PHONEME DURATION MODELLING FOR SPEAKER VERIFICATION

CHARL JOHANNES VAN HEERDEN

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

PHONEME DURATION MODELLING FOR SPEAKER VERIFICATION

By

Charl Johannes van Heerden

Submitted in partial fulfilment of the requirements for the degree

Master of Engineering (Computer)
in the
Faculty of Engineering, the Built Environment and Information Technology
at the
UNIVERSITY OF PRETORIA

Advisor: Professor E. Barnard

April 2008

SUMMARY

PHONEME DURATION MODELLING FOR SPEAKER VERIFICATION

by

Charl Johannes van Heerden

Advisor: Professor E. Barnard

Department of Electrical, Electronic and Computer Engineering

Master of Engineering (Computer)

Higher-level features are considered to be a potential remedy against transmission line and cross-channel degradations, currently some of the biggest problems associated with speaker verification. Phoneme durations in particular are not altered by these factors; thus a robust duration model will be a particularly useful addition to traditional cepstral based speaker verification systems. In this dissertation we investigate the feasibility of phoneme durations as a feature for speaker verification.

Simple speaker specific triphone duration models are created to statistically represent the phoneme durations. Durations are obtained from an automatic hidden Markov model (HMM) based automatic speech recognition system and are modeled using single mixture Gaussian distributions. These models are applied in a speaker verification system (trained and tested on the YOHO corpus) and found to be a useful feature, even when used in isolation. When fused with acoustic features, verification performance increases significantly.

A novel speech rate normalization technique is developed in order to remove some of the inherent intra-speaker variability (due to differing speech rates). Speech rate variability has a negative impact on both speaker verification and automatic speech recognition. Although the duration modelling seems to benefit only slightly from this procedure, the fused system performance improvement is substantial.

Other factors known to influence the duration of phonemes are incorporated into the duration model. Utterance final lengthening is known be a consistent effect and thus "position in sentence" is modeled. "Position in word" is also modeled since triphones do not provide enough contextual information. This is found to improve performance since some vowels' duration are particularly sensitive to its position in the word.

Data scarcity becomes a problem when building speaker specific duration models. By using information from available data, unknown durations can be predicted in an attempt to overcome the data scarcity problem. To this end we develop a novel approach to predict unknown phoneme durations from the values of known phoneme durations for a particular speaker, based on the maximum likelihood criterion. This model is based on the observation that phonemes from the same broad phonetic class tend to co-vary strongly, but that there is also significant cross-class correlations. This approach is tested on the TIMIT corpus and found to be more accurate than using back-off techniques.

Keywords: speaker verification, phoneme durations, duration modeling, prosodic features, hidden Markov models, Gaussian mixture models, eigen vectors, maximum likelihood, speech rate normalization.

OPSOMMING

FONEEM LENGTE MODELLERING VIR SPREKER VERIFIKASIE

deur
Charl Johannes van Heerden
Adviseur: Professor E. Barnard
Departement Elektriese, Electroniese en Rekenaar-Ingenieurswese
Meester in Ingenieurswese (Rekenaar)

Hoër-vlak kenmerke work beskou as ’n moontlike oplossing vir transmissie lyn - en kruis kanaal effekte, wat tans van die vernaamste probleme is wat met spreker verifikasie verbind word. Foneem lengtes in besonder word nie deur hierdie faktore beïnvloed nie; dus sal ’n robuuste lengte model ’n handige toevoeging wees vir tradisionele akkoestiese spreker verifikasie stelsels. In hierdie verhandeling ondersoek ons die moontlikheid van foneem lengtes as ’n kenmerk vir spreker verifikasie.

Eenvoudige spreker spesifieke trifoon modelle word geskep om die foneem lengtes statisties voor te stel. Die lengtes word verkry vanaf ’n versteekte Markov model gebasseerde automatiese spraak herkenningstelsel en word gemoddelleer deur enkel mengsel Gauss verspreidings. Hierdie modelle word dan prakties toegepas in ’n spreker verifikasie stelsel (opgelei en getoets op die YOHO korpus) en daar word gevind dat foneem lengtes ’n handige kenmerk is, self al word dit as die enigste kenmerk gebruik. Wanneer hierdie kenmerk egter met akkoestiese kenmerke gekombineer word, neem die spreker verifikasie akkuraatheid heelwat toe.

’n Nuwe spraak tempo normalisering tegniek word ook ontwikkel met die doel om van die intra-spreker variansie (as gevolg van verskillende spraak tempos) te verwyder. Variërende spraak tempos het ’n negatiewe impak op beide spreker verifikasie asook spraak herkenning. Alhoewel dit blyk dat tempo normalisering die lengte kenmerke min verbter, is die verbetering in die gekombineerde stelsel veelseggend.

Ander faktore wat foneem lengtes beïnlvoed word ook in die model geïnkorporeer. Dit is welbekend dat uiting finale verlenging ’n konstante effek is en dus word “posisie in die sin” in ag geneem. “Posisie in die woord” word ook voorgestel aangesien trifone nie genoeg kontekstuele inligting verskaf nie. Daar word gevind dat die stelsel heelwat beter vaar as gevolg van die feit dat veral sekere vokale se lengtes baie sensitief is ten opsigte van hul posisie in ’n woord.

Data skaarsheid word ’n probleem wanneer spreker spesifieke lengte modelle gebou word. Die data skaarsheid probleem kan tot ’n mate aangespreek word deur onbekende foneem lengtes te skat vanaf beskikbare data. Om hierdie rede ontwikkel ons dus ook ’n nuwe benadering, gebasseer op die maksimum waarskynlikheids metode, ten opsigte van voorspelling van onbekende foneem lengtes deur lengtes van bekende foneme te gebruik. Die model is gebasseer op die opmerking dat foneme van dieselfde breë fonetiese klas geneig is om sterk te kovarieër, maar dat daar ook beduidende tussen-klas korrelasies bestaan. Hierdie benadering word getoets op die TIMIT korpus en daar word gevind dat hierdie metode onbekende lengtes beter kan voorspel as huidige “terugval” metodes.

Kernwoorde: spreker verifikasie, foneem lengtes, lengte modellering, prosodiese kenmerke, versteekte Markov modelle, Gauss mengsel modelle, eievektore, maksimum waarskynlikheid, spraak tempo normalisering.

ACKNOWLEDGMENTS

This research was made possible by the generous studentship awarded to me by the Human Language Technologies (HLT) Research group at the Meraka Institute. The guidance and supervision I received from Professor Etienne Barnard and Dr. Marelie Davel was instrumental in enabling me to complete this work.

The results of the first frame-level cepstral speaker verification system I built was improved significantly when I conducted the same tests on YOHO using the Gaussian Mixture Model system developed by Neil Kleynhans during his Masters studies. The subsequent work we did together with an HLT intern, Reinier Coetzer enabled us to improve this system to be close to, if not the best GMM system yet tested on YOHO. Their hard work and enthusiastic support is greatly appreciated.

Last but definitely not least, the support of Albert Gazendam from the CSIR's Opteron Cluster Computing Center must be mentioned. Without the use of C4, the experiments I conducted on YOHO would have taken several years. I was on many occasions one of the most frequent users of C4's CPU time and Albert was always willing to assist.

TABLE OF CONTENTS

CHAPTER ONE

INTRODUCTION

1.1 CURRENT SPEAKER VERIFICATION PROBLEMS

Speaker verification (SV) is a biometric access control technique where a user's voice is used to classify him as being either who he claimed to be or an impostor. It is a valuable biometric for several reasons; users consider it non-intrusive compared to other more accurate biometrics such as iris or retinal scanning [1], it is the most natural way for humans to communicate and it is a commercially viable biometric since the infrastructure for large scale implementation is already in place (telecommunications).

There are significant problems that limit the accuracy and subsequent large scale implementation of speaker verification though. These problems arise largely due to the traditional features (cepstral/acoustic) that have been used for SV. Transmission line degradations are probably the single most detrimental factor preventing the high accuracies obtained on clean (microphone recorded) speech from being obtained on telephone speech [2]. Performance is also negatively influenced by cross-channel degradation which occurs when enrollment and verification takes place on different channels [3]. The possibility of recording attacks [4], where an impostor replays a previously recorded voice sample from a legitimate speaker, is also a big concern. Finally intra-speaker variability [5], some sources of which are discussed in section 2.3 complicates the verification process.

1.2 HIGHER-LEVEL FEATURES AS A POSSIBLE SOLUTION

A possible solution to all of the above mentioned problems is to incorporate so called "higher-level" features into the verification process, in particular phoneme durations. Phoneme durations are much less susceptible to transmission line and cross-channel degradation than traditional acoustic or cepstral features. Unless the signal to noise ratio (SNR) is so low that audibility is impaired, the physical duration of the pronounced phoneme will not change when spoken over a telephone or channel different from that used during enrollment.

Much research has been done on these higher-level features and will be briefly discussed here as an in depth overview can be found in chapter 2. The use of higher-level features for speaker verification started receiving considerable interest when NIST held a speaker recognition workshop in 1998 [5].

Speech rate normalization (SRN) was investigated in [6],[7] but has not been addressed satisfactorily. This research was also conducted along automatic speech recognition (ASR) lines and it remains to be seen how it could benefit speaker verification. Phoneme durations were also modeled in [8],[9],[10],[11]. Data scarcity, which poses a big problem to the use of prosodic information [12], was addressed by backing off to more general models (eg from triphone to monophone models). We want to argue that accurate phoneme prediction can be used as a more sophisticated backoff technique that will ultimately improve overall system performance.

Phoneme durations is not yet an established feature that can be robustly incorporated into any speaker verification system. To this end, accurate duration models need to be developed that take all the interacting factors which influence phoneme durations [13] into account. Also, rather than backing off to more general models, data scarcity needs to be addressed by utilizing existing information to predict unknown durations.

We propose that accurate duration modeling which incorporates factors mentioned in [13] will improve speaker verification performance. Furthermore we believe that performing SRN will improve SV since some intra-speaker variability is removed. In addition it will be interesting to see whether known phoneme durations could be utilized to predict unseen durations such that these predicted durations are a closer match than current back-off techniques.

1.3 OVERVIEW OF DISSERTATION

The goal of this dissertation is to analyze the utility of phoneme durations as a feature in text-dependent speaker verification, to determine if speech rate normalization can be useful in the SV process and finally to develop a model able to predict unseen phoneme durations such that they are more accurate than the traditional back-off approach.

The results of research conducted towards these goals are reported in this dissertation and will be presented as follows:

- A comprehensive overview of speaker verification and what has been achieved with regard to incorporating higher-level features in particular are given in chapter 2.
- Chapter 3 discusses the baseline cepstral system that was built. A novel duration model and its application to the speaker verification process is also discussed.
- A novel SRN procedure is proposed in chapter 4. Results obtained by applying this SRN on the YOHO corpus is also reported.
- Factors known to influence phoneme durations were incorporated in the duration model. Chapter 5 reports on this refinement of the model.
- Chapter 6 discusses introductory work done in building a model that could predict unseen phoneme durations.

CHAPTER TWO

BACKGROUND

2.1 INTRODUCTION

Biometric user identification is a term used to define the use of unique intrinsic physical human traits to recognize or identify human beings from a database of known identities. Speaker verification is such a biometric access control technique where a speaker claims to be a specific identity. In this case, the speaker's voice is used to classify him as either the claimed identity or an impostor.

In order to perform speaker verification, features need to be extracted from the speaker's voice. Typically, one needs enrollment data (data to generate models for a speaker) and testing data (data to evaluate the accuracy of the models generated with the training data). There are many existing algorithms for feature extraction, each with its own advantages and disadvantages. In order to understand the existing features and design and implement a new feature, a broad overview and understanding of speaker verification in general is needed. This chapter provides an overview of the literature related to the main aspects addressed in this dissertation. In particular, the following fields will be covered:

- Section 2.2 discusses the different categories and related aspects of speaker verification as well as the current trend of this field
- Section 2.3 gives some insight into what has already been done with regard to duration modelling and how it has been applied to improve state of the art speaker recognition systems

- Section 2.4 provides an overview of other so-called “higher level features”, generally consisting of various types of prosodic information.

2.2 THEORY AND CONCEPTS

2.2.1 BIOMETRIC USER IDENTIFICATION

As mentioned above, biometric user identification can be defined as a technique for identifying or verifying the identity of a human being from some distinguishing human characteristic or trait.

There are several different techniques for user identification. The two techniques considered most accurate are retinal and iris scanning. Although very accurate, these techniques are considered intrusive and are not widely accepted by the public [1]. Speech, on the other hand, is considered non-intrusive by users, but has the drawback of being less accurate than the former two methods.

Biometric identification comprises two main fields: user identification and user verification. Identification also comprises two fields, namely open-set and closed-set identification. In the closed-set scenario a user will always be granted access. Such systems are typically used in a setup where everyone who uses the system is trusted and impostors are not considered a threat. The user will simply be identified as the entity in the database to which he is most similar. In open-set identification, a user may be rejected as the system considers the possibility of impostors. Verification on the other hand is the process whereby a user claims to be an identity by using for example a smart card or a pin number. The system then decides by some threshold value if he is indeed the claimed identity or an impostor. Typically, a score is awarded to the claimant based on a match to the claimed model. If the score is above/below a predetermined threshold, the user is accepted/rejected.

