since it does not take into account the separate resonances of the oral and nasal tracts, the effects of the tongue, or constrictions along the vocal tract.

The ‘loseless tube’ model of the vocal apparatus may be extended to the model of the speech production system shown in Figure 2.5.

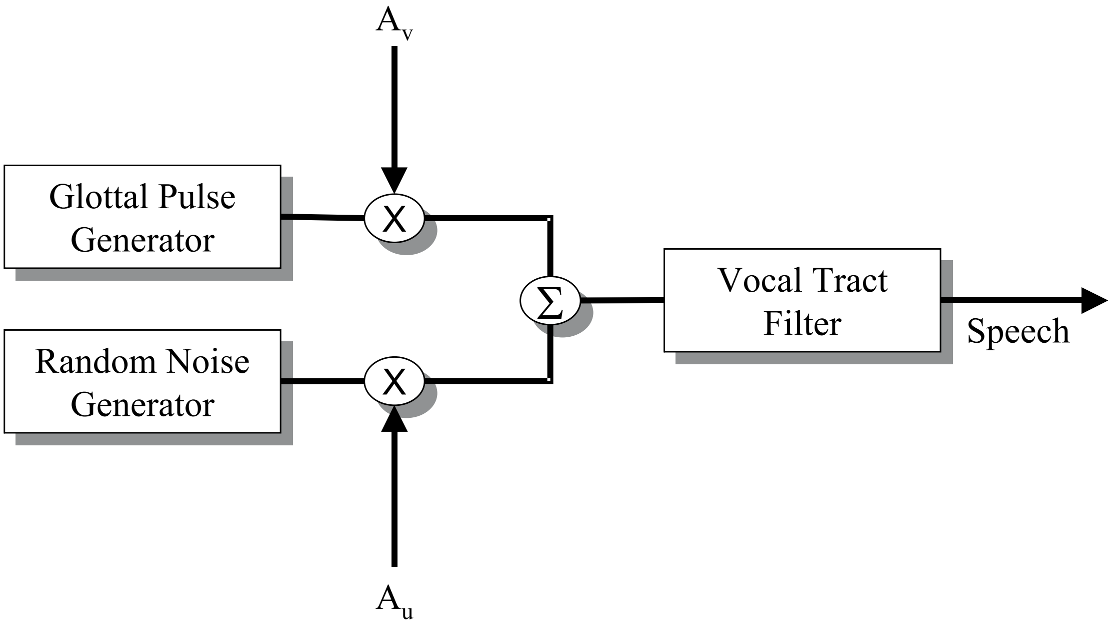


Figure 2.5: Source-filter model of vocal tract.

This ‘source-filter’ model approximates human speech production by modeling it as a signal source modified by a variable transfer function filter. The source represents the glottal excitation (either periodic pulses in the case of voiced sounds or noise in the case of unvoiced) and the time varying filter is equivalent to the vocal tract. The gain controls A_u and A_v control the amplitude of the voiced and unvoiced sources. By specifying the gain values and the transfer function, each of the sounds which the vocal apparatus may produce can be approximated.

This is also an over simplification however, since fricative sounds are not filtered by all the resonances of the vocal tract (since the sound is produced at a constriction somewhere along it). It also assumes that the source and filter are independent and linearly separable, which is not true since the vocal chord vibrations are affected by pressure within the vocal tract. These points are usually ignored however and the source-filter is assumed an adequate representation of the speech production process.

The model is also of particular importance in computational speech processing, since the signal processing technique of linear predictive analysis is capable of producing estimates of the source and filter. The filter, being a representation of the vocal tract, may then be used in the classification of both speech and speakers.

Chapter 3

Speech Processing and Recognition Techniques

3.1 Introduction

There are essentially two methods for performing automatic speech recognition. The first (referred to in [67] as the 'acoustic-phonetic' approach) exploits a set of rules derived from the fields of phonetics and linguistics to interpret the speech signal. The alternative is the use of statistical pattern classification [14, 58, 67, 69] in which mathematical pattern matching techniques are used to perform the recognition. The acoustic phonetic approach exploits a large body of information which relates characteristics of speech sounds such as voicing, nasality, fortis/lenis, to higher level linguistic units such as phonemes. These relationships are, however, highly complex and as yet it is not well understood how to deal with the large variations between individual realisation of sounds which are identified as the same phonetic unit. It is also unclear how such rules could be incorporate into a computational framework such as would be required to perform ASR.

The statistical pattern classification approach ignores most linguistic knowledge of the speech signal (or rather, it is usually too difficult to incorporate such knowledge into the mathematical framework it uses). Instead it gains its 'knowledge' by example — training mathematical models on large amounts of training data. This approach has many advantages — firstly, a set of mathematically rigorous techniques exist which guarantee to optimize the performance of the model for a given set of training data. Secondly, since no knowledge of the signal is assumed, the techniques are equally applicable to a wide range of speech units — word, syllable or phoneme models may all be generated with little modification to the basic system. Thirdly, this model can easily be extended to incorporate a model of language, and the choice of vocabulary, syntax and task for which the recogniser is developed have no effect on the implementation. However, the performance of such systems are highly sensitive to the quantity of training data used for creating the acoustic models. Often many hours of speech is required. The models are also highly sensitive to the noise and environmental conditions present when the training speech was recorded since this is modelled along with the required speech signal. When used under good conditions the recognition accuracy of such systems is very high, and the ease with which they may be implemented in a computational framework means they are the preferred method of speech recognition used in all commercially available recognisers.

Figure 3.1 shows a block diagram of a typical statistical speech recogniser.

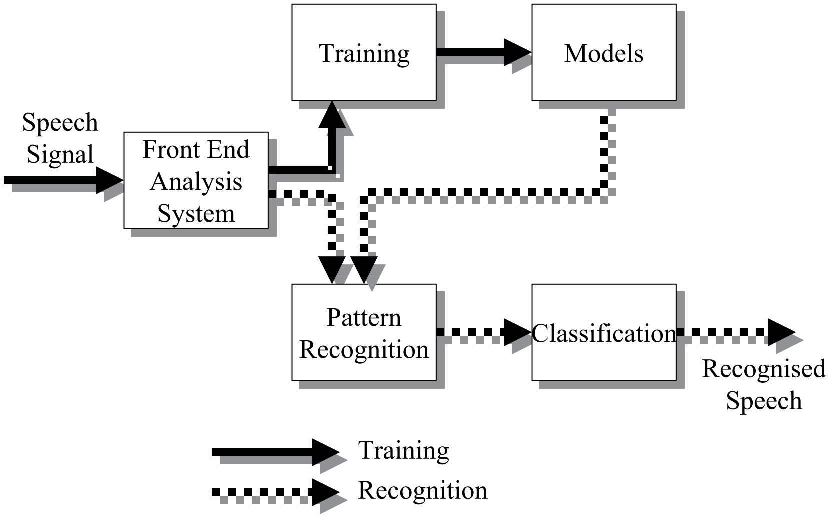


Figure 3.1: Block diagram of a typical statistical speech recognition system.

In this chapter, the techniques used for the front end signal analysis, statistical modelling and pattern recognition will be described in detail.

3.2 Front End Analysis System

The purpose of the front end processing stage is to parameterise the incoming speech signal. The reason for this is two fold: firstly, to represent the signal in a more compact form and secondly, to extract relevant acoustic features from the speech signal to be used in the recognition process.

In all the experiments to be described, a Mel Frequency Cepstral Coefficient (MFCC) [16] front end was used. Other methods of parameterisation have been used for speech recognition and have been found to give good recognition performance [53], however MFCCs allow some computationally efficient techniques for speaker characterisation and normalisation to be incorporated directly in the parameterisation stage. Figure 3.2 shows the components of such an analysis scheme, each of which will be described in detail.

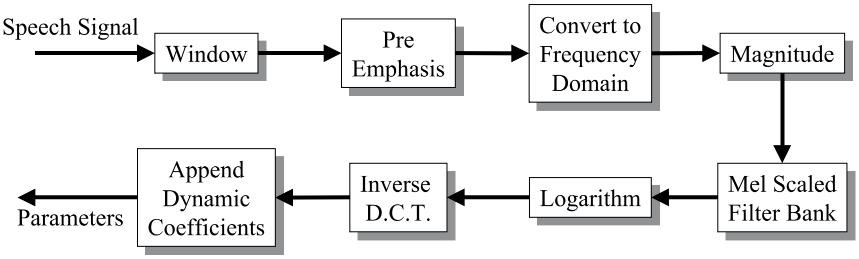


Figure 3.2: MFCC parameterisation scheme.

3.2.1 Windowing

The first procedure in the parameterisation scheme is to window the incoming speech into blocks. As shown by Figure 3.3, the speech signal is continually varying when observed over long periods, while over periods of 20-30 ms the signal is, to a reasonable approximation, stationary. The signal is stationary over this time due to physiological limitations of the speech articulators — the various organs involved in speech production are unable to move fast enough to change their output in a shorter time span [58]. The speech is therefore parameterised in overlapping blocks as shown by Figure 3.4 and the signal in each block is assumed to be stationary. In all the experiments to be presented, a block length of 25.6 ms with 15.6 ms overlap between blocks was used.

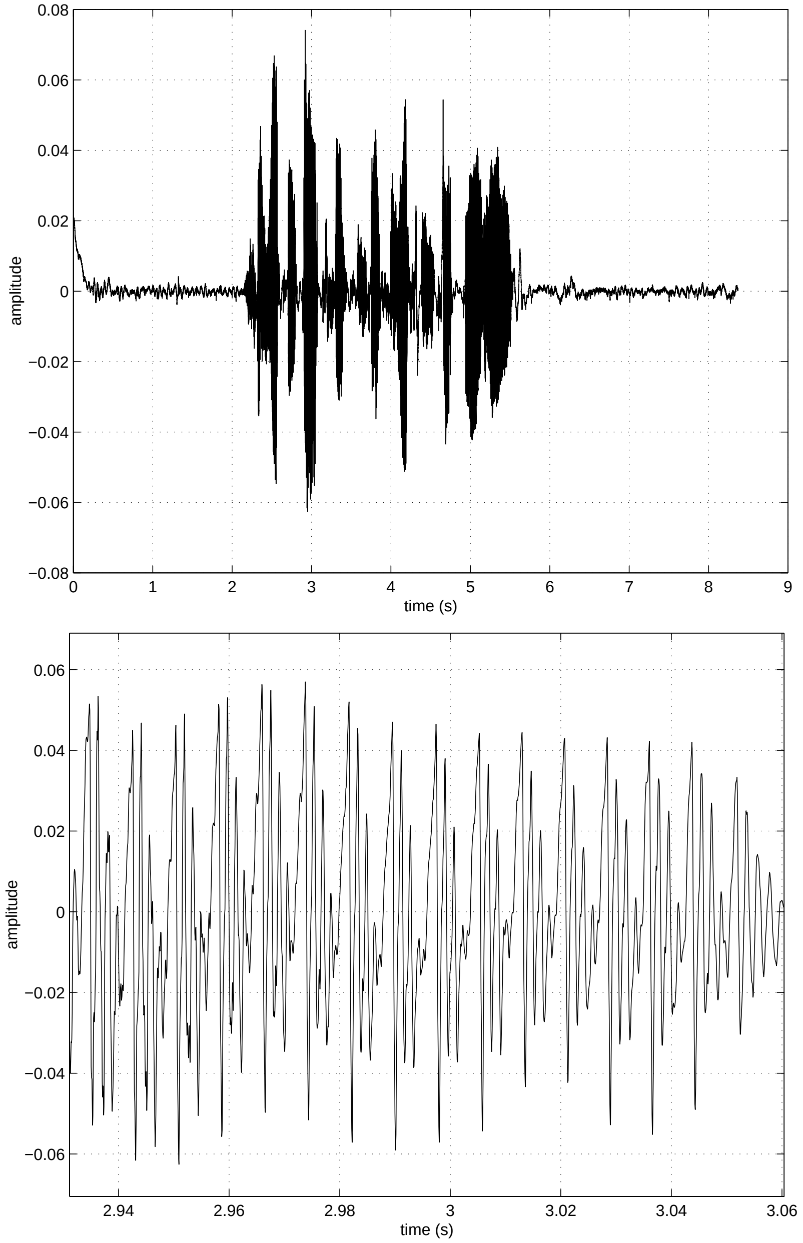


Figure 3.3: A typical speech signal. Top plot shows the signal continually varying over the length of an utterance. Bottom plot shows that the signal is approximately stationary over the duration of several analysis frames.

Windowing using a rectangular window would introduce artifacts into the frequency response of the signal due to the sharp discontinuities the window edges [18]. A Hamming window, as shown in Figure 3.5, which tapers at its edges rather than having a sharp discontinuity, introduces fewer artifacts and is therefore used.

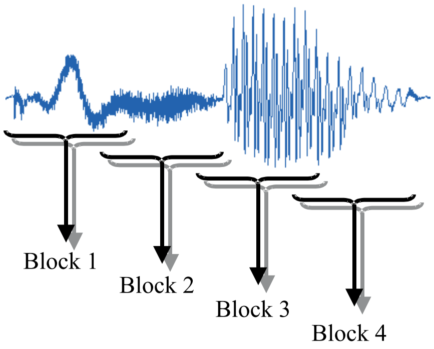


Figure 3.4: The speech is processed in overlapping blocks. The signal within each block is assumed to be stationary.

3.2.2 Pre-Emphasis Filter

As discussed in more detail in Section 2.2, The vocal apparatus may be modelled as a pipe (the vocal tract) open at one end (the lips) and with a sound source at the other (the larynx). The excitation source has a high frequency roll off of -12db/octave while radiation at the lips may be approximated by a 6db/octave spectral lift, resulting in a combined spectral tilt of -6db/octave. It is desirable to have a constant dynamic range across the entire frequency spectrum [58] and the speech is therefore processed to give a 6db/octave lift. This process is usually performed by a first order digital high pass filter. The transfer function of the filter used in the experiments was $H(z) = 1 - 0.96z^{-1}$. The frequency response of the filter is shown in Figure 3.6.

3.2.3 Conversion to Frequency Domain

In Section 2.2 it was shown that the vocal apparatus may be modelled as the output of a sound source being convolved with a time varying filter. The speech sound being produced is characterised by the configuration of the articulators. In the source-filter

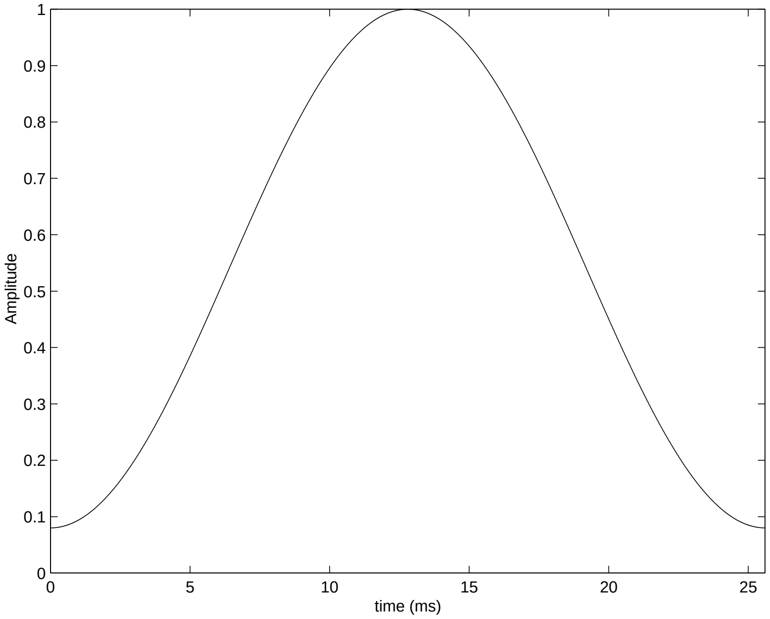


Figure 3.5: Weighting function of Hamming window.

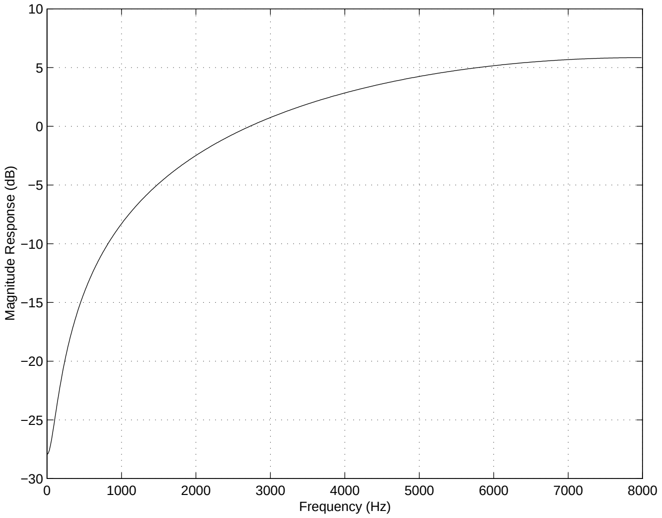


Figure 3.6: Frequency response of pre-emphasis filter, $H(z) = 1 - 0.96z^{-1}$, giving 6db/octave spectral lift.

model this is described by the kind of excitation used and the frequency response of the filter. Since we wish to identify the sound being produced, we may go some way to achieving this by modelling the frequency domain characteristics of the filter. The majority of parameterisation techniques are therefore frequency domain based, and hence the speech signal is converted to its spectral representation. Two methods for deriving the frequency characteristics of the signal were used in this study:

Discrete Fourier Transform

The Discrete Fourier Transform (D.F.T.) is a standard signal processing technique for obtaining the frequency response of a signal. The D.F.T. of a frame from a typical speech signal is shown in Figure 3.7. A discussion of the D.F.T. may be found in many introductory signal processing texts (eg [36, 65, 88]) and a 'C' language implementation of the fast Fourier transform (which is a computationally efficient method of calculating the Fourier transform) is given in [64]. It will not be discussed further, other than to say that it produces a reliable estimate of the spectrum of a signal which may be used for subsequent processing.

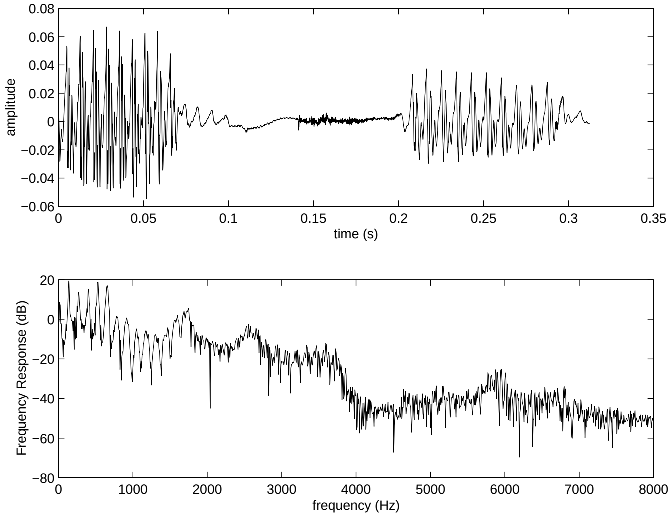


Figure 3.7: Top: A single frame from a typical speech signal. Bottom: Frequency spectrum of the signal obtained using the D.F.T.

Linear Predictive Analysis

The technique of linear prediction is based upon the assumption that sample values of speech may be approximated by a linear combination of the preceding p samples. Mathematically,

$$\begin{aligned}\tilde{x}[n] &= a_1x[n-1]+a_2x[n-2]+a_3x[n-3]\cdots a_px[n-p] && (3.1)\\ &= \sum_{k=1}^{p} a_kx[n-k] && (3.2)\end{aligned}$$

where $\tilde{x}[n]$ is the predicted sample at time n and $a_1, a_2\ \ldots\ a_p$ are the predictor coefficients. Generally it will not be possible to exactly predict the signal, leading to an error $e[n]$ for each sample :

$$e[n] = x[n] - \tilde{x}[n]. \qquad (3.3)$$

The coefficients are determined by solving a set of linear simultaneous equations so as to minimise the mean squared error, E, between the predicted signal and the actual signal.

$$E = \sum_n e^2[n] = \sum_n [x[n]-\tilde{x}[n]]^2 = \sum_n \left[x[n] - \sum_{k=1}^{p} a_kx[n-k]\right]^2 \qquad (3.4)$$

where n is the number of samples over which the error is to be minimised. We need to find a_k such that

$$\begin{aligned}\delta E/\delta a_j = -2\sum_n x[n-j]\cdot\left[x[n]-\sum_{k=1}^{p} a_kx[n-k]\right] &= 0 && (3.5)\\ j &= 1,2,\ldots,p\end{aligned}$$

which gives

$$\begin{aligned}\sum_{k=1}^{p} a_k \sum_n x[n-j]\cdot x[n-k] &= \sum_n x[n]\cdot x[n-j] && (3.6)\\ j &= 1,2,\ldots,p\end{aligned}$$

a set of p linear equations for the set of p unknowns a_k. The choice of p is a compromise between modelling accuracy and computation time — In general, one pair of poles is required to model each of the formants, plus a residual 4-6 poles to model possible zeros and general spectral trends in the signal [57]. p is generally therefore between 10-15, and solving this system of equations is not trivial. Two efficient methods exist for finding the solution — The auto-correlation method and the covariance method. Again these are both covered in most signal processing texts and will not therefore be covered here.

Once the predictor coefficients are known they may be used to estimate the vocal tract response.

The error signal may be calculated if the predictor coefficients are known

$$e[n] = x[n] - \sum_{f=1}^{p} a_f x[n-f] \tag{3.7}$$

and it follows that the original signal may be reconstructed if the error signal and predictor coefficients are known:

$$x[n] = e[n] + \sum_{f=1}^{p} a_f x[n-f]. \tag{3.8}$$

Taking z-transforms

$$\begin{aligned} X(z) &= E(z) + \left[\sum_{f=1}^{p} a_f z^{-f}\right] X(z) && (3.9)\\ X(z) &= E(z) / \left(1 - \sum_{f=1}^{p} a_f z^{-f}\right) && (3.10)\\ &= E(z)H(z) && (3.11) \end{aligned}$$

where $E(z)$ and $X(z)$ are the z-transforms of $e(n)$ and $x(n)$. $H(z)$ is the transfer function of an all pole filter and equation 3.11 shows that the speech signal may be viewed as the output of this filter when the error signal, $E(z)$ is input. From a physical point of view, $E(z)$ describes the vocal tract excitation and $H(z)$ the response of the vocal tract — a precise analogy to the source-filter model. An approximation of the vocal tract response may be obtained by substituting $z = e^{j\omega T}$ in $H(z)$:

$$H(\omega) = 1/\left(1 - \sum_{f=1}^{p} a_f e^{-j\omega f T}\right) \tag{3.12}$$

and evaluating $\mid H(\omega) \mid$ at various values of ω as shown in Figure 3.8. This is directly analogous to the source filter model described in Section 2.2.

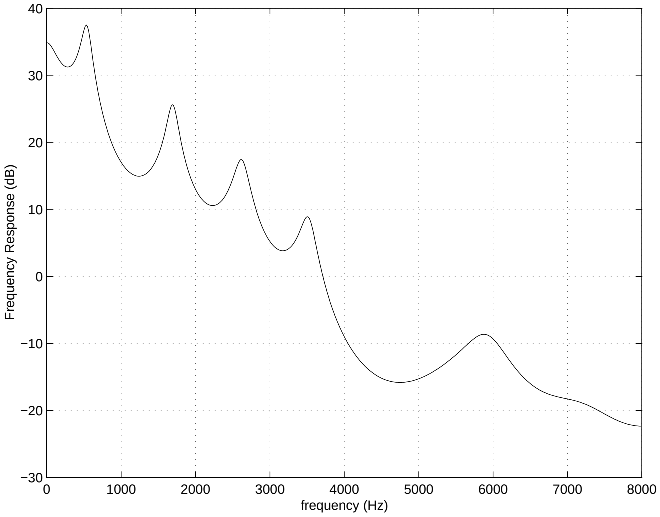


Figure 3.8: Approximation of the vocal tract frequency response obtained using LPC analysis.

3.2.4 Magnitude

The phase of the signal carries little useful information for recognition [57] and would increase the amount of computation required for subsequent processing. The phase is therefore discarded to leave the log magnitude spectrum of the signal.

3.2.5 Mel Filter Bank

The human auditory system does not resolve sound equally at all frequencies; rather the response of the system can be considered to be split into frequency bands known as ‘critical bands’ [37, 57, 91]. If two sounds are played with frequencies within a

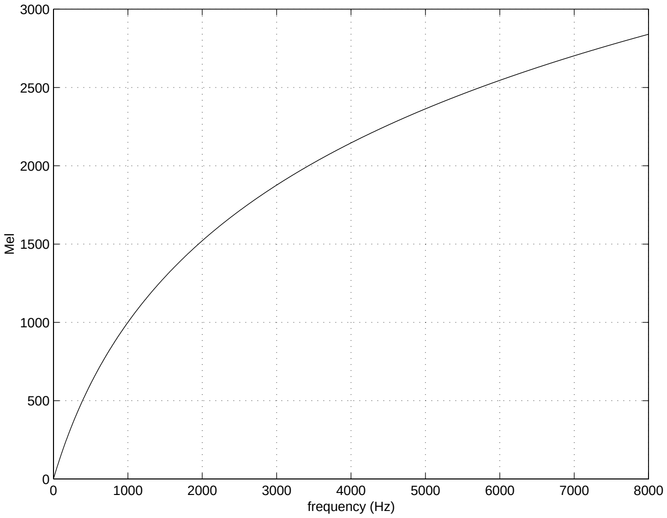


Figure 3.9: Relationship of Mel scale to frequency.

single critical band, the signal with higher energy will mask the other. Experiments have shown that the bandwidths of the critical bands are roughly linearly related to frequency below 1kHz and approximately logarithmically related to frequency above 1kHz. Hence low frequency sounds have better resolution than high frequencies. Studies have shown that emulating this performance in the front end processing stage can result in improvements in recognition performance [16]. This may be implemented by using a filter bank, with non-linear spacing of the filters across the frequency range. The filter bank used in the study follows the Mel-scale where

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3.13}$$

As Figure 3.9 shows, the Mel scale is approximately linear from 0 to 1kHz, and logarithmic thereafter. Triangular filters, linearly spaced along the Mel scale give rise to the required variation in frequency resolution as we go from low to high frequency as shown in Figure 3.10.

At each point, the frequency spectrum of the signal is multiplied by the filter weight at that frequency. The output of each filter bank channel is then the sum of these weighted frequency components. In the experiments to be described, 26 filters were used, giving 26 Mel frequency coefficients as the output of the filter bank.

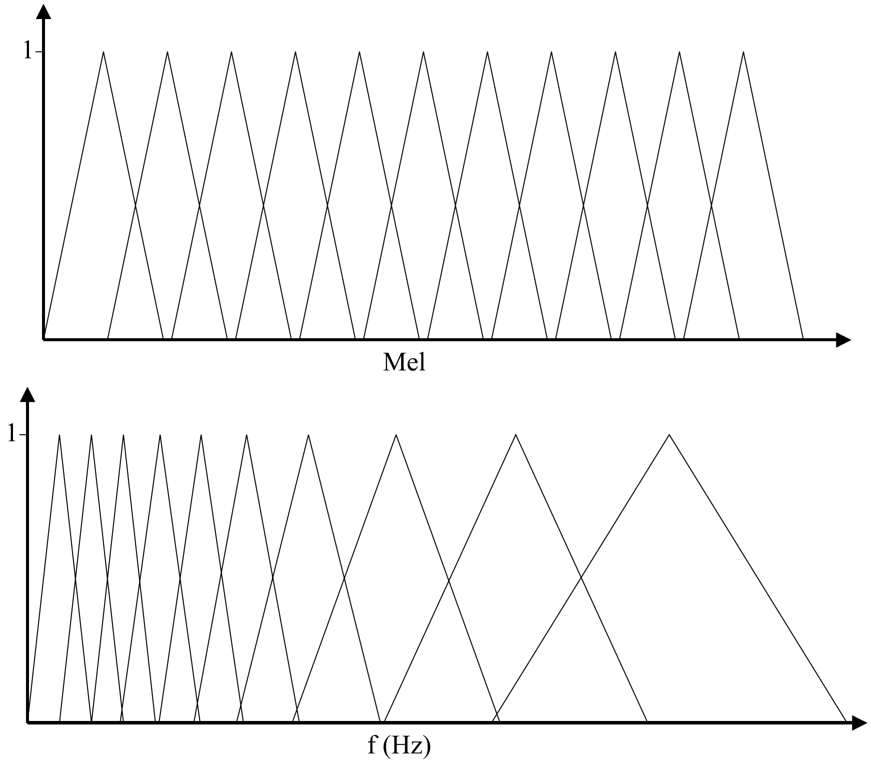


Figure 3.10: Placement of filter banks to emulate critical band behaviour of human auditory system. (Top) Filters are linearly placed along the Mel scale giving rise to the required non linear spacing in the frequency domain (Bottom).

3.2.6 Log

If we assume the frequency response of the speech signal, $H(\omega)$ is the product of the spectra due to the source $S(\omega)$ and the vocal tract $F(\omega)$,

$$|H(\omega)| = |S(\omega)| \times |F(\omega)| \tag{3.14}$$

as the source filter model suggests, then taking logs gives us :

$$\log_{10}|H(\omega)| = \log_{10}|S(\omega)| + \log_{10}|F(\omega)| \tag{3.15}$$

i.e. in the log magnitude spectrum, the contribution from each of the components of the model are summed. The contribution from the vocal tract tends to be slowly varying (low frequency) while that from the excitation source is of higher frequency. Hence the contributions are separable by means of a linear filtering operation on the log magnitude spectrum.

3.2.7 Inverse DCT

Taking the inverse transform of the log magnitude spectrum gives the cepstral coefficients of the speech signal. The component due to the periodic excitation source may be removed from the signal by simply discarding the higher order coefficients. In this study, 12 coefficients were retained after the DCT.

The inverse DCT also serves to decorrelate the coefficients, an assumption which is made in the modelling technique to be described.

3.2.8 Addition of Dynamic Coefficients

In the recognition methods to be described, no use is made of the fact that consecutive frames of speech are likely to be highly correlated, since the articulators may only move a limited distance in the 10 ms gap between frames [85]. Dynamic features, that is, values which attempt to explain the way in which the speech signal is varying between successive frames, such as those presented in [25, 28, 52] are therefore appended to the static coefficients. The following was used to calculate the first order dynamic coefficients, known as velocity, or delta coefficients:

$$d_t = \frac{\sum_{\theta=1}^{N} \theta(c_{t+1} - c_{t-1})}{2\sum_{\theta=1}^{N} \theta^2} \tag{3.16}$$

where d_t is the delta coefficient at time t and c_t is the static coefficient at time t and N is the width of the window [86]. Since this formula relies on the current samples preceding and subsequent samples, at the beginning and end of the speech the first and last parameters are copied to fill the required regression window. Second order, known as acceleration, or delta-delta coefficients are obtained by applying the same formula to the delta coefficients. In this study the window size used was two.

Finally, the log spectral energy has also been shown to be a useful feature for discrimination, and it, along with its dynamic coefficients is also appended to the feature vector.

In the experiments presented here we therefore have a 39 component feature vector comprising of :

- 12 Mel frequency cepstral coefficients
- 12 delta coefficients
- 12 delta-delta coefficients
- log energy
- delta log energy
- delta-delta log energy

As the frame advance rate is 10ms and each frame consists of 39 coefficients, there are 3900 coefficients per second. If each coefficient is represented at 16 bit precision, this leads to a bit rate of 62.4 kBits/s — reduced over the raw speech signal. More importantly, the techniques described extract features which are relevant to the classification of the speech signal and may be used to generate accurate models of the speech sounds.

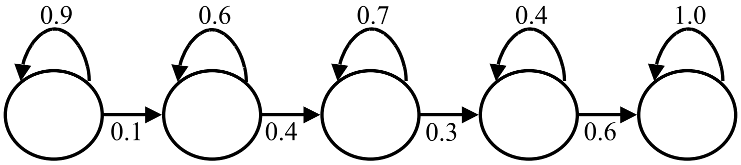
3.3 Hidden Markov Models (HMMs)

Any method of modelling speech must account for the fact that the information in the signal is carried by the temporal ordering of the sounds. The model must also be able to describe the variation within sounds, while identifying the differences between them. A stochastic process is able to perform both these requirements. Such a method, and one which has become extremely popular in the modelling the speech data is that of hidden Markov modelling.

Here a brief description of the general principles behind the method is given, followed by a discussion of how Hidden Markov Models may be used in the classification of unknown speech signals. More detailed descriptions are given in [14, 45, 54, 67, 69]

3.3.1 Description of an HMM

Figure 3.11 shows an example hidden Markov model



$$A = [a_{ij}] = \begin{bmatrix} 0.9 & 0.1 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.6 & 0.4 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.7 & 0.3 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.4 & 0.6 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

$$B = [b_{jk}] = \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0.1 & 0.8 & 0.1 & 0.0 \\ 0.0 & 0.2 & 0.8 & 0.0 \\ 0.7 & 0.0 & 0.0 & 0.3 \\ 0.2 & 0.0 & 0.0 & 0.8 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix} \end{array}$$

$$\pi = \begin{bmatrix} 0.4 & 0.6 & 0.0 & 0.0 & 0.0 \end{bmatrix}$$

Figure 3.11: A 5 state left right, discrete HMM with 4 output symbols.

The model consists of a number of *states*, shown as the circles in Figure 3.11. At time t the model is in one of these states and outputs an *observation* (A, B, C or D). At time $t+1$ the model moves to another state, or stays in the same state and emits another observation. The transition between states is probabilistic and is based on the transition probabilities between states which are given in the state transition matrix, A, where A_{ij} is the probability of being in state i at time t and moving to state j at time $t+1$. Notice that in this case A is upper triangular. While in a general HMM transitions may occur from any state to any other state, for speech recognition applications transitions only occur from left to right i.e., the process cannot go backwards in time, effectively modeling the temporal ordering of speech sounds. Since at each time step there must always be a transition from a state to a state each row of **A** must sum to a probability of 1.

The output symbol at each time step is selected from a finite dictionary. This process is again probabilistic and is governed by the output probability matrix B, where B_{jk} is the probability of being in state j and outputting symbol k. Again since there must always be an output symbol at time t, the rows of $\mathbf{B}$ sum to 1.

Finally, the entry probability vector, π, is used to describe the probability of starting in each of the i states of the model — π_i being the probability of starting in state i.

The model is fully described by the parameter set $\mathbf{M} = [\pi, \mathbf{A}, \mathbf{B}]$.

3.3.2 The Markov Source

Such a model may be used in conjunction with a random number generator (RNG) to produce an observation vector. Initially, the starting state is determined using π and the output of the number generator. Then at each time interval t, the output symbol is chosen based on B and the RNG, and A is used to determine the next state. The process continues until state 5 is reached. This state has a self transition probability (that is, a probability of returning to itself) of 1, and outputs only a single dummy symbol, D. After this state is reached, all output symbols will therefore be D. A typical output sequence may be AAAABBBABBBBBBCCCCCCBBBCD. At each time instant we know only the output of the model, but not which state we are in. The state is effectively ‘hidden’ from us (though, as will be shown in Section 3.5.2, it is possible to calculate the most likely state sequence).

3.4 Application of HMMs to Speech Recognition

If we make the assumption that the speech articulators, while generating a given sound are moving between a series of target positions, and at each position they generate a characteristic output for a varying length of time, the correlation between this and the hidden Markov model is clear. Each ‘target position’ becomes a state in the model, and the ‘sound generated’ (actually the vector output by our front end at that time frame) is represented by the output symbol. With our present example the outputs must be discretised into a finite number of symbols by, for example, vector quantisation of the speech vectors. The model can however be extended to allow for a continuous set of output symbols defined by a probability density function, which is more appropriate for modelling speech sounds.

There are two problems associated with the application of HMMs to speech recognition :

- *The training problem:* Given a set of utterances, labelled at some level of speech unit, generate a set of models (i.e. estimate the values of A, B and π) each of which represents one of the units of speech.
- *The recognition problem:* Given a sequence of speech frames whose classification is unknown, and a set of well trained models, identify the most likely model for each input vector

The training problem is the more difficult of the two. However an algorithm exists (the Baum-Welch algorithm) which guarantees to produce a locally optimal model for a given set of training data. This procedure is covered in detail elsewhere [14, 67, 86] and will not be discussed here. We will assume that a well-trained set of models exist for the speech we wish to recognise. The recognition problem may be solved by means of maximum likelihood classification. That is we find the model, or series of models which has the highest probability of having produced the given unknown observation sequence.

3.5 The Classification Problem

For isolated word recognition, given the unknown observation sequence, $O_1, O_2, O_3, \ldots, O_T = \mathbf{O}$ and a set of S models, $\mathbf{M} = m_1, m_2, m_3, \ldots, m_S$, each with N states and M discrete output symbols, we wish to find the class, C such that

$$C = \underset{i=1,2,\ldots,S}{\arg\max} P(\mathbf{O}|m_i) \tag{3.17}$$

It is unrealistic to estimate $Pr(\mathbf{O}|\mathbf{M})$ by evaluating every possible state sequence which could have generated the observation, since in general there are some N^T such sequences. Instead a recursive method utilising the ‘forward probabilities’ is used.

3.5.1 Baum-Welch Classification

The forward probabilities, $\alpha_t(j)$, are defined as the joint probability of emitting the partial observation sequence, $O_1, O_2, \dots, O_t$ and being in state s_j at time t, i.e.

$$\alpha_t(j) = Pr(O_1, O_2, \dots, O_t, s_j@t|\mathbf{M}). \tag{3.18}$$

Then the required probability is given by

$$Pr(\mathbf{O}|\mathbf{M}) = P^{BW} = \sum_{j=1}^{N} \alpha_T(j) \tag{3.19}$$

The forward probability at the next time instant, $t+1$ for some state j depends only on the current forward probabilities, the transition probabilities from the current state to the next state and the probability of outputting the next observation from the current state. Hence the forward probabilities may be calculated recursively

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(O_{t+1}) \qquad t = 1, 2, \dots, T-1 \tag{3.20}$$

The starting conditions are given by

$$\alpha_1(j) = \pi(j) b_j(O_1) \tag{3.21}$$

and hence using equation 3.19 we may calculate the required probability.

3.5.2 Viterbi Classification

The Viterbi algorithm allows the most likely state sequence through the model to be identified. The summation in equation 3.20 is replaced by a maximum operator resulting in a ’best path’ search. A recursive algorithm similar to equation 3.20 is used to calculate the probabilities at each time step. The recursion is

$$\phi_{t+1} = \max_{i=1,2,\dots,N} \left[\phi_t(i) a_{ij}\right] b_j(O_{t+1}) \qquad t = 1, 2, \dots, T-1 \tag{3.22}$$

with starting conditions

$$\phi_1 = \pi(j) b_j(O_1). \tag{3.23}$$

The final probability of emitting the sequence $\mathbf{O}$ for a given model is found by maximising over all states,

$$\overline{Pr(\mathbf{O}|\mathbf{M})} = P^V = \max_{i=1,2,\dots,N} \phi_T(i) \tag{3.24}$$

In order to recover the most likely state sequence, at each time instance, t, and for each model state, i, we record the state at time t which maximised equation 3.22 at time $t+1$:

$$\psi_{t+1}(j) = \underset{i=1,2,\dots,N}{\arg\max}[\phi_t(i) a_{ij}] \qquad t = 1, 2, \dots, T-1 \tag{3.25}$$

The most likely state at time T, given by equation 3.24 is used to recover the state sequence by back tracking :

$$k = \psi_T(j) = \underset{i=1,2,\dots,N}{\arg\max}[\phi_T(i)] \tag{3.26}$$

k, the most likely state at time $T-1$ is used to find the most likely state at $T-2$ from $\psi_{T-1}(k)$, and so on until the most likely state at $t = 1$ is found. Given the optimal state sequence for a set of frames we may use the information to learn about the structure of the model, or use the data to re-estimate the parameters of the models [86]. The Viterbi algorithm is also computationally less expensive than Baum-Welch since we do not have to perform the summation in equation 3.20, and a trellis structure may be used to provide an efficient implementation.

3.6 Extensions to Basic HMMs

The discussion above has given a basic background into the use of HMMs for statistical pattern matching. Several extensions to the basic model are used in practical recognition systems which will now be discussed.

3.6.1 Continuous Density HMMs

The models described in Section 3.3.1 generate their output from a finite library of observations. The parameterised speech data is a continuously varying quantity and as such would have to be quantised to one of these observation values using vector quantisation techniques [47,68] to be used with such a system. This results in additional errors being introduced due to quantisation noise. An alternative is to replace the discrete output probabilities with a continuous probability distribution of observations as shown in Figure 3.12. The multivariate Gaussian distribution is the most widely used, because a weighted mixture of Gaussians may model, arbitrarily closely, any probability density function [46].

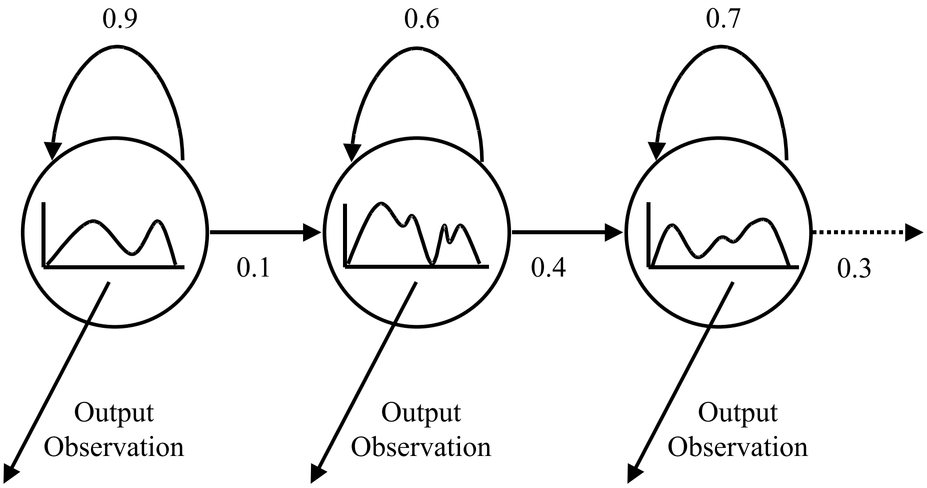


Figure 3.12: Continuous density HMM

The rows of B are then replaced by the parameters of the PDF, and the output probability is modelled by,

$$b_j(O_t) = \sum_{m=1}^{X} c_{jm}\mathcal{N}(O_t, \mu_{jm}, \Sigma_{jm}) \qquad j = 1, 2, \dots, N \tag{3.27}$$

where X is the number of mixtures, c_{jm} is the weight of mixture m in state j and $\mathcal{N}(O_t, \mu_{jm}, \Sigma_{jm})$ is the probability of observation vector O_t from multivariate Gaussian distribution with mean vector μ_{jm} and covariance Σ_{jm},

$$\mathcal{N}(O,\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} e^{\left(-\frac{1}{2}(O-\mu)'\Sigma^{-1}(O-\mu)\right)} \tag{3.28}$$

The probability distribution for each state in every model has its own set of means and covariances, though in practice, to reduce computation times, each of the components in the feature vector are assumed to be uncorrelated and as such a diagonal covariance matrix is used.

3.6.2 Semi Continuous HMMs

In a semi continuous HMM, shown in Figure 3.13, all models share a large common pool of distributions or ‘modes’ and the output probabilities for a given state are weighted sums of this common pool of modes [31, 32]. The output probability matrix for each state is reduced to a vector of weights for each of the modes. This particular HMM topology reduces the amount of space required for the storage of the models (the pool needs only be saved once) and also allows for finer modelling of the feature vector distributions, as a large mixture of Gaussians may be used to model the distributions, but may result in a loss in generality for the output distributions — it is possible that the probability distribution for a certain state may have components that are not present in any of the modes and can not therefore be accurately modelled. The SCHMM is therefore a compromise between the computational complexity and high storage requirements of a continuous HMM and the simplicity but lack of generality of a discrete HMM.

3.6.3 Sub Word Modelling

The previous discussion of HMMs concentrated on the use of whole word models, that is, each HMM represents a single word in the required recognition vocabulary. For isolated word recognition it is then sufficient to pick the model with the highest output probability as the recognised word. In practice, in anything other than a small vocabulary system (i.e. one with fewer than about 100 words) it is unlikely that there will be sufficient training data to train whole word models. It would also be difficult to

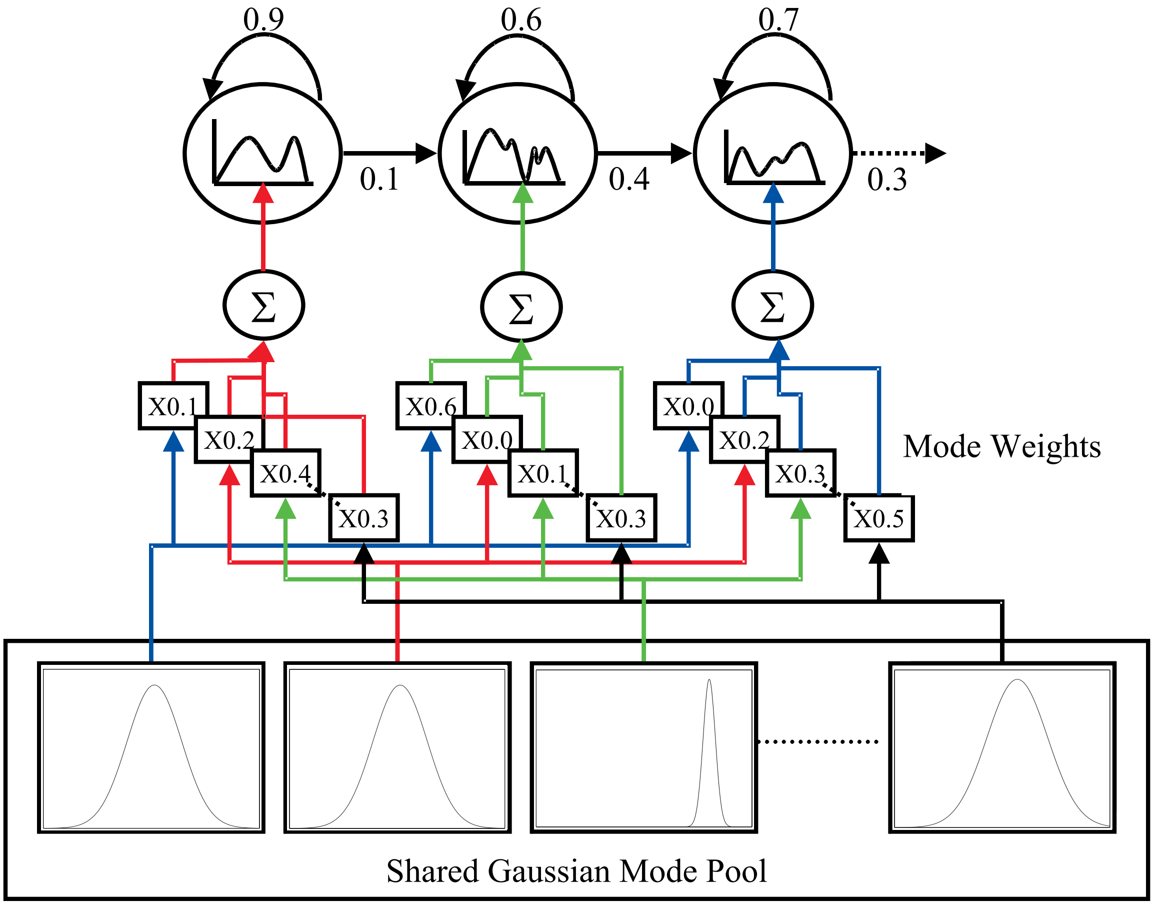


Figure 3.13: Semi Continuous HMM.

add new words to such a system since each new word would require a large amount of data. Instead, sub word models are built for each sound or ‘phoneme’ in the language, and words are described as sequences of these units. Since many of the IPA symbols are difficult to represent in a computer system, they are encoded to the ARPABET symbols as shown in Table 3.1.

At recognition time the best path through concatenated sequences of these phones is used to determine the identity of the input utterance. The output probability for each series of phones is calculated and the sequence with the highest output probability chosen.

	IPA	ARPABET	Example		IPA	ARPABET	Example
	symbol	symbol			symbol	symbol	
Vowels	i	iy	lead	Consonants	p	p	pin
	ɪ	ih	pit		b	b	but
	e	eh	pet		t	t	ten
	æ	ae	pat		d	d	den
	u	uw	boon		k	k	can
	ʊ	uh	good		g	g	game
	ʌ	ah	putt		f	f	full
	ɒ	oh	pot		v	v	very
	ə	ax	about		θ	th	thin
	ɜ	er	burn		ð	dh	then
	ɔ	ao	born		s	s	some
	ɑ	aa	barn		z	z	zeal
Diphthongs	eɪ	ey	bay		ʃ	sh	ship
	aɪ	ay	buy		ʒ	zh	measure
	ɔɪ	oy	boy		tʃ	ch	chain
	ɑʊ	aw	now		dʒ	jh	jane
	əʊ	ow	load		m	m	man
	ɪə	ia	peer		n	n	not
	ɛə	ea	pair		η	ng	long
	ɔə	ua	pore		l	l	like
	ʊə	ua	poor		r	r	run
					j	y	yes
					w	w	went
					h	hh	hat

Table 3.1: IPA symbols for the phonemes used in RP English transcription with examples of their use.

3.6.4 Lexical Decoding

At recognition time it is impossible to evaluate the output probability for *all* the possible sequences of concatenated phonemes so lexical decoding is used to place constraints on the sequences of phones which are evaluated. A dictionary is used which maps words within the recognition lexicon to the sub word units being used as shown in Table 3.2.

Word	Pronunciation	Probability
barter	b aa t ax sp	0.5
barter	b aa t ax r sp	0.5
bartered	b aa t ax d sp	1.0
barterer	b aa t ax r ax sp	0.4
barterer	b aa t ax r ax r sp	0.6
barterers	b aa t ax r ax z sp	1.0
bartering	b aa t ax r ih ng sp	1.0
barters	b aa t ax z sp	1.0
bartes	b aa t s sp	1.0

Table 3.2: Extract from typical pronunciation dictionary. Each word in the lexicon is associated with one or more sequences of phonemes describing its pronunciation

Only sequences of phones which correspond to words within the lexicon are investigated at recognition time. Multiple pronunciations of a single word may be included in the dictionary, and a probability of occurrence associated with each distinct pronunciation. Adding a new word or pronunciation to such as system is achieved by simply including it in the dictionary (and adjusting the probabilities of the pronunciations if necessary [63].

3.6.5 Syntactic Analysis

Syntactic analysis imposes further constraints on the network of sub word HMMs to be searched. Only paths for which the corresponding words are in a proper sequence based on the task grammar are investigated. The grammar may consist of a finite state network which explicitly defines all word combinations which are acceptable to the recogniser [56, 63].

Alternatively a statistical grammar may be used — A trigram language model for instance gives the probability of sets of 3 words occurring which is then incorporated into the final probabilities for each path investigated [10, 60].

It is within this framework of acoustic analysis and stochastic modelling with lexical and grammatical constraints that the techniques for speaker characterisation and adaptation must be incorporated. To improve ASR accuracy however, we require some idea of the nature of the differences between speakers' speech and how they may be identified from the acoustic signal.

Chapter 4

Interspeaker Variation

Figure 4.1 shows a very basic model of the way in which we generate speech sounds. We have some concept of a target sound that we wish to produce in order to communicate an idea, and via our vocal apparatus we attempt to make that sound. There are two important effects which cause speakers to produce different realisations of a given word. These are variation in the target sound to be produced caused by learned speaking styles, and variation in the realisation of the sound caused by differences in vocal apparatus. In this chapter these variations are described, as are ways to account for them. We neglect many other factors which cause variations in the acoustic signal, such as the desired rate of speaking, the emotional state of the speaker, the effect of illnesses such as colds, etc.

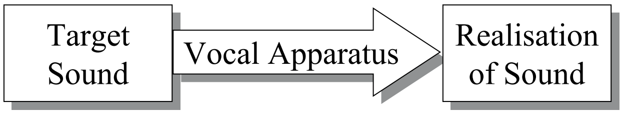


Figure 4.1: Simple speech production model

4.1 Variation in Target Sound

Differences in what we perceive as the target sound in a particular context will manifest themselves in the accent and dialect used when speaking. A distinction should be made between the two — dialects consist of variations in the syntactic structure of the language, the vocabulary used, and the associated pronunciations. Accents differ only in their pronunciations [82]. The speech data used during this study is derived from databases recorded by speakers reading from scripts, rather than spontaneous speech. As such, the syntax and lexicon are predefined and any dialect the talker has will not be represented in the speech. Differences in pronunciation will be apparent in the utterances, and we therefore concentrate on variations in accents between talkers.

4.1.1 Sources of Accent Variation

Wells [82] suggests many factors which influence the accent used when speaking :

- *Geographical region:* Accents frequently indicate the geographical region from which the talker originates. For native speakers, the precision with which we may place a speaker depends largely on our familiarity with the region. At a coarse level most people may easily differentiate between British and American English speakers, while someone from England would find it easy to distinguish between Northern and Southern British, or Liverpudlian and Mancunian, while having difficulty differentiating two American accents - a task readily performed by an American. For non-native speakers, Flege [20–22] identifies several factors with influence the degree of accent identified in a talkers speech. These are shown to be related to both the talker (such as the age at which the second language was learnt and the length of time spent in a country where the second language is the native language) and the listener (for instance with their familiarity to the sentence being uttered).

- *Socio-Economic class:* In British English, there is is often a wide variation between the accent used by differing social classes in a given region. It is also generally found that the amount of variation between regional dialects is a function of class as shown in Figure 4.2. Speakers in class IV or V as defined in [70] show a much wider variation in accents between regions, whilst between those in class I the variation is less pronounced. RP (received pronunciation) English, traditionally characterised as the accent used by the ‘upper classes’, is generally taken to be the ‘standard’ pronunciation of British English to which variations in other accents are referred. This is because it is non-localised, and also because a

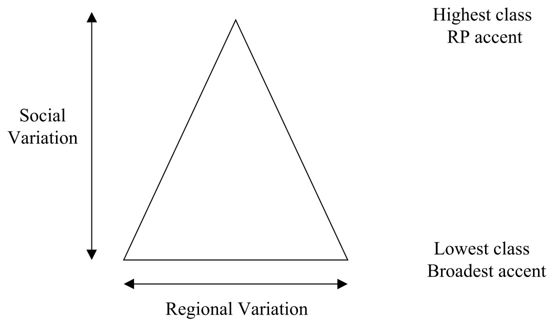


Figure 4.2: Amount of variation in regional accent as a function of social class. After [82]

large amount of research has been conducted into this accent.

- *Age:* There is often a difference between the accents used by older and younger people. Accent is generally learnt up to the age of around 11, and it is the influence of ones peers who are largely responsible for influencing accent. As such it is children who tend to introduce and proliferate changes in accent.

- *Style:* Accent changes as the style of speech alters. In normal, conversational speech with friends or family accents are likely to be broad since we are not monitoring our speech. Formal speech, when talking to strangers or being interviewed is generally less likely to have such large accent effects. The accent decreases further when reading aloud and further still when reading a list of words or phrases.

4.1.2 Result of Accent Variation

The used accent may alter the way in which a person talks in several ways [82]

- *Phonetic Realization:* These are differences in the way in which we produce a certain phoneme, for instance the degree of lip rounding in the diphthong /əʊ/ as in ‘coat’, or the starting point of the phoneme /ɑʊ/ as in ‘out’. Differences may also occur depending on the surrounding phonemes of the sound begin produced - an effect known as co-articulation.

- *Phonotactic Distribution:* These are differences in the allowed sequences of phonemes in the accent. The rhotic English accents (Scottish, Irish, parts of the

West country, many of the American accents) allow the phoneme /r/ in a variety of contexts including pre-consonantal (eg farm) and absolute-final position (eg far). The non-rhotic accents do not.

- *Phonemic Systems:* These are differences in the number and identity of phonemes available to the speaker. RP English for instance has two close-back vowels, /u/ as in boon and /ʊ/ as in good. Scottish English does not have the short version, and as such both these phonemes are represented by /u/.

4.1.3 Consequence of Accent Variation to Automatic Speech Recognition

Variations in accent cause certain problems when performing automatic speech recognition. As was discussed in Section 3.1, most successful speech recognition systems have acoustic models which represent the various sounds in the language. Since the phonetic transcriptions of the speech used to train the models are not of sufficient detail to identify the differences in phonetic realization between two accents, a single model usually accounts for all the different realisations. This leads to higher variances within the model parameters since they are modelling a broader set of acoustic parameters resulting in greater overlap between models and reduced recognition performance, as shown in [17, 75].

The manner in which the acoustic models may be ordered to produce words in the lexicon is described in the pronunciation dictionary. Since a single dictionary is used for all speakers, it must cover the lexical distribution used by speech with all accents which the recogniser may encounter. This leads to a large increase in the number of pronunciations in the dictionary, increasing search times and decreasing performance due to the added possibility of confusions [41].

A further point is that, while it is highly unlikely that a speaker will change accent mid conversation, their is no reason for the recogniser to consistently choose pronunciations from a given accent group. That is, a word may be output based on its southern English pronunciation, immediately after one with a Scottish pronunciation. As has been described, the phonotactic distribution of these two accents are different and this should not occur. Additional constraints could be put on the recognition system, restricting outputs to pronunciations of words from a single accent group, which would reduce the decoding time and improve performance since the search space would be reduced.

4.1.4 Use of Accent Specific Information to Improve ASR Accuracy

If the accent group used by a speaker could be accurately and quickly identified then the problems outlined previously could be reduced.

To account for the variations in lexical distribution, a pronunciation dictionary for each of the accent groups could be produced. Once a speaker is identified as having a particular accent, pronunciations from other groups could either be removed or their probability of occurrence greatly reduced by means of a weighting factor. In [34] and [72], methods of automatically generating pronunciation dictionaries which could be used in such a system are presented.

The techniques used for language identification [12, 29, 84, 89, 90] could be used in the classification of accent, since accent identification is essentially the same task on a finer level. In [90], four popular methods of language identification are compared. Gaussian mixture models are shown to have the lowest performance, with the advantage of requiring no labelled training data, and running in real time. The other three systems are based on comparing the output probabilities of phone recognisers from one or more languages combined with syntactic models from each language. In ‘Phone recognition followed by language modeling’ (PRLM) the output of a phone recogniser from a single language is decoded using grammar models from several different languages. The grammar model which gives the highest output likelihood is chosen as the most probable language. An extension of this is ‘Parallel phone recognition followed by language modeling’ (P-PRLM). Here the output of several phone recognisers is decoded by each of the language models and the results combined - this allows for languages with different phone sets to be classified. In ‘Parallel phone recognition’ (PPR), the language is identified using multiple recognisers with acoustic and grammar models from a single language. These three techniques are shown in Figure 4.3.

Such systems have much higher accuracies than Gaussian mixture models, at the cost of greatly increased computation time. Despite the fact that P-PRLM has been shown to be successful in automatic accent classification [7], this technique would be

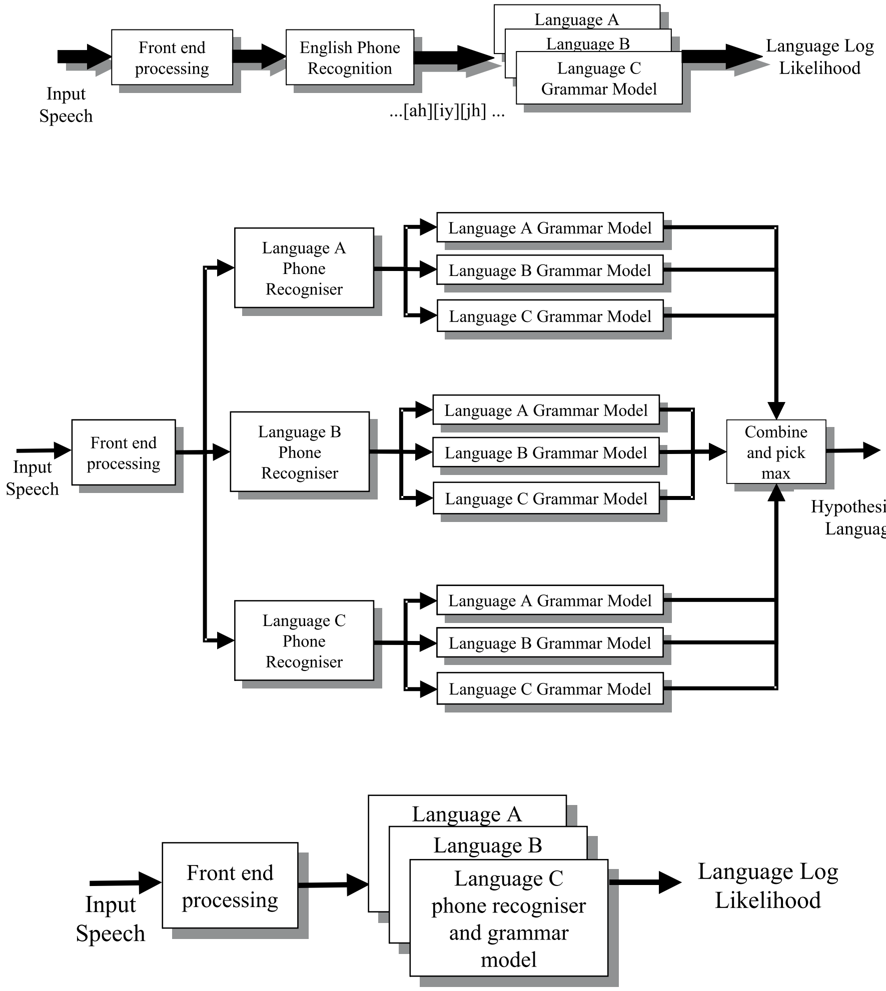


Figure 4.3: Methods for automatic language identification. (Top) PRLM. (Middle) P-PRLM. (Bottom) PPR

too computationally expensive to run in conjunction with a real-time recogniser.