2.2.2 SPEECH IN SECURITY APPLICATIONS

Speech is an acoustic wave or signal that is transformed at several levels to produce a distinguishing biometric feature. The transformations occur at semantic, linguistic, articulatory and acoustic levels [14].

As mentioned above, speech-recognition is considered a non-intrusive biometric verification method. For a system to be a commercial success it has to be accepted by the public [1] and thus speech is one of the most promising features for large-scale implementation in providing access control to sensitive computer and communication systems.

Another reason speech is considered to be commercially viable, especially in telephone banking, is that the infrastructure for collecting speech samples, such as telephones, is already available and in place. The implication is that the cost of a speaker verification system may only be in the software, as the hardware is already in place.

2.2.3 SPEAKER VERIFICATION

Speaker verification is a one-to-one mapping between a user's voice and a claimed identity's voice. The user typically indicates the identity he claims to be by entering a pin or using a smart card [14]. A user would have to go through an enrolment session during which a model will be created representing his voice. Several different approaches to creating a model of a speaker's voice exist and will be discussed later.

A typical speaker verification system consists of five steps [14]; digital speech data acquisition, feature extraction, pattern matching, making an accept/reject decision and an enrolment session.

In order to be implemented in high security access control systems, speaker verification has to be robust against errors. Consequently most of the research over the past decade has been focused on improving speaker verification systems to be robust against some factors known to be detrimental to speaker verification accuracy; ambient room noise, voice reflections, fading over telephone channels and differing emotional states of users [15].

Another important aspect of speaker verification systems mentioned by [15] is the types of errors they make. These constitute FA (false acceptance) and FR (false rejection) errors. FA errors are the number of users falsely accepted by the system while FR errors are the number of users falsely rejected. The two types of errors typically display an inverse relationship to each other. They occur because of a partial overlap between the probability density functions of the models of different speakers and can thus not be eliminated completely. The decision of how much of either one to allow is system-dependent and is physically realized by setting a decision threshold.

2.2.3.1 CLASSES OF SV SYSTEMS

Several different approaches are used in speaker verification, based on what the user is required to say in order to be verified. The different pattern-matching methods will be discussed later, but it is worth mentioning here that text-independent systems perform best when GMM (Gaussian mixture models) are used while text-dependent systems perform better with HMMs (Hidden Markov models) [16].

- *Text-dependent systems.* In a text-dependent system, the user is required to say a predetermined phrase such as a pin or password. The phrase to be said is already stored in a database to be matched [17]. A user will need to be cooperative in order to be verified.

- *Text-independent systems.* Text-independent systems are systems where the user can say what he wants in order to be verified [17]. They are commonly used for background identification where a user may not even know that he is being verified. These systems also differ from text-dependent systems in that the user need not be cooperative to be identified.

- *Text-prompted systems.* A text-prompted system is the most attractive for use in deflecting recording attacks [14]. The main idea is that the system dictates what needs to be said in order to be verified. A well-known example of such a system was proposed by [4] and works on the principle of combination-lock phrases. The system will typically have models of e.g. “twenty”, “thirty”, “one”, “nine” etc. The user is then prompted with random combinations of these lock-phrases, e.g. “twenty-seven”. Prompting a user to say a few of these makes it almost impossible to guess beforehand exactly what combination of phrases will be prompted. Another important aspect addressed by this proposal is that of sufficient training data for robust model estimation. With too little training data, reliable models cannot be trained. For a system to be a commercial success though, a user needs to spend as little time as possible during enrollment, thus resulting in a trade-off between too little training data and alienating your users. Using the approach proposed by [4], relatively little speech data can be used to train relatively robust models for subword units such as “twen”, “thir”, “for” and “ty”.

2.2.4 FEATURES USED FOR SV

Features are extracted from a user’s voice in order to perform pattern matching to the stored model of a speaker. A common mistake people make is to want to include as many different features as possible [14]. The problem that arises owing to this is called “the curse of dimensionality”. The more features one uses, the larger the feature dimensions become and consequently the strain on computing. That is why it is very important to understand the advantages and disadvantages of the different features and to use only the ones most relevant to the problem at hand.

Several methods do exist to reduce dimensionality. Traditional methods include principal component analysis and factor analysis.

A technique that can be used to determine whether a feature is a good one for speaker verification is ANOVA (analysis of variance) [14]. It comprises measuring Fisher's F-ratio

$$F = \frac{variance\ of\ speaker\ means}{average\ intraspeaker\ variance} \tag{2.1}$$

between sample probability density functions of different features. To be a good feature, a high F-ratio is desirable.

2.2.4.1 FILTERBANK-BASED CEPSTRAL PARAMETERS

This approach is based on how the human ear perceives frequencies over the audio spectrum. This is done non-linearly. It has been found [18] that implementing a similar front-end improves recognition. This method thus provides an easier way to obtain the desired non-linear resolution.

The first processing done on the signal is applying a pre-emphasis filter to the speech signal in order to boost the higher frequencies. The filter has the transfer function

$$x_p(t) = x(t) - a \times x(t-1) \tag{2.2}$$

where $x_p(t)$ is the pre-ephasized sample, $x(t)$ is raw signal sample at time t and a is the pre-emphasis coefficient which lies in the interval $[0.95, 0.98]$.

The signal is analyzed locally [17]. This is done by applying a window of which the duration is much shorter than the signal. The typical duration of a window is $20 - 30ms$. The windows are also chosen to overlap partially, usually by $10ms$. Three window functions that are popular today are the Hamming, Hanning and rectangular windows. Rectangular windows are the simplest windowing functions, but also have the highest amount of spectral leakage. Hamming and Hanning windows in constrast have much less spectral leakage and are thus useful for analyzing signals that are longer than the window length. These two windows are very similar and differ only in a single parameter α, which is 0.5 for the Hanning window and 0.54 for the Hamming window.

The resulting spectrum usually contains much redundant information, such as fluctuations in frequency. Only the envelope of the spectrum is of interest though. Because of this need and the fact that some important distinguishing high frequencies are naturally attenuated by the vocal tract, a series of localized filters are applied to the spectrum in order to obtain an approximately equal resolution on the Mel-scale. The Mel-scale is an auditory scale that is similar to the frequency scale of the human ear. It is defined as

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.3}$$

where f is the variable frequency.

Mel frequency cepstral coefficients are then calculated from the log filterbank amplitudes by using the discrete cosine transform:

$$C_n = \sum_{k=1}^{K} S_k \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], n = 1, 2, ..., L \tag{2.4}$$

where K is the number of log-spectral coefficients, S_k are the log-spectral coefficients and L is the number of cepstral coefficients required $(L \leq K)$ [18].

2.2.4.2 LPC-BASED CEPSTRAL PARAMETERS

This method attempts to model speech based on a linear model. The model is based on mapping the speech production organs to filters. The model assumes four speech production "modules": the glottal source, the vocal tract, the nasal tract and the lips. Each is represented as follows: the glottal source is represented by a low pass filter, the vocal tract by an AR (auto regressive) filter, the nasal tract by an ARMA (auto regressive moving average) filter and the lips by an MA (moving average) filter. Thus the whole speech production system can be characterized globally as an ARMA filter and representing the speech signal is in effect obtaining the filter coefficients of the ARMA filter [17]. In order to simplify this often complicated problem, the ARMA filter is estimated by an AR filter.

Thus, in effect the vocal tract can be estimated by the all-pole filter with transfer function

$$H(z) = \frac{1}{\sum_{i=0}^{p} a_i z^{-i}} \tag{2.5}$$

where p is the number of poles and $a_0 \equiv 1$. The filter coefficients a_i are chosen so that they minimize the mean square filter prediction error summed over the analysis window [18].

The autocorrelation of the windowed speech samples are then calculated, followed by the recursive calculation of the filter coefficients using sets of auxiliary coefficients. These auxiliary coefficients can be seen as reflective coefficients of an acoustic tube similar to the vocal tract being modelled. A comprehensive mathematical analysis has been done by [18].

The result is a set of LPC (linear prediction coefficients) that can be interpreted as the coefficients of the filter modelling the vocal tract.

2.2.4.3 DELTA, ACCELERATION AND THIRD DIFFERENTIAL COEFFICIENTS

Durational information can be used to enhance the performance of a speech-recognition system substantially [18].

Delta coefficients are calculated using the following regression formula [18]

$$d_t = \frac{\sum_{\theta=1}^{\Theta} (c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{2.6}$$

where d_t is a delta coefficient at time t computed in terms of the corresponding static coefficients $c_{t-\theta}\ to\ c_{t+\theta}$.

2.2.5 PATTERN-MATCHING TECHNIQUES FOR SV

The pattern-matching section of speaker verification involves computing a match score by matching the features of the current speaker to those of the claimed identity. The techniques used to create the speaker models dictate which features can be successfully used and how strongly they feature in the verification process. Some of the best known techniques are DTW (dynamic time warping), VQ (vector quantization), NN (nearest neighbour), HMMs, artificial neural networks, MLP (multilayered perceptron), NTN (neural tree networks) and nasal co articulation. Recently GMMs and SVMs (support vector machines) have also gained popularity. Only three will be briefly discussed.

There are two categories of models: stochastic and template models [14]. Stochastic models are based on probabilities and likelihoods, while template models are deterministic. In the latter case, the observation is assumed to be an imperfect copy of the stored model. The alignment of observed frames to template frames is selected in such a way as to minimize the distance between them. A mathematical analysis of this category is beyond the scope of this study.

DTW and VQ are template-based techniques while HMMs are stochastic models. Only a short discussion of the first two techniques will be given, since HMMs were extensively used during the research.

2.2.5.1 DYNAMIC TIME WARPING

Before HMMs gained popularity as the de facto method for speech recognition, DTW was a popular alternative. This method is popular due to its ability to compensate for speech rate variability. It is typically text-dependent and tries to match the sequence of training templates to the test sequence. It overcomes time mismatches by “warping” the test template in the time domain so as to maximize it’s alignment to the training template. This is achieved by performing some dynamic programming in the form of the DTW algorithm to minimize the distance between the templates. The speaker score is then computed as the minimum distance between the two templates. An accept/reject decision is made based on whether this score is above or below some threshold[14].

2.2.5.2 *VECTOR QUANTIZATION*

This method is a simplified version of DTW since it ignores temporal information in a voice. The advantage of VQ is simplicity, but for some applications, important information contained in speech rate variability will be lost by using this method.

A VQ codebook is designed by using standard clustering procedures. The centroids of these clusters are represented in the codebook. Pattern matching is then done by calculating the distance between the input vector and the minimum distance codeword in the code book [14]. A score is calculated by finding the distance between a claimed speaker's model and the test vectors. An accept/reject decision is then made, as mentioned above, by comparing this score to some threshold.

2.2.5.3 *HIDDEN MARKOV MODELS*

HMMs have, according to Campbell [14], been found to outperform all other methods when used in text-dependent systems and to be at least equal in performance to VQ when used in text-independent systems.

It works on the principle of states (where every state is a deterministically observable event), which are connected by a transition network. The corresponding state transition probabilities are $a_{ij} = p(s_i|s_j)$. Baum-Welch decoding can be used to determine the probability that a sequence of speech frames was generated by any particular model [19].

The likelihood is given by the score for L frames of input speech given the model

$$p\left(x(1;L)|model\right) = \sum_k \prod_{i=1}^{L} p\left(x_i|s_i\right) p\left(s_i|s_{i-1}\right) \tag{2.7}$$

where s is the state, x the unknown variable and k all state sequences.

2.2.6 ATTACKS ON SPEAKER VERIFICATION SYSTEMS

SV systems can fail in various ways. Statistical failures can occur as well as planned attacks, some of which are briefly discussed below.

- *Mimicry.* Mimicry is the act of a speaker mimicking the voice of a claimed identity. Such attempts have been shown to be able to thwart some speaker verification systems.

- *Recordings.* Recording attacks are another very real threat that has not been given much attention yet. This attack consists of the impostor playing a recording of a legitimate user's voice to gain unauthorised access to a speaker verification-protected system. It is a very sophisticated attack though, since the impostor will have to be able to record the pin number or whatever is to be prompted beforehand in order to gain access to the system. Text-independent systems are thus extremely vulnerable to such attacks, since any recording of a user's voice will allow access to the system.

SV systems are specifically designed to prevent unauthorized speakers from gaining access to particular systems. Understanding the type of attacks that can occur is thus important when collecting enrollment data for and designing such a system. The YOHO corpus, which was the main corpus used during the reasearch described in this dissertaion, (corpus described in appendix 1), was specifically proposed to counter recording attacks while minimizing the required increase in enrollment data.

2.2.7 SPEAKER VERIFICATION IMPLEMENTATIONS

Using standard corpora has proved to be very valuable with regard to making progress in SV research and development [20]. It allows different research groups to compare their systems on the same data using similar test protocols. A good overview of the most popular corpora are given in [20]. Since the focus of the SV research described in this dissertation is text-dependent, systems which have been tested on the YOHO corpus will be discussed. A proper testing protocol for YOHO as well as results others have been able to obtain on it are described in [15].