In [8, 33, 50], methods of determining regional accent are given, however they require the user to utter specific words or sentences designed to highlight the differences between accents. This would be unacceptable in many applications. We therefore investigated methods of accent identification as a by product of the recognition process - that is with virtually no computational overhead, utilising phonotactic knowledge of each accent (Section 7).

A solution to the problem of variations in phonetic realization is to build separate model sets for speakers with similar realisations. In [7] and [75] it was shown that this approach can provide significant improvements in recognition accuracy.

4.2 Variation in Vocal Apparatus

Once a speaker has decided upon the sound he or she wishes to make in a given situation, the speaker attempts to arrange his or her vocal apparatus to produce an acoustic realisation that is as close as possible to the target.

4.2.1 Sources of Vocal Apparatus Variation

As with all areas of human anatomy there are significant differences between the vocal apparatus of individuals. The length of the vocal tract varies between male and female speakers, from about 13cm to 18cm; the nasal cavity size can vary and the characteristics of the vocal cords change from speaker to speaker. Such changes, unlike accent variation, are specific to a given individual, rather than to a group of talkers.

4.2.2 Result of Vocal Apparatus Variation

Variations in vocal apparatus result in measurable differences to the speech signal produced, even if two speakers wish to produce the same phonetic realization. The position of the formants will change as an approximately linear function of vocal tract length. The pitch varies as a function of the vocal chord characteristics and the bandwidth and inclination of the signal spectrum as a whole will also vary from speaker to speaker.

4.2.3 Consequence of Variation to ASR

As with differences in phonetic realization, speaker dependent variations for a given phoneme result in increases in the variances of the models and a subsequent reduction in recognition accuracy. Several methods have been proposed to attempt to overcome the problem of variation between speakers' realisations of a given phoneme:

4.2.4 Speaker Dependent Recogniser

The simplest method of dealing with variation in speakers, and one which accounts for differences in phonetic realization as well as changes in vocal apparatus, is to build a speaker dependent (SD) recogniser. The acoustic models in an SD recogniser are trained on the speech of a single speaker and therefore have very much lower variations than those trained on speech from many talkers. There are however several disadvantages to this approach. Firstly, a large amount of training data is required to successfully estimate the parameters of the models - many hours of speech - and it would be impractical to collect this whenever a new user wanted to use a system. Secondly, once trained for an individual speaker, performance for other users is normally extremely poor since their speech is unlikely to match precisely that of the user on which the models were built. As a consequence, each user would have to have their own set of models which would require a large amount of storage space if the system were to be used by many speakers.

4.2.5 Speaker Adaptation and Speaker Normalisation

The terms 'speaker adaptation' and 'speaker normalisation' have been used loosely and in different ways in the literature. The distinction we make here is that 'adaptation' means adapting the speech models to become 'closer' to the speaker, whereas 'normalisation' means adapting the speaker's data to some standard or canonical talker.

Two distinctions should be made when discussing speaker adaptation and normalisation techniques :

- *Supervised vs Unsupervised:* Supervised adaptation/normalisation requires the new user to read a given passage to the system before adaptation takes place. This passage is often chosen to highlight differences in pronunciation, and to cover as large a number of phonemes as possible. Alternatively the speaker may be asked to correct the output of the recogniser during recognition. This technique is only appropriate when the speaker is to be using the system for a significant period of time and the time used to read the passage or correct the recogniser is insignificant compared with the time spent using the system. Un-supervised adaptation generally occurs without the user being required to utter a specific phrase. The recogniser is used to identify which models to adapt with each utterance.

- *Batch vs Incremental:* Batch adaptation is performed by collecting a large sample of the user speech and subsequently processing it. Incremental adaptation is performed at recognition time and the recogniser should be seen to improve in accuracy as the speaker uses the system.

From a user's perspective, unsupervised, incremental adaptation is preferable, however this method is significantly more difficult than the batch, supervised method — problems may arise in unsupervised techniques if, for instance, the recogniser misrecognises a section of speech and subsequently adapts the wrong model.

Speaker Adaptation

Speaker adaptation avoids the need for large amounts of training data from a single speaker, by allowing a set of speaker independent (SI) models to be adapted to the new user. The initial set of SI models may be trained using data from a database containing many hours of speech and model the general characteristics of each speech unit. Speech from individual talkers is then used to adapt the models such that they more closely fit the individual acoustic properties of the talker.

A variety of methods have been proposed to perform speaker adaptation:

- *Bayesian adaptation [9,23,40]:* This technique combines the output distribution parameters from a set of speaker independent models with new parameters generated from the new speaker's adaptation data. The combination takes the form of a weighted sum — as more data is collected from the talker, the weighting for the prior models is decreased until the models are equivalent to the speaker dependent case. The method is easily incorporated into the hidden Markov model recognition framework [40], but does have the disadvantage that adaptation of a given model may only be performed once an example of the sound represented by that model has been given. Also, the output distributions produced by the technique can only be single Gaussian mixtures which can result in the adapted models having lower performance than multiple mixture, speaker independent models.

- *Transformations of model parameters [5,15,27,35,44]:* Adaptation algorithms of this type seek to estimate a set of transformations (linear or non-linear) which may be used to transform the parameters of the original models to better match those of the adaptation data. One of the most popular schemes is MLLR (Maximum Likelihood Linear Regression), in which a linear affine transform, initially of the means of the models [44] and later of both the means and variances [27], is found which maximises the likelihood of the adaptation data. This method has several advantages. Firstly it fits rigorously within the hidden Markov model framework, secondly, a single transformation may be applied to all the models, hence an example of each phoneme is not required before adaptation can take place. As more adaptation data is acquired, specific transforms for different classes of phoneme are generated until a transform for each model is obtained. In [15] correlations between sounds are calculated and used to predict linear transforms for unseen phonemes from those for which adaptation data is available. The use of non-linear transforms [5] [35] have shown some improvements in recognition accuracy. The transform is usually implemented by a feed-forward multi layer perceptron. The topology of such systems is arbitrarily chosen and as such often has a large number of parameters to estimate, requiring large amounts of adaptation data and as such, linear techniques are currently more popular.

- *VQ prototype modification:* An alternative to continuous density HMMs is to vector quantize the parameterised input speech and then perform recognition using discrete HMMs trained on the quantised values. The codebooks used for the vector quantisation may be modified to better match the adaptation data from a new speaker using either Bayesian techniques [71] or by estimating a transform which maps the SI codebook to the new speaker's input vectors [26]. Such models have been largely abandoned however due to the increased use of continuous density HMMs instead of VQ and discrete ones.

- *Speaker Clustering [26, 48, 59, 77]:* Speaker clustering reduces the variances in the model set by identifying speakers within the training set who have similar speech characteristics. A separate set of models, which will have smaller variances than truly speaker independent models, is then generated for each of the speaker groups. During recognition, the goal is to associate the test speaker with one of the clusters, and use the models for that cluster to recognise their speech, thus improving recognition performance. Several clustering procedures have shown to give increases in recognition performance. In [48] speakers are clustered depending on parameters relating to their vocal tract dimensions, while in [77] the dialect of the talker is identified and used to cluster similar speakers.

Speaker Normalisation

Speaker normalisation describes techniques in which the input vectors from a speaker are adapted in some way so as to reduce the variation between speakers [11, 13, 19, 42, 74,79,80,87]. The normalisation may be applied to speakers in the training set, resulting in acoustic models with smaller variances and less overlap [42]. This is not due to the model parameters being explicitly changed as in a speaker adaptation scheme, it is simply a result of all the training speakers appearing acoustically more similar. If the same normalisation procedure is then applied to the test speakers, their data will better match the models, leading to a decrease in recognition errors. Rose and Lee [43] perform the adaptation by moving the positions of the filter bank channels in an MFCC front end so as to expand or compress the frequency spectrum of the signal. They present two methods of determining the required normalisation factor for each speaker — the first requires performing a probabilistic alignment of the utterance, parameterised at each warp factor, to a transcription . The second requires decoding the utterance using several Gaussian mixture models — one for each normalisation factor. Both these procedures are computationally too expensive to be used in a real time recognition system. Burnett and Fanty [11] use Brents algorithm to find the optimum normalisation factor (i.e. that which minimizes the output log likelihood of the recogniser), however this still typically takes 8-10 passes over the adaptation data. Eide and Gish select the correct warp factor based on the ratio of the speakers median third formant frequency to the median third formant frequency of all the speakers. Zhan and Westphal [87] extend this by comparing the use of the median of the first, second and third formant frequencies for determining the normalisation factor. All these methods estimate the median formant position over a large number of frames representing several different phonemes. Since many of the phonemes have different values for the formant frequencies, averaging over several will result in a poor estimate of the actual formant value. In Chapter 5 we present a new method of normalisation which address many of the criticisms of the current methods. The normalisation factor is determined without evaluating recognition performance over a range of normalisation factors and is therefore computationally more efficient than the methods of Lee and Rose. The position of the first and second formants is used in calculating the normalisation factor — the test speakers formants are matched to distributions of the formants of all speakers. The distributions are estimated individually for each phoneme, removing the problem of averaging formant positions over different sounds.

Chapter 5

Speaker Normalisation

Differences in vocal tract length account for much of the variation in the realisation of a target sound between talkers. Automatic Speech Recognisers trained with speech data from a large number of different talkers must model this variability, leading to models with higher variances and significant overlap. This in turn leads to poor recognition performance. Similarly, mismatches between training and test set speakers leads to increased confusion at the recognition stage. If these physiological differences between talkers may be identified, either explicitly through estimation of the vocal tract length [61, 83] or implicitly through estimation of some parameter related to vocal tract length (such as formant positions) then it may be possible to use this information to remove some inter speaker variability by simple signal processing techniques. In this chapter a rapid, unsupervised method of speaker normalisation is developed and used to reduce the mismatch between speakers, thereby improving recognition accuracy

5.1 Preliminary Investigation - Spectral Matching

A preliminary investigation was conducted to determine whether the variation between speakers may be reduced by mapping the frequency spectrum of the test speaker's utterance to that of a canonical speaker using simple linear transformations of the frequency spectrum of the speech. If the normalisation were sufficiently powerful, a speaker dependent system trained on the utterances of a 'canonical' speaker would then, give similar performance for all the normalised utterances.

5.1.1 Data

The speech data used for the investigation consisted of the central 25.6ms frame of 5 vocoids, (/ae/ /er/ /iy/ /oy/ /uw/) from 49 speakers (18 female and 31 male) selected from dialect region 1 of the TIMIT database (Section A.2.1). As shown in Section 3.2.3, linear prediction coefficients may be used to generate a close approximation to the frequency response of the vocal tract. A power spectrum derived from the LPC coefficients is preferred to direct estimation of the spectrum because the inherent smoothness of the LPC-derived spectrum results in more clearly defined peaks at the formant frequencies, as shown in Figure 5.1. 20th order linear prediction was used to model each of the vowel segments and 800 point normalised frequency response plots were generated for each utterance.

5.1.2 Selecting the Canonical Speaker

In [76] it was shown that there are significant differences between vocal tract frequency responses of male and female speakers. Females were generally found to have higher formant and fundamental frequencies because of their shorter vocal tracts. It was therefore decided to perform the normalisation in a gender dependent manner, using two different canonical speakers, one male and one female so that the normalisation was not simply removing the gross differences between speakers related to their gender.

A distance between a given speaker, i and the other speakers was defined to be:

$$\begin{aligned} E(i) &= \sum_{m=1}^{S}\sum_{l=1}^{V}\sum_{n=1}^{F}(x_l^n(i) - x_l^n(m))^2, \\ & i = 1, 2, 3, \cdots S \\ & i \neq m \end{aligned} \tag{5.1}$$

where V is the number of vowel sounds being compared (in this case $V = 5$), S is the number of speakers (18 for the female talkers and 31 for the males) and F is the upper frequency limit of the spectrum (8000). $x_l^n(m)$ is the amplitude of the spectrum for speaker m, vowel l at frequency n. The canonical speaker was chosen as the speaker with the minimum $E(i)$, that is, the speaker with the smallest squared differ-

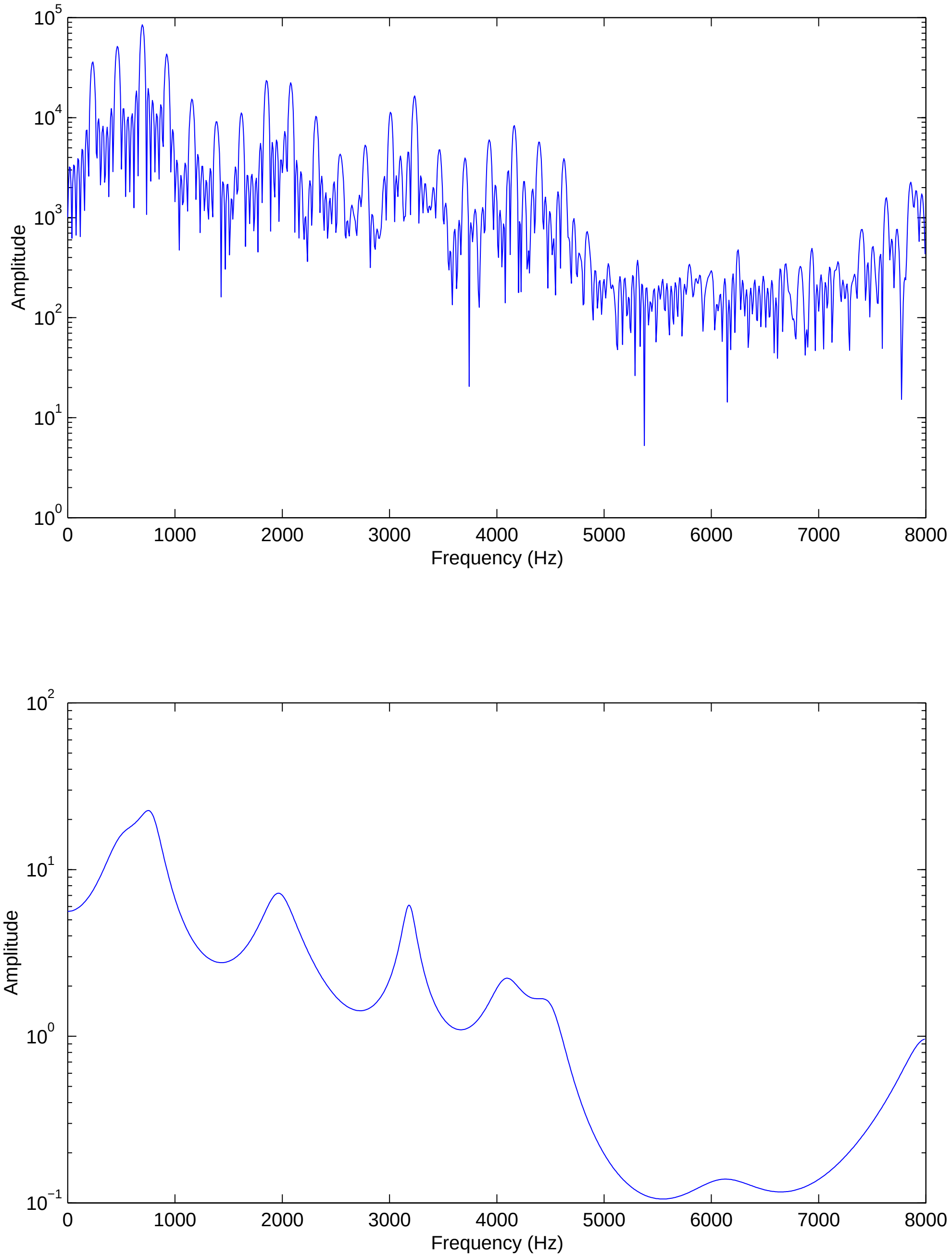


Figure 5.1: Frequency spectra for a single 25.6ms utterance of the vocoid /ae/. Top: FFT derived spectrum. Bottom: LPC derived spectrum.

ence between the frequency spectrum of their vowels and those of all the other talkers. It should be noted that this distance is calculated using only the single central frame of speech for each vowel.

5.1.3 Normalisation Of Frequency Spectra

Two different methods of transforming the frequency spectra in order to normalise them with those of the canonical speaker were investigated.

Frequency Offset

A frequency offset such that $X'(f) = X(f + s)$ where $X(f)$ is the amplitude of the frequency response of utterance X at frequency f was implemented. This shifts the spectrum up or down the frequency scale depending on the value of s. The effect of the shift is to move the absolute and relative positions of the formants while leaving their bandwidth unchanged. Values of s from -200 to +200 Hz in increments of 10 Hz were used and the squared error between the offset signal and the canonical speaker's utterance calculated. Figure 5.2 shows, for each of the five vowels of a typical speaker, the change in error as s is varied. The optimum offset for each utterance is defined as that which minimises the squared error. It is encouraging to note from Figure 5.2 that the optimum offset is negative for all the vowels and approximately the same value (between 65Hz and 110 Hz) for four of these. This implies that in all cases the spectrum has to be shifted down the frequency axis to match that of the canonical talker i.e. the speaker has a shorter vocal tract (leading to higher frequency resonances) than that of the canonical talker. The result of applying the optimum offset to the frequency response of the speaker's utterance is shown in Figure 5.3. The shifted spectrum, together with the original unshifted response and the canonical response for the vowel /er/ are shown. The effect is to align the first formants of both speakers. Plotting the optimally offset responses of a single vowel for all talkers, Figure 5.4, shows that the major effect of the shift is to align the first formants. This is due to the fact that the first formant is the highest energy feature of the response and as such aligning them will result in the smallest squared error between the test speakers and the canonical speaker.

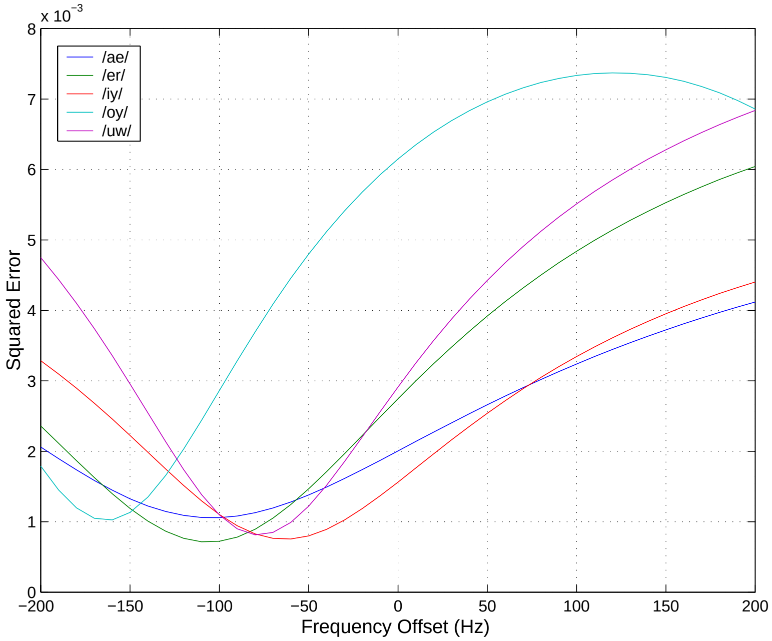


Figure 5.2: Error verses frequency offset for the 5 vowels of a single talker. Each curve represents a different vowel.

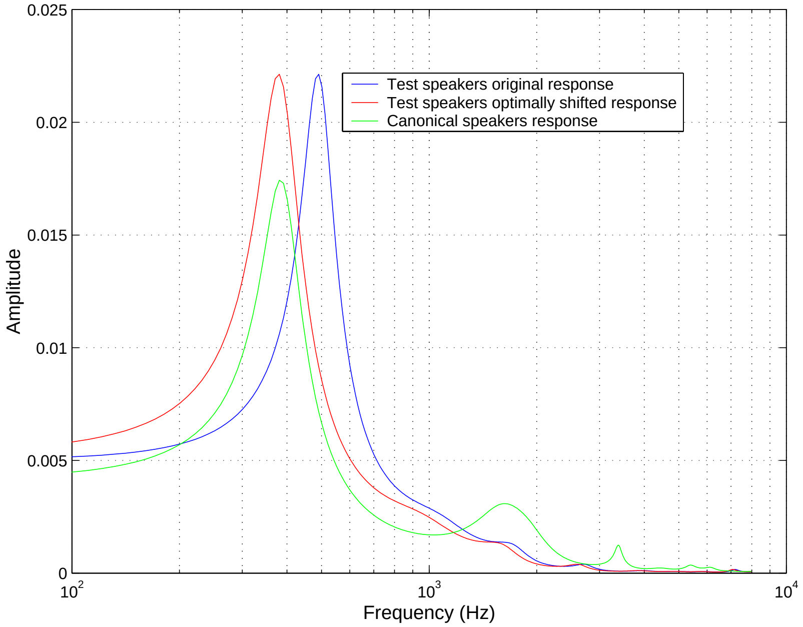
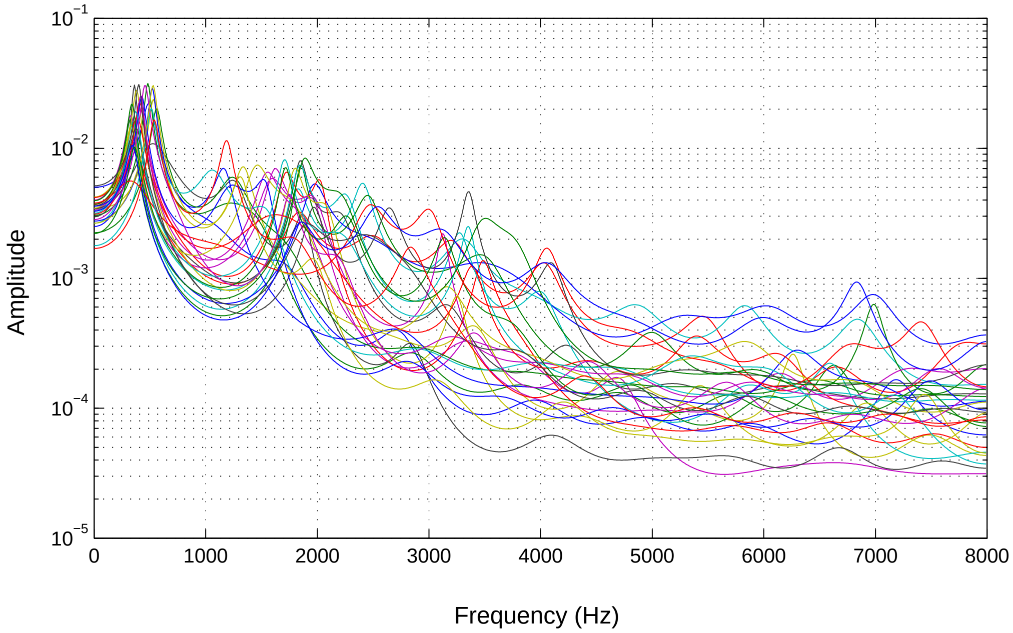


Figure 5.3: Effect of applying the optimal offset to the frequency spectrum of a single utterance.



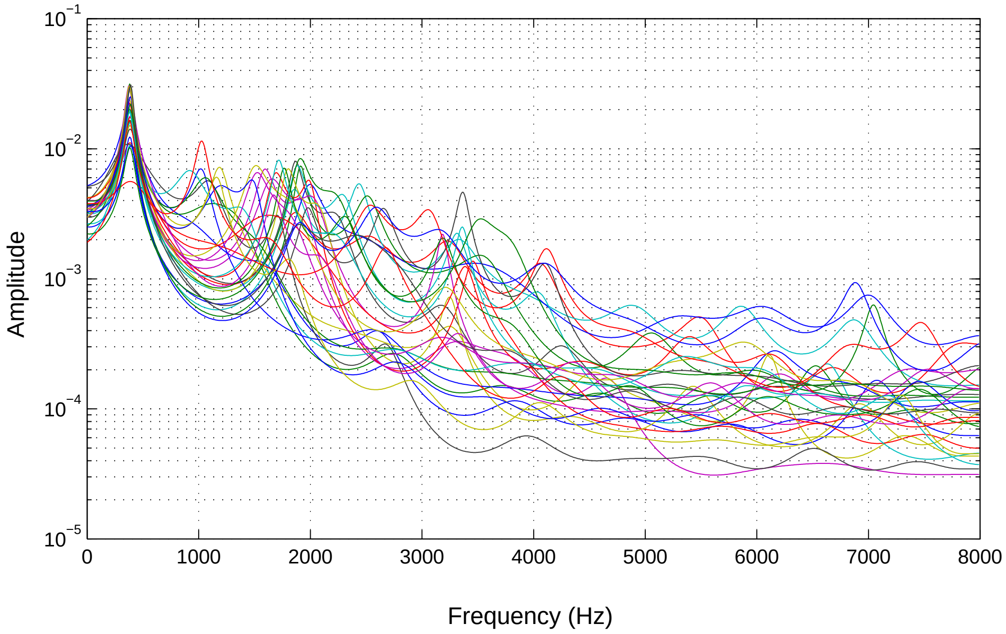


Figure 5.4: Frequency spectra of vowel /er/ for all male speakers, unshifted. Bottom: Frequency spectra of vowel /er/ for all male speakers, optimally offset.

Frequency Scaling

A scaling of the frequency axis such that $X'(f) = X(fm)$ was implemented. This expands or compresses the response about the 0 Hz point, altering the formant band widths and absolute positions, while not effecting their relative positions. In terms of the ‘uniform tube’ model of the vocal tract as described in Section 2.2 this is equivalent to scaling the length of the tube. If the tube is scaled by a value a, the frequency of the tube’s resonances are scaled by a factor $\frac{1}{a}$. Values of m from 0.25 to 1.75 in increments of 0.05 were investigated. Figure 5.5 shows the resulting variation in error for each vowel of a single speaker as the scaling factor is varied. Figure 5.6 shows the effect of the optimum scaling on the response of a single vowel, together with the canonical speaker’s response for that vowel. Again the overall effect of the shift for all speakers is to align the first formants and thereby reduce the squared error, as shown in Figure 5.7.

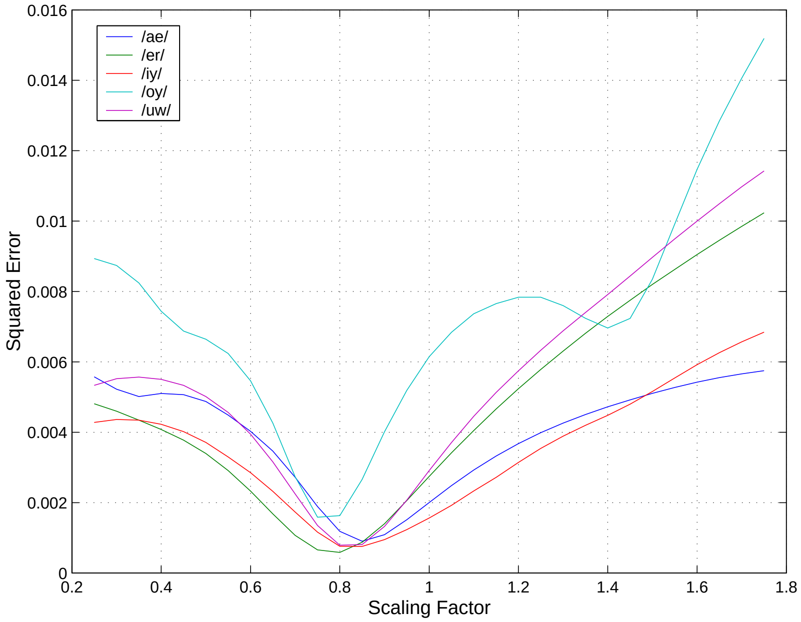


Figure 5.5: Error versus frequency scaling factor for the five vowels of a single talker. Each line represents a different vowel.

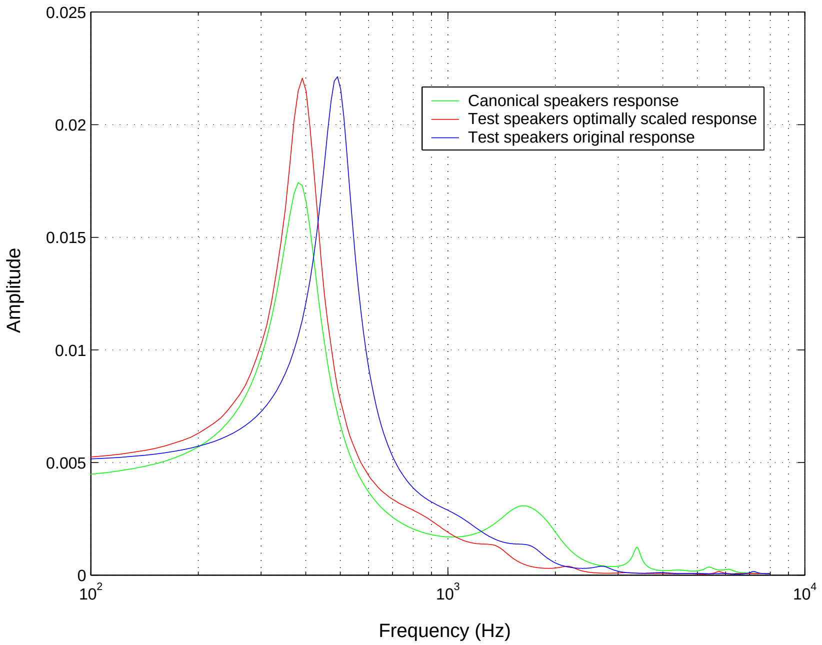


Figure 5.6: Effect of applying the optimal scaling the frequency spectrum of a single utterance.

5.1.4 Classification Experiment

To investigate the usefulness of the transformations in reducing inter-speaker variability, a simple classification experiment was conducted. The 'city block' distance, D_{cb}, between utterances from each speaker and the canonical speaker was calculated, where:

$$D_{cb} = \sum_{f=1}^{N} \mid C(f) - T(f) \mid \tag{5.2}$$

$C(f)$ is the amplitude of the canonical speaker's utterance at frequency f, $T(f)$ is the amplitude of the test speaker's utterance at frequency f and N is the upper frequency limit of the response. The test utterance was then assigned to the vowel class which gave the smallest D_{cb}.

$$Class = \arg\min_{v=1}^{5} \sum_{f=1}^{N} \mid C_v(f) - T(f) \mid \tag{5.3}$$

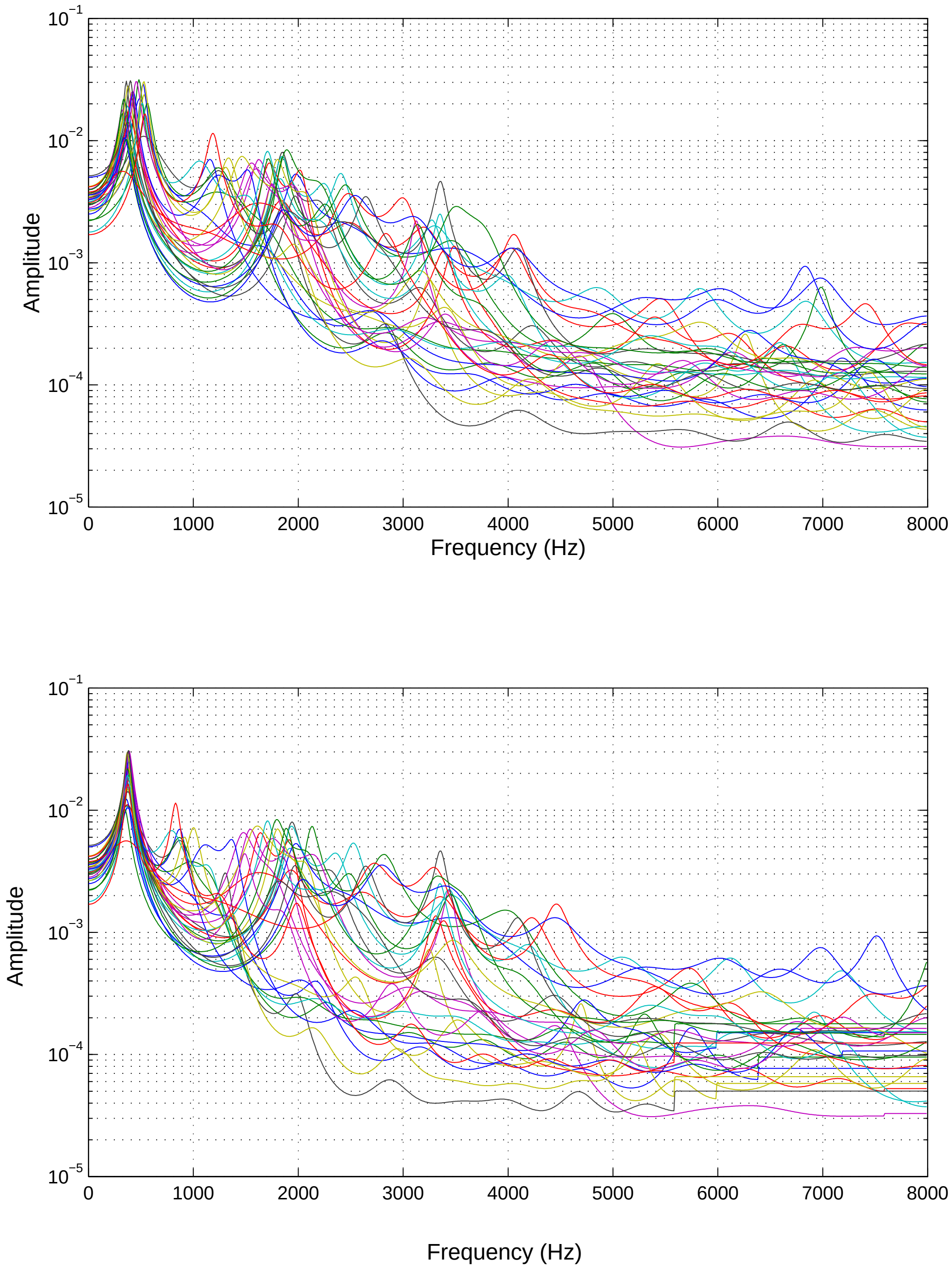


Figure 5.7: Top: Frequency spectra of vowel /er/ for all male speakers, unscaled. Bottom: Frequency spectra of vowel /er/ for all male speakers, optimally scaled.

where v is the vowel class and $c_v(f)$ is the canonical response for vowel v

Table 5.1 shows the percentage of correctly classified vowels for the original, optimally offset and optimally scaled spectra. The alignment of the first formant by the translations has provided a significant increase in classification accuracy implying that simple linear transforms in the frequency domain can increase discrimination between certain vowel classes. The frequency offset shows better improvements than the frequency scaling, even though the later is more consistent with the acoustic theory related to changes in vocal tract length. An explanation for this is that the offset was evaluated at increments of 10 Hz and as such the first formants would be aligned to within 10 Hz of each other. The scaling was evaluated at increments of 0.05, and since the first formant is at approximately 300Hz, they are only aligned to within 15Hz of each other $(300\text{Hz} \times 0.05 = 15\text{Hz})$. Hence improvements using the scaling are slightly less than those obtained using the offset.

The method of exhaustively searching for the optimal transformation is, however, computationally highly expensive. Transforming the frequency response by changing the LPC's directly, rather than in the frequency domain would reduce the required number of new parameters which need to be calculated from 800 to just 20, reducing the computational burden. The technique must also be shown to work over all speech sounds rather than just a small subset of the vowels.

Spectra Type	/ae/	/er/	/iy/	/oy/	/uw/	Average
Female Original	70.6	41.2	58.8	94.1	70.6	67.1
Female Optimal Offset	82.3	70.6	76.5	88.2	94.1	82.3
Female Optimal Scaling	70.6	58.8	100	88.2	100	81.2
Male Original	86.7	23.3	53.3	73.3	63.3	60
Male Optimal Offset	86.7	53.3	53.3	93.3	73.3	72
Male Optimal Scaling	86.7	40	53.3	96.7	70	69.3

Table 5.1: Results of classification experiments for unnormalised, optimally offset and optimally scaled frequency spectra (% Correctly Classified).

5.2 Normalisation by LPC Pole Matching

The previous transform showed that aligning test speakers' first formants to those of a canonical speaker could provide an increase in recognition accuracy for a simple vowel classification task. In Section 3.2.3 it was shown that the transfer function described by the linear prediction coefficients may be interpreted as the vocal tract filter of the source-filter model. In [74] it was shown that the apparent identity of a speaker may be modified by directly altering the LPC pole positions to those of a canonical speaker, however the results were not used to improve recognition accuracy and were only evaluated by human listening tests. Here we attempt to extract the value of the first formant directly from the vocal tract transfer function then normalise the test speaker directly to the canonical speaker in the LPC domain. This removes the need to evaluate the transfer function explicitly and should increase the computational efficiency of the normalisation.

5.2.1 Data

The preliminary experiment was conducted on frequency spectra derived from the LPC coefficients of a single frame from each vowel of every speaker. There was no assurance that these spectra were representative of the vowel sound in general, rather than just the short segment observed. The roots of the 20 LPCs for 5 contiguous 25.6ms frames with 10 ms overlap, taken from the center of the sound were observed for several speakers. Figure 5.8 shows a plot of the roots of the LPCs in the z-plane over 5 frames for a typical vowel utterance. The clustering of poles near the unit circle indicates that the LPCs vary little from frame to frame and it was concluded that the LPCs were providing a stable representation of the sounds. For the subsequent experiment, the values of the LPCs were averaged across the 5 frames. The data was also extended to included examples of all seventeen vocoid sounds in the TIMIT transcriptions rather than just five.

5.2.2 The Transform

The original transform relied on exhaustively searching for the optimum offset or scaling which minimised the squared error between the frequency response of the test speaker and a reference speaker. This was seen to effectively align the first formants. The LPC representation of the speech signal provides a method of directly evaluating the formant frequencies which can then be matched to those of the reference speaker.

As shown in Section 3.2.3, linear prediction approximates the vocal tract response as an all pole filter

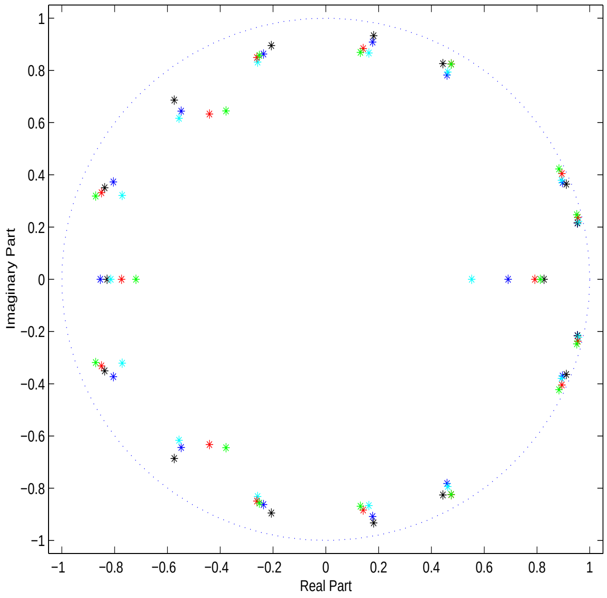


Figure 5.8: Variation of LPC pole placement over 5 contiguous frames from a typical utterance of the vowel /ao/. Clustering of poles near unit circle indicates that the LPCs are providing a stable representation of the sound.

$$H(z) = \frac{1}{\sum_{i=1}^{m} a_i z^{-i}} \tag{5.4}$$

$$= \frac{1}{A(z)} \tag{5.5}$$

Where a_i are the predictor coefficients and m is the analysis order. Estimates of the resonances of the vocal tract (the formants) are given by the roots of the predictor polynomial, $A(z)$. For each root, r_i, the frequencies and bandwidths are given by:

$$F_i = \frac{\theta_i f_s}{2\pi} \tag{5.6}$$

$$B_i = \frac{-\ln |r_i| f_s}{\pi} \tag{5.7}$$

where f_s is the sampling frequency and θ_i and $|r_i|$ are the angle and magnitude of r_i respectively.

The roots representing the formants typically have very small bandwidths (by observation, typically $|r_i| > 0.9$) and low frequencies. The first and second formants for each utterance were therefore located by sorting the roots into order of ascending frequency, and extracting the two lowest frequency roots with $|r_i| > 0.9$. Figure 5.9 shows a plot of the roots of the polynomial, with the formant bandwidth decision threshold. Four pairs of poles lie outside the threshold and the formant frequencies associated with these are shown on the Fourier transform of the signal in Figure 5.10, along with the lpc derived spectrum.

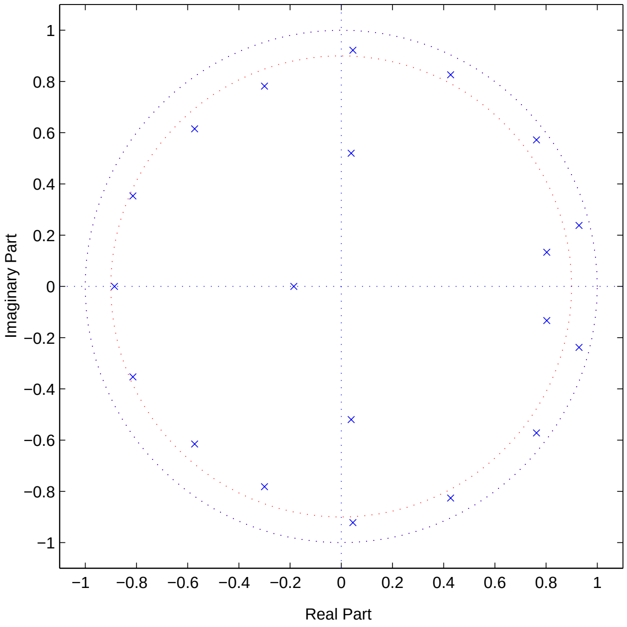


Figure 5.9: LPC pole placement in the z-plane for the vowel /oy/ showing the bandwidth threshold (red) used for identifying the speech formants

The accuracy of the formant finding algorithm is sometimes compromised, particularly for high pitched speech where the first and second formants merge to a single peak, and for the third and fourth formants where the bandwidths are often lower than the threshold. In the majority of cases however it was reliably able to identify the first and second formants and is also computationally far cheaper than other methods such as those presented in [30] and [81]

Having located the formants from the LPC coefficients, they may be directly transformed to more closely match those of the canonical speaker. This is computation-

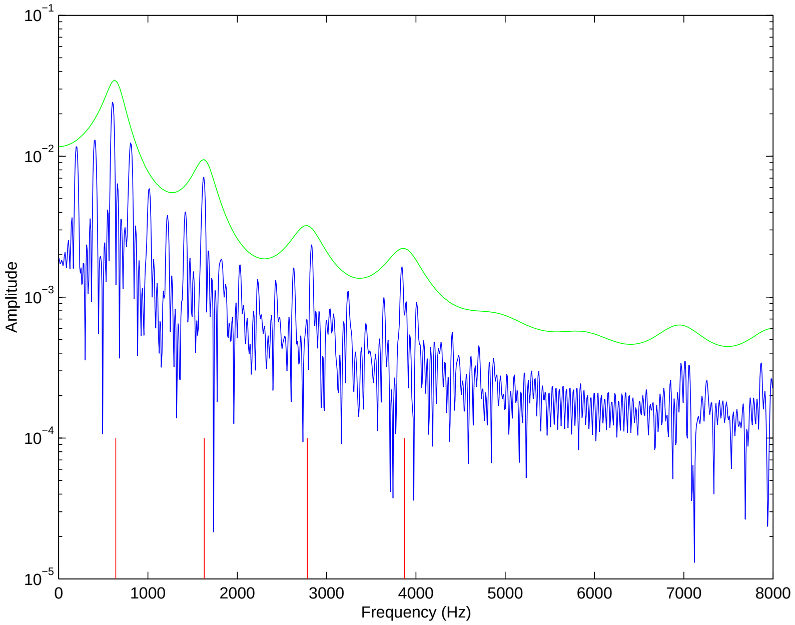


Figure 5.10: FFT and LPC derived frequency spectra for the vowel /oy/ showing the candidate formant locations estimated from the LPC roots

ally more efficient than deriving the frequency response and performing an exhaustive search over various scaling factors.

The new transform is defined such that if θ_n^1 and θ_n^2 are the angles of the new speaker's first and second formants for a given vowel, and θ_r^1 and θ_r^2 are those of the reference speaker for the same vowel, then the transformed angles θ_t^1 and θ_t^2 are given by

$$\theta_t^1 = a\theta_n^1 \tag{5.8}$$

$$\theta_t^2 = a\theta_n^2 \tag{5.9}$$

where a is found so that

$$(\theta_r^1 - \theta_t^1)^2 + (\theta_r^2 - \theta_t^2)^2 \tag{5.10}$$

is minimised. Setting

$$Y = (\theta_r^1 - a\theta_n^1)^2 + (\theta_r^2 - a\theta_n^2)^2 \tag{5.11}$$

and differentiating with respect to a gives

$$\frac{\partial Y}{\partial a} = 2(a((\theta_n^1)^2 + (\theta_n^2)^2) - (\theta_r^1\theta_n^1 + \theta_r^2\theta_n^2)). \tag{5.12}$$

For $\frac{\partial Y}{\partial a} = 0$

$$a = \frac{\theta_r^1\theta_n^1 + \theta_r^2\theta_n^2}{(\theta_n^1)^2 + (\theta_n^2)^2} \tag{5.13}$$

Using similar notation for the bandwidths leads to:

$$|B_t^1| \quad = \quad b|B_n^1| \tag{5.14}$$

$$|B_t^2| \quad = \quad b|B_n^2| \tag{5.15}$$

where

$$b = \frac{|B_r^1||B_n^1| + |B_r^2||B_n^2|}{(|B_r^1|)^2 + (|B_r^2|)^2} \tag{5.16}$$

It should be noted that this is no longer a simple linear transformation in the frequency domain. Instead we are finding a multiplicative factor which minimises the distance between the poles of the canonical and test speakers in the z-plane, thereby normalising both the frequency and bandwidth of the formants.

5.2.3 Results

Figure 5.11 shows the results of applying the shift to a single utterance. On the top row, the unnormalised roots of the test and canonical speaker's LPCs are shown in the z-plane, alongside the LPC derived spectra. On the bottom row the normalised roots and spectra are shown. The roots associated with the first and second formants have been aligned in the z-plane, matching the test and canonical speakers' formants.

The effect of applying the transform to the spectra of several speakers for the vowel /ih/ is shown on the bottom of Figure 5.12; the unnormalised spectra are shown on the top. The variance around the first and second formants has been significantly reduced by the transform, as would be expected. For each speaker's utterance of the vowel /ih/,

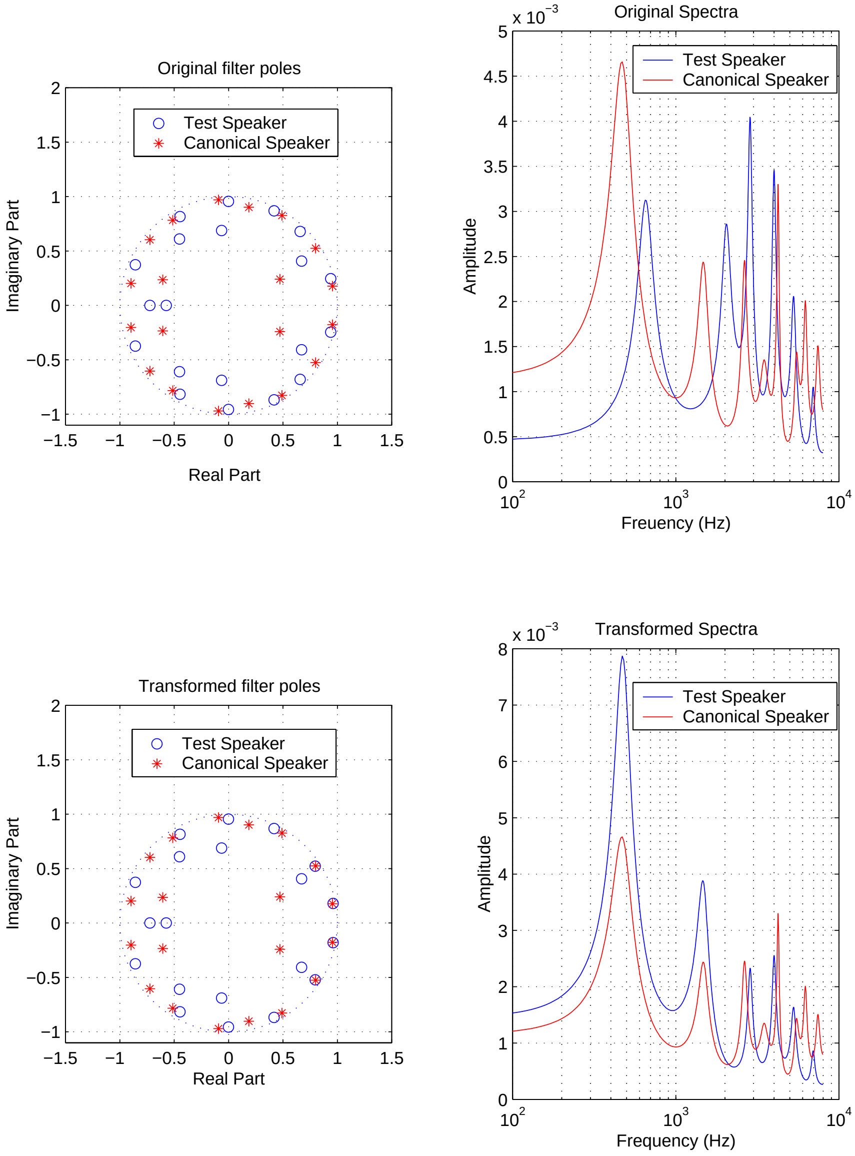


Figure 5.11: Effect of LPC transform on single utterance. Top: Unnormalised spectra and filter pole positions. Bottom: Normalised spectra and filter pole positions

Figure 5.13 shows the position of the first and second formants when plotted against each other. The effect of the normalisation is to cluster talkers with the same ratio of first to second formant frequency to a single point within the f1-f2 plane. Before the normalisation speakers are randomly scattered in the formant plane; after the shift, the speakers are ordered along an ellipse, the axes of which are given by:

$$y = \frac{1}{2}\sqrt{(f_1)^2 + (f_2)^2} + \frac{f_1}{2} \tag{5.17}$$

$$x = \frac{1}{2}\sqrt{(f_1)^2 + (f_2)^2} + \frac{f_2}{2} \tag{5.18}$$

where f_1 and f_2 are the frequency of the reference speaker's formants, as is shown in appendix B. The ratio of the speaker's first and second formants is therefore preserved by the transform, while their absolute positions is normalised toward that of the reference speaker. Formant ratio theory [51] states that the ratio of the lower formants is of significantly more importance than their absolute positions in defining the perceived identity of vowel sounds. This is demonstrated by Figure 5.14 which shows the distribution of formant ratios for the vowels /ih/ and /aa/ after normalisation. The two vowels show distinct peaks at different points along the formant ratio axis indicating that the F1-F2 formant ratio is a good discriminator of these sounds. This has been shown to be true for many other of the vowel sounds [76]. The ability to reduce inter-speaker variance, while retaining the ratio of the lower formants is of significant importance since this will preserve the discriminative information between the sounds, while reducing the within class variance.

5.2.4 Recognition Test

To investigate the usefulness of the transform, a simple classification experiment was performed. 12 MFCC's were generated from both the original and normalised utterances and a multivariate normal classifier was used to assign the utterance to the vowel class with the highest probability. Assuming the features to be uncorrelated (i.e. covariance matrix is diagonal), the probability of x being in class c is given by

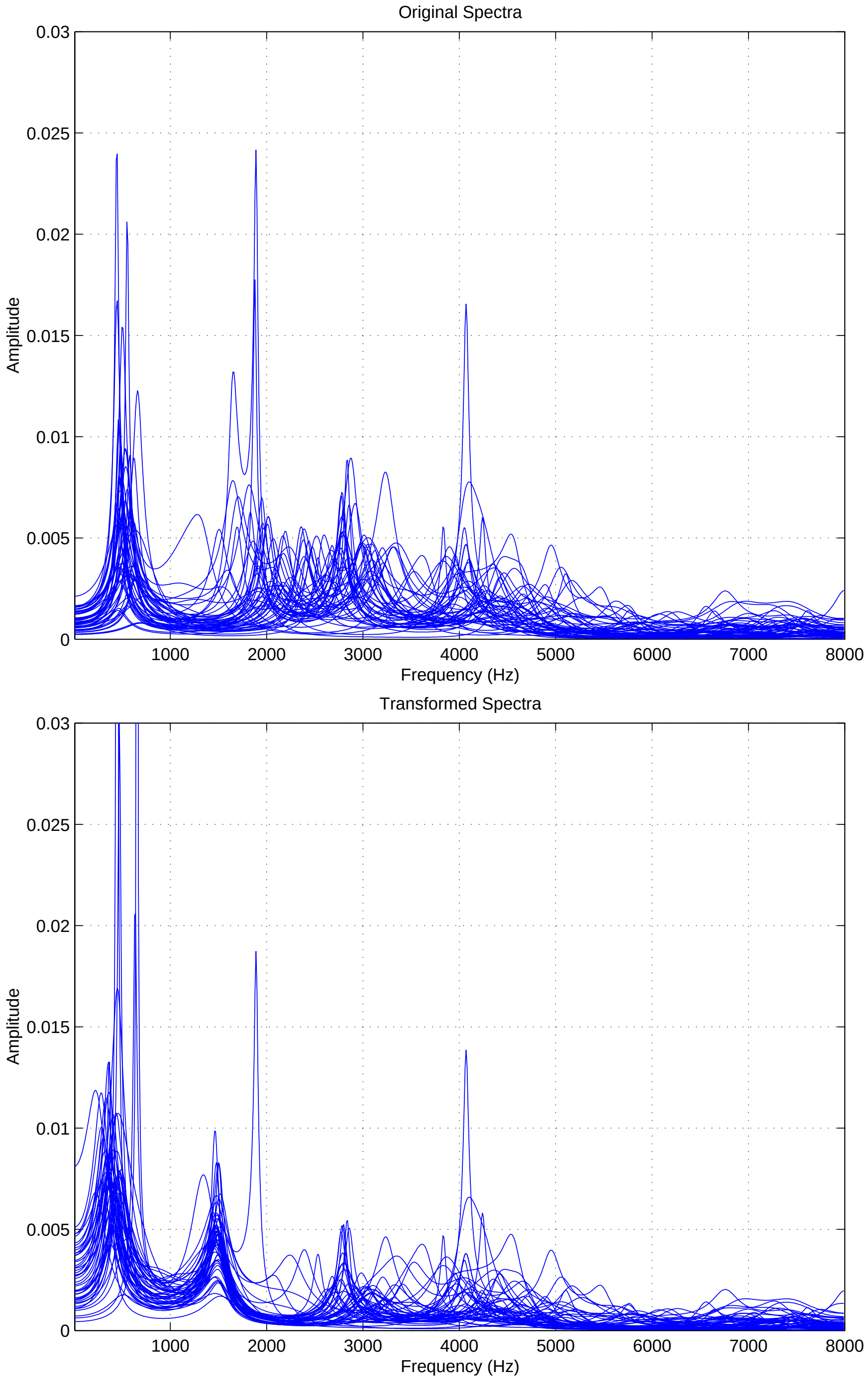


Figure 5.12: Effect of LPC transform on the spectra of the vowel /ih/ from 50 speakers. Top: Unnormalised spectra. Bottom: Normalised spectra.

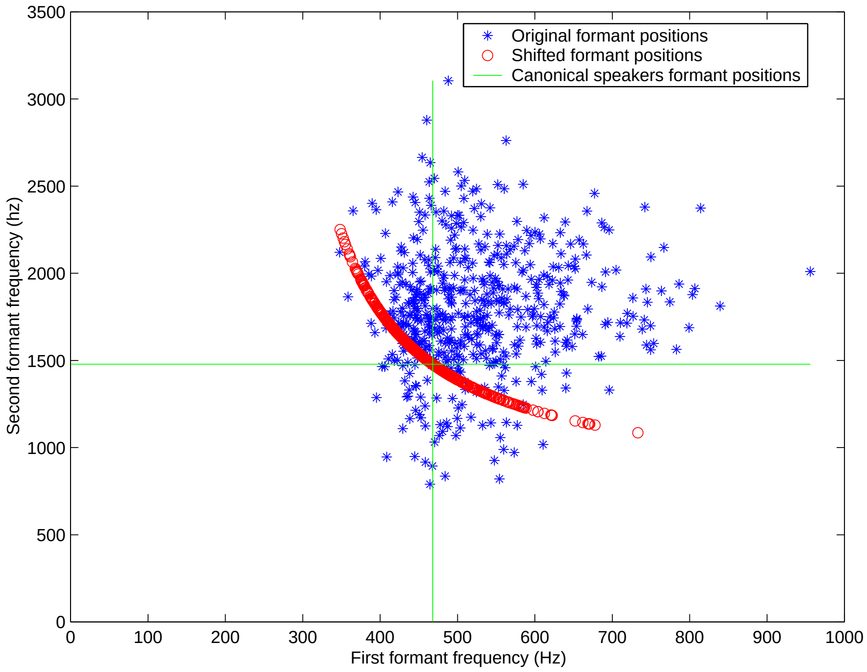


Figure 5.13: Position of speakers formants' in the F1-F2 plane for the vowel /ih/. Speakers with the same formant ratio cluster to a single point on an ellipse after the normalisation.

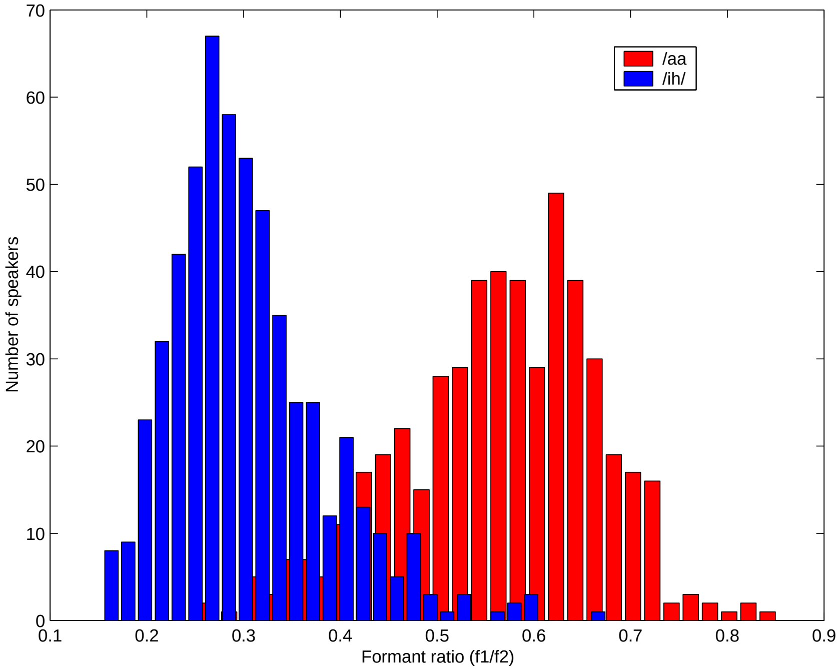


Figure 5.14: Histogram of number of speakers with a given first/second formant ratio for two vowels, showing the discriminative information available from the formant ratio.

$$P(x) = \frac{1}{(2\pi)^{d/2}|\sigma^c|^{1/2}} \exp\left(\left[-\frac{1}{2}\sum_{k=1}^{d}\frac{(x_k-\mu_k^c)^2}{(\sigma_k^c)^2}\right]\right) \tag{5.19}$$

where σ^c is the diagonal of the covariance matrix for class c, μ^c is the vector of means for each feature in class c, d is the dimensionality of the data (in this case 12) and x is the test utterance.

The results of the recognition experiment are given in Table 5.2 and show that the transform provides a significant improvement in classification accuracy.

Phoneme	Recognition rate (%) before normalisation	Recognition rate (%) after normalisation
iy	58.3	51.5
ih	12.3	43.9
eh	12.4	25.9
ey	46.2	54.6
ae	41.9	55.6
aa	18.6	54.5
aw	44.0	59.7
ay	40.0	60.6
ah	5.4	44.4
ao	25.5	63.8
oy	65.2	72.8
ow	27.4	57.2
uh	4.0	72.4
uw	15.8	32.6
er	62.4	61.2
ax	21.3	45.5
ix	20.5	56.1
Average	30.7	53.7

Table 5.2: Results of multivariate normal classification experiments (% phone recognition accuracy) before and after normalisation by matching test speakers' vowels to those of a canonical speaker by LPC pole matching.

5.2.5 HMM Based Recognition

Given the observed improvement in accuracy using a simple classifier when normalisation was used, an HMM monophone recogniser for the entire TIMIT database using normalised and unnormalised data was constructed. Separate male and female models were produced. The data comprised of 3260 training and 1120 test sentences for the male model, and 1360 training and 560 test sentences for the female model. Figure 5.15 shows the method used to generate the MFCC's from the raw speech data. A window duration of 25.6 mS with a frame period of 10 ms was used.

The models were created using HTK [86]. A 3 state, single Gaussian mixture, left right with no skips, diagonal covariance matrix topology was used. The models generated with the unnormalised training data were tested with unnormalised test data to provide a baseline performance measure. The normalized models were then tested with normalised test data , giving an upper bound on the performance increase available using the normalisation technique. The results of the recognition tests, given in Table 5.3, show significant improvements in the percentage of correctly identified vowel segments for both the male and female cases.

5.2.6 Conclusions

Direct transformation of a speaker's utterance in the LPC domain has proved successful in reducing inter-speaker variability and increasing recognition performance. However the technique still relies upon calculating a separate normalisation factor for every frame of the utterance. While this provides the maximum improvement in accuracy, it is still computationally to expensive to implement in a current real-time recogniser architecture. The transform is also highly dependent upon the selected canonical speaker and normalisation is only performed on the vowel sounds since the formant estimator is only able to provide candidate formant frequencies for vocoid sounds. The normalisation still requires supervision, since the segment labels are required to determine which of the canonical speaker's vowels to normalise to. A new transform was implemented in order to address these issues.