The next sections will discuss various implementations and approaches taken by several authors to improve on previous speaker verification results. In order to compare different speaker verifications systems, they have to be tested on some common database of speakers. The availability and use of standard speech corpora over the past 10 years is one of the major reasons why speaker verification has improved so much over this period [20]. A comprehensive overview of different corpora that is publicly available and intended for use in speaker recognition development and evaluation was done by [20]. The salient features with respect to application to the above mentioned purposes are described. The available corpora are listed below with the institute responsible for its distribution mentioned in brackets.

- TIMIT and derivatives (LDC)
- SIVA (ELRA)
- PolyVar (ELRA)
- PolyCost (ELRA)

- King (LDC)
- YOHO (LDC)
- Switchboard I-II including NIST Evaluation Subsets (LDC)
- Speaker Recognition Corpus (OGI)

2.2.7.1 TEXT-INDEPENDENT APPROACHES

Text-independent speaker recognition is the process whereby a speaker is recognised by text considered to be unknown. This approach to speaker recognition was first mentioned by [21] in 1974. Tests were conducted for text-dependent speaker verification in order to determine which features provided the highest accuracy. Cepstral coefficients were found to perform the best with 98% accuracy for only 1 second of speech. The test set consisted of only 10 speakers though, each having spoken 60 sentences. The feasibility of text-independent speaker verification was first investigated here, the motivation being that humans can distinguish between speakers even if they say different texts. An important drawback that is still applicable today was mentioned by [21] in that additional variability is introduced due to the differences in text. Tests were conducted using the same distance measure as was used for the text-dependent case and an accuracy of 93% for 2 seconds of speech was observed. This observation resulted in the formulation of a new branch of speaker recognition that has since then seen considerable research conducted to make it commercially viable.

20 years later, [22] conducted a thorough study on the state of text-independent speaker verification up to date. Two important aspects of text-independent speaker verification were mentioned; the identification and importance of addressing detrimental channel features and the commercial potential of text-independent speaker verification. Research was done to address the former by utilizing channel invariance properties of certain features. In particular, a segmental approach was used that utilized normalized segment scores as input to the speaker verification system. The speaker verification system was based on a probabilistic approach and among others investigated the use of Gaussian mixture models (GMMs) for use as speaker models. In addition an interesting approach was used in that Bayesian probability theory was used to determine a confidence score on tests which in turn enabled improved recognition accuracy since scores with a low confidence could be rejected. This is an important concept that can eventually be incorporated into a system that uses phone durations in a text-independent context. Tests were conducted on the Switchboard database.

In 1995 [2] presented a high performance text-independent speaker identification and verification system based on Gaussian mixture models. The identification system was based on a maximum likelihood classifier while the verification system was a likelihood ratio hypothesis tester. The verification system utilized background speaker normalization. [2] tested the performance of the recognition systems on the TIMIT, NTIMIT, Switchboard and YOHO corpora.

The GMM approach models the distribution of the extracted feature vectors from a speaker's voice. Mathematically, for D-dimensional feature vector x for a given speaker s, the model is represented as

$$p(x|\lambda_s) = \sum_{i=1}^{M} p_i^s b_i^s(x) \tag{2.8}$$

where M is the number of mixtures and $b_i^s(x)$ is a uni-modal Gaussian density given by

$$b_i^s(x) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i^s|^{\frac{1}{2}}} \exp -\frac{1}{2}(x - \mu_i^s)' (\Sigma_i^s)^{-1} (x - \mu_i^s) \tag{2.9}$$

and μ is the mean and p in 2.8 is the mixture weights which are subject to the constraint that

$$\sum_{i=1}^{M} p_i^s = 1 \tag{2.10}$$

Furthermore, [2] used the notation $\lambda_s = (p_i^s, \mu_i^s, \Sigma_i^s)$ to denote all the parameters of the model of a single speaker.

The well known expectation maximization (EM) algorithm was used to train the models. Cohorts were used for score normalization. A cohort set is a small selection of speakers other than the true speaker, which are used to normalize the speaker's score. During normalization, this cohort set is used to represent the alternative hypothesis, or impostor model. That is, to determine whether the true speaker ($P(\lambda_s|X)$) or an impostor ($P(\lambda_c|X)$) is speaking an utterance X, one computes the likelihood ratio given by

$$LL = \frac{P(\lambda_s|X)}{P(\lambda_c|X)} \tag{2.11}$$

or alternatively given by

$$\Lambda(X) = \log\left(p(\lambda_s|X)\right) - \log\left(p(\lambda_c|X)\right) \tag{2.12}$$

when working in the log domain. The decision of whether to accept the speaker as the true speaker or an impostor is then made by choosing the former if $\Lambda > \theta$, where θ is some predetermined threshold value, and the latter otherwise.

The reasoning followed in choosing the cohort speakers is largely dictated by the eventual application of the speaker verification system. Firstly, the background set can be chosen in order to represent impostors that sound similar to the speaker, referred to as dedicated impostors [2]. Another approach is to select a random set of speakers as the background set, thus expecting casual impostors who will try to represent a speaker without consideration of sex or acoustic similarity. By selecting the dedicated impostor background set, in contrast, the system may be vulnerable to speakers who sound very different from the claimed speaker [2]. The selection of the background set on a per speaker basis using the dedicated impostor approach is now described; first the $N = 20$ closest speakers to a speaker were determined using pair-wise distances between the speaker and all others. The pair-wise distance between speakers i and j with corresponding models λ_i and λ_j is

$$d(\lambda_i, \lambda_j) = \log \frac{p(X_i|\lambda_i)}{p(X_i|\lambda_j)} + \log \frac{p(X_j|\lambda_j)}{p(X_j|\lambda_i)} \tag{2.13}$$

where $\frac{p(X_i|\lambda_i)}{p(X_i|\lambda_j)}$ is a measure of how well speaker i scores with his/her own model relative to how well speaker j scores with speaker i's model. The ratio becomes smaller as the match improves.

These N speakers are known as the close cohort set, which is denoted by $C(i)$ for speaker i. The final background set consists of the $B = 10$ maximally spread speakers from $C(i)$, denoted $B(i)$. To determine $B(i)$, the closest speaker to i is moved to $B(i)$ and B is set to 1 (1 speaker in the background set). The next speaker c from those left in $C(i)$ to be moved to $B(i)$ is then selected as

$$c = \arg\max \frac{1}{B'} \sum_{b \in B(i)} \frac{d(\lambda_b, \lambda_c)}{d(\lambda_i, \lambda_c)} \tag{2.14}$$

where $c \in C(i)$. This procedure is repeated until $B' = B$. According to [2], the maximal spread constraint is to prevent "duplicate" speakers from being in the cohort set. The importance of obtaining a more general cohort set is that speakers that are very close as well as speakers that are very far (using the distance measure form 2.13) gives better results than simply choosing the 10 closest speakers.

The procedure followed by [2] proved to be extremely efficient and very good results were obtained. Among others, an equal error rate (EER) of 0.51% was achieved on the YOHO corpus. As a final note, [2] commented on the limiting factors of transmission degradations which include noise and microphone variability with regard to the possibility of further improvements. It was suggested that front-end processing and robustness techniques would play an important role in addressing these issues. In [23], the emphasis was on the difference in performance between the TIMIT (clean speech recorded with a high quality microphone) which produced very good results and NTIMIT (telephone speech version of

TIMIT) which produced significantly worse results; 99.8% for males and 99.0% for females on TIMIT compared to 62.5% and 56.5% for males and females respectively on NTIMIT. The limiting performance factor was thus positively identified as the corruption of a speech signal due to transmission line effects rather than the crowding of the feature space.

In [16], Reynolds addressed some of the above mentioned issues. GMMs were used, but this time using a single universal background model (UBM) and testing this approach on the 1999 NIST speaker recognition evaluations (SRE) corpora. The advantage of using a well trained UBM compared to different background models as in [2], [23] is that there needs to be only one alternative model; a new background set does not need to be calculated every time a new speaker is added to the system and the overall processing requirements is reduced since the alternative model is a single model that needs to be tested in comparison to 10 or 20 as used in [2], [23]. The problem of transmission degradations was addressed in that a handset detector and score normalization technique (HNORM) was developed to compensate for the use of different microphones in the recording of training and test data. Waveform compensation was also applied, but it was found that there is still much room for improvement between matched and mismatched conditions [16].

Transmission degradations or channel variability can of course cause intersession variability where the speech by the same speaker differ between sessions. Two techniques to address this problem in SVM speaker recognition have been proposed; within-class covariance normalization (WCCN) [24] and nuisance attribute projection (NAP) [25]. Intersession variability is modeled using within-speaker covariance matrices, but the two techniques differ in the way in which the eigen vectors are weighted.

It was noted that speaker recognition systems up until 1999 utilized only low-level acoustic features. [16] felt that there lies a lot of potential in searching for high-level features (such as durations of phonemes and prosody) that is naturally used by humans in order to discriminate among speakers. The high-level features by themselves are not expected to produce good results, but the improvement would become evident when the high-level features were fused with the lower level ones. Efficient techniques for both identifying the high-level features and fusing them with the lower-level acoustic features still needed to be developed.

At the same time, research addressing the same problems of signal degradation over transmission channels was done by [26]. Different normalization techniques were investigated using the 1998 NIST SRE corpora; that of UBMs to normalize test scores, the use of cohort models to normalize test scores and also handset normalization. For the cohort set, a novel method called T-norm was created. T-norm comprises estimation of mean and variance parameters at test time that is then used in the normalization of the test score. It was found that this new algorithm provided better results than the use of a UBM. A disadvantage was mentioned in that the T-norm method does not perform well when a different language is used by the cohort speakers to that of the claimed speaker. The UBM on the other hand can be trained to compensate for these cases. Other interesting observations by [26] were that there was a huge improvement in accuracy when the cohort set was enlarged from 10 to 20, but no significant increase thereafter.

An improved version of the T-norm normalization technique was presented by [27]. The technique was adapted to a speaker adaptive T-norm. The traditional T-norm estimated parameters on scores obtained during testing. The estimated mean and standard deviation scores are then used to normalize the target speaker score S for an observation O as

$$S_{tgt|tnorm}(O) = \frac{S_{tgt}(O) - \mu_{tnorm}}{\sigma_{tnorm}} \tag{2.15}$$

Tests were conducted on the 2004 NIST SRE corpora with the Switchboard corpus used for training.

2.2.7.2 *TEXT-DEPENDENT APPROACHES*

The text-dependent approach was the first approach used in speaker recognition. Speakers typically have to say a predetermined word or phrase such as a password to be recognized. The additional variable of “unknown text” is removed, resulting in a more accurate and easy to implement verification system. Because of the recording attack described in section 2.2.6, a user is sometimes prompted to say one of several pre-recorded phrases. With the advent of modern computers, even this has become insufficient in guarding against recording attacks. In order to counter such attacks, a variant of text-dependent speaker verification called randomized phrase prompting was first proposed by [4] in 1991. [4] reasoned that a small amount of phrases for enrolment data provides a short enrolment time and high system accuracy but at the cost of being predictable and thus potentially vulnerable to attack. On the other hand, a large amount of phrases for a high accuracy system will result in a much longer enrolment time, which is not commercially viable. Thus [4] proposed recording a small amount of phrases that can be combined to provide the required amount of randomness to make recording attacks unlikely to succeed. The specific approach proposed was that of using combination-lock phrases. Combination-lock phrases consist of sets of three words, the words being anything between $21 - 97$, with teens, decades, double digits and words ending in 8 being eliminated for various reasons. This results in $56^3 = 175616$ possible phrases. This technique is also the one on which the YOHO corpus is based.

It was also observed that performance could be improved substantially when breaking up the phrases into partial words such as “Twen”, “Thir”, “ty” etc. The reason for this is that it allows each word model to be trained on more data, for example every word's "ty" can be used to train a robust "ty" model, all ones $(1-9)$ can be used for training 9 models instead of single models for "31" for example. This important observation paved the way for the use of phonemic verification systems. The system was also the first to make use of log-likelihood scoring, which is the most common method of scoring today. It was found that the likelihood scoring technique reduced the EER by a factor of 4 to 6 times. Template matching was used as a means of determining the distance between a speaker and the claimant.

Modern approaches to text-dependent speaker verification rely mostly on continuous density hidden Markov models (HMMs) to model a speaker's characteristics [28]. HMMs are usually trained with a maximum-likelihood approach, such as the previously mentioned EM algorithm. When little training data is available, the model tends to be under-trained, which means that the variance is not representative of the true distribution of the speaker's speech, but rather is representative only of the very small amount of training data. Techniques to counteract this phenomenon in text-dependent speaker verification systems were proposed by [28] by way of variance flooring. This approach sets a lower limit on the variances. The difficulty lies in determining what a good floor will be. This difficulty is addressed in two ways. One option is to take the variance of a non-client multi-speaker model and use it as a fixed variance during the EM algorithm, thus only updating the means and mixture weights. An alternative approach is to floor the variance after every iteration of the EM algorithm to prevent it from becoming too small. This flooring can be applied in three dimensions; vector index, time and feature space.