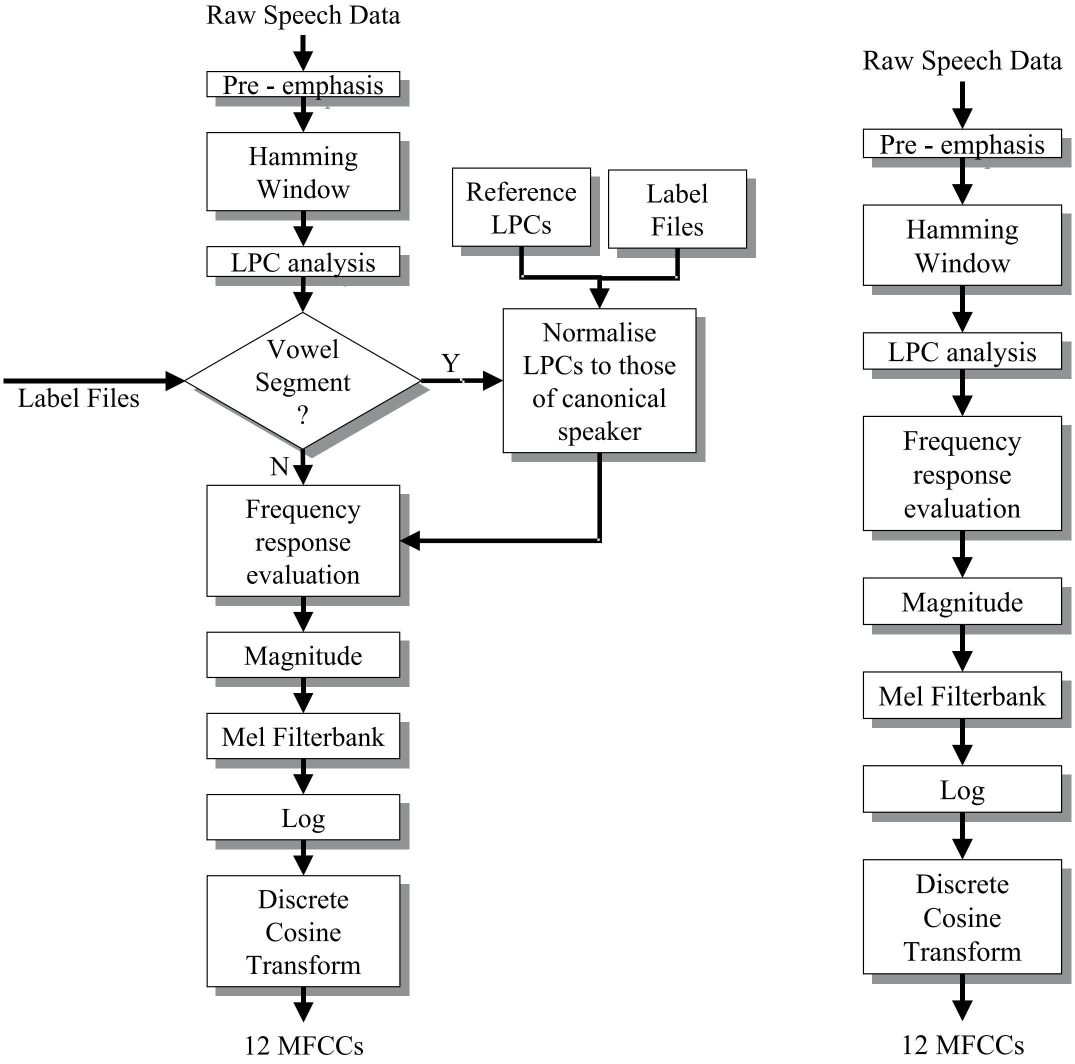


Figure 5.15: Parameterisation of data for HMM recogniser (Unnormalised (right) and normalised (left) cases.)

	Male		Female	
Phoneme	Recognition Rate before normalisation	Recognition Rate after normalisation	Recognition Error before normalisation	Recognition Rate after normalisation
iy	49.8	90.3	52.7	57.8
ih	24.6	40.1	23.0	53.7
eh	34.5	43.4	41.3	57.3
ae	64.2	66.5	62.2	83.9
ax	29.3	54.2	27.6	44.0
ah	36.6	59.0	39.0	58.5
uw	46.9	72.3	45.5	66.5
uh	34.0	44.3	37.5	45.6
ao	48.6	88.8	39.0	64.5
aa	36.7	61.0	39.0	77.5
ey	56.8	60.7	59.9	64.5
ay	46.0	77.1	53.4	74.3
oy	39.1	87.5	34.1	87.2
aw	49.2	59.9	47.5	41.0
ow	40.0	60.0	27.9	60.7
er	61.3	83.3	55.8	75.3
Average	37.4	61.7	40.31	59.5

Table 5.3: Results of HMM classification experiments (% phoneme recognition accuracy) before and after normalisation by matching test speakers to those of a canonical talker by LPC pole matching.

5.3 Normalisation to Canonical Distribution

Other work ([19, 43, 80]) has shown that a single normalisation factor applied to all frames of a speaker's utterance can provide a useful reduction in error rate. A method of combining the normalisation factors for all the frames into a single normalisation for each speaker was therefore developed and tested. The new normalisation also discards the concept of a 'canonical speaker', replacing it with a statistical representation of all the speakers to which each test speaker is normalised.

5.3.1 The Transform

The previous transform was defined such that, if θ_n^1 and θ_n^2 are the angles of the poles representing the test speaker's first and second formants (related to the frequencies of the formants by $F_i = \frac{\theta_i f_s}{2\pi}$) and θ_r^1 and θ_r^2 the angles of the reference speaker's formants, then for frame i the transformed poles, $\theta_t^1(i)$ and $\theta_t^2(i)$ are given by :

$$\theta_t^1(i) = a(i)\theta_n^1(i) \tag{5.20}$$

$$\theta_t^2(i) = a(i)\theta_n^2(i) \tag{5.21}$$

where

$$a(i) = \frac{\theta_r^1\theta_n^1(i) + \theta_r^2\theta_n^2(i)}{(\theta_n^1(i))^2 + (\theta_n^2(i))^2} \tag{5.22}$$

To remove the need for a canonical talker, estimates (by the lpc root finding method discussed in Section 5.2.2) of the first and second formants for each frame of each vowel in the TIMIT training set are calculated. A uni-variant Gaussian distribution is then used to model each formant in each vowel class.

$$\theta_r^1 = \frac{1}{\sigma_1} \exp -\frac{1}{2}\frac{(x-\mu_1)^2}{(\sigma_1)^2} \tag{5.23}$$

$$\theta_r^2 = \frac{1}{\sigma_2} \exp -\frac{1}{2}\frac{(x-\mu_2)^2}{(\sigma_2)^2} \tag{5.24}$$

$a(i)$ is then found such that :

$$
\begin{aligned}
a(i) &= \arg\max_a L(i) &(5.25)\\
a(i) &= \arg\max_a L(a\theta_n^1(i), a\theta_n^1(i)) &(5.26)\\
a(i) &= \arg\max_a \Pr(a\theta_n^1(i) \mid \theta_r^1) \Pr(a\theta_n^2(i) \mid \theta_r^2) &(5.27)
\end{aligned}
$$

$$a(i) = \arg\max_a \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\frac{(a\theta_n^1(i) - \mu_1)^2}{{\sigma_1}^2}\right\} \exp\left\{-\frac{1}{2}\frac{(a\theta_n^2(i) - \mu_2)^2}{{\sigma_2}^2}\right\} \tag{5.28}$$

That is, $a(i)$ is chosen so as to maximise the likelihood of the transformed formants having come from the two formant distributions for that vowel class. Figure 5.16 demonstrates this — the two original vowels, shown by the lines f1 and f2 are scaled by a factor ’a’ such that they are closer to the means of distributions F1 and F2. This is an extension to the work of Eide and Gish [19] who calculate a normalisation factor from the ratio of the mean value of the third formant for the test speaker, to the mean value of the third formant for all speakers.

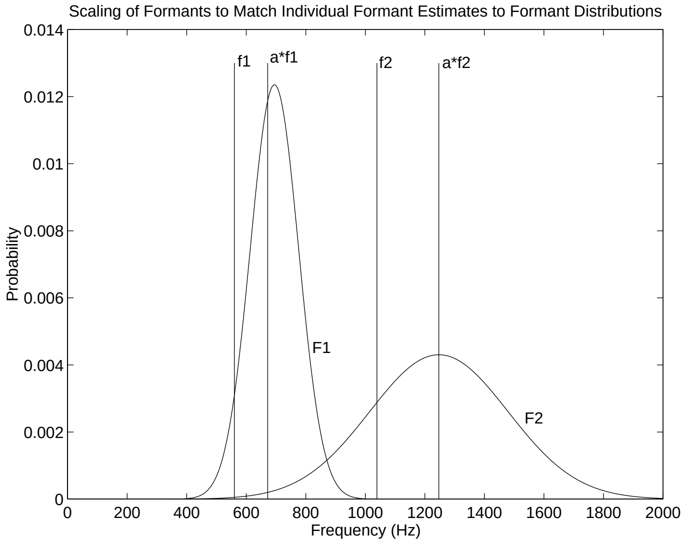


Figure 5.16: Normalisation of formant estimates to formant distributions.

A closed form solution to Equation 5.25 may be found by differentiating the logarithm of the likelihood equation:

$$L(i) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\frac{(a(i)\theta_n^1(i) - \mu_1)^2}{{\sigma_1}^2}\right\} \exp\left\{-\frac{1}{2}\frac{(a(i)\theta_n^2(i) - \mu_2)^2}{{\sigma_2}^2}\right\}. \quad (5.29)$$

$$\ln(L(i)) = \ln\left(\frac{1}{2\pi\sigma_1\sigma_2}\right) - \frac{1}{2}\left(\frac{\left(a(i)\theta_n^1(i) - \mu_1\right)^2}{{\sigma_1}^2} + \frac{\left(a(i)\theta_n^2(i) - \mu_2\right)^2}{{\sigma_2}^2}\right) \quad (5.30)$$

which may be expanded to give

$$\ln(L(i)) = \ln\left(\tfrac{1}{2\pi\sigma_1\sigma_2}\right) - \tfrac{1}{2}\left(\frac{(a(i)^2\theta_n^1(i)^2 - 2a(i)\theta_n^1(i)\mu_1 + {\mu_1}^2}{{\sigma_1}^2}\right.$$

$$\left. + \frac{(a(i)^2\theta_n^2(i)^2 - 2a(i)\theta_n^2(i)\mu_2 + {\mu_2}^2}{{\sigma_2}^2}\right). \quad (5.31)$$

Differentiating with respect to $a(i)$ results in

$$\frac{\partial(\ln(L(i))}{\partial(a(i))} = \frac{-a(i)\theta_n^1(i)^2 + \theta_n^1(i)\mu_1}{{\sigma_1}^2} + \frac{-a(i)\theta_n^2(i)^2 + \theta_n^2(i)\mu_1}{{\sigma_2}^2} \quad (5.32)$$

and setting this to zero and rearranging gives

$$a(i) = \frac{\dfrac{\theta_n^1(i)\mu_1}{(\sigma_1)^2} + \dfrac{\theta_n^2(i)\mu_2}{(\sigma_2)^2}}{\left(\dfrac{\theta_n^1(i)}{\sigma_1}\right)^2 + \left(\dfrac{\theta_n^2(i)}{\sigma_2}\right)^2} \quad (5.33)$$

Values of $a > 1.3$ or $a < 0.7$ are ignored since this has generally been observed to indicate a failure of the formant picking algorithm to locate the correct formants (e.g. matching the third formant to the second formant distribution).

The normalisation factors, $a(i)$, each have an associated likelihood, $L(i)$, where

$$L(i) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\frac{(a(i)\theta_n^1(i) - \mu_1)^2}{\sigma_1^2}\right\} \exp\left\{-\frac{1}{2}\frac{(a(i)\theta_n^2(i) - \mu_2)^2}{\sigma_2^2}\right\}. \quad (5.34)$$

These are then used to calculate a single normalisation factor, $N(I)$, for each speaker :

$$N(I) = \frac{\sum_{i=1}^{j} a_i^I L_i^I}{\sum_{k=1}^{j} L_k^I} \tag{5.35}$$

where j is the total number of vowel frames for speaker I and L_x^I is $L(x)$ for speaker I. This alleviates the need for a separate normalisation for each frame, greatly reducing the computational overhead of the technique.

Normalising the estimate of each $a(i)$ by the its associated likelihood has a further advantage. If the estimate of the formant picking algorithm for a particular frame is poor, even the optimum value of $a(i)$ will still result in the transformed formants having a poor match to the distributions. This being the case, their value of $L(i)$ will also be low and they will receive a relatively small weighting in the calculation of $N(I)$

5.3.2 Experimental Studies

A series of recognition experiments were conducted using the new transformation to compare it to the previous techniques.

Dialect Independent Normalisation

The first experiment was a direct continuation of the previous work on LPC normalisation (Section 5.2.2). Gender independent formant distributions for each of the vowels in the TIMIT data base were generated from the training set. These were then used in equations 5.25— 5.35 to find a normalisation factor, $N(I)$ for each speaker I. Normalisation was then performed on the LPCs for each of the vowel frames. The angle of the poles representing the first and second formants were multiplied by the normalisation factor, scaling their frequency by a factor $N(I)$. The normalised LPC's were then used to generate the MFCC's used as the input vectors for the HMM recognition experiment. Figure 5.17 shows the parameterisation procedure including the normalisation technique.

Separate male and female models were generated using the normalised data and HTK. Identical recogniser topologies and raw data sets to those mentioned in 5.2.5 were used. The models were tested using the normalised test data and the recognition

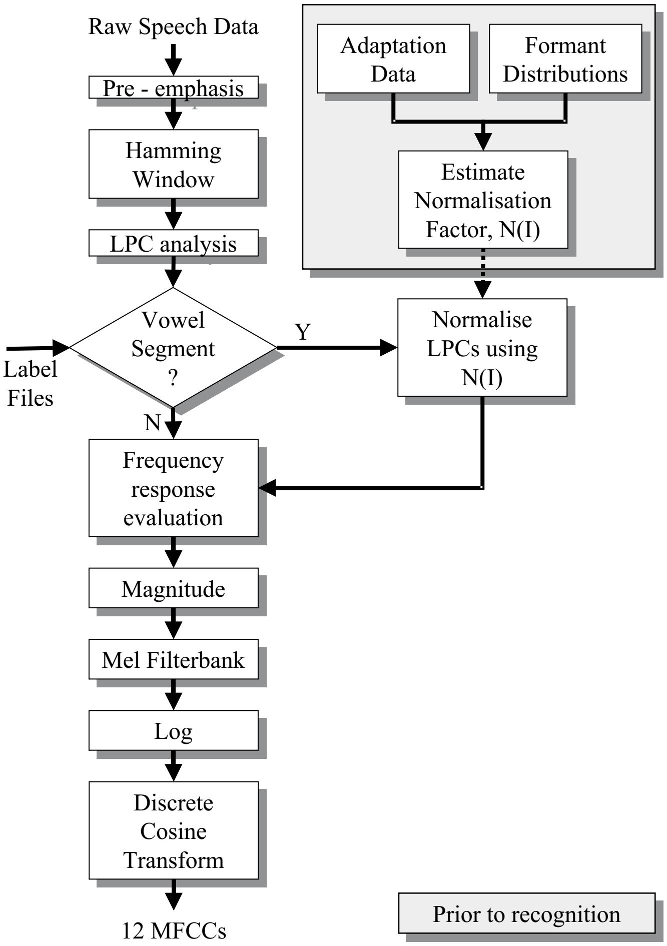


Figure 5.17: Data parameterisation incorporating normalisation of vowels by LPC warping. Warping factor estimated by matching each speaker's formants to a distribution.

error rates and reference error rates (calculated using un-normalised data) are shown in Table 5.4.

The results given in Table 5.4 were some what disappointing, showing a lower improvement than that obtained by optimally normalising every frame. The overall recognition rate (including all phonemes rather than just vowels) increased by just 0.31% (from 41.57% to 41.88%) for the male model, and 0.46% (from 42.70% to 43.16%) for the female case. This was caused by the accuracy of the unnormalised phonemes (the non vowel frames) decreasing thereby offsetting the improvement in vowel recognition.

In order to try and resolve this problem a second experiment was conducted in which *all* the speech was normalised rather than just the vowel frames. Since unvoiced speech has no clearly defined formants, normalisation cannot be performed on these segments by simply shifting the roots of the LPCs by the normalisation factor. In [43], Lee and Rose perform speaker normalisation by warping the Mel filter bank channels (Figure 5.18 — Compressing the filter bank effectively expands the spectrum shifting the formants up in frequency. Expanding the filter bank compresses the spectrum shifting the formants down. Their method of determining the correct warping factor consisted of performing an alignment of the utterance parameterised at several normalisation factors and selecting the one which gave the highest output likelihood from the recogniser. Here we use the new method of selecting the correct warping factor, $N(I)$, and scale the filter bank by a factor of $\frac{1}{N(I)}$. This does not require performing the alignment at a number of different warp factors and is therefore computationally more efficient than the work presented in [43].

The HMMs used identical topology to that mentioned in Section 5.2.5 and were trained and tested using the normalised data set. Overall results were again disappointing, the recognition rate increased by just 0.23% over the vowel only normalisation for males, and 0.4% over the vowel only case for females.

A possible explanation for this result is given in [39] in which the effects of accent on vowel formant position is discussed. The transformation is effective only if the ratio of the test speaker's first and second formants is close to the ratio of the means of the f1 and f2 reference distributions. In [39] it is suggested that the effect of accent in terms of phonetic realisation is to adjust the spacing of the formants and therefore the f1-f2 ratio. This would adversely effect the performance of the normalisation. To investigate this effect, a series of accent independent experiments were conducted.

	Male		Female	
Phoneme	Recognition Rate before normalisation	Recognition Rate after normalisation	Recognition Rate before normalisation	Recognition Rate after normalisation
iy	49.8	50.5	52.7	52.5
ih	24.6	27.0	23.0	27.2
eh	34.5	36.8	41.3	39.2
ae	64.2	68.1	62.2	55.1
ax	29.3	28.3	27.6	26.9
ah	36.6	37.3	39.0	43.5
uw	46.9	46.7	45.5	46.0
uh	34.0	34.6	37.5	41.0
ao	48.6	49.0	39.0	39.6
aa	36.7	39.4	39.0	39.9
ey	56.8	58.4	59.9	65.7
ay	46.0	46.9	53.4	53.1
oy	39.1	41.6	34.1	41.7
aw	49.2	45.5	47.5	53.3
ow	40.0	39.6	27.9	32.0
er	61.3	83.3	61.1	62.5
average	43.6	45.8	43.2	45.0

Table 5.4: Results of hmm classification experiments (% phoneme recognition accuracy) before and after normalisation by matching test speakers to a distribution of formants.

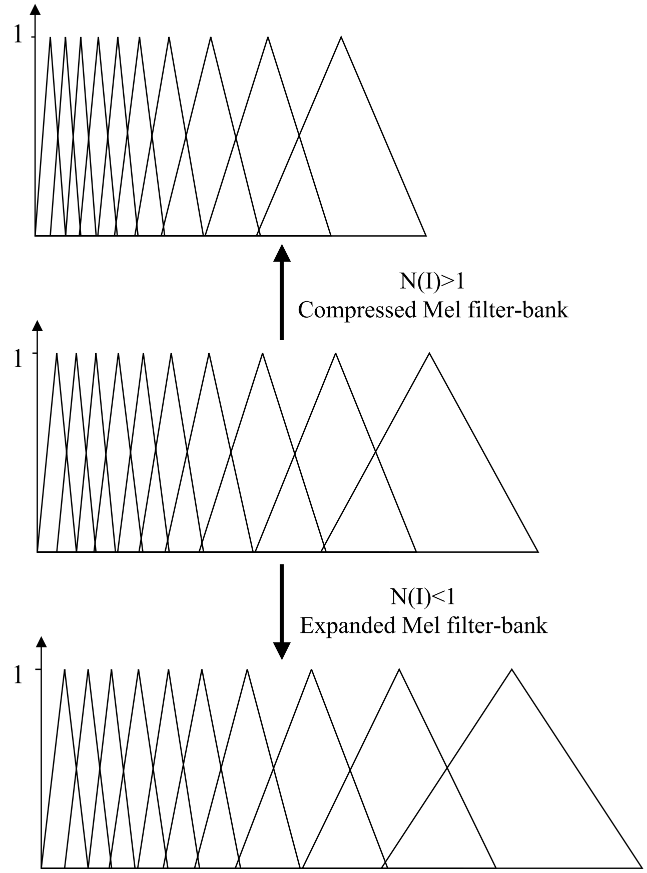


Figure 5.18: Warping of Mel filter-bank dependent upon normalisation factor.

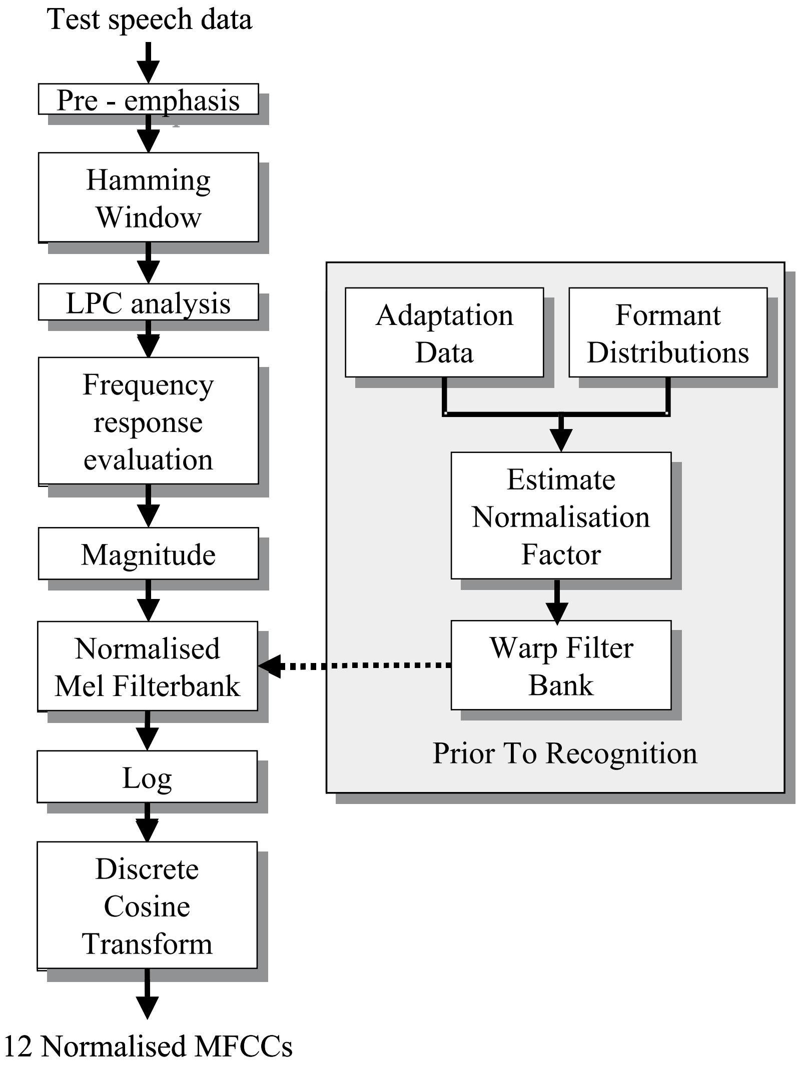


Figure 5.19: Waveform parameterisation incorporating normalisation by filter-bank warping.

Dialect Dependent Normalisation

The TIMIT database is divided into 8 distinct dialect regions, and therefore to test the effect of dialect variation, gender independent distributions for each of the vowels were generated within each dialect region. The change from gender dependent to gender independent was necessary due to the reduction in the size of the data set available for each of the models. Data set sizes varied from 330 sentences (220 training and 110 test) for dialect region 8 to 1020 sentences (760 training and 260 test) for dialect region 2.

A normalisation factor for each speaker was calculated using the vowel distributions from the speaker's accent group and equations 5.25- 5.35. The distribution of selected warp functions is shown in Figure 5.20. There is a clear distinction between the normalisations for male and female speakers - Female speakers generally have a normalisation factor less than one, while males have a factor greater than one. The warping compresses the frequency response of the female speakers and expands it for the males. This is what would intuitively be expected since, in general, women have shorter vocal tracts and correspondingly higher formants than men.

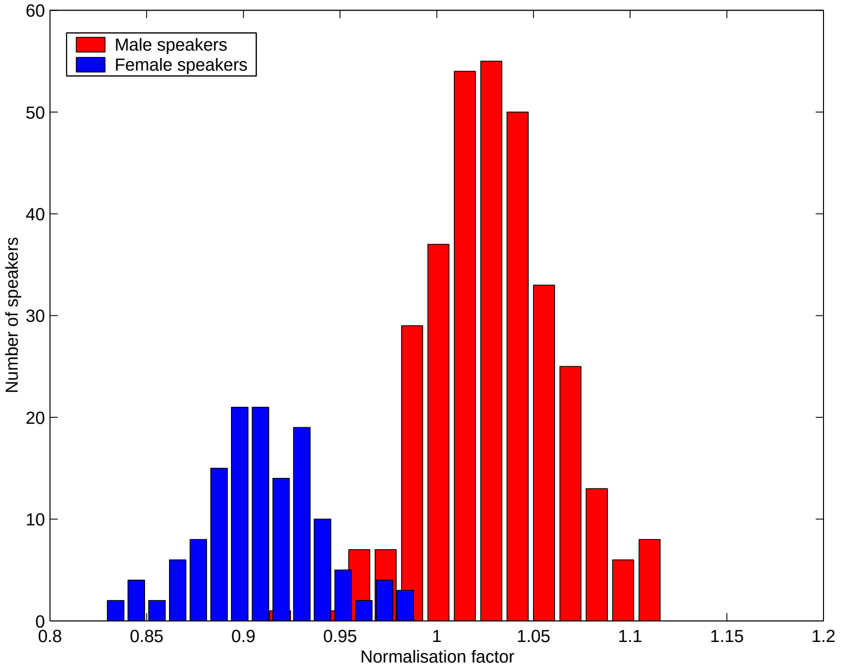


Figure 5.20: Distribution of warping factors

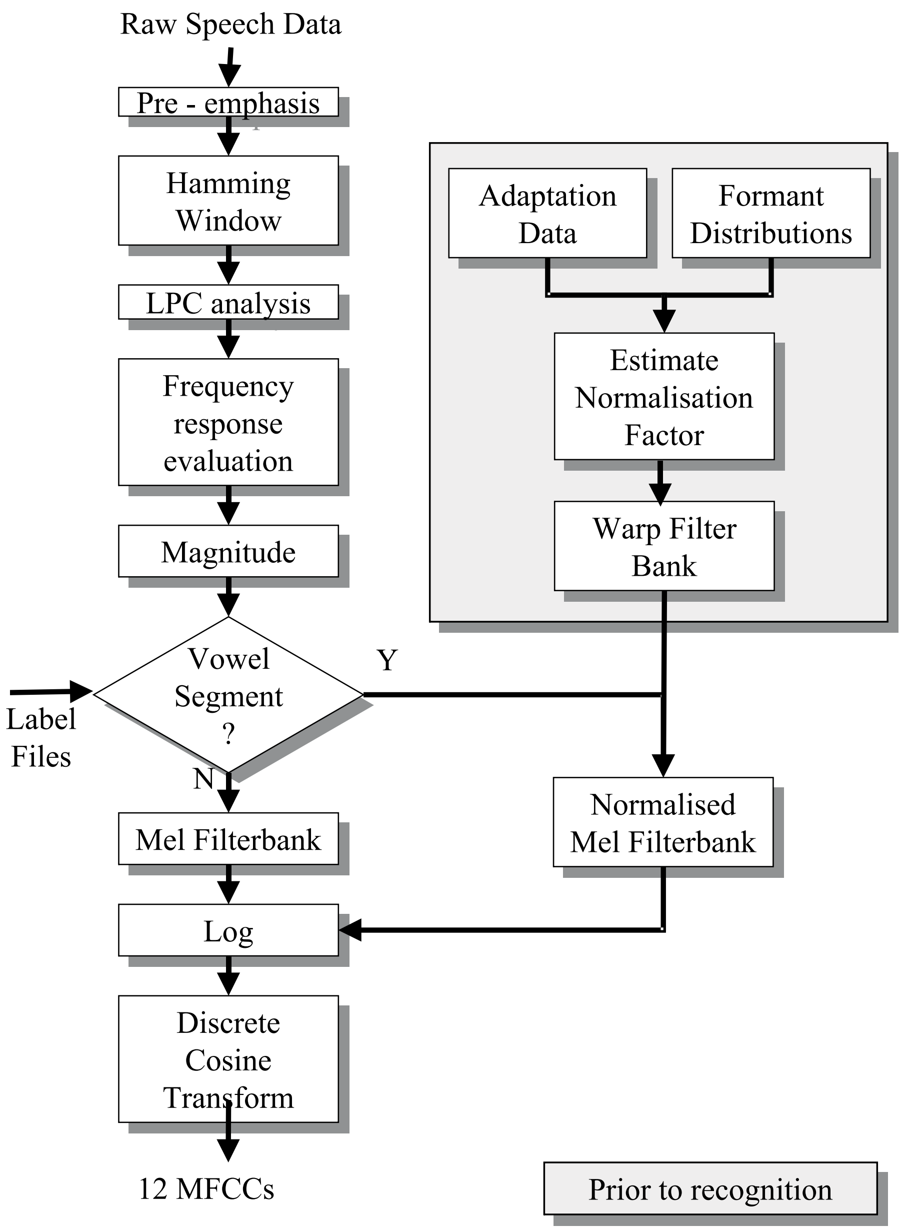


Figure 5.21: Waveform parameterisation including normalisation of vowels by filter bank warping

Three recognition experiments were performed on the data from each dialect region. In each case the previously used HMM topology was again implemented.

No warping The scheme shown in Figure 5.15(right) was used to parameterise the waveform data from each class. Here the warp factor is not used, and the results represent a baseline dialect dependent result to which the improvements provided by the normalisation may be compared.

Vowel warping In this scheme, shown in Figure 5.21 the vowel segments of each utterance are coded using the warped Mel-Filter bank, while the other segments are coded as before.

Complete warping Here every frame in the utterance is coded using the warped Mel filter-bank as shown in Figure 5.19. This represents the maximum improvement available from the technique and was performed to investigate whether normalising unvoiced segments of data provided a significant improvement in recognition accuracy.

The results of the three experiments are given in Table 5.5. Overall recognition results averaged across the dialect regions improved from 38.95% for the reference case to 40.66% for vowel normalisation, and finally to 41.72% for the fully normalised case. Warping of the Mel filter bank by a single normalisation value for each speaker can provide reasonable reductions in error rate for a low computational overhead. Further, the results show that although the normalisation is derived purely from the vowel segments of the speech, applying the same normalisation to all speech sounds provides further improvement than only normalising the vowels.