2.3 DURATION MODELING

The use of higher-level features was mentioned as a possible source of improvement during the 1998 NIST speaker recognition evaluation. A comprehensive overview of the goals, methodologies used and results obtained are given by [5]. NIST started these workshops in order to improve the understanding and implementation of current speaker recognition technology and also to help facilitate the commercial implementation thereof.

Interesting applications were mentioned, one being how to classify a speaker recognition task. The question was asked: "Who is to benefit, the speaker or someone else?" This is considered a useful split since it has a significant impact of the task definition, performance evaluation and the system design [5]. A good overview of the different definitions and operating modes of the speaker recognition task is also given. A good distinction is made between open-set and closed-set speaker identification systems in than the former is the scenario where the actual speaker may be none of the candidate speakers and the latter where the actual speaker will be one of the candidate speakers. The speaker identification task is classified as an N-class problem while the speaker verification task is a 2-class problem. Furthermore, the speaker recognition task can be either cooperative or non-cooperative.

Some sources of variability in a speaker's voice are also discussed [5]. These are

- *Session.* When training and test sessions are not conducted as part of the same session, performance starts to drop
- *Health.* Bad health can be detrimental to speaker recognition systems. Laryngitis is considered being the worst state of health with regard to speaker recognition. Emotional- and metabolic states were also found to play a role.
- *Educational level and intelligence.* This affects cooperative systems to a larger extent than non-cooperative systems. In a cooperative environment, a speaker may have to read phrases of arbitrary difficulty, or be required to remember pass phrases.
- *Speech effort and speaking-rate.* This affects the speech signal in a complex way. The Lombard effect is mentioned (where people tend to talk louder when exposed to loud auditory noise).
- *Experience.* The more a user interacts with the system, the better the user is able to use it and the models also become more accurate.

A detailed account is also given on some factors already discussed, such as degrading transmission line effects and normalization techniques.

During the 2001 NIST SRE, the EER on the corpus used was reduced by 71% to 0.2% by using high-level features such as pronunciation models, prosodic dynamics, pitch and duration features, phone streams and conversational interactions in the SuperSID project [29]. One of the most promising advantages that these high-level features may provide is more robustness against acoustic degradation due to transmission channel effects. The environment decided on was that of text-independent speaker verification.

The features used were compared to a baseline performance generated by evaluating the system using a GMM-UBM system. The results ranged from an EER of 3.3% to 0.7% as the amount of training data increased from a single conversation to 8 conversations per speaker. The other high-level features used will be listed with the EER obtained using 8-conversational training, displayed in brackets.

Prosodic features used included pitch and energy distributions (16.3%), pith and energy track dynamics (14.1%, dropping to 9.2% when combined with the former) and prosodic statistics (15.2%). Phone features used were among others phone Ngrams (4.8%), phone binary trees (3.3%), cross-stream phone modelling (4.0%, dropping to 3.6% when fused with temporal systems) and pronunciation modelling (2.3%). Conversational features produced an EER of 26%. Future work proposed by [29] were among others to determine confidence intervals that will enable one to determine if the features can be reliably used in that particular case.

The work that initiated at least some of the above research was that done by [6]. The viewpoint was from the speech recognition angle and not the speaker recognition one. Words were modelled using the durations of the phonemes that make up the word. Issues directly addressed were the prediction of durations of unseen words, the "pre-pausal" effect (as also mentioned above) and the differing speech rates among speakers. The rate-of-speech (ROS) was computed to be the average number of phonemes per second over an utterance. This ROS measure was then used to normalize the durations of the phones. It was found that normalization at speaker level in contrast to hypothesis level yielded consistent and better results. The "back off" strategy was also first mentioned by [6] and found to improve performance. The word duration modelling and subsequent normalization proved to increase word recognition substantially.

The speech rate normalization investigated by [6] was taken a step further by [7]. The relationship between speech rate variation and intrinsic phone durations to speakers were investigated. It was found that even after the application of a novel speech rate normalization technique, the variance explained by the speaker and phone type remained constant, leading to the theory that intrinsic phone durations are speaker-specific. The speech rate was computed as

$$rate_{LR} = \frac{\frac{S_{L+1}-\omega_L}{S_{l+1}-S_L} + \frac{\omega_R-S_r}{S_{r+1}-S_r} + r - l - 1}{S_{l+1} - \omega_L + \omega_R - S_r + \sum_{i=l+1}^{r-1} S_{i+1} - S_i} \qquad (2.16)$$

where S_i is a speech unit mark falling in a window ω of length $625ms$, ω_L the left window boundary and ω_R the right window.

It was claimed that speech rate perception are functions of both syllable rate and phone rate, with the two constituents having a correlation r of only 0.6. The speech rate was normalized by applying the inverse of the curve produced by for example (2.16) to every $100ms$ of speech. The results showed that phone duration variation for phonemes was reduced from 2.48% to 1.48% for vowels (with the variable speaker "omitted" with the normalization technique) and from 1.73% to 0.42% for consonants. An important result from the speaker verification point of view is that the variation between speaker combined with phone-type interaction was increased from 2.01% to 2.57%. It was also observed that phone durations were consistently longer towards the end of an utterance.

A different approach to utilizing the information inherent to duration was taken by [30].

A Viterbi algorithm was implemented and tested on the YOHO database. The state-transition probabilities were changed to temporal constraints, with state-durations being modelled for every speaker. The basic idea was to change the state-transition probabilities if

$$a_{i,i}^{\tau} = \begin{cases} 1 & \text{if } \tau < t_{min_i} \\ 0 & \text{if } \tau > t_{max_i} \\ \frac{D_i(\tau)-d_i(\tau)}{D_i(\tau)} & \text{if } t_{min_i} < \tau < t_{max_i} \end{cases} \tag{2.17}$$

$$a_{i,i+1}^{\tau} = \begin{cases} 0 & \text{if } \tau < t_{min_i} \\ 1 & \text{if } \tau > t_{max_i} \\ \frac{d_i(\tau)}{D_i(\tau)} & \text{if } t_{min_i} < \tau < t_{max_i} \end{cases} \tag{2.18}$$

with τ being the number of frames in state i up to time t; with t_{min} and t_{max} being the minimum and maximum observed times for that phone, $d_i(\tau)$ is the probability of state duration equal to $D_i(\tau) = \sum_{t=\tau} d_i(\tau)$. The gamma and geometric distribution were used to model the state distributions and were compared with each other. The use of the gamma distribution proved to provide the best results with an EER of 0.33% achieved when 97 speakers were used as impostors.

An approach to complement durational information was proposed by [31]. They tried to model statistical pronunciations across multiple phone streams and referred to this approach as phonetic information in the cross-stream (cross-language) dimension. Ngram language modelling techniques were used to directly estimate speaker-language dependent phone models from the speaker's available training data. The exact methodology is not as important here as the fact that a significant improvement was found when a simple linear combination of features from both the time and crossstream dimensions were done. This confirmed that there exist complementary information in both dimensions in the way in which phonemes are pronounced by different speakers.

While [29] focused on many different high-level features, [8] investigated duration features only. Results in accordance with the prediction of [16] and the results of [29] were found in that results were average when only duration features were used, but when combined with traditional features such as cepstral coefficients, the EER was reduced by more than 50%. An interesting observation was also made by [8] that the improvement of the EER using traditional features seems to saturate after a few minutes of speech while the improvement of duration features continues to increase. Also, [8] found that duration features are much less sensitive to noise than traditional features and can thus prove crucial in maintaining the high accuracy of handset speaker verification systems when ported to telephone-based systems.

Each word and each phone was modelled by [8] in terms of its duration and context. Experiments were also done with specific word-usage of certain speakers. For the duration study, three types of feature vectors were used; word features, where the sequences of phone durations in a word is considered the features for that word, with different pronunciations of words having different associated feature vectors; 1-component phone features, which are 1-dimensional features consisting of the duration of the specific phone only; 3-component phone features, which are sequences of 3-state HMMs of the phone state durations. [8] predicted that modelling phone durations separately and not as part of the 3-component phone features may yield better results.

The training procedure followed by [8] consisted of extraction of all training data, creating UBM GMMs for all features as mentioned above and then adapting the UBM to the specific speaker using a small amount of adaptation data. Different models were also trained for phones of words just preceding a pause (silence longer than $200ms$) since it was found that a pause alters the normal duration of phones/words. Another important strategy used was a back-off strategy. This approach was taken to avoid the use of poorly adapted models. Thus, only models for which more than 5 samples were available for adaptation were used to score test words. This approach yielded better results than when all test words were scored. In particular, the baseline system (MFCC features only) yielded an EER of 0.90%. When combined with duration features, the EER dropped to 0.40% and when this result was combined with lexical features as well, the result dropped to 0.29%.

The results of several systems exploiting high-level features were described by [29], as mentioned above. [32] gives an insightful account of how the fusion of the high-level features was accomplished. The 9 best techniques were selected for fusion, with the best being the traditional GMM-cepstra approach and the other 8 high-level features. The two fusion techniques used were the perceptron classifier with no hidden layers and the GMM. The perceptron was thus similar to a linear discriminant with the output function being a sigmoidal one. The perceptron classifier was used to combine all 8 higher-level features. This approach surpassed the GMM-cepstra approach in performance and when all 9 features were fused, the best performance to date on the NIST 2001 SRE corpus was achieved.

The importance of higher-level features such as prosodic features sequences was also investigated by [9]. A support vector machine (SVM) was used to model the features. In addition to other work already described, it was found that pitch features are the most useful high-level feature, followed by duration and energy features. The most important pitch features are the ones that represent pitch level. For energy features, the rising and falling patterns are considered the most important and for duration modelling on a syllable basis, it was found that the nucleus was more important than onset or coda of the syllable.

The use of SVMs was also incorporated into the approach taken by [33] to do speaker and language recognition. A detailed discussion is given on exactly how the support vector machine was built and how the kernel was designed. In particular, a new sequence kernel was developed which was called the generalised linear discriminant sequence (GLDS) kernel. An interesting observation from this paper was that SVMs were shown to contain complementary information to GMMs in that a substantial improvement was observed when the two methods were fused for a language recognition application as well as a speaker recognition application.

Higher-level features operate at longer time spans than frame level cepstral coefficients [34] and consequently more training data is needed for reliable model estimation. The most recent application of such higher-level features in speaker verification was that by [11] during the NIST 2006 Speaker Recognition Evaluation (SRE) workshop. An SVM was used to model prosodic feature sequences; in particular pitch, energy and duration features. The pitch features included mean, minimum, maximum and slope within a syllable. The duration features were made text-independent by normalization based on mean and variance from background data. Since these prosodic feature sequences are usually variable in length, a new kernel was developed to measure similarity between two sequences. In addition three specific areas of potential improvement were investigated:

- *feature parameterization*
- *intersession variability compensation*
- *conditions features based on text and part-of-speech constraints*

A “soft bin transform” was used to transform prosodic features in order to obtain a fixed length vector which is in turn easier to model with an SVM. This was done by using VQ to train a GMM on some held out training data. A sample was then transformed to a vector of posterior probabilities from each of the Gaussians in the original GMM. The features were then normalized to a uniform distribution.

Within-speaker variability was compensated for by using nuisance attribute projection (NAP), as proposed by [25]. This method was developed specifically for intersession variability compensation in speaker verification when using SVM’s. A matrix

$$P = I - UU^T \tag{2.19}$$

is created, which projects the features onto a different subspace. This new subspace is supposed to be more resistant to intersession variability. The intersession variability subspace for the prosodic features was determined by [11] by using principal component analysis of the within-class covariance matrix.

Another interesting technique used by [11] was to use predetermined constraints to identify specific parts of speech which were more likely to have greater across-speaker than within-speaker prosodic variability. A simple HMM-based tagger trained on Penn Treebank-3 data was used for this task. Constraints based on POS reflecting discourse categories rather than grammatical categories were found to work better. It was also interesting to note that from the syntactic-based constraints, personal pronouns seemed to be more usefull than for example adjectives, nouns, verbs etc. This prosodic based system was subsequently fused with a frame-based cepstral MLLR system.

2.4 PROSODY

Prosody is defined as the physical phonetic effects that a speaker uses to express his intentions during speech. It can also be defined as consisting of four cues used by a listener to interpret the speaker's intentions [35]:

- *Pauses.* To distinguish between phrases
- *Pitch.* Fundamental frequency temporal rate of change
- *Rate/relative duration.* Phoneme durations
- *Loudness.* Amplitude

In contrast to the traditional acoustic features, prosodic features are considered learned traits that may provide speaker specific information at a higher level [29]. Since the focus of this dissertation is on the durational component, we focus mainly on duration.

Duration modeling has in addition to speaker recognition, also received considerable interest from two other fields; text-to-speech and speech recognition.