The normalisation procedure is still supervised however, since labelled utterances from each of the test speakers is required so that the formant frequencies are normalised to the correct set of vowel distributions. This is of little use in a realistic recognition scenario where labelled data from a new test speaker is unlikely to be available. An investigation was conducted into methods of performing the normalisation in an unsupervised manner, that is, without prior knowledge of the test transcription.

	Normalisation method		
Dialect region	None	Vowel only	All
DR1	38.56	39.58	41.57
DR2	41.17	42.87	43.30
DR3	40.15	41.95	42.83
DR4	38.12	40.26	41.15
DR5	37.81	40.08	40.49
DR6	37.63	38.78	40.0
DR7	40.15	41.41	42.79
DR8	38.04	40.37	41.65
Average	38.95	40.60	41.72

Table 5.5: Recognition results for dialect dependent, filter bank warping schemes.

5.4 Unsupervised Normalisation

5.4.1 Speaker Adaptation Scheme

Each speaker in the TIMIT database says two identical ‘speaker adaptation’ sentences (the so called ‘sa’ sentences). In this experiment, the normalisation factor for each of the test speakers was derived from the data for just these two sentences. Normalisation was then performed on all ten of the speaker’s utterances (including the ‘sa’ sentences). Transcriptions for just the two ‘sa’ sentences are therefore required rather than for all the test material - this is equivalent to a type of ‘speaker enrolment’ system, where new talkers are asked to say a few predefined sentences before continuing their interaction with the system. Results of the system are given in Table 5.6.

The results are comparable to those shown in Table 5.5 in which all the test speakers’ utterances were used to calculate the normalisation factor, showing that only a limited amount of data is required to accurately estimate the warping factor for each talker.

5.4.2 Two Pass Recognition

In this experiment, two recognition passes on each of the speaker’s utterances is made. The first pass uses the un-normalised data from a test speaker to generate a set of recognition files for the utterance. These are then used to calculate the normalisation for that speaker. The data is then re-parameterised using the calculated normalisation factor and

Dialect Region	Phoneme Recognition Accuracy
DR1	42.06 %
DR2	43.61 %
DR3	42.90 %
DR4	41.14 %
DR5	40.15 %
DR6	40.10 %
DR7	42.65 %
DR8	41.38 %
Mean	41.74 %

Table 5.6: Recognition results for speaker enrolment scheme

Dialect Region	Phoneme Recognition Accuracy
DR1	41.30 %
DR2	42.49 %
DR3	41.55 %
DR4	39.62 %
DR5	38.51 %
DR6	37.75 %
DR7	41.40 %
DR8	41.03 %
Mean	40.46 %

Table 5.7: Recognition results for two pass recognition scheme

the scheme shown in Figure 5.19. This data is then recognised and the output taken as the final recognised transcription of the utterance. Recognition rates for the method are given in Table 5.7.

The results of this method are lower than those obtained using the enrolment procedure, largely due to the fact that the transcriptions generated in the first pass, and subsequently used to calculate the normalisation factor are only approximately 40% correct. It does however represent an entirely unsupervised adaptation scheme which would be fast enough to be performed in a real time recognition system.

5.5 Summary and Conclusions

In this chapter a simple method of normalising the frequency response of a speaker's utterances was introduced. The method was systematically extended to produce an unsupervised method of normalisation using filter bank warping techniques.

The distribution of normalisation factors shows a marked difference between male and female speakers, and this indicates that the method is normalising variations in vocal tract length. The results of the dialect independent recognition tests suggest, however, that the system cannot normalise for differences caused by varying accents.

Chapter 6

Speaker Clustering

In this chapter, work is presented which studies a method for automatically clustering speakers. The procedure is a data driven technique utilising semi-continuous HMMS which is initially used as a method for automatically classifying accent, and is then used as a means of dividing the available data set into acoustically similar clusters of talkers.

6.1 Accent Identification Using SCHMMs

6.1.1 Introduction

Different talkers may use several different realisations of the same phonetic unit while speaking. As mentioned in Section 4, speaker clustering is a method of reducing the variance of the recognition models by training multiple sets of models on speakers with acoustically similar realisations of the same target sound. At recognition time the task is to quickly and accurately assign the unknown test speaker to one of the clusters. The models for that cluster should provide a better match to the subjects speech patterns, thus reducing recognition errors.

Speaker clustering differs from the previously presented work on speaker normalisation in that no attempt is made to alter the speech sounds from the talker, they are simply classified as being from one of a number of distinct groups, and training and recognition is then done within group. The method works at a level between signal processing of the speech signal (be it in raw or parameterised form) such as vocal tract normalisation techniques and the phonetic level such as the methods presented in [34] and Chapter 7.

While it would be possible to directly cluster the parameterised input speech, this would be highly computationally expensive owing to the large amount of data from each talker (one new vector every 25.6ms). This approach would also be difficult to implement in the context of a typical recognition system. Our premise is that speakers who are acoustically similar will tend to use the same distributions within a semi continuous HMM when their speech is recognised. This provides a mechanism for identifying clusters of speakers. This approach has the advantage that clustering may be done as a by-product of the recognition process with little additional computation.

The technique is based on the assumption that if speech from talkers in all accent groups has been used to train an HMM recogniser, then the modes in the mixture distributions will separately model the variations in the realisation of a particular target sound for a particular accent. Speech spoken with a particular accent will therefore occupy a distinct set of regions within the pattern space. This will undoubtedly not hold for some sounds and different accents may well share many of the same regions. However, if enough sounds are available, these effects should average out to make classification possible using regions in which the assumption is good. By estimating and recording at training time the regions of the pattern space used by speakers with known accents, classification of a new speaker's accent may be performed by observing which part of pattern space (that is, which modes of the mixture distribution) they utilise.

If a small number of modes were used to model the data within a state, as is normal in a continuous density HMM(Section 3.6.1), this approach would be too coarse and the distinction between accent groups too small to accurately model. In addition, a recogniser which uses triphone models might have several thousand models, each of which has several states with an associated mixture distribution. This is clearly unmanageable. In a SCHMM (Section 3.6.2) the distributions are shared between all the models, each state having a different weight on each distribution. This is useful for our purposes for two reasons: firstly, it restricts the number of distributions to a manageable number; secondly, it means that each sound is quite finely modelled. Hence we use a semi continuous HMM in which a large number of modes are available in every state and the distinction between different accents should be better defined.

6.1.2 Method

In order to cluster the speakers the regions of the pattern space used by a speaker were identified and a distance measure between speakers was calculated based on the sub space used by them. We then cluster based on this 'speaker dissimilarity' measure.

Identifying the Speakers Pattern Space

In order to identify the subspace of the pattern space used by a speaker, a set of semi-continuous speaker independent phoneme models were generated. The models share a common pool of 256 multivariate Gaussian modes which cover the acoustic space of all accents present in the training data. The model topology was 3 states per model, left-right with no skips. The models consist of the state transition matrices plus a set of mode weights for each state. The mode weights describe how the pool of Gaussians are combined to form the mixture distribution for that state.

The models were generated using HTK's parameter tying facilities — an original set of continuous models were trained, and the mixtures tied across all models to produce a semi continuous topology. The weights and mode pool were then updated using embedded Baum Welch re-estimation.

In order to identify which of the modes in the speaker independent mode pool were used by a given speaker, the SI models were used to recognise data from each of the training speakers. Table 6.1 shows a fragment of the data recorded during recognition — for each frame, the number of the mode which best matched the frame is recorded, along with the most likely model and state for that frame. From this information, a set of mode utilisation vectors as shown in Table 6.2 are generated — For each model state during recognition, the mode which has the highest likelihood most frequently is associated with that state. For each speaker we therefore have a vector of 132 (44 models * 3 states per model) modes. States which were not represented in the test data and therefore have no mode assigned a dummy mode (-1). Such models are not used in calculating the speaker dissimilarity measure. The mode utilisation vector for each speaker may be interpreted as coarse representation of the parameter space used by that speaker.

Frame	1	2	3	4	5	6	7	8	9	10	11	12	13	14	...	1027
model	sh	sh	sh	sh	sh	sh	ae	ae	ae	ae	ae	ae	d	d	...	sp
state	1	2	2	2	3	3	1	1	2	2	3	3	1	1	...	1
mode	27	28	28	37	37	38	10	10	12	12	19	10	2	2	...	130

Table 6.1: Example of recorded data for a speaker's utterance. For each frame, the most likely mode, model and state is recorded.

Model	aa			ae			ah			...	zh		
State	1	2	3	1	2	3	1	2	3	...	1	2	3
Speaker 1 Modes	10	11	13	7	8	9	27	14	15	...	169	170	171
Speaker 2 Modes	10	11	17	-1	-1	-1	32	14	15	...	168	171	172
Speaker 3 Modes	10	12	13	7	8	9	13	15	16	...	-1	-1	-1
⋮													
Speaker N Modes	10	12	13	-1	-1	-1	13	14	15	...	189	190	191

Table 6.2: Example Mode utilisation vectors. The mode which occurred most frequently for each model state during recognition is associated with that state.

Model	aa			ae			ah			...	zh		
State	1	2	3	1	2	3	1	2	3	...	1	2	3
Speaker 1 Modes	10	11	13	7	8	9	27	14	15	...	169	170	171
Speaker 2 Modes	10	11	17	-1	-1	-1	32	14	15	...	168	171	172
$Distance(1,2) = 4/9 = 0.44$													
Speaker 3 Modes	10	12	13	7	8	9	13	15	16	...	-1	-1	-1
Speaker 4 Modes	10	12	13	-1	-1	-1	13	14	15	...	189	190	191
$Distance(3,4) = 2/6 = 0.33$													

Table 6.3: Calculation of a simple dissimilarity measure. The dissimilarity between 2 speakers is the number of states observed in the test data for which their modes are different, normalised by the number of observed states.

Generating a Speaker Dissimilarity Measure

To perform the clustering, a dissimilarity measure between pairs of speakers is required so that similar speakers may be clustered together. A simple method would be to compare, on a state by state basis, pairs of mode utilisation vectors as shown in Table 6.3. The dissimilarity is calculated as number of non identical pairs of modes. States with a mode of -1 for either of the speakers (i.e. the model state was not represented in the test data and therefore an estimate of the most used mode could not be made) are ignored. The final sum is then normalised by the number of pairs of states compared (i.e. those in which neither speaker recorded -1).

While this method provides a simple method of generating a speaker dissimilarity measure, it does not take into account the relative similarity of the modes : Given the three example modes shown in Figure 6.1, it is clear that modes 1 and 2 are extremely similar, and that two speakers utilising those modes in the same state should have a lower dissimilarity measure than two using modes 1 and 3.

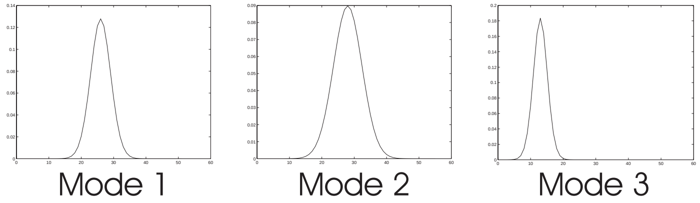


Figure 6.1: Example Modes

In order to account for this, a dissimilarity measure between the modes, MD, was calculated :

$$MD_{ij} = \sum_{k=1}^{N} \frac{(\mu_k^i - \mu_k^j)}{(\sigma_k^i + \sigma_k^j)} \tag{6.1}$$

where MD_{ij} is dissimilarity between mode i and j; N is the number of components in the mode; μ_k^i is the kth mean value for mode i; σ_k^i is the kth variance value for mode i.

The dissimilarity between two speakers can now be calculated as the sum of the dissimilarities between the modes, normalised by the number of comparisons as shown in Table 6.4.

Model	aa			ae			ah			. . .	zh		
State	1	2	3	1	2	3	1	2	3	. . .	1	2	3
Speaker 1 Modes	10	11	13	7	8	9	27	14	15	. . .	169	170	171
Speaker 2 Modes	10	11	17	-1	-1	-1	32	14	15	. . .	168	171	172
$Distance(1,2) = \frac{MD(13,17)+MD(27,32)+MD(169,168)+MD(170,171)+MD(171,172)}{9}$													
Speaker 3 Modes	10	12	13	7	8	9	13	15	16	. . .	-1	-1	-1
Speaker 4 Modes	10	12	13	-1	-1	-1	13	14	15	. . .	189	190	191
$Distance(3,4) = \frac{MD(15,14)+MD(16,15)}{6}$													

Table 6.4: Example dissimilarity measure incorporating mode dissimilarity

Given this dissimilarity matrix between speakers, a clustering method can be used to obtain groups of similar speakers.

The Clustering Algorithm

To cluster the training speakers, a variation on K-means clustering was used.

i For the entire data set, the two maximally separated points (ie the two speakers with the largest dissimilarity) are found and all the speakers assigned to the nearest of these points to form an initial pair of clusters.

ii The centroid of each cluster is then calculated, were the centroid is defined as the speaker with the minimum - maximum distance to any of the other speakers within that cluster.

iii Speakers are allocated to their closest centroid.

iv New clusters are formed from the allocations of speakers in [iii].

v Repeat ii — iv until iteration converges (ie no speakers change clusters) or a pre-defined number of iterations have been completed.

vi If number of clusters is less than the required number, find the current cluster with maximum separation between any pair of speakers (ie the widest spread cluster). Take the maximally separated points within this cluster and assign the rest of the points to the closer of the two, effectively splitting the cluster into two new smaller clusters.

viii Find the minimax centroids of those two new clusters and repeat from iii.

This process then provides a number of cluster centroids which are generated without reference to the speakers' accents. The expectation is that each of the training clusters would contain a majority of speakers from a single accent group.

To cluster the test data, dissimilarity measures between each test speaker and the cluster centroids were calculated using the method described for generating dissimilarity measures between training speakers, ie the normalised sum of the mode dissimilarities. The test speakers are then assigned to the cluster with the most similar centroid.

6.1.3 British v American English Classification

The Data

The technique was tested on its ability to discriminate British and American accented English speech. The WSJCAM0 database was used to provide the British English data, and WSJ1, the American English. Speech from 98 speakers from the training sets of both databases were used to train the models, providing a total of 8596 sentences. The speech was parameterised to provide a 12 component MFCC vector, augmented with velocity, acceleration and log energy coefficients. Cepstral mean normalisation was applied to each sentence processed to compensate for differences in the recording procedure of each of the databases. The clustering of the training set to determine the cluster centroids was performed on a subset of 29 speakers from each of the two databases. For testing, speech from a set of 40 speakers from WSJ and 19 speakers from WSJCAM0 were used.

Separation of Training Data

As mentioned in Section 6.1.2 the clustering of the training data gives an initial indication of whether the method is applicable to accent classification. The 58 speakers were clustered into two groups. The first contained 29 American and 13 British speakers, while the second was composed entirely of the remaining 16 British talkers - a reasonable grouping for classification purposes (Table 6.5). The gender distributions of the two clusters showed no split between male and female talkers, indicating that the accent differences between speakers have a greater effect on mode usage than speaker gender.

Cluster	1	2
British Talkers	13	16
American Talkers	29	0

Table 6.5: Distribution of training speakers using SCHMM based clustering

Results on Original Databases

The test data was classified using the procedure given in Section 6.1.2. Speakers assigned to the centroid of cluster 1 were designated American, while those assigned to cluster 2 were designated British. Classification accuracy was tested after 1, 2, 3, . . . , 8 sentences of speech had been recognised. The results are shown in Figure 6.2.

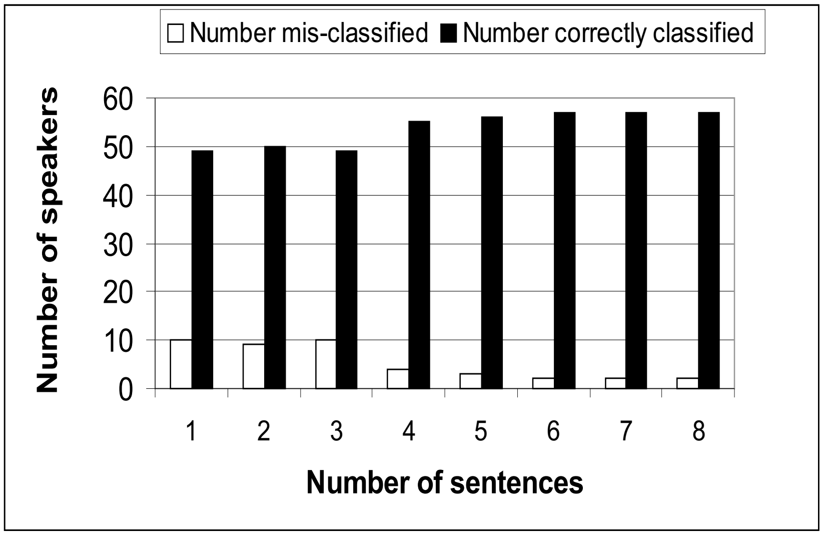


Figure 6.2: Results of accent classification experiment on WSJ1 and WSJ-CAM0 data using a SCHMM technique.

10 of the 59 speakers are misclassified after 3 sentences are available but this falls to 4 speakers after 4 sentences are available and 2 speakers after 6.

Results on Independent Databases

The American and British accented speech was derived from two separate databases recorded under different conditions. Cepstral mean normalisation was used on the data in an attempt to alleviate any overall spectral differences between the two datasets, but we were concerned that the "accent recognition" demonstrated here might be no more than identification of two sets of data which differed in their acoustic characteristics and which were represented in both the training and the test data. We therefore ran an experiment to verify the techniques on an independent set of data. Sentences from twenty speakers from the American-accented TIMIT database (dialect region one) were tested using the same method as described in Section 6.1.2. Results are shown in Figure 6.3. The same pattern of fewer unclassified and misclassified speakers as more data becomes available is shown and the final classification performance is comparable to that achieved on non-independent data. This result shows that the accent classification is independent of the conditions used when recording the database.

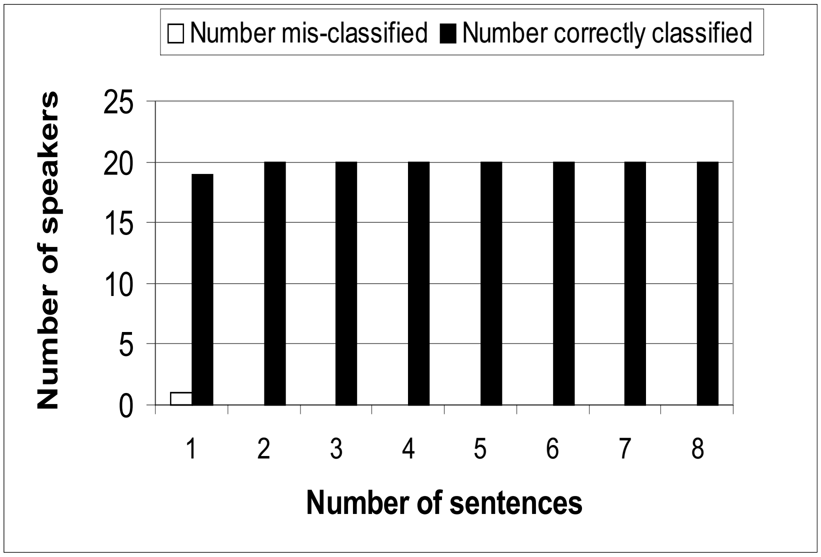


Figure 6.3: Results of accent classification experiment on TIMIT data using a SCHMM technique.

6.1.4 Regional Accent Classification

Having shown that the technique is capable of discriminating between British and American accented speech, two experiments were performed to test the methods ability to differentiate between British regional accents and American regional accents.

British English Accent Discrimination

The Subscriber database [73] has accent classifications for each of the talkers in the test and training set A.1.2. A SCHMM was built using the training set and the method described above used to divide the training talkers in to various numbers of clusters. Test speakers were then assigned to a cluster as before. The distribution of accents, and the gender of speakers for various numbers of clusters are shown in Tables 6.6 to 6.8

These results show some clustering of talkers into their accent groups, the London and Liverpudlian groups for example. However most clusters contain talkers with a number of different accents. Also, there seems to be little clustering of talkers by gender. This is most significant in the 2 cluster results where, if the accents were not greatly different, it would be expected for the split in talkers to be dominated by the variation between males and females.

An explanation for the lack of accent discrimination comes from the fact that the Subscriber database was recorded over telephone channels. It is possible that the clustering is showing variations in telephone handsets or line conditions rather than in the talkers' speech.

American English Accent Discrimination

In order to investigate whether the telephone channel conditions were responsible for the poor results observed in the British regional accent discimination task, the TIMIT database (Section A.2.1) was used in an identical experiment. TIMIT is a clean speech database and as such will have no associated channel effects. It is also approximately the same size as Subscriber and each of the talkers is labelled with respect to their accent. The dialect and gender distributions for each of the generated clusters is shown in Tables 6.9 to 6.11

In these experiments, most of the clusters tend to contain speakers of a single gender, though there still seems to be little discrimination in terms of the annotated accent.

	Cluster		
Dialect Region	Train-1	Train-2	Total
Northern British	64 (38.5%)	169 (36%)	233 (36.6%)
Southern British	31 (18.6%)	88 (18.7%)	119 (18.7%)
Liverpudlian	0 (0%)	6 (1.3%)	6 (0.9%)
Welsh	3 (1.8%)	8 (1.7%)	11 (1.7%)
London	3 (1.8%)	30 (6.4%)	33 (5.2%)
Irish	6 (3.6%)	45 (9.6%)	51 (8.0%)
Scottish	34 (20.5%)	67 (14.3%)	101 (15.9%)
West Country	25 (15.0%)	57 (12.1%	82 (12.9%)
Total	166	470	636

	Cluster			
Dialect Region	Train-1	Train-2	Train-3	Total
Northern British	49 (28.9%)	37 (40.2%)	147 (35.2%)	233 (36.6%)
Southern British	25 (19.8%)	8 (8.6%)	86 (20.5%)	119 (18.7%)
Liverpudlian	0 (0%)	0 (0%)	6 (1.4%)	6 (0.9%)
Welsh	2 (1.6%)	2 (2.2%)	7 (1.7%)	11 (1.7%)
London	3 (2.4%)	2 (2.2%)	28 (6.7%)	33 (5.2%)
Irish	4 (3.2%)	12 (13%)	35 (8.4%)	51 (8.0%)
Scottish	25 (19.8%)	18 (19.5%)	58 (13.8%)	101 (15.9%)
West Country	18 (14.3%)	13 (14.1%)	51 (12.2%)	82 (12.9%)
Total	126	92	418	636

Table 6.6: British accent distribution - Top: 2 clusters. Bottom: 3 clusters

	Cluster									
Dialect Region	Train-1		Train-2		Train-3		Train-4		Total	
Northern British	45	(38.8%)	34	(40.5%)	26	(51%)	128	(33.4%)	233	(36.6%)
Southern British	21	(18.1%)	7	(8.3%)	14	(27.4%)	77	(20.1%)	119	(18.7%)
Liverpudlian	0	(0%)	0	(0%)	1	(2%)	5	(1.3%)	6	(0.9%)
Welsh	2	(1.7%)	2	(2.4%)	1	(2%)	6	(1.6%)	11	(1.7%)
London	3	(2.6%)	2	(2.4%)	0	(0%)	28	(6.7%)	33	(5.2%)
Irish	4	(3.4%)	11	(13.1%)	1	(2%)	35	(9.1%)	51	(8.0%)
Scottish	24	(18.1%)	18	(21.4%)	5	(9.8%)	54	(14.1%)	101	(15.9%)
West Country	17	(14.6%)	10	(11.9%)	5	(9.8%)	50	(13.1%)	82	(12.9%)
Total	116		84		53		383		636	

Table 6.7: British accent distribution - 4 clusters

	Cluster		
Gender	Train-1	Train-2	Total
Male	75 (45.2%)	234 (49.8%)	309 (48.6%)
Female	91 (54.8%)	236 (50.2%)	327 (51.4%)
Total	166	470	636

	Cluster			
Gender	Train-1	Train-2	Train-3	Total
Male	68 (45.2.5%)	19 (20.7%)	225 (53.8%)	309 (48.6%)
Female	58 (54.8%)	73 (79.3%)	193 (46.2%)	327 (51.4%
Total	126	92	418	636

	Cluster				
Gender	Train-1	Train-2	Train-3	Train-4	Total
Male	64 (55.2%)	17 (20.2%)	9 (17.0%)	219 (57.2%)	309 (48.6%)
Female	52 (44.8%)	69 (79.8%)	42 (83.0%)	164 (42.8%)	327 (51.4%)
Total	116	84	53	383	636

Table 6.8: British gender distribution. 2 - 4 clusters

	Cluster		
Dialect Region	Train-1	Train-2	Total
DR1	11 (12.2%)	27 (7.4%)	38 (8.3%)
DR2	13 (14.4%)	62 (16.9%)	75 (16.4%)
DR3	12 (13.3%)	62 (16.9%)	74 (16.2%)
DR4	8 (8.9%)	60 (16.3%)	68 (14.9%)
DR5	18 (20%)	51 (13.8%)	69 (15.1%)
DR6	10 (11.1%)	24 (6.5%)	34 (7.4%)
DR7	12 (13.3%)	65 (17.7%)	77 (16.8%)
DR8	6 (6.7%)	16 (4.3%	22 (4.8%)
Total	90	367	457

	Cluster			
Dialect Region	Train-1	Train-2	Train-3	Total
DR1	8 (15.1%)	6 (7.7%)	24 (7.4%)	38 (8.3%)
DR2	8 (15.1%)	10 (12.8%)	57 (17.5%)	75 (16.4%)
DR3	6 (11.3%)	18 (23.1%)	50 (15.3%)	74 (16.2%)
DR4	6 (11.3%)	8 (10.3%)	54 (16.6%)	68 (14.9%)
DR5	9 (17.0%)	15 (19.2%)	45 (13.8%)	69 (15.1%)
DR6	8 (15.1%)	5 (6.4%)	21 (6.4%)	34 (7.4%)
DR7	6 (11.3%)	11 (14.1%)	60 (18.4%)	77 (16.8%)
DR8	2 (3.8%)	5 (6.4%)	15 (4.6%)	22 (4.8%)
Total	53	78	326	457

Table 6.9: American accent distribution - Top: 2 clusters. Bottom: 3 clusters

	Cluster									
Dialect Region	Train-1		Train-2		Train-3		Train-4		Total	
DR1	5	(12.8%)	8	(13.3%)	24	(7.7%)	1	(2.2%)	38	(8.3%)
DR2	5	(12.8%)	8	(13.3%)	53	(17.0%)	9	(19.6%)	75	(16.4%)
DR3	6	(15.4%)	7	(11.7%)	52	(16.7%)	9	(19.6%)	74	(16.2%)
DR4	5	(12.8%)	4	(6.7%)	53	(17.0%)	6	(13.0%)	68	(14.9%)
DR5	10	(25.6%)	10	(16.7%)	39	(12.5%)	10	(21.7%)	69	(15.1%)
DR6	3	(7.7%)	9	(15.0%)	19	(6.1%)	3	(6.5%)	34	(7.4%)
DR7	4	(10.2%)	9	(15.0%)	60	(18.3%)	7	(15.2%)	77	(16.8%)
DR8	1	(2.6%)	5	(8.3%)	15	(4.8%)	1	(2.2%)	22	(4.8%)
Total	39		60		312		46		457	

Table 6.10: Americn accent distribution - 4 clusters

	Cluster		
Gender	Train-1	Train-2	Total
Male	4 (4.4%)	319 (86.9%)	323 (70.7%)
Female	86 (95.6%)	48 (13.1%)	134 (29.3%)
Total	90	367	457

	Cluster			
Gender	Train-1	Train-2	Train-3	Total
Male	3 (5.7%)	12 (15.4%)	308 (94.5%)	323 (70.7%)
Female	50 (94.3%)	66 (84.6%)	18 (5.5%)	134 (29.3%)
Total	53	78	326	457

	Cluster				
Gender	Train-1	Train-2	Train-3	Train-4	Total
Male	2 (5.1%)	1 (1.7%)	296 (94.9%)	24 (52.2%)	323 (70.7%)
Female	37 (94.9%)	59 (98.3%)	16 (5.1%)	22 (47.8%)	134 (29.3%)
Total	39	60	312	46	457

Table 6.11: American gender distribution. 2 - 4 clusters

Discussion

While the method shows good ability to discriminate between English and American speakers who have large differences in their accents, it seems unable to identify the smaller variations between regional American and British accents. In this case however, the method is being judged against the subjective assessment of the talkers accent given by the annotator of the database. The lack of success in identifying accents may be due to the fact that the processes of discretising a continually varying set of speaking styles into one of 8 accent classes means that talkers in a single accent group may have widely different speaking styles which the method is unable to identify. In effect, the system may be classifying similar talkers together, but the similarities are due to effects other than those associated with their accent. Indeed, the results of the American regional accent classification task have shown that the method is capable of identifying some structure in that it can discriminate between male and female talkers. While the method can not be used to determine an unknown speaker's regional accent, it is possible that by building separate recognition models within each cluster, and testing with test talkers assigned to that cluster, recognition accuracy may be improved.

6.2 Clustering Recognition Experiments

6.2.1 Annotated Accent Clustering Experiments

An initial experiment was conducted to determine whether in fact building models for each of the annotated accent groups and testing within group would produce an increase in recognition accuracy over the accent independent case. That is, even if the accent classification technique were 100% accurate, would it make a large improvement in recognition accuracy? The TIMIT database was used for the investigation and the speech data parameterised as before.

Method

Phoneme level Continuous HMM recognisers were generated using HTK. A 3 state, single Gaussian mixture, left right with no skips, diagonal covariance matrix topology was used for the experiments. Initially a dialect independent recogniser was built using all 462 speakers in the training set. This was then tested using all 168 test speakers. Individual sets of models for each of the 8 dialect regions were then trained, and the test data for the appropriate region used to evaluate their recognition accuracy. The results of the recognition tests are given in Table 6.12.

Dialect Region	Phoneme Recognition Accuracy
Dialect Independent	59.17%
dr1	58.41%
dr2	61.39%
dr3	60.51%
dr4	57.59%
dr5	57.30%
dr6	58.04%
dr7	59.62%
dr8	58.67%
Mean	58.94%

Table 6.12: Dialect dependent recognition results

Discussion

The results show that the use of the annotated accent groups as a means of clustering speakers provides no improvement in recognition accuracy over the dialect independent case. It is possible that any improvement in accuracy is offset by the reduction in the amount of training data, though this is largely mitigated by the results for the DR8 accent group. This so called ’army brat’ group consists of speakers who moved around during their childhood and as such have no clearly defined accent. This group also has the smallest number of speakers and is, as such, equivalent to a dialect independent model trained with less data. The difference in recognition accuracy between this and the dialect independent model is only 0.5% implying that both sets of models are fully trained.