Developing accurate phoneme duration models has been a topic of discussion for several years, especially with regard to the potential benefits for automatic speech recognition (ASR) [13]. In [36],[37] we showed that accurate phoneme duration models can significantly improve state of the art speaker recognition (SR) systems in a text-dependent environment. For practical applications of both ASR and speaker recognition, duration models have to be developed for text-independent speech. This is not a trivial problem as there are many factors influencing the duration of phonemes in text-independent speech, such as position in word, position in sentence, stress, preceding and following phonemes, speech rate etc. Although the work done in [36] was in a text-dependent environment, it did confirm earlier findings by [7] that phoneme durations are also speaker-specific to a large extent, which adds another dimension to the model estimation. All these factors contribute to making data scarcity a significant obstacle to characterizing phoneme durations accurately. This obstacle, which was first identified in 1988 [12], remains arguably the most significant one to the more general use of phoneme durations. While ASR virtually ignores prosody [38], work by [39],[40] strongly suggests that it may be beneficial to ASR to explicitly incorporate prosody models.

Speech rate variability has been found to be detrimental to recognition accuracy in ASR, especially when it deviates greatly from the training data [41] or in general for speakers who talk faster than average [42]. In order to reduce the error, this speech rate variability needs to be removed by ways of speech rate normalization [43]. Several normalization algorithms have been developed, with some recent proposals such as [7],[42].

An attempt to estimate the individual contributions of the abovementioned factors to the total variance was made by [13]. A hierarchical analysis of variance was performed and it was found that much of the variance can indeed be explained by these factors. Because of the type of ANOVA performed, it was not possible to examine interactions among the factors, which may omit important information. Duration patterns were also modelled in [8],[9],[10],[11] in order to improve speaker recognition performance. It was observed that significant improvements in accuracy can be achieved by separately modelling word durations, single phoneme durations and state durations using 3-state hidden Markov models (HMMs). Data sparseness was addressed in all cases by a back-off technique, through which word-models would be backed off to triphone models and the latter to single phoneme models. This ignores the effect of the specific factor being addressed on the particular phoneme. Rao Gadde [10] also performed a simple speech rate normalization. The speech rate was calculated as the number of phonemes per second. By applying this simple normalization technique, a consistent improvement in word recognition was observed over several databases.

Text-to-speech synthesis is another field that will benefit from accurate duration models. Two popular methods used to date include Sums-of-Products(SoP) [44] and classification and regression tree methods (CART) [45]. The aim of this field with regard to duration modeling is to capture and represent the “naturalness” of human speech with mathematical models [38]. It was found that there are too many factors influencing the durations of phonemes to be covered by any realistic database [46]. On the other hand it has also been shown that the less frequent factor combinations occur frequently enough to guarantee that on average at least one such example will occur in an utterance [44]. For text-to-speech synthesis the model has to be able to generalize even for such infrequent events, thus most of the research over the past couple of years have focused on among others to developing models with good generalization capabilities.

Bayesian networks were used to model vowel segment duration for text-to-speech synthesis by [47]. Discrete and continuous nodes were used to model linguistic factors that can influence segment duration. The Bayesian networks outperformed both CARTs as well as SoPs, confirming that the factor interactions are both complex and important in duration modeling.

The complicated factor interaction was also identified by [48], [49] when attempting to use duration as a key indicator in models attempting to automatically identify English accent. While useful, the effect of the other factors influencing phoneme durations, such as prepausal lengthening and speech rate, complicated the automatic identification of accent.

Taken together, these studies are strong evidence that accurate phoneme duration models can greatly benefit both ASR and speaker recognition. However, no sophisticated model exists yet because of data scarcity (which limits the number of factors that can be modeled), the many different factors which have an influence on the duration of phonemes and the fact that interaction effects between the different factors are not incorporated into the models.

CHAPTER THREE

CONTEXT-DEPENDENT TRIPHONE DURATIONS AS A FEATURE

3.1 INTRODUCTION

To address some of the issues affecting SV accuracy, as discussed in chapters 1 and 2, a new class of features based on temporal information in spoken utterances was proposed in [8] (for text-independent speaker recognition) and [50] (for text-dependent speaker verification). In [50] preliminary tests demonstrated the value of these features in addressing problems due to noise and recordings. The database of speakers in [50] was very small and the claims of temporal information improving the equal error rate (EER) of SV systems had to be verified on a larger corpus of data. Here, we report on a set of experiments using the YOHO corpus (see appendix 1), which has been widely used to evaluate SV systems [15] and has a structured methodology for performing comparative tests [20]. Another reason why the YOHO corpus was chosen for evaluation as opposed to more recent corpora (such as the NIST speaker recognition evaluation corpora), is the text-dependent nature of the current approach to duration modeling. It will be an interesting extension of the current research to see how this approach scales to a text-independent environment. This chapter concludes with a discussion of a number of ways in which more sophisticated models may be used to further enhance the accuracy of the duration model. Some preliminary results are presented which indicates that such a model may indeed provide a more accurate model of phoneme durations.

3.2 EXPERIMENTAL FRAMEWORK

As mentioned in chapter 2, it was predicted in [16] that higher-level features would need to be fused with lower level features (such as frame based cepstral features) in order to be useful. This has been confirmed in [11]. In order to reliably compare the results of phone durations on speaker verification to results from other SV approaches, a frame level cepstral based system would need to be constructed that is comparable to other acoustic systems mentioned in table 3.1.

Table 3.1: Equal error rates obtained on the YOHO database by other researchers. The superscripts distinguish between experiments that use impostors that have been seen^1, and those that use impostors not yet seen^2

Research group	Type of system	EER
ITT [4]	Continuous Speech Recognition (CSR)	1.7%
ITT [15]	Neural Network (NN)	0.5%
MIT/LL's [2]	Gaussian Mixture Model	0.51%
Rutgers [51]	Neural Tree Network (NTN)	0.65%
Reynolds [2]	Gaussian Mixture Model	0.58%
Wan & Campbell [52]	SVM, normalized polynomial kernel	$0.34\%^1$, $0.59\%^2$
Campbell & Assaleh [53]	Polynomial classifier	$0.18\%^1$, $0.31\%^2$

3.2.1 FRAME LEVEL CEPSTRAL BASED HMM SYSTEM

An HMM-based ASR system was constructed using the HTK 3.2.1 toolkit [18]. A basic triphone recogniser was trained on the TIMIT corpus in accordance with the guidelines for building a conventional speech recognition system in [18]. 3-State HMMs were constructed with one mixture per state. The triphone recognizer was then retrained using a restricted dictionary (limited to the 56 combination lock phrases as described in appendix 1) and also using all the YOHO enrollment data. The resulting set of models were used as the UBM. When one has a well trained speaker independent model, it is common practice to adapt this model to a particular speaker in order to better model the characteristics of that speaker [18]. New models were thus estimated for each individual speaker using his/her enrollment data and applying supervised adaptation to the UBM, firstly by means of MLLR and secondly using MAP techniques.

This ASR system could then be used to perform verification by using Baum-Welch decoding to determine the probability that a sequence of speech frames was generated by any particular model. To determine whether a particular speaker or an impostor spoke an utterance, Eq. (2.7) (using both the UBM and a particular speaker's model) could be used in conjunction with (2.12) to calculate a LL score. A speaker was then accepted or rejected if their LL score was above or below some empirically predetermined threshold. The exact testing procedure is described in section 3.3.

The ASR system was also utilized for determining phoneme duration boundaries or segmentation. The speech recognition system was applied by means of forced alignment to obtain the segmented boundaries. A restricted pronunciation dictionary was used in that no pronunciation variations were allowed.

3.2.2 FRAME LEVEL CEPSTRAL BASED GMM SYSTEM

In addition to the HMM-based system, a GMM-based system developed by Kleynhans [54] was also trained on the YOHO corpus. K-means clustering was first employed to generate good initial values as input to the EM algorithm. Much work was done together with an intern student in finding optimal values for the GMM-based system, such as the optimal number of mixtures to use, the number of training iterations, best value for variance flooring etc. It was found that 512 mixtures works well and a UBM with this number of mixtures was trained. Individual models were then adapted from the UBM for each speaker using their enrollment data.

3.2.3 PROPOSED PHONEME DURATION MODELING

3.2.3.1 CHOOSING PARAMETRIC MODELS

Since the duration of a phoneme is known to depend on its acoustic context, we model the durations of context-dependent phonemes (from here on referred to as triphones). These durations are obtained by forced alignment of each YOHO utterance, using the known transcription and the speaker-specific acoustic model described above. Only one pronunciation per word was allowed, thus resulting in 53 triphones. To decide which parametric model to use for the duration density functions of the triphones, several parametric forms were fitted to the triphone durations obtained in this fashion. Typical results for the speaker-specific and speaker-independent distributions are shown in Figures 3.1 and 3.2, respectively. The histogram density estimates, as shown by the bar graphs in these figures, are consistently unimodal, suggesting that a single parametric component will be sufficient. For the speaker-specific density function of Figure 3.1, the Birnbaum-Saunders and Gamma distributions seem to fit the data best, whereas the normal distribution provides a better fit for the speaker-independent case (Figure 3.2). However, all these differences are fairly small, and we have therefore used a normal distribution in all the experiments described below.

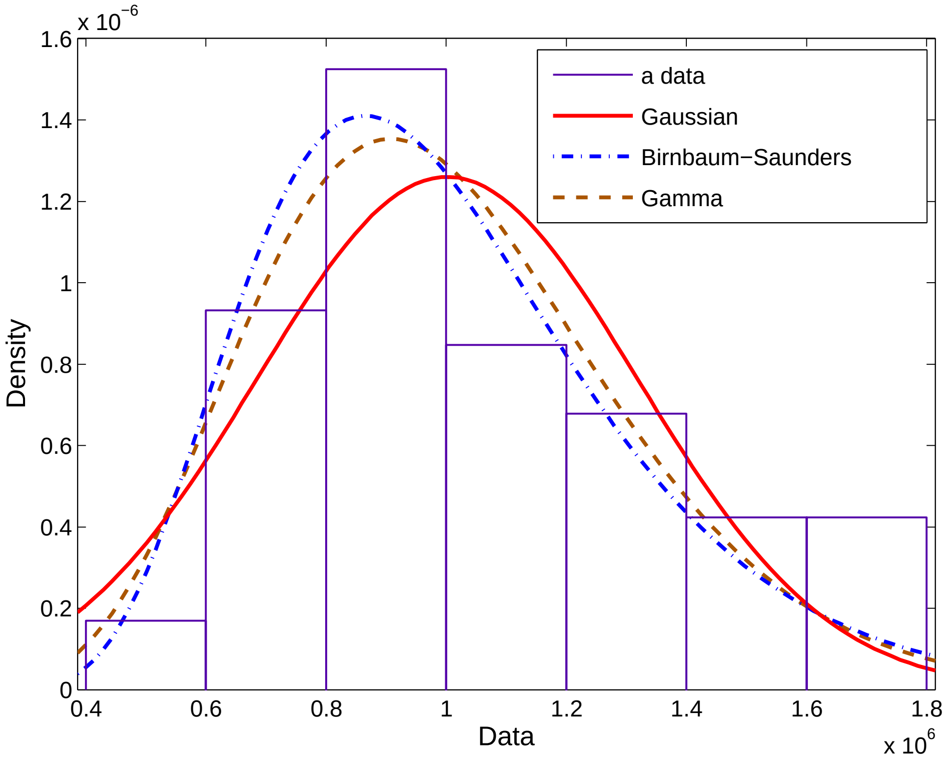

Figure 3.1: Distribution of durations of the triphone“s-eh+v” for a single speaker with different distribution functions fitted to it.

3.2.3.2 DETAILS OF TRIPHONE DURATION MODELS

The models used in our tests were constructed for each triphone k by calculating the sample mean

$$\overline{x} = \frac{1}{M}\sum_{n=1}^{M} x_n \tag{3.1}$$

where M is the number of observations of the triphone and x_n is the duration of the n'th observation. An unbiased estimate of the sample variance σ^2 was also calculated as

$$s_{M-1}^2 = \frac{1}{M-1}\sum_{n=1}^{M} (x_n - \overline{x})^2 \tag{3.2}$$

Every speaker thus has 53 duration models of the form (μ, σ^2). The duration models were constructed by using all the extracted durations from the 4 enrollment sessions. Testing was then performed by first extracting the durations of the triphones in the test session and then calculating a score

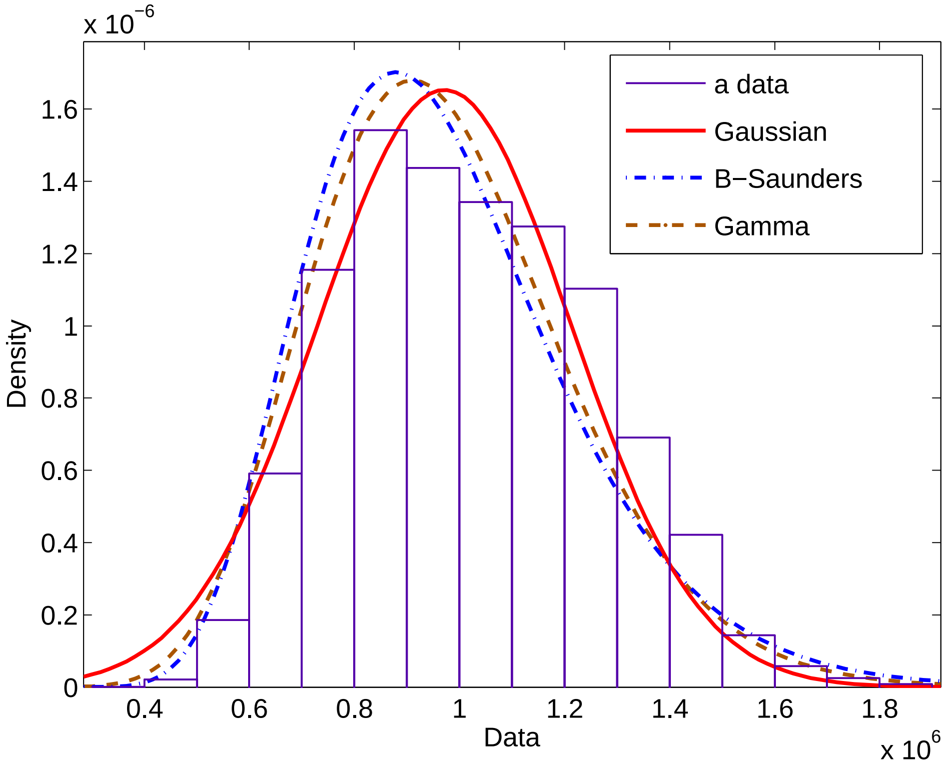

Figure 3.2: Distribution of durations of the triphone“s-eh+v” for 168 speakers with different distribution functions fitted to it.

$$P(x|\overline{x}, s^2_{M-1}) = \frac{1}{\sqrt{2\pi s^2_{M-1}}} e^{-\frac{(x-\overline{x})^2}{2s^2_{M-1}}} \tag{3.3}$$

where x is the observed duration of a specific triphone. The evaluation of Equation 3.3 yields a value that occurs on the normal distribution with parameters $(\overline{x}, s^2_{M-1})$. This value is normalized by evaluating the normal distribution with the same value, but using the universal background model (UBM) parameters (which are the means and variances of the appropriate context-dependent triphone, calculated across all training sessions by all speakers). A score is then generated for a speaker i as

$$Score_i = \frac{1}{L}\sum_{l=1}^{L} log(P(x_l|\lambda_c)) - \frac{1}{L}\sum_{l=1}^{L} log(P(x_l|\lambda_{UBM})) \tag{3.4}$$

where L is the number of observed triphones in the test session. Tests were again performed on a rotating scheme as before, where one speaker is the claimed “client” and all speakers excluding the (acoustic) cohort set are tested using the claimed speaker’s models. Once all scores have been obtained, they were again put in an ordered list and the EER was determined.

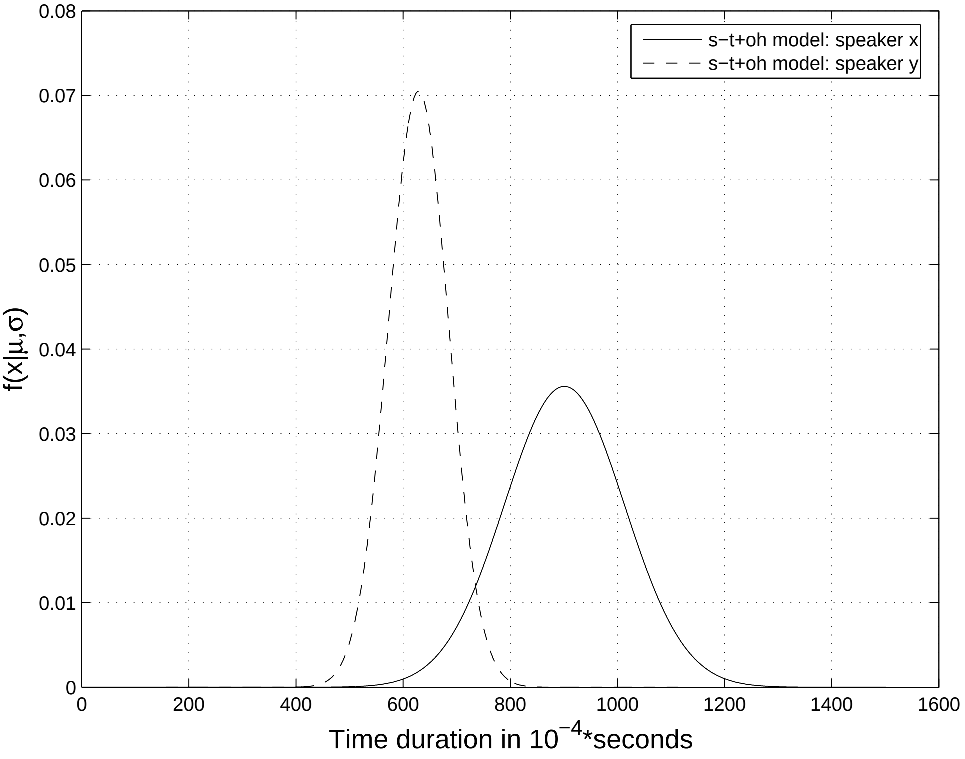

Figure 3.3: Probability distributions of a triphone that provides good discrimination between a pair of speakers.

Figures 3.3 and 3.4 illustrate typical distributions of durations observed in our tests. Figure 3.3 shows an example of a triphone that provides good discrimination between two speakers; in other words, phoneme durations of speaker x matched to the model of y would produce a poor score and the general UBM would be chosen, resulting in a correct reject decision of the impostor. Figure 3.4 illustrates a bad example of a triphone to use, since speaker x's durations would match speaker y's durations well, resulting in a good score for an impostor.

Since both cases are observed in our data, we used an empirical methodology to determine whether durations are useful for the task of speaker verification.

3.3 TESTING PROCEDURE

In order to compare the performance of our proposed speaker verification system to that of other speaker verification systems, a standard testing procedure was employed, similar to

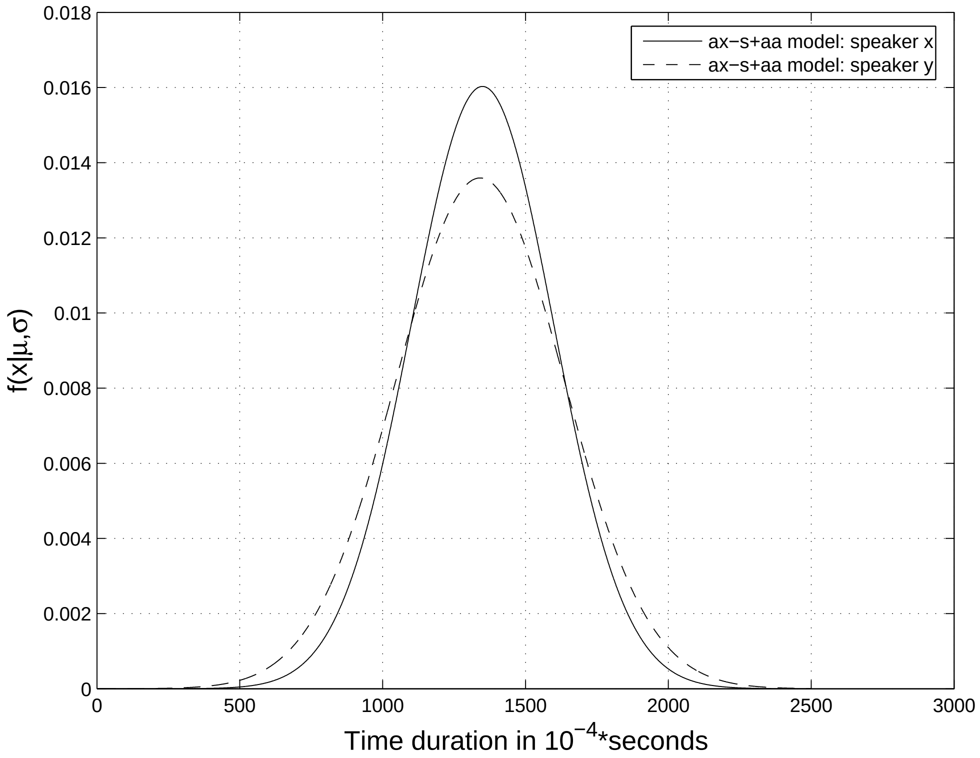

Figure 3.4: Probability distributions of a triphone that does not provide good discrimination between a pair of speakers.

that used by others on the same corpus (see [2], [51], [55]). The exact test procedure is most clearly described by Reynolds [2] and is described in section (described in Section 3.3.1).

Table 3.2 summarizes the results of several different tests that were performed by Reynolds [2] on the YOHO corpus. (In this table, *msc* denotes "maximally-spread close" and *msf* "maximally-spread far"; these are two different approaches to selecting cohort speakers – see below.)

Table 3.2: Equal error rates reported in [2] for different experimental conditions.

Test	**YOHO(eer)**
M(10 msc)	0.20
M(5 msc, 5 msf)	0.28
F(10 msc)	1.88
F(5 msc, 5 msf)	1.57
M+F(10 msc)	0.58
M+F(5 msc, 5msf)	0.51

The test *M+F(10 msc)* was used as basis for our comparison, the only difference being that all four enrollment sessions were used for enrolling the speakers. (Reynolds used the fourth session for cohort selection).

In order to perform comparable tests using the temporal features, we had to adapt the use of cohorts for score normalization. A cohort set is a small selection of speakers other than the true speaker, which are used to normalize the speaker's score. That is, to determine whether the true speaker ($Pr(\lambda_c|X)$) or an impostor ($Pr(\lambda_{\overline{c}}|X)$) is speaking, we compute the likelihood ratio:

$$likelihoodratio = \frac{Pr(\lambda_c|X)}{Pr(\lambda_{\overline{c}}|X)} \tag{3.5}$$

In Equation 3.5 X denotes the spoken utterance, λ_c the claimed speaker model and $\lambda_{\overline{c}}$ the cohort (also known as background or impostor) model. By applying Bayes' rule and discarding the constant prior probabilities for claimant and impostor speakers (they are accounted for in the decision threshold) [2] and working in the log domain, Equation 3.5 can be rewritten as

$$\Lambda(X) = \log p(X|\lambda_c) - \log p(X|\lambda_{\overline{c}}) \tag{3.6}$$

The speaker is accepted as the claimed speaker if $\Lambda(X) > \theta$ and rejected as an impostor if $\Lambda(X) < \theta$ where θ is an appropriate threshold [2]. θ can be speaker specific (which is computationally more expensive, but also more accurate) or global. The determination of the EER in our test used a global threshold approach, as in [2].

This standard approach to normalization works well if only one type of feature is employed. However, the choice of cohort speakers dictates a group of speakers that cannot be tested as possible impostors, which complicates the procedure when a second feature set is to be used. (If the cohort speakers are based on acoustic features only, they will not necessarily be a good model when using the durational features.) We therefore chose to normalize the temporal features using a universal background model (UBM) rather than a cohort set, and thus to also use a UBM approach for the acoustic feature. This approach complicates the comparison of results to that of a dedicated-impostor cohort approach, since in the latter approach some of the most likely impostors are eliminated from the test set.

3.3.1 DETAILED TEST DESCRIPTION

The HTK 3.2.1 toolkit [18] was used to construct the speaker verification system. MFCCs were used as input features together with delta and acceleration coefficients. Hidden Markov Models (HMMs) with one Gaussian mixture per state were created for all context-dependent triphones occurring in the restricted grammar set.

A cohort set of 10 speakers were selected for every speaker in the database in accordance with the procedure in [2]. Choices that arise with background speakers are the choice of specific speakers and the number of speakers to employ. The selection can be viewed from two different points of view. Firstly, the background set can be chosen in order to represent impostors that sound similar to the speaker, referred to as dedicated impostors [2]. Another approach is to select a random set of speakers as the background set, thus expecting casual impostors who will try to represent a speaker without consideration of sex or acoustic similarity. By selecting the dedicated impostor background set, in contrast, the system may be vulnerable to speakers who sound very different from the claimed speaker [4].

The selection of the background set was done on a per speaker basis and it was decided to use the dedicated impostor approach [2]. First the $N = 20$ closest speakers to a speaker were determined using pair-wise distances between the speaker and all others. The pair-wise distance between speakers i and j with corresponding models λ_i and λ_j is

$$d(\lambda_i, \lambda_j) = \log \frac{p(X_i|\lambda_i)}{p(X_i|\lambda_j)} + \log \frac{p(X_j|\lambda_j)}{p(X_j|\lambda_i)}, \tag{3.7}$$

where $\frac{p(X_i|\lambda_i)}{p(X_i|\lambda_j)}$ is a measure of how well speaker i scores with his/her own model relative to how well speaker j scores with speaker i's model. The ratio becomes smaller as the match improves.