The clustering procedure used in Section 6.1.4 was able to identify the variation between male and female speakers whilst being unable to distinguish between accent groups. This suggests that the effects of accent could be masked by variations in vocal tract and if this is the case, removing variations in vocal tract may allow accent to be more easily identified,

6.2.2 Effects of Vocal Tract Variation on Regional Accent

To investigate whether the effects of vocal tract variation masked those due to regional accent, we used vocal tract normalisation of speakers prior to model building. With differences in vocal tract between speakers removed, we can see whether significant improvements in recognition accuracy are gained by building dialect dependent models.

Method

The investigation was conducted as follows :

1. Select two distinct accent groups from the 8 in the TIMIT database. This was done by listening to several of the talkers from each dialect group and selecting two which sounded very different.

2. Train dialect independent models using both sets of training data from the selected accent groups. Measure the performance on both sets of test data.

3. Train Dialect dependent models for each of the accent group and test within group.

4. Use vocal tract normalisation on all utterances and retrain dialect independent models on normalised data. Measure recognition accuracy on all normalised test data. Performance should rise slightly over the unnormalised case.

5. Use normalised data to build dialect dependent models and test within group on normalised test data. If dialect effects are masked by vocal tract variation, these models should show significant improvements in recognition accuracy over the unnormalised case. The improvements between the dialect dependent cases should be greater than that seen between the dialect independent case.

The vocal tract normalisation procedure used is described fully in 5.

Results

The selected dialect regions were dr2 (northern) and dr5 (southern). The results of the recognition experiments are given in Table 6.13.

Training Data	Test Data	Phoneme Recognition Accuracy
Both - unnormalised	Both - unnormalised	53.12%
dr2 - unnormalised	dr2 - unnormalised	54.56%
dr5 - unnormalised	dr5 - unnormalised	52.06%
Both - normalised	Both - normalised	55.42%
dr2 - normalised	dr2 - normalised	56.88%
dr5 - normalised	dr5 - normalised	54.26%

Table 6.13: Dialect dependent recognition results—normalised and unnormalised cases.

The improvement in using normalised models is approximately 2% in both dialect independent and dependent cases.

Discussion

The results show that the use of dialect dependent models in the recognition system, even after vocal tract effects have been removed, provides little improvement in recognition accuracy. An explanation for this result could come from the phonetic transcriptions used in generating the models. The label files for the TIMIT database are hand annotated, fine level phonetic transcriptions. If the variation in pronunciation of a phrase between two different dialect regions is large (for instance the difference between the word ’bath’ for southern talkers who use /ɑ/ and northern British talkers, who use /æ/) then the difference would result in a different transcription of the phrase in each case. If the different pronunciation is consistent for all talkers with a given dialect then the accent variation will already have been accounted for in the labelling. In effect, the dialect independent model consists of a shared set of models for phonemes common to all accent groups plus separate subsets used by only a few of the dialect groups. Hence the dialect independent system has near identical performance to the dialect dependent case. To investigate this effect, an experiment was conducted using phone level label files generated from a standard pronunciation dictionary, rather than the supplied transcriptions.

6.2.3 Use of Standard Pronunciation Dictionary to Generate Label Files

The purpose of this experiment was to investigate recognition accuracy if pronunciation differences are not accounted for in the phonetic transcriptions.

Generating Label Files

In addition to the phonetic transcriptions, the TIMIT database also includes word level transcriptions of each sentence. A standard pronunciation dictionary (also supplied with the database) with a single pronunciation per word, was used to construct new, ‘standard pronunciation’ transcriptions for each of the files. This was performed by simply replacing each word in the transcription with its corresponding pronunciation from the dictionary.

Experiment

A set of speaker independent monophone hidden Markov models were generated using the new label files, as well as dialect dependent models for each of the 8 TIMIT dialect regions. Model topology in all cases was three state, left right with no skips, diagonal covariance matrix. The speaker independent models were tested using data from all accent groups. Accent dependent models were tested ‘within group’ on a single dialect region.

Phoneme recognition results are given in Table 6.14. Again, little improvement is gained by using dialect specific models over the speaker independent case.

Conclusions

Even with labelling differences removed, the lack of improvement suggests that there is little systematic variation in pronunciation of phonemes between American accent groups. It is suggested in [82] that differences between local accents are largest in countries which have been English speaking for longest. American accents tend to be far less variable than British ones and as such the modeling technique may not be powerful enough to identify the small variations between groups.

Dialect Region	Phoneme Recognition Accuracy
Dialect Independent	50.84%
dr1	50.90%
dr2	53.47%
dr3	52.65%
dr4	49.66%
dr5	48.69%
dr6	50.55%
dr7	52.82%
dr8	49.25%
Mean	51.00%

Table 6.14: Dialect dependent recognition results using label files generated from a pronunciation dictionary

System	% Correct	% Accuracy
SI	33.72	26.08
2 Cluster	33.85	26.26
3 Cluster	34.02	26.33
4 Cluster	34.17	26.44
5 Cluster	34.16	26.48
6 Clusters	33.95	26.09

Table 6.15: Results of recognition experiment for clustered Subscriber data

6.2.4 Data Driven Clustering Recognition

Since the method of regional accent classification is based on a purely data driven clustering procedure, it is likely that there are some similarities between the speakers in each cluster. In order to establish whether the clustering method may be used to increase the recognition accuracy, the clusters generated previously were used to build cluster models. The same topology models as those used for the clustering (256 mode SCHMM, 3 state, left right with no skips) were implemented. The test utterances were then recognised using the cluster models to which the test speaker had been allocated. Results for the speaker independent case and 2 to 6 clusters for subscriber and 2 to 4 clusters for TIMIT, are given in tables 6.15 and 6.16.

While the improvements appear small, it should be noted that the same amount of training data is being used in all cases, even though the number of parameters being es-

System	% Correct	% Accuracy
SI	49.95	44.01
2 Cluster	51.89	45.70
3 Cluster	51.90	45.60
4 Cluster	51.62	45.20

Table 6.16: Results of recognition experiment for clustered TIMIT data

timated is increasing proportional to the number of clusters, therefore any improvement is significant. The improvements are also larger than those given by clustering based on the annotated accents (Table 6.12), indicating that, as suggested, the clustering method identifies similarities between speakers which are different from those used to identify the accent.

6.3 Conclusions

The accent classification experiments have shown that clustering speakers based on their use of the model parameter space is capable of identifying gross differences between the accents of different talkers. The observed failure of the method to distinguish between regional accents may be due to the following reasons :

- It is possible that the clustering of British and American English accents may have been due to effects such as speaking rate, line conditions or spectral slope present in the databases and not specifically on the effect that we perceive as ‘accent’. As many of these effects would be identical in a single database, the method would fail to discriminate between regional accents.

- There are no clear definitions as to what constitutes a given accent - two speakers described as ‘northern British’ may have considerably different accents and it is not therefore unreasonable to find the clusters containing speakers with many accent classifications.

- The large differences between accents - that of one phoneme being substituted for another - will already have been accounted for in the phonetic labelling. As such, the models will not contain accent specific information in these cases.

The fact that building accent specific models does not improve recognition accuracy, even if vocal tract differences are removed, also indicates that the accent labelling given in the databases covers too wide a variation of speaking styles to be useful in reducing model variance and therefore improving recognition accuracy.

Clustering speakers using a purely data driven technique does, however, give some improvement in accuracy, despite the reduction in training data for each model. Since the clusters do not correlate with the labelled accent, we may conclude that there are variations between speakers which are more useful in reducing recognition accuracy than those which manifest themselves as ‘accents’.

Chapter 7

Phonotactic Models for Accent Classification

It has been shown [7, 75]that the *gross* differences between accents, such as those between British and American English may be overcome by the use of the use of accent specific model sets, lexicons and grammars. To use this method effectively we must be able to quickly and accurately classify the speaker's accent so as to know which model set etc. to use. In this chapter we present a method of accent classification which models the accent using higher level phonetic features (diphones) rather than the acoustic signal as was used in the method presented in the previous chapter. Hence we are using the *phonotactics* of the accent rather than the *phonetic realisations* to model the differences between speakers.

7.1 Introduction

Languages and dialects each have a set of rules which describe how the sounds which make up the language may be combined to form words. In English for instance, although the sounds /p/ and /f/ are available, the word 'Pfropf' is not a valid word since /p//f/ is not a allowed sequence of phonemes — the *phonotactic* rules of English do not allow it. The rules of German are different however — the word means 'stopper'. Phonotactic rules can be extended to describe the likelihood of a certain ordering of sounds occurring during speech from a certain language. Previous studies have shown that the phonotactic rules can be used to identify a given language [90] or regional accent [38]. Typically, in these systems, the probability of occurrence of diphones (that is, pairs of phonemes) for a given accent is estimated from the output of a recogniser when recognising speech from that accent. The output of the recogniser for speech from an unknown talker is used in conjunction with these probabilities to determine the talkers' accent.

There are, however, certain problems with this approach which are addressed in the technique described here. Inconsistent recognition errors in the training phase (i.e. if identical input phone sequences were decoded differently on different occasions) would introduce errors into the model. The occurrence of 'preferred' error patterns, that is if the recogniser frequently output a particular incorrect phone sequence regardless of the accent or language, would also introduce incorrect information into the model. Both of these forms of error would result in a reduction in the performance of the system when the model was subsequently used in the classification of an unknown talker. The accuracy of the phoneme recogniser used in the previous experiments was only approximately 45%. Hence, only about 20% of diphones available from the output would be correct — the errors in the phonotactic model would be large if this output were used to generate it. Instead, a pronunciation dictionary for each of the accents to be identified was used to generate the model which removes the problem of incorrect recogniser output. However, it does mean that the technique relies on the dictionary transcription for each entry to be correct.

7.2 Method

The premise of the technique is that the phonotactic information about a language which is contained within a pronunciation dictionary can be modelled, and that this model may then be used to classify the output of the recogniser as being from a certain language. For example, if a diphone occurs frequently in a dictionary for language A and infrequently for language B, then the occurrence of that diphone in the output of the recogniser is a strong indicator that the speaker is of language A. The model is built by measuring the amount of information supplied by a particular diphone to the classification task. This is done by calculating the mutual information of a given diphone.

7.2.1 Mutual Information

Mutual information [6, 78] is a measure of the reduction in the uncertainty of a source gained by observing a certain output. In general, if we have a zero memory source which may be one of K classes and let $\Pr(S_k)$ describe the output distribution for the kth class, we then have K a priori distributions $\Pr(S_1), \Pr(S_2), \ldots, \Pr(S_k)$. We define an information unit to be

$$I(S_k) = \log \frac{1}{\Pr(S_k)}. \tag{7.1}$$

The base of the logarithm defines the units (base 2 implies bits, 10 implies Hartleys, etc). We now define the average information provided by the source, the *source entropy* to be

$$H(S) = \sum_{k=1}^{K} \Pr(S_k) I(S_k) \tag{7.2}$$

or

$$H(S) = -\sum_{k=1}^{K} \Pr(S_k) \log \Pr(S_k) \tag{7.3}$$

To illustrate the idea of mutual information, consider a pattern recognition task in an R dimensional pattern space, in which the rth dimension has been quantised to L_r levels, $q_r(l)$, where $l = 1, 2, \ldots, L_r$. We wish to know the amount of information about the classification supplied by each of the dimensions. Suppose we are told the value of the rth dimension — to consider what we have learned about the kth class, S_k, we require the conditional probability $\Pr(S_k|q_r(l))$ with associated information $-\log(\Pr(S_k|q_r(l))$. The source entropy given this observed feature value is then :

$$H(S|q_r) = -\sum_{k=1}^{K} \sum_{l=1}^{L_r} \Pr(S_k, q_r(l)) \log(\Pr(S_k|q_r(l)) \tag{7.4}$$

$H(S|q_r)$ is known as the *equivocation* of the source given that we may observe dimension r. The amount of information provided by dimension r is simply :

$$I(S, q_r) = H(S) - H(S|q_r) \tag{7.5}$$

that is, how much has the uncertainty of the source been reduced by observing the dimension. This quantity is known as the *mutual information* and may be represented by :

$$\begin{aligned} I(S, q_r) &= \sum_{k=1}^{K} \sum_{l=1}^{L_r} \Pr(S_k, q_r(l)) \log \frac{\Pr(S_k|q_r(l))}{\Pr(S_k)} && (7.6) \\ &= \sum_{k=1}^{K} \sum_{l=1}^{L_r} \Pr(S_k, q_r(l)) \log \frac{\Pr(S_k, q_r(l))}{\Pr(S_k)\Pr(q_r(l))} && (7.7) \end{aligned}$$

Returning to the accent classification problem, the amount of information supplied for the discrimination task by observing a certain output from the recogniser (that is, the mutual information of a given diphone) is calculated as follows: the probabilities of occurrence of diphone d_i in American accented speech

$$\Pr(d_i|A) \tag{7.8}$$

and in British accented speech

$$\Pr(d_i|B) \tag{7.9}$$

were estimated directly from the entries in the dictionary:

$$\Pr(d_i|A) \approx \frac{\text{Number of occurrences of } d_i \text{ in American dictionary}}{\text{Total number of diphones in American dictionary}} \tag{7.10}$$

and

$$\Pr(d_i|B) \approx \frac{\text{Number of occurrences of } d_i \text{ in British dictionary}}{\text{Total number of diphones in British dictionary}}. \tag{7.11}$$

The amount of information $I(d_i)$ for discrimination of the accent supplied by diphone d_i can be estimated as follows:

$$I(d_i) = \sum_{j=1}^{2} \Pr(A_j, d_i) \log_2 \frac{\Pr(A_j, d_i)}{\Pr(A_j) \Pr(d_i)} \qquad \text{bits,} \tag{7.12}$$

where $A_1 = A$ (American accent) and $A_2 = B$ (British accent). This is simply the mutual information when the observation vector has a single dimension.

7.2.2 Pronunciation Dictionaries

The classification task was that of identifying British and American accented English speech. To build the model, British and American pronunciation dictionaries were required. The BEEP dictionary [1] provided the British English pronunciations and CMUDICT [2] the American. The BEEP dictionary provides phonemic transcriptions for over 250000 words, while CMUDICT contains approximately 100000 pronunciations.

7.2.3 The Model

The technique has the advantage that any diphone not occurring in either pronunciation dictionary has $I(d_i) = 0$. Although such diphones may be frequently output by the recogniser, they will contribute nothing to the classification. If the recogniser output were used to train the model, this would not be the case and incorrectly decoded diphones would contribute spurious information to the classification resulting in increased errors. Also diphones which are incorrectly classified but are legal (i.e. diphones which occur in the dictionary) will contribute noise to the classification which should average to zero if enough diphones are used.

The distribution of the diphones in the dictionaries is highly skewed, some diphones occurring thousands of times and some a handful. Hence the estimates of the probabilities of occurrence of diphones are subject to a large variance. This variance is calculating by modelling the distribution of diphones as a multinomial. In this case, the variance associated with diphone d_i is given by

$$V_i = \Pr(d_i)(1 - \Pr(d_i))/N \tag{7.13}$$

where N is the total number of diphones in both dictionaries and $\Pr(d_i)$ is approximated by its estimate,

$$\Pr(d_i) \approx \frac{\text{Number of occurrences of } d_i \text{ in both dictionaries}}{\text{Total number of diphones in both dictionaries}}. \tag{7.14}$$

In order to alleviate the problem of poor estimates of $\Pr(d_i)$ caused by infrequently occurring diphones (which could have spuriously high information associated with them), the variance of $I(d_i)$ is calculated as V_i and $I(d_i)$ is normalised by dividing by $\sqrt{V_i}$. Hence the normalised information for phoneme d_i is given by

$$I_n(d_i) = I(d_i)/\sqrt{V_i} \tag{7.15}$$

A high value for $I_n(d_i)$ implies that d_i supplies a high amount of information about the identity of the accent. It does not, however, tell us which accent is more likely should that diphone be output by the recogniser. Hence we define a new signed information value, $J(d_i)$ where

$$J(d_i) = \text{sgn}(\Pr(d_i|B) - \Pr(d_i|A))I_n(d_i). \tag{7.16}$$

$J(d_i)$ is positive for any diphone that occurs more frequently in British accented speech than in American and negative if the situation is reversed. Since the two dictionaries do not have identical phone sets, it was necessary to construct a new phone set to cover both sets of pronunciations. This was simply the union of the sets used in each dictionary. One disadvantage of this approach is that the method is highly reliant upon the labelling used in the construction of the dictionaries, and we may be modeling the method which was used to label a particular sound in a given dictionary , rather than the effects of the accent.

The 10 diphones which provide the maximum information for classification of each accent are given in Table 7.1 along with the percentage of information for the accent which they contribute.

The high information content associated with diphones such as /l/ /er/ and /n/ /er/ for American speech is due to the fact that most American accents are rhotic and these diphones appear ‘word final’. Most English accents are non rhotic and as such do not

Rank	American Diphone	Percentage info. from diphone	British Diphone	Percentage info. from diphone
1	/m/ /ah/	2.46	/sh/ /n/	5.18
2	/ah/ /n/	2.36	/y/ /uh/	2.65
3	/ah/ /l/	2.26	/ih/ /z/	2.38
4	/sh/ /ah/	1.99	/r/ /z/	2.28
5	/n/ /ah/	1.80	/uh/ /l/	2.25
6	/iy/ /ah/	1.78	/t/ /t/	1.98
7	/er/ /z/	1.34	/ay/ /z/	1.86
8	/l/ /er/	1.32	/ih/ /ae/	1.81
9	/n/ /er/	1.31	/b/ /l/	1.70
10	/d/ /er/	1.28	/t/ /ih/	1.70

Table 7.1: Diphones with highest information measure for classification of accent.

often have these diphones. The distinction in the dictionaries between words such as ‘accumulate’ ([/ah/ /k/ /y/ /uw/ /m/ /y/ /ah/ /l/ /ey/ /t/] in the American dictionary and [/ax/ /k/ /y/ /uw/ /m/ /y/ /uh/ /l/ /ey/ /t/] in the British) shows the reason for the high information content for the /ah/ /X/ diphones in American and /uh/ /X/ in British. This models the longer vowel sounds typically associated with an American ‘drawl’.

If the technique is successful, the output of the recogniser for American accented speech will consist of diphones with values of $J(d_i)$ which are mostly negative, and diphones with values of $J(d_i)$ which are mostly positive for British accented speech.

7.2.4 Classification

To classify the accent of an unknown speaker, a phone recogniser is trained on speech from speakers with both accents. The phone level label files used when training the models were generated from word level transcriptions of the training sentences, and the appropriate dictionary (Beep for WSJCAM0 and CMUDICT for WSJ1). Again it is possible that here we are modelling the differences between the dictionaries use of a particular symbol to represent a given sound, rather than actual differences between the accents.

The output of the recogniser for speech from the unknown speaker is concatenated into diphones. A sequential technique is then used to perform the accent classification — a decision on the speaker’s accent is made when at time T a score J_T is outside one of two thresholds. J_T is derived as follows: A null hypothesis $\mathcal{H}_0$ is proposed, that the speaker is “mid-Atlantic” i.e. that the frequency of his/her diphone usage is taken in equal proportions from American and British accented speech. Define:

$$I_k = J(d_{f(k)}) \tag{7.17}$$

where $f(k)$ gives the index of the k’th diphone in the sequence of diphones output by the recogniser. We accumulate the values of I_K such that at time T

$$C_T = \frac{1}{T}\sum_{k=1}^{T} I_k \tag{7.18}$$

Under $\mathcal{H}_0$, for a random sequence of diphones output by the recogniser, the expected value of C_T is the mean of J and the variance $Var(C_T)$ of C_T is given by

$$Var(J_T) = {\sigma_I}^2/T \tag{7.19}$$

where σ_I^2 is the variance of the set of values of $J(d_i)$. Hence if at time T, the value C_T is outside $\pm 2 * SD(J_T)$ where

$$SD(J_T) = \sqrt{Var(J_T)}, \tag{7.20}$$

then with 95% confidence, the accent is British if C_T is positive and American if C_T is negative.

Figure 7.1 shows the value of C_T for typical british and American accented sentences. The two 95% confidence thresholds (which follow a $1/\sqrt{T}$) curve) are shown as dotted lines. It can be seen that in the case of the American speaker, the lower threshold is exceeded after about 30 diphones have been processed indicating that the accent is American. For the British talker, the upper threshold is exceeded after about 20 diphones, indicating that the speaker is British. Classification is done by noting the duration for which the score C_T lies outside each of the two 95% confidence thresholds over the entire length of the test utterance. The accent is classified as the accent whose threshold was exceeded for the longest period.

7.3 Results

The technique was evaluated by classifying the speaker's accent after 1, 2, 3, ..., 8 sentences of speech had been processed. In practice, very few speakers produced values of C_T which lay outside *both* thresholds and the most commonly-observed behaviour was for C_T to exceed one of the thresholds and then remain outside it (as shown in Figure 7.1). However, if the score remained within the thresholds after all the diphones have been seen, the result is “unclassified”. The results in Figure 7.2 show that when there is only a small amount of data available, the technique may produce the result “unclassified” since the diphones observed do not contribute sufficient information to confidently identify the accent. After 3 sentences are available however, there are no unclassified or misclassified speakers.

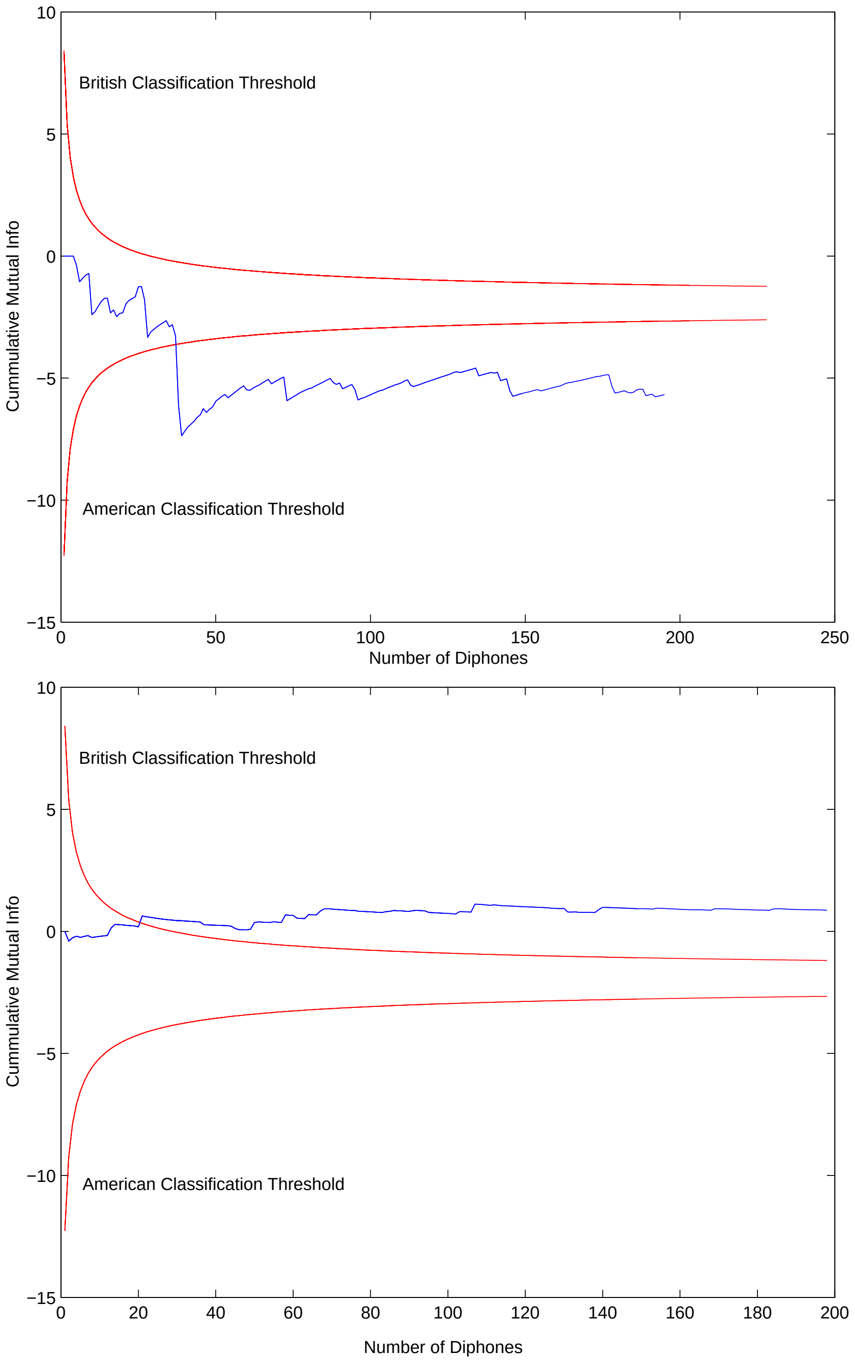


Figure 7.1: Value of C_T for (top) American-accented sentence and (bottom) British accented sentence.

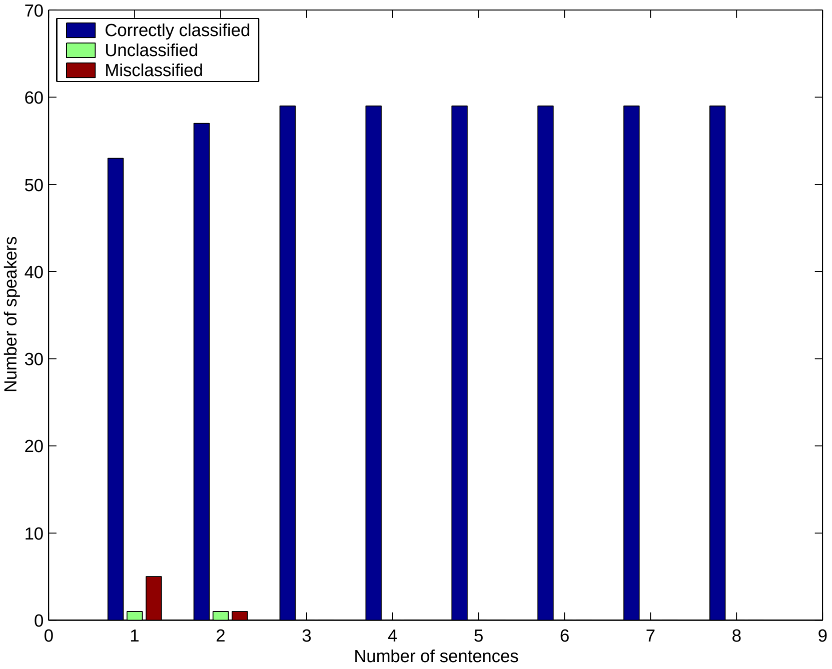


Figure 7.2: Results of accent classification using phonotactic models on original data.

As with the accent identification using clustering technique, it was possible that the accent identification being seen was simply the identification of 2 databases recorded under different conditions. Again, the technique was tested using the independent, American accented, TIMIT database. If differences between the databases were being modelled (rather than differences between the accents of the speakers within them) the classification performance would be significantly lower than those seen previously . The accent classification results are shown in Figure 7.3. The same pattern of initially “unclassified” results, followed by correct classification as more test data is made available is observed using the independent database. This indicates that the discrimination being demonstrated by the technique is indeed *accent* rather than *database* identification.

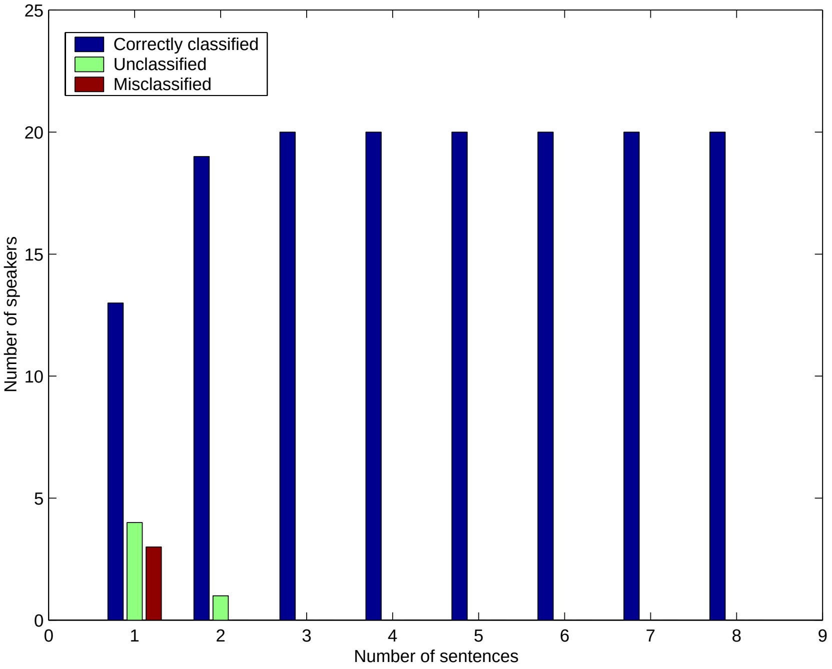


Figure 7.3: Results of accent classification using phonotactic models on TIMIT data

7.4 Conclusions

In this chapter a method of using the phonotactic differences between British English and American English as a method of accent classification was presented. The results showed good classification performance after only a small amount of data was available for classification. Importantly, classification performance was also maintained for an independent data set not used in the training procedure, indicating that genuine ‘accent’ identification was being performed rather than database identification. In comparison to the clustering procedure given in Section 6, this procedure has the advantage that it does not rely on a special recogniser topology (the clustering technique was based around a SCHMM). The output of *any* phoneme recogniser may be concatenated into diphones and used as the input to the classifier - as phoneme recognition performance improves, classification will require less data since the number of ‘information providing’ diphones will increase. The procedure is also computationally efficient, requiring only a simple ’lookup’ procedure to obtain the information value for the given diphone and an accumulation of the value of C_T. Extending the method to discriminate between more than two accents would require a change in classification strategy — while the information between a diphone and any accent may be calculated, a problem arises when weighting the information measure for more than two accents, since at present the information is made positive or negative depending on the more likely accent. This could be solved by associating the information measure for that diphone with the most likely accent, and accumulating separate scores for each accent. When one of the accent scores crosses a given threshold, the speaker may then be classified as being from that accent.

Chapter 8

Conclusion and Further Work

The aims of the thesis were to classify speakers based on characteristics of their speech, and to identify methods which may use this knowledge to improve the accuracy of automatic speech recognition systems. Since the methods were to be applicable to ‘interactive speech systems’ rather than ‘dictation systems’ further constraints were to be meet. Firstly, they must be computationally efficient, secondly they must be unsupervised, and finally, they must require very small amounts of adaptation data to provide improvements in accuracy.

8.1 Summary

After introducing the process of human speech production and the signal processing techniques currently used to extract information from the signal useful for the recognition task, the current preferred method of performing automatic speech recognition was discussed. This provided a general description of the environment in which any scheme for improving recognition accuracy would have to fit. A detailed description of the manner in which the speech from different talkers may vary was then given. These variations were identified as being either due to learned differences in speaking style, such as those due to geographical origin and social class, or physiological differences between speakers such as vocal tract length. These differences, if not accounted for in some manner, lead to a significant reduction in recognition accuracy.