These N speakers are known as the close cohort set, which is denoted by $C(i)$ for speaker i. The final background set consists of the $B = 10$ maximally spread speakers from $C(i)$, denoted $B(i)$. To determine $B(i)$, the closest speaker to i is moved to $B(i)$ and B' is set to 1 (1 speaker in the background set). The next speaker c from those left in $C(i)$ to be moved to $B(i)$ is then selected as

$$c = \arg \max_{c \in C(i)} \left\{ \frac{1}{B'} \sum_{b \in B(i)} \frac{d(\lambda_b, \lambda_c)}{d(\lambda_i, \lambda_c)} \right\} \tag{3.8}$$

This procedure is repeated until $B' = B$. According to [2], the maximal spread constraint is to prevent "duplicate" speakers from being in the cohort set.

For speaker i, all other speakers (excluding i's cohort set of 10 speakers) were then used as impostors and tested using Equation 3.6. Speaker i's verification data was also tested using Equation 3.6, resulting in 1270 impostor attacks and 10 true attempts to gain access to the system (since every speaker has 10 verification sessions). This process was repeated for all speakers in the corpus, resulting in 175260 impostor attacks and 1380 true attempts.

In particular, Equation 3.6 was evaluated as follows using the cohort set and the claimed speaker model: First, $\log p(X|\lambda_c)$ was evaluated as

$$\log p(X|\lambda_c) = \frac{1}{T}\sum_{t=1}^{T} \log p(x_t|\lambda_c) \tag{3.9}$$

where T is the number of frames in the utterance and $\frac{1}{T}$ is used to normalize the score in order to compensate for different utterance durations.

$\log p(X|\lambda_{\bar{c}})$, the probability that the utterance was from an impostor was calculated using the claimed speaker's cohort set as

$$\log p(X|\lambda_{\bar{c}}) = \log\left\{\frac{1}{B}\sum_{b=1}^{B} p(X|\lambda_b)\right\} \tag{3.10}$$

where $p(X|\lambda_b)$ was calculated as in Equation 3.9.

The UBM that was used to normalize the phoneme durations in actual fact consists of 49 independent triphone models. It was constructed by simply calculating the mean and variance for each of the 49 triphones, using all observations in the enrollment set over all speakers.

The EER was then calculated by creating a list of all the likelihood ratios, sorting it and finding the threshold point where the percentage of true speakers below the threshold is equal to the percentage of false speakers above the threshold.

3.4 RESULTS

The results obtained using conventional acoustic scores, temporal features, and a combination of the two types of features, are summarized in table 3.3, and the corresponding DET curves can be seen in figure 3.5. The combined EER was obtained by taking a linear combination of the likelihood ratios obtained using phoneme duration and MFCC features, with the weighting constant determined empirically.

Table 3.3: Equal error rates obtained on the YOHO database (M+F, msc).

Feature set	EER
MFCCs	0.68%
Time	9.2%
MFCCs and phoneme durations	0.32%

Several tests have been performed on the YOHO database [15], the results of which can be seen in table 3.1. Only the test in [2] can be directly compared to the system described here, since the other tests were performed under different conditions.

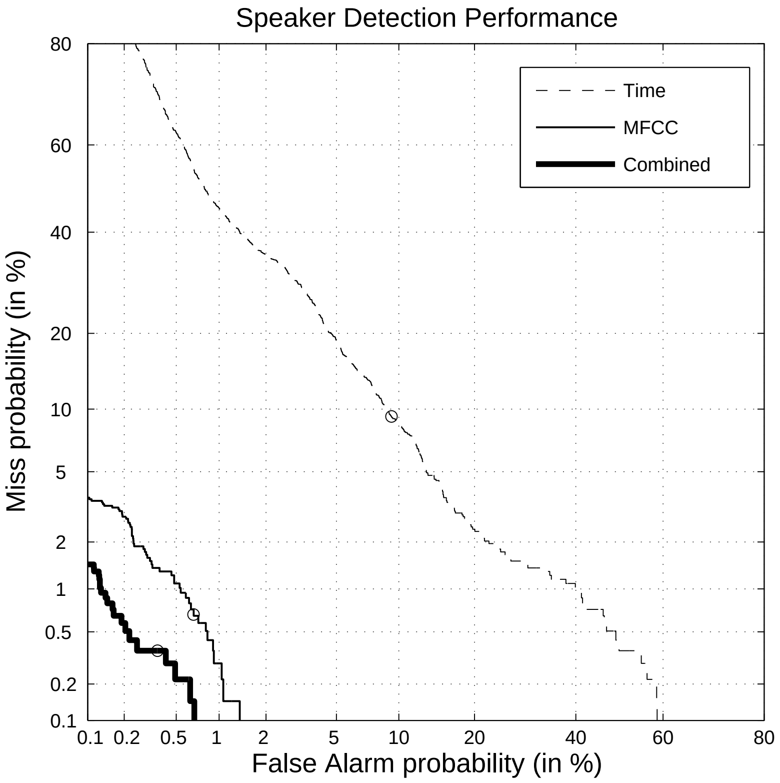

Figure 3.5: DET curves for phoneme durations, MFCC and combined features

Our results with the acoustic (MFCC) features are seen to be comparable to those achieved by other researchers. The temporal features by themselves are significantly less reliable than the acoustic features, but reduce the error rate by a factor of approximately two when combined with those features. This suggests that the temporal features are reasonably uncorrelated with the acoustic features, and the scatter plot in figure 3.6 confirms this impression. (For clarity, only 400 randomly-selected pairs of acoustic and temporal scores are shown in the figure). The correlation coefficient between the scores using the two types of features was found to be 0.201.

3.5 DISCUSSION

It has been shown that durations of context-dependent triphones constitute a feature set that can improve the accuracy of speaker verification systems to a significant degree. Although our results were obtained with an HMM in a text-dependent application, it seems likely that

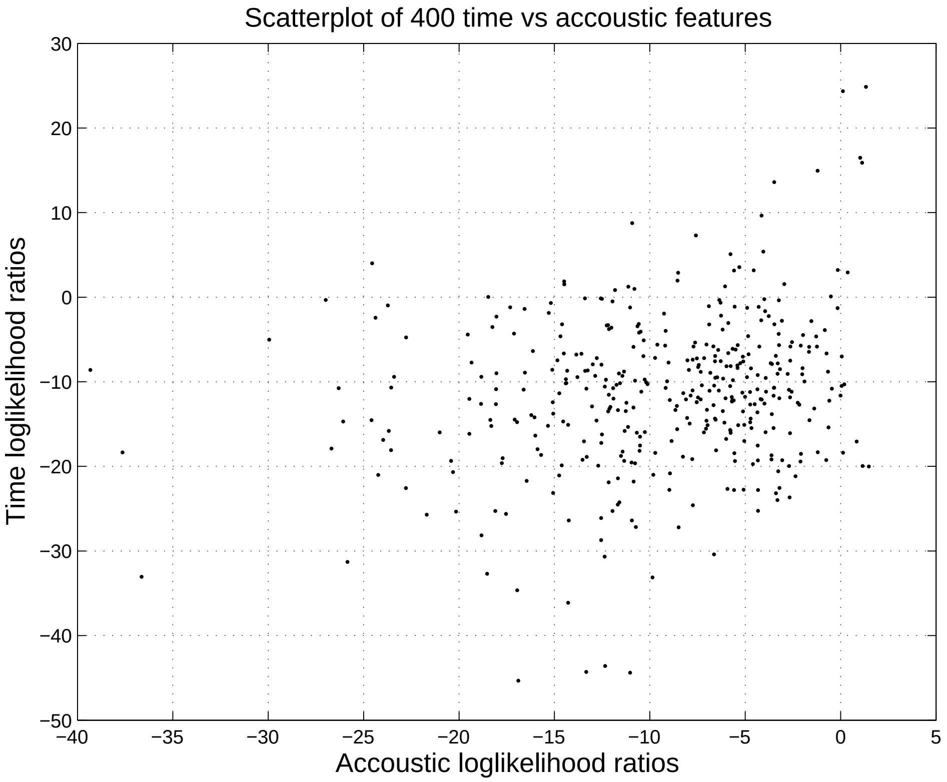

Figure 3.6: Correlation between the temporal and acoustic features

an equally low correlation between acoustic and temporal scores will be found with other classes of SV systems. This was indeed true for the text-independent speaker recognition system in [8]. We are therefore confident that similar improvements will be obtained in other SV systems.

Our current system uses the temporal features in a fairly crude fashion: all triphones are modelled with independent Gaussian distributions, and all triphone scores are combined with equal weight. It will be interesting to see how much improvement can be obtained with more sophisticated models (which, for example, assign greater weight to more discriminative triphones or those which have been observed more frequently, or consider correlations between the different triphone durations).

Initial experiments with more sophisticated duration models (section 4.3) suggest that accounting for effects such as speech rate should further improve the discriminative power of duration models. Modelling effects such as the position of the phoneme in the utterance should produce additional improvements and will be discussed in more detail in section 5.

Another promising area for further research is related to the relative robustness of temporal and acoustic features to factors such as channel variation and speaker condition. In [50] temporal information was found to be more robust against channel interference than MFCCs, but that result needs to be tested on a more substantial corpus. (Unfortunately, YOHO is not suitable for this purpose, since variable recording conditions were not part of the YOHO protocol.)

Since triphone durations are a very compact descriptor of an utterance, this feature set may also be useful in detecting and deflecting replay attacks. A database of durations during previous verification sessions may be maintained conveniently. One can then calculate the probability of a specific triphone or a sequence of triphones having the same (within some small threshold) duration, setting a threshold for an acceptable probability and rejecting the speaker as an impostor launching a replay attack if the probability is lower than the threshold. Overall, it seems as if triphone durations are likely to be a useful addition to almost any toolbox for SV system development.

3.6 CONCLUSION

A text-dependent speaker verification system based on Hidden Markov Models was presented. A set of features, based on the temporal duration of context-dependent phonemes, is used in order to distinguish amongst speakers. Our approach was tested using the YOHO corpus; it was found that the HMM-based system achieved an equal error rate (EER) of 0.68% using conventional (acoustic) features and an EER of 0.32% when the phoneme duration features were combined with the acoustic features. This compares well with state-of-the-art results on the same test, and shows the value of the temporal features for speaker verification. These features may also be useful for other purposes, such as the detection of replay attacks, or for improving the robustness of speaker-verification systems to channel or speaker variations. Our results confirm earlier findings obtained on text-independent speaker recognition [8] and text-dependent speaker verification [50] tasks, and contain a number of suggestions on further possible improvements.

Chapter Four

Speech Rate Normalization of Phoneme Durations

4.1 INTRODUCTION

Speech is the most natural way for humans to communicate with each other. Over the past decade, much work has been done in man-machine communications in order to incorporate speech as a new modality in multimedia applications [2]. We are interested in two particular disciplines which have received considerable interest: speech recognition, in which the aim is for the machine to extract and understand the linguistic message in the speech, and speaker recognition, where the goal is to identify, recognize or verify the speaker responsible for producing the speech. As mentioned in chapter 3, there are several factors that have limited the success of integrating speech into machine communications such as transmission line degradations, channel mismatch and speech rate variability [41]. Speech rate variability has been found to be significant in increasing the error rate of speech recognition, especially when it deviates greatly from the training data [41]. Several ways have been proposed to remove this speech rate variability by way of speech rate normalization, with some recent proposals such as [7].

In chapter 3 and in [50] we have shown that phoneme durations is a useful high-level feature that can be used effectively in speaker verification. Since our goal is to create a phoneme duration model that can be useful in any speaker verification setup, we continue to investigate factors that can be addressed to make our model more robust and more accurate. In this chapter, we demonstrate that speech-rate normalization can be applied to further improve the accuracy of this feature. The refinement of the duration model is described in detail, in particular the way in which it was constructed. The model is then applied in a novel speech rate normalization technique on the YOHO corpus and the resulting normalized durations are submitted to a speaker verification system. The equal error rate (EER) using the normalized durations is then compared with the EER using unnormalized durations, and also with the EER when duration information is not employed.

4.2 REFINEMENT OF PREDICTED PHONEME DURATION

In the preceding sections we assumed that the duration of a particular phoneme spoken by a given speaker is described by a normal distribution, independently of the durations of other phonemes in the utterance. This is clearly not realistic - for example, the speaking rate will tend to influence all the phonemes in an utterance in a correlated manner. It is therefore interesting to ask whether a more detailed duration model can be developed, to account for such influences on phoneme durations. A more complete model could also include factors such as the position of the phoneme in the word or utterance, but for now we have concentrated on the influence of speaking rate.