8.1.1 Identifying and Compensating for Physiological Differences

Current techniques for compensating for physiological differences were shown to take two general forms - *speaker adaptation* in which the recognition model parameters are transformed to more closely match those of the speaker; and *speaker normalisation* in which the input speech (or, more precisely, the method of parameterisation) is altered in some way so as to make the parameters more closely match the correct model. The former has the problem that, due to the large number of parameters in the recognition models, a significant amount of data from the unknown speaker is required to generate the transforms. This would be un-acceptable in systems where the speaker is only using the system for a very limited time. The speaker normalisation techniques reviewed gave significant increases in performance without lengthy enrolment times, but most required an exhaustive search over some parameter space to identify the ‘best’ normalisation factor for a given speaker. Since the definition of ‘best’ frequently took the form of ‘highest recogniser output probability’, multiple recognition passes were usually performed to determine the correct parameter. This did not fit our requirement of ‘little computational overhead’ and so a method was sought to identify the normalisation parameter without an exhaustive search.

Chapter 5 describes the development of this technique — initially a method of normalising the input parameters based on transforming the frequency spectrum of the test speakers to those of a canonical speaker was presented. This gave significant improvements in a simple vowel classification task. The method was shown to effectively align the first formant of the test speaker to that of the canonical speaker, however it still required an exhaustive search over all normalisation factors to find the correct one. To overcome this, a method of estimating the formant locations and aligning the speakers directly in the LPC domain was developed and again shown to give significant improvement in a recognition task.

The problem of aligning to a single canonical reference speaker who may not have been representative of the speech as a whole was addressed next. A uni-variant distribution was calculated from estimates of the first and second formants for each vowel sound from many speakers. A normalisation was found which maximised the likelihood of the test speaker’s formants having been taken from these distributions. A closed form solution to the maximisation equation was derived, and a method of combining the normalisation factors for each speech frame into a single factor for each speaker was presented. This combination had the advantage that it accounted for inaccuracies in the formant picking algorithm by weighting normalisation values depending on how well the transformed formants fitted the appropriate distribution. Once calculated, the normalisation was used in two different ways. Initially, the pole locations of the LPCs of the test speaker for vowel sounds were shifted. The recognition accuracy improvements observed using this method were small since, despite vowel recognition accuracy improving, recognition accuracy for the contoid sounds decreased. In order to normalise all sounds, rather than just the vocoids, the method proposed in [43], that of moving the positions of the filter-bank, was used. Results showed better improvements in recognition accuracy however the method was still supervised since labelled speech data was required so as to identify which of the distributions to normalise the test speaker's formants to.

In order to overcome this, initially a speaker enrolment method was tested in which 2 sentences of speech was used to derive the normalisation factor. This showed similar performance to using all the test data. Finally, a two pass scheme was implemented in which an initial transcription generated from recognising the un-normalised data was used to calculate the normalisation factor. This was then applied and a second recognition pass made. This scheme again gave similar performance to using labelled speech.

8.1.2 Identifying and Compensating for Learned differences

Currently, techniques for compensating for learned differences in speech, particularly accent effects, rely on using a separate acoustic model set for each of the accents likely to be observed. At recognition time the correct model set is then used. This technique has shown to give good improvements for gross accent differences such as those between native and non-native speakers of a language. The method of identifying the correct model set to use is, however, often either computationally highly expensive, requiring the use of multiple recognisers running in parallel, or requires the user to utter a particular sentence designed to highlight accent differences. Again this is of little use in the task described above. To overcome these problems we have developed 2 methods of accent classification which do not rely on multiple recognisers or accent id utterances.

In Chapter 7 a technique based on modeling the phonotactics of the accents to be identified was proposed. The model was based on the mutual information between the occurance of a particular diphone and the accent, the information scores being calculated from accent specific pronunciation dictionaries. Information scores were accumulated and accent classification made dependent on the period for which a confidence threshold was exceeded. Classification performance was shown to be excellent for even limited amounts of input data (Figure 7.2) and also for an independent database (Figure 7.3).

In Chapter 6 a new data driven clustering technique was introduced which used semi continuous HMMs to identify clusters of talkers within the pattern space. The premise was that speakers with similar accents would cluster to similar regions within the pattern space, and at recognition time the test speaker could be assigned to one of the clusters and their accent identified. The technique was shown to perform well for identifying British and American accented English, again even for an independent database. The technique was then used to identify *regional* accents within these two groups. The method failed at this task, even after the effects of vocal tract variation had been removed. A recognition experiment showed however, that building dialect specific models for regional accents, and testing within group provided no improvement in recognition accuracy. This suggests that either the accent classifications provided with the databases are inaccurate, or that the variation within a regional accent group is as significant as across them. The latter is more likely to be the case since the accent groups cover a wide geographical area, and also contain speech from all social classes.

The clustering technique was then used to try and improve recognition accuracy by identifying clusters of similar talkers without reference to their accent. The premise was that building models for speakers with similar characteristics would reduce the variances within the model parameters and therefore increase recognition accuracy. Results showed a small improvement over the speaker independent case, despite the reduction in data for estimating the model parameters caused by splitting the training set.

8.2 Conclusions

The speaker normalisation scheme of Chapter 5 fulfills the requirements of low computational overhead by using information which may be efficiently generated as part of the parameterisation scheme (i.e. the formant locations) to calculate the normalisation factor. Using the first pass output of the recogniser to provide transcriptions for the adaptation meant that the normalisation was unsupervised, and it was shown that improvements in accuracy after 2 adaptation sentences had been processed were comparable to those gained using much greater amounts of adaptation data. The experiments do show that there is a cost in meeting these constraints. Computational efficiency was improved by calculating a normalisation factor for each speaker rather than each frame, and this was seen to significantly reduce the effectiveness of the procedure. In the future this may be overcome to some extent by the the fact that the increase in computational power of the systems running A.S.R. technology will allow the use of more computationally expensive methods. The move to unsupervised adaptation was also seen to reduce the improvements available, and it is unclear how to alleviate this problem — speakers are always likely to object to enrolment procedures.

The speaker clustering scheme presented in Chapter 6 was also unsupervised and shown to be able to differentiate between British and American talkers with only very small amounts of adaptation data, however it showed an inability to differentiate between regional accents. This may have been due to the fact that the method was being compared to the subjective decisions of a human listener about each speaker's accent. There is no reason to suppose that these decisions are consistent and accurate, or should correlate with clusters identified by a data driven approach such as this. It should also be noted that Wells [82] suggests that the effects of accent diminish with the style of speaking - the more formal or contrived the situation, the less we use accent specific pronunciations. It is difficult to imagine a more contrived situation for a member of the public than reading a list of Wall Street Journal sentences to a computer!

It is possible that the clusters may represent groups of speakers with similar speaking characteristics which are related to effects other than accent. The recognition experiments based on clusters of speakers generated by the procedure gave increases in recognition accuracy in excess of those obtained by clustering based on the annotated accent, indicating that this may indeed be the case. We must beware of discarding data driven techniques such as this, which may be useful in improving recognition accuracy, simply because they do not fit with the results we expect given our perception different speaking styles.

The accent identification method of Chapter 7 again met the requirements of low computational overhead, and was shown to accurately classify speakers after very small amounts of data had been processed. There was, however, some question as to whether the information used to identify the accent came directly from differences between the accents, or from differences between the way in which the dictionaries had been labelled for each accent. Care should be taken to ensure that what appears as the automatic identification of speech characteristics which we perceive to be similar (ie British and American accents) is not the identification of some other correlated effect such as the labelling of the two dictionaries.

8.3 Further Work

There are several ways in which the techniques may be extended or further investigation in a particular are made :

- The normalisation scheme currently only models the formant distributions as a single Gaussian. It is likely that using a *multiple mixture component* Gaussian distribution will give a better match to the observed data, and thereby improve the accuracy of the normalisation estimate.

- The results of the speaker enrolment experiment showed that estimating the normalisation factor using 2 labelled sentences provided similar improvements to using 10 sentences. An investigation should be made to determine how much labelled data is required to accurately estimate the normalisation factor.

- The normalisation method should be implemented in a real time recognition system to investigate whether the observed increase in recognition accuracy result in noticeable improvements in system performance.

- The speaker clustering scheme currently only associates a single Gaussian component with each model state. This could be extended to a *distribution*, better modeling the feature space occupied by each speaker and improving the clustering of similar talkers

- The phonotactic method should be extended to identify multiple accents using dictionaries generated from the methods described in [34].

- Both the clustering and phonotactic methods of accent classification should be validated on an independent British English database.

With the development of interactive speech systems proceeding at a tremendous rate, it is likely that speech recognition technology will be used in increasingly diverse situations, with ever more complicated tasks and larger numbers if users. As this occurs, if the shortcomings of present recognition technology are not to result in increasing numbers of frustrated users, the problems of recognising speech from diverse speaker populations must continue to be addressed, and techniques such as those presented here improved upon.

Appendix A

Speech Databases

A.1 British English Databases

A.1.1 WSJCAM0

The WSJCAM0 database is a British English equivalent of a subset of the American English WSJ0 database. It is a clean speech database recorded using two different microphones (one head mounted and one desk mounted) and sampled at 16kHz, 16 bits/sample. All included talkers are native English speakers, recruited from the Cambridge area of the U.K. (though this is not necessarily their regional accent). The database consists of a 92 speaker training set and a 48 speaker test set. This test set is then subdivided into two evaluation and one development set. Automatically aligned phone level transcriptions for all the sentences are provided in addition to word level transcriptions. Full details of the recording and transcription procedure are given in [24]

Table A.1 shows the gender distribution for each of the sets of talkers

	Number of Speakers		
Data Set	Male	Female	Total
Training	46	46	92
Development Test	10	8	18
Evaluation Test (1)	7	7	14
Evaluation Test (2)	7	7	14

Table A.1: Gender distribution of training and test sets in WSJCAM0

The talkers are each labelled with one of nine accent classifications shown in Table A.2

Northern
Southern
Eastern
Western
Midlands
Welsh
Scottish
Irish
other

Table A.2: WSJCAM0 Accent categories

A.1.2 Subscriber

Subscriber [73] is a British English database collected over the British telephone network and as such is subject to transmission effects from the network. The database consists of 1017 speakers split into a training and test sets. Age, accent and gender information for is speaker is recorded. The bandwidth is limited from 300Hz-3.4kHz and there is significant noise on many of the recordings. Also there is no control over the microphone used by the speaker (it is the one supplied with their telephone handset) or the telephone line they use when making the call to the automated recording system. The database is supplied with phonetic trascriptions for each of the utterances. Table A.3 shows the gender distribution for the training and test sets. The talkers in subscriber are also classified as having one of the nine accents shown in Table A.4. The accent classification is made based on the pronunciation of the two "Shiboleth" sentences included in every talkers prompting script.

	Number of Speakers		
Data Set	Male	Female	Total
Training	309	327	636
Test	187	194	381

Table A.3: Gender distribution of training and test sets in Subscriber

Dialect Region	Geographical Region
SBS	Southern British Standard (RP)
LON	London Area
R-WEST	West of England (Rhotic)
WAL	Wales
NB-LIV	Liverpool Area
NB	North of England
R-LANCS	Lancashire (Rhotic)
R-IRISH	Ulster (Rhotic)
R-SCOTS	Scotland (Rhotic)

Table A.4: Subscriber accent categories

A.2 American English Databases

A.2.1 TIMIT

The TIMIT database [4] is a clean speech (16KHz, 16 bit sampled, little or no background noise) database consisting of 6300 sentences read by American English talkers, 10 sentences each from 630 talkers. The database is subdivided into a training set, consisting of 502 talkers, and a test set of 128 talkers. The data is then subdivided into 8 dialect regions given in Table A.5. The gender split for both the training and the test set across each of the dialect regions is given in Table A.6

Dialect Region	U.S. Geographical Region
dr1	New England
dr2	Northern
dr3	North Midland
dr4	South Midland
dr5	Southern
dr6	New York City
dr7	Western
dr8	Army Brat (moved around)

Table A.5: Dialect regions in TIMIT Database

	Number of Speakers		
Data Set	Male	Female	Total
Training	366	136	502
Test	112	56	128

Table A.6: Gender distribution of training and test sets in TIMIT

A.2.2 WSJ1

The WSJ1 continuous speech recognition corpus is a clean speech (recorded with head mounted microphone in quiet office conditions) database. The speech is sampled at 16KHz, 16 bits/sample. The training data consists of 77800 utterances read by 245 speakers and the generic test set contains 8200 utterances read by 30 speakers. More detailed information on the database is given at [3].

Appendix B

Effect of LPC Transform in f1 - f2 Plane

Let the reference speaker's formant frequencies be f_1 and f_2, new speaker's formant frequencies be g_1 and g_2, transformed formant frequencies be h_1 and h_2. From 5.13

$$h_1 = ag_1 \tag{B.1}$$

and

$$h_2 = ag_2 \tag{B.2}$$

where

$$a = \frac{f_1 g_1 + f_2 g_2}{(g_1)^2 + (g_2)^2} \tag{B.3}$$

Letting $p = \frac{g_2}{g_1}$ gives

$$h_1 = \frac{f_1 + pf_2}{g_1 + pg_2} g_1 \tag{B.4}$$

and

$$h_2 = \frac{f_1 + pf_2}{g_1 + pg_2} g_2 \tag{B.5}$$

solving B.4 for g_1 gives

$$g_1 = \frac{ph_1 g_2}{f_1 + pf_2 - h_1} \tag{B.6}$$

substituting g_1 in B.5

$$\frac{ph_1 h_2}{f_1 + pf_2 - h_1} + ph_2 = f_1 + pf_2 \tag{B.7}$$

which may be simplified to

$$h_1 + ph_2 = f_1 + pf_2 \tag{B.8}$$

But

$$\frac{h_2}{h_1} = \frac{ag_2}{ag_1} = p \tag{B.9}$$

Hence

$$\begin{aligned} h_1 + \frac{(h_2)^2}{h_1} &= f_1 + \frac{h_2}{h_1} f_2 \qquad \text{(B.10)} \\ (h_1)^2 + (h_2)^2 &= f_1 h_1 + f_2 h_2 \qquad \text{(B.11)} \end{aligned}$$

let $h_1 = r \sin\alpha$ and $h_2 = r\cos\alpha$, therefore $r^2 = (h_1)^2 + (h_2)^2$. Substituting in B.11 gives

$$\begin{aligned} r^2 &= f_1 r \sin\alpha + f_2 h_2 \qquad \text{(B.12)} \\ h_2 &= \frac{r}{f_2}\left(r - f_1 \sin\alpha\right) \qquad \text{(B.13)} \end{aligned}$$

similarly

$$h_1 = \frac{r}{f_1}(r - f_2 \cos\alpha). \tag{B.14}$$

Substituting $r = \frac{h_2}{\cos\alpha}$ in B.13 and simplifying leads to :

$$h_2 = \cos\alpha(f_2\cos\alpha + f_1\sin\alpha) \tag{B.15}$$

Similarly, in B.14

$$h_1 = \sin\alpha(f_2\cos\alpha + f_1\sin\alpha) \tag{B.16}$$

But,

$$f_2\cos\alpha + f_1\sin\alpha \equiv R\sin\alpha + \theta \tag{B.17}$$

where $R = \sqrt{(f_1)^2 + (f_2)^2}$ and $\tan\theta = \frac{f_2}{f_1}$. Hence :

$$h_1 = \sin\alpha R\sin(\alpha + \theta) \tag{B.18}$$

$$h_2 = \cos\alpha R\sin(\alpha + \theta) \tag{B.19}$$

using $\sin A\sin B = \frac{1}{2}[\cos(A - B) - \cos(A + B)]$:

$$h_1 = \frac{R}{2}\left[\cos\theta - \cos(2\alpha + \theta)\right] \tag{B.20}$$

but $R\cos\theta = f_1$, hence :

$$h_1 = -\frac{R}{2}\cos(2\alpha + \theta) + \frac{f_1}{2} \tag{B.21}$$

Using $\sin A\cos B = \frac{1}{2}[\sin(A + B) + \sin(A - B)]$

$$h_2 = \frac{R}{2}[\sin(2\alpha + \theta) + \sin\theta] \tag{B.22}$$

but $R\sin\theta = f_2$, hence

$$h_2 = \frac{R}{2}\sin(2\alpha + \theta) + \frac{f_2}{2} \tag{B.23}$$

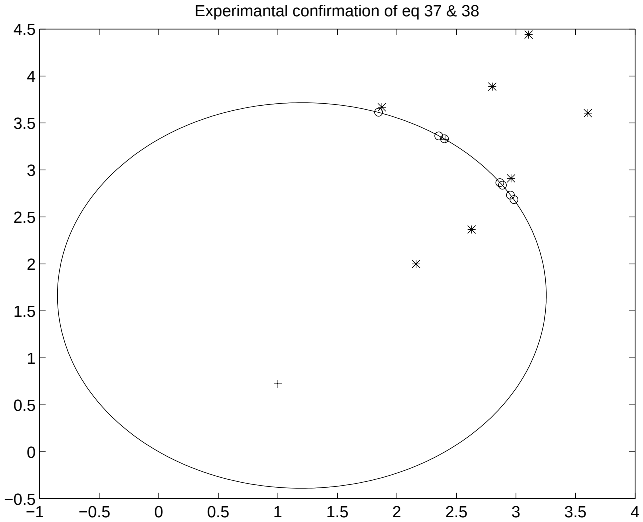


Figure B.1: Experimental confirmation of ellipse.

Equations B.21 and B.23 are a parametric form of equation of an ellipse with axes given by :

$$\frac{1}{2}\sqrt{(f_1)^2 + (f_2)^2} + \frac{f_1}{2} \tag{B.24}$$

$$\frac{1}{2}\sqrt{(f_1)^2 + (f_2)^2} + \frac{f_2}{2} \tag{B.25}$$

This is confirmed experimental by Figure B.1 which shows some randomly generated f1 - f2 pairs normalised to a reference speaker, and also the ellipse defined by the reference speaker's formants. The transformed poles clearly lie on the locus of the ellipse.

Bibliography

[1] The British English Example Pronunciation (BEEP) dictionary is available from ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/.

[2] Available from http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[3] Information available at http://www.itl.nist.gov/iaui/894.01/corpora/wsj1.htm.

[4] *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus - Reference Manual*.

[5] V. Abrash, A. Sankar, H. Franco, and M. Cohen. Acoustic adaptation using non-linear transformations of HMM parameters. In *Proc. ICASSP'96*, pages 729–732, Atlanta, GA, May 1996.

[6] H. Andrews. *Introduction To Mathematical Techniques in Pattern Recognition*. Robert E. Krieger, 1983.

[7] L. Arslan and J. Hansen. Language accent classification in American English. *Speech Communication*, 18:353–367, 1996.

[8] W. Barry, C. Hoequist, and F. Nolan. An approach to the problem of regional accent in automatic speech recognition. *Computer Speech and Language*, pages 355–366, 1989.

[9] P. Brown, C. Lee, and J. Spohrer. Bayesian adaptation in speech recognition. In *Proc. ICASSP'83*, pages 761–764, 1983.

[10] F. Brugnara and M. Federico. Techniques for approximating a trigram language model. In *Proc. ICSLP'96*, volume 4, pages 2075–2078, Philadelphia, PA, Oct. 1996.

[11] D. C. Burnett and M. Fanty. Rapid unsupervised adaptation to children's speech on a connected-digit task. In *Proc. ICSLP'96*, volume 2, pages 1145–1148, Philadelphia, PA, Oct. 1996.

[12] C. Corredor-Ardoy, J. L. Gauvain, M. Adda-Decker, and L. Lamel. Language identification with language-independent acoustic models. In *Proc. Eurospeech'97*, pages 55–58, Rhodes, Greece, Sept 1997.

[13] S. Cox. Speaker normalisation in the MFCC domain. Private Communication, 1998.

[14] S. J. Cox. Hidden markov models for automatic speech recognition. *British Telecom Technical Journal*, 6(2):105–115, April 1988.

[15] S. J. Cox. Predictive speaker adaptation in speech recognition. *Computer Speech and Language*, 9:1–17, 1995.

[16] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics Speech and Signal Processing*, 28:357–366, Aug 1980.

[17] C. de la Torre, J. Caminero-Gil, J. Alvarez, C. M. del Alamo, and L. Hernández-Gómez. Evaluation of the Telefónica I+D natural numbers recogniser over different dialects of Spanish from Spain and America. In *Proc. ICSLP'96*, volume 4, pages 2032–2035, Philadelphia, PA, Oct 1996.

[18] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan, 1993.

[19] E. Eide and H. Gish. A parametric approach to vocal tract length normalization. In *Proc. ICASSP'96*, Atlanta, GA, May 1996.

[20] J. Flege. The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, 76(3):692–707, Sept 1984.

[21] J. Flege. Factors affecting degree of percieved foreign accent in English sentences. *Journal of the Acoustical Society of America*, 84(1):70–79, July 1988.

[22] J. Flege and K. Fletcher. Talker and listener effects on degree of percieved foreign accent. *Journal of the Acoustical Society of America*, 91(1):370–389, Jan 1992.

[23] E. Frangoulis. A novel speaker adaptation approach for continuous density HMMs. In *Proc. ICASSP'91*, pages 861–864, May 1991.

[24] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young. WSJCAM0 corpus and recording description. Technical Report CUED/F-INFENG/TR192, Cambridge University Engineering Department, 1994.

[25] S. Furui. Speaker independent isolated word recognition using dynamic features of the speech spectrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34:52–59, Feb 1986.

[26] S. Furui. Unsupervised speaker adaptation method based on hierarchical spectral clustering. In *Proc. ICASSP'89*, volume 1, pages 286–289, 1989.

[27] M. Gales and P. Woodland. Mean and variance adaptation within the MLLR framework. *Computer Speech and Language*, 10:249 – 264, 1996.

[28] N. Harte, S. V. Vaseghi, and B. Milner. Dynamic features for segmental speech recognition. In *Proc. ICSLP'96*, volume 2, pages 933–936, Philadelphia, PA, Oct. 1996.

[29] T. J. Hazen and V. Zue. Segment-based automatic language identification. *Journal of the Acoustical Society of America*, 18(4):2323–2331, Apr. 1997.

[30] J. Holmes, W. Holmes, and P. Garner. Using formant frequencies in speech recognition. In *Proc. Eurospeech'97*, pages 2083–2086, Rhodes, Greece, Sept 1997.

[31] X. Huang. Phoneme classification using semicontinuous hidden Markov models. *IEEE Transactions on Signal Processing*, 40(5):1062–1067, 1992.

[32] X. D. Huang, K. F. Lee, and H. W. Hon. On semi-continuous hidden-Markov modeling. In *Proc. ICASSP'90*, pages 689–692, Alburquerque, NM, Apr. 1990.

[33] A. Huggins and Y. Patel. The use of shibboleth words for automatically classifying speakers by dialect. In *Proc. ICSLP'96*, volume 4, pages 2017–2020, Philadelphia, PA, Oct 1996.

[34] J. Humphries and P. Woodland. Using accent-specific pronunciation modelling for improved large vocabulary speech recognition. In *Proc. Eurospeech'97*, pages 2367–2370, Rhodes, Greece, Sept 1997.

[35] F. Johnasen and M. Johnsen. Non-linear input transformations for discriminative HMMs. *Proceedings ICASSP'94*, I:225 – 228, April 1994.

[36] E. Kamen and B. Heck. *Fundamentals of Signals and Systems Using Matlab*. Prentice Hall, 1997.

[37] D. Kewley-Port and Y. Zheng. Auditory models of formant frequency discrimination for isolated vowels. *Journal of the Acostical Society of America*, 103(3):1654–1666, 1998.

[38] K. Kumpf and R. King. Automatic accent classification of foreign accented Australian speech. In *Proc. ICSLP'96*, pages 1740–1743, Philadelphia, PA, Oct 1996.

[39] P. Ladefoged. *A Course in Phonetics*. Harcourt Brace, 1993.

[40] C. Lee, C. Lin, and B. Juang. A study on speaker adaptation of the parameters of continuous density hidden Markov models. *IEEE Transactions on Signal Processing*, 39(4):806–813, April 1991.

[41] K. Lee. *Automatic Speech Recognition—the Development of the SPHINX System*. Kluwer Academic Publishers, 1989.

[42] L. Lee and R. Rose. A frequency warping approach to speaker normalization. *IEEE Transactions on Speech and Audio Processing*, 6(1):49–60, Jan 1998.

[43] L. Lee and R. C. Rose. Speaker normalization using efficient frequency warping procedures. In *ICASSP'96*, 1996.

[44] C. Leggetter and P. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, Apr. 1995.

[45] S. Levinson, L. Rabiner, and M. Sondhi. An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell System Technical Journal*, 62(4):1035–1074, Apr 1983.

[46] L. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions on Information Theory*, 28(5):729–734, 1982.

[47] J. Makhoul, S. Roucos, and H. Gish. Vector quantization in speech coding. *Proceedings IEEE*, 73(11):1551–1588, Nov 1985.

[48] Y. S. Masaki Naito, Li Deng. Speaker clustering for speech recognition using the parameters characterizing vocal-tract dimensions. In *Proc. ICASSP'98*, volume 2, pages 1889–1893, 1998.

[49] I. Matthews. *Features for Audio Visual Speech Recognition*. PhD thesis, University of East Anglia, 1998.

[50] D. R. Miller and J. Trischitta. Statistical dialect classification based on mean phonetic features. In *Proc. ICSLP'96*, volume 4, pages 2025–2027, Philadelphia, PA, Oct. 1996.

[51] J. D. Miller. Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, 85(5):2114 – 2134, May 1998.

[52] B. Milner. A generalised approach for the inclusion of temporal information into features for speech recognition. *Proc. IOA Autumn Conference*, 2:217–224, 1996.

[53] B. Milner and S. Vaseghi. Speech modelling using cepstral-time feature matrices and hidden Markov models. In *Proc. ICASSP'94*, volume I, pages 601–604, Adelaide, Austrailia, Apr. 1994.

[54] R. Moore. Recognition—the stochastic modelling approach. In C. Rowden, editor, *Speech Processing*. McGraw-Hill, 1992.

[55] J. O'Connor. *Phonetics*. Penguin Books, 1973.

[56] J. Odell, D. Ollason, V. Valtchev, and D. Whitehouse. *The HAPI Book*. Entropic Cambridge Research Laboratory, 1997.

[57] D. O'Shaughnessy. *Speech Communication - Human and Machine*. Addison Wesley, 1987.

[58] F. Owens. *Signal Processing Of Speech*. Macmillan New electronics. Macmillan, 1993.

[59] M. Padmanabhan, L. Bahl, D. Nahamoo, and M. Picheny. Speaker clustering and transformation for speaker adaptation in large-vocabulary speech recognition systems. In *Proc. ICASSP'96*, pages 701–704, Atlanta, GA, May 1996.

[60] A. Paeseler and H. Ney. Continuous speech recognition using a stochastic language model. In *Proc. ICASSP'89*, pages 719–722, May 1989.

[61] A. Paige and V. W. Zue. Calculation of vocal tract length. *IEEE Transactions on Audio and Electroacoustics*, AU-18(3):268 – 270, Sept 1969.

[62] G. Peterson. Parameters of vowel quality. *Journal of Speech and Hearing Research*, 4:10–29, 1961.

[63] K. Power. The listening telephone—automating speech recognition over the PSTN. *British Telecom Technology Journal*, 14(1):112–126, 1996.

[64] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in 'C'*. Cambridge University Press, 1996.

[65] J. G. Proakis and D. G. Manolakis. *Introduction to Digital Signal Processing*. Macmillan, 1988.

[66] G. Pullum and W. Ladusaw. *Phonetic Symbol Guide*. University of Chicago press, 1986.

[67] L. Rabiner and B. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1992.

[68] L. Rabiner, S. Levinson, and M. Sondhi. On the application of vector qauntization and hidden markov models to speaker-independent isolated word recognition. *The Bell Systems Technical Journal*, 62(4):1075–1105, 1983.

[69] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, February 1989.

[70] D. Rose. Official social classification in the U.K. *Social Research Update*, July 1995. Available at http://www.soc.surrey.ac.uk/sru/SRU9.html.

[71] D. Rtischev, D. Nahamoo, and M. Picheny. Speaker adaptation in a large-vocabulary speech recognizer via vq prototype modification. Technical report, IBM Research Division, July 1989.

[72] F. Schiel. A new approach to speaker adaptation by modelling pronunciation in automatic speech recognition. *Speech Communication*, 13:281–285, 1993.

[73] A. Simons and K. Edwards. Subscriber - a phonetically annotated telephony database. *Proceedings of the Institute of Acoustics*, 14(6):9–16, 1992.

[74] J. Slifka and T. R. Anderson. Speaker modification with LPC pole analysis. In *Proc. ICASSP '95*, pages 644–647, Detroit, MI, May 1995.

[75] C. Teixeira, I. Tancoso, and A. Serralheiro. Recognition of non-native accents. In *Proc. Eurospeech'97*, pages 2375 – 2378, Rhodes, Greece, Sept 1997.

[76] C. Tuerk. *Automatic Speech Synthesis Using Auditory Transforms and Artificial Neural Networks*. PhD thesis, Cambridge University, 1992.

[77] D. Van-Compernolle, J. Smolders, P. Jaspers, and T. Hellmans. Speaker clustering for dialectic robustness in speaker independent recognition. In *Proc. Eurospeech'91*, pages 723 – 726, 1991.

[78] J. van der Lubbe. *Information Theory*. Cambridge university press, 1997.

[79] H. Wakita. Normalization of vowels by vocal-tract length and its application to vowel identification. *IEEE Transactions on Acoustics, Speech and Signal Processing.*, 25:183–192, Apr 1977.

[80] S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalisation on conversational telephone speech. In *Proc. ICASSP'96*, volume 1, pages 339–343, Atlanta, GA, May 1996.

[81] L. Welling and H. Ney. A model for efficient formant estimation. In *Proc. ICASSP'96*, pages 797–800, Atlanta, GA, May 1996.

[82] J. Wells. *Accents of English 1*. Cambridge University Press, 1995.

[83] D. Wen, N. Campbell, and H. Norio. Fast and robust joint estimation of vocal tract and voice source parameters. In *Proc. ICASSP'97*, pages 1291–1294, Munich, Germany, Apr. 1997.

[84] Y. Yan, E. Barnard, and R. Cole. Development of an approach to automatic language identification based on phone recognition. *Computer Speech and Language*, 10:37—54, 1996.

[85] S. Young. Large vocabulary speech recognition. *Accoustics Bulletin*, pages 5 – 12, September / October 1995.

[86] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University Technical Services Ltd., 1997.

[87] P. Zhan and M. Westphal. Speaker normalization based on frequency warping. In *Proc. ICASSP'97*, pages 1039–1042, Munich, Germany, Apr. 1997.

[88] R. Ziemer, W.Tranter, and D. Fannin. *Signals and systems : Continuous and discrete*. Macmillan, third edition, 1993.

[89] M. A. Zissman. Automatic language identification using gaussian mixtures and hidden Markov models. In *Proc. ICASSP'93*, volume 2, pages 399–402, Apr. 1993.

[90] M. A. Zissman. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing*, 4(1):31–44, January 1996.

[91] E. Zwicker. Subdivision of the audible frequency band into critical bands (frequenzgruppen). *Journal of the Acoustical Society of America*, 33:248–260, 1961.