To do this, we developed a model for predicting the duration of a phoneme of the form:

$$t(ms) = \begin{bmatrix} t_{f,s} & \chi_{w,s} \end{bmatrix} \cdot \lambda_{\mathbf{w,f,s}}^{\mathbf{T}} \tag{4.1}$$

where $t_{f,s}$ is the speaker-specific mean estimate of the phoneme duration for phoneme f and $\chi_{w,s}$ is the “stretch factor” for a specific word w spoken by s. This is determined as

$$\chi_{w,s} = \frac{\tau - \hat{\tau}}{\sum \sigma_n} \tag{4.2}$$

Here τ is the true word length, $\hat{\tau}$ is the estimated word length that was determined by summing the means of the phonemes constituting the word and $\sum \sigma_n$ is the sum of the standard deviations of these phonemes. Finally, $\lambda_{w,f,s}$ is the vector of parameters obtained from a General Linear Model (GLM) in order to model the effect of the speech rate on the specific phoneme. The GLM has the form

$$y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k \tag{4.3}$$

and the coefficient vector $\mathbf{b}$ is determined as:

$$\mathbf{b} = \left(X'X\right)^{-1} X'Y \tag{4.4}$$

This novel speech rate normalization technique was then applied to the testing procedure described above as follows. For every test session, the parameters $\lambda_{f,s}$ and $\lambda_{w,s}$ of the claimed speaker were used to normalize the session's phonemes with regard to speech rate and produce

$$t_{norm} = \frac{t_{measured} \cdot t_{f,s}}{\left[\begin{array}{cc} t_{f,s} & \chi_{w,s} \end{array}\right] \cdot \lambda^{\mathbf{T}}_{\mathbf{w,f,s}}} \tag{4.5}$$

4.2.1 THE EFFECT OF NORMALIZATION ON THE CORRELATION WITH ACOUSTIC SCORES

When two features are fused in order to obtain a new feature, the performance of the feature will only improve if the two features are uncorrelated to a certain degree. High-level features have received considerable interest over the past couple of years and have been shown to contain valuable uncorrelated (to Mel Frequency Cepstral Coefficients (MFCCs)) information with regard to speaker identity [29]. It was thus interesting to note that the correlation between our duration features and the MFCCs was only 0.24. After normalizing, the correlation decreases even further to 0.19 - the exact reason for this is an interesting field for further research. Scatterplots of the features can be seen in Figures 4.1 and 4.2.

4.3 PROCEDURE AND RESULTS

The same testing procedure as described in section 3.3 was used to test the efficiency of the proposed speech rate normalization procedure, apart from the fact that the claimed identity's ASR model was used for segmentation. This contributed to the 2.1% improvement in duration model EER as described in chapter 3.

The corpus used was YOHO, as described in appendix 1. In addition, a test was conducted to measure the accuracy of the models compared to two other approaches: (a) the duration of each triphone is assumed to be constant and (b) the duration of each triphone is assumed to scale linearly with the stretch factor. The results are summarized in Table 4.1, and clearly show the importance of rate normalization in accounting for triphone durations. The second column of Table 4.1 contains the mean-squared difference between the actual and predicted phoneme durations (averaged over all test utterances), and the third column contains the standard error of this estimate (that is, the standard deviation of all differences

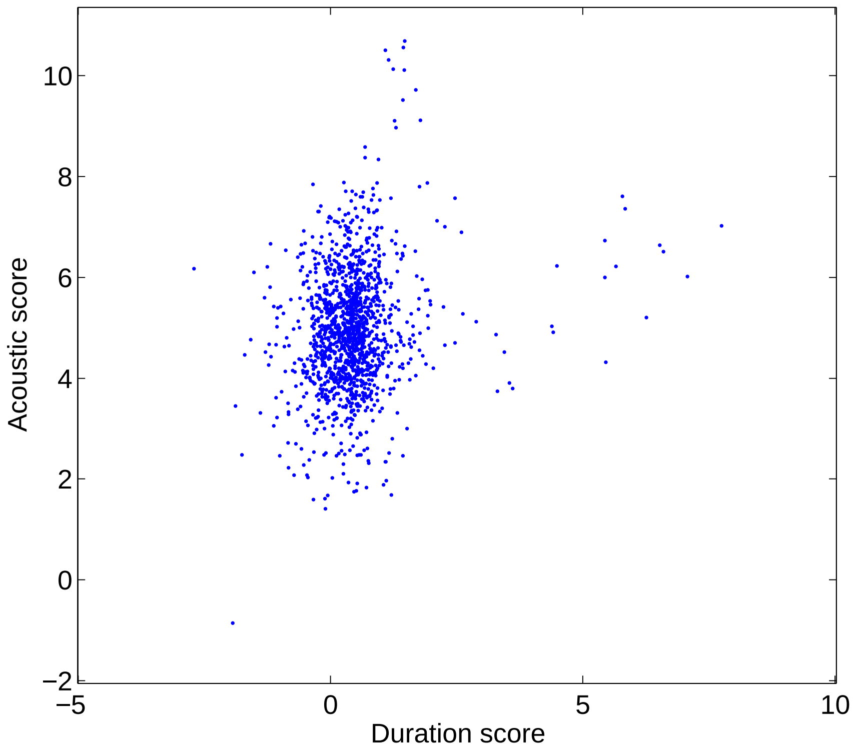

Figure 4.1: Correlation between the normalized temporal and acoustic features.

divided by the square root of the total number of phonemes in these utterances). Our general linear model is also seen to be significantly more accurate than a simple linear scaling of phoneme durations.

Table 4.1: Comparison of three approaches to the modelling of speech rate.

Model	**MSE (msec)**	**Standard error of MSE (msec)**
Constant speaker-specific duration per phoneme	777.25	65.94
Linear scaling of phoneme durations	522.33	25.46
General linear model of phoneme durations	430.59	16.86

Speaker verification tests were also conducted with and without the duration normalization procedure, giving the results in Table 4.2. Despite the significant improvement in modelling accuracy achieved with speech-rate normalization, such normalization gives only a small reduction in error rate when duration information is used by itself for speaker verification (7.1% to 6.8%). However, in combination with the acoustic scores, the normalization

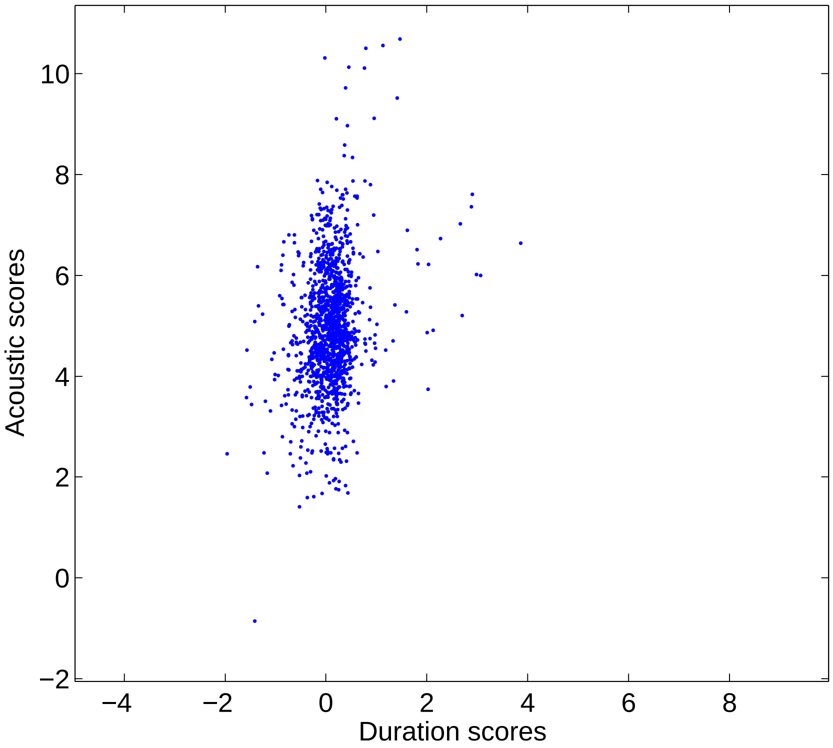

Figure 4.2: Correlation between the unnormalized temporal and acoustic features.

procedure is shown to be highly efficient. This difference is also illustrated in the detection error trade-off (DET) curves shown in Figure 4.3.

Table 4.2: Equal error rates obtained on the YOHO database (M+F, msc).

Feature set	**EER before normalization**	**EER after normalization**
MFCCs	0.68%	0.68%
Time	7.1%	6.8%
MFCCs and time	0.29%	0.21%

Our results (0.21% EER) with the combined acoustic (MFCC) and temporal features are seen to compare favorably with those achieved by other researchers (table 3.1 and [15]). The temporal features by themselves are significantly less reliable than the acoustic features, but reduce the error rate by a factor of approximately three when combined with those features. The duration features were linearly combined with the acoustic features, the parameters determined by exhaustive testing and optimization. This phenomenon was foreseen by Reynolds et al. [16] when they mentioned that higher-level features need to be investigated and that these would probably not give good performance on their own (as we experienced),

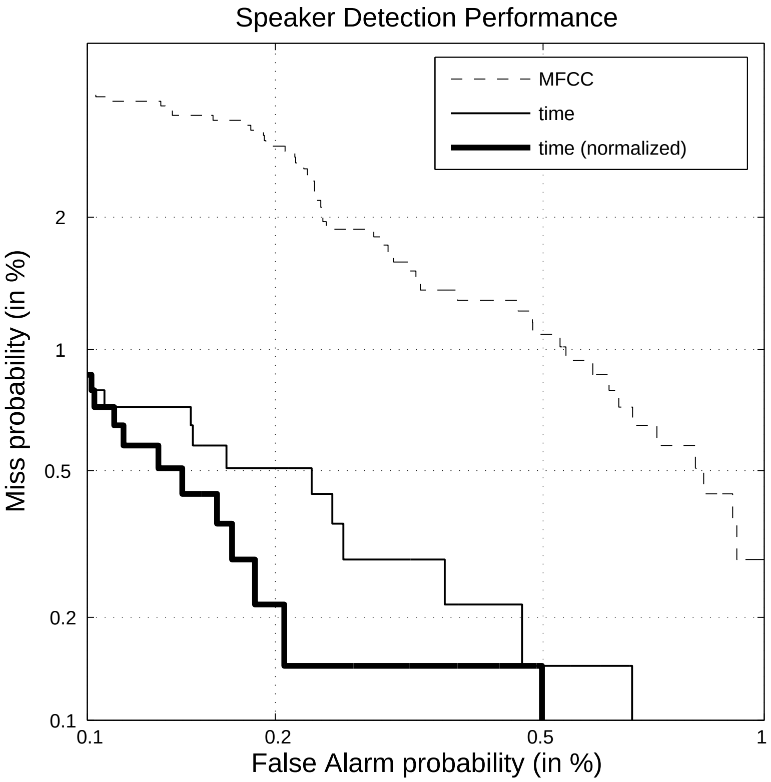

Figure 4.3: DET curves for MFCC and combined features (time unnormalized and normalized).

but that they could beneficially be fused with more conventional features to obtain good performance. Similar observations were made in [29] when EERs as high as 26% were observed, but the fusion with conventional features produced an improvement of 71%. Our observations suggest that the duration features are reasonably uncorrelated with the acoustic features, and the scatter plots in Figures 4.1and 4.2 confirm this impression.

4.4 CONCLUSION AND FUTURE DIRECTIONS

A novel approach to speech rate normalization was presented in this chapter. Models were constructed to model the way in which speech rate variation of a specific speaker influences the duration of phonemes. The models were evaluated in two ways. Firstly, the mean square error in phoneme duration based on our normalization was compared to the same error when such normalization is not applied. The second evaluation used the durations of context-dependent phonemes in speaker verification. Both methods showed that this approach to normalization is indeed effective to counteract the effect of variable speaking rates.

In particular, we have shown that phoneme durations constitute a speaker trait that can improve speaker recognition systems. Durations are subject to various influences, such as changes in speaking rate. Tests on the YOHO corpus have confirmed that speech rate normalization can improve the robustness and accuracy of phoneme durations as a feature in speaker recognition. Speech recognition will also benefit from speech rate normalization, as has been proposed in [7]. Further research should be done on other corpora: differing speech rates were not part of the YOHO protocol, and we expect that normalization will be even more significant with more variable corpora.

It was also noted that occasional failures of the automatic alignment process (especially erroneous boundary detection for phonemes at the beginning and end of phrases) contributed significantly to the errors that occur when using temporal information by itself. Rectifying this problem is expected to enhance the power of duration features significantly. A possible remedy for this problem is to incorporate acoustic scores in weighting the duration models. A low acoustic score will then indicate that the particular phoneme has not been recognized with high reliability and can thus be discarded or assigned a lower weight than durations detected with high reliability. This approach can also be used if duration features are to be used in a text-independent application, which will be the next step towards a practical implementation of this feature.

Although this research has been done with an HMM-based text-dependent speaker verification system, results such as those obtained with the text-independent system from [8] suggest that the low correlation observed between temporal and acoustic scores can be beneficial in other classes of speaker verification systems.

The duration models we have used are still rather crude since all triphones are assigned equal weights and are modelled by independent Gaussian distributions. The models can probably be improved by considering other factors such as the frequency of observation of triphones, the acoustic reliability of the observation, correlation between triphones and giving greater weight to more discriminative triphones.

It will also be interesting to investigate the relationship between an increase in the number of free parameters and the subsequent increase in accuracy.

Overall, it seems as if triphone durations are likely to be a useful addition to almost any toolbox for speaker verification system development.

CHAPTER FIVE

REFINEMENT OF DURATION MODELS

5.1 INTRODUCTION

Phoneme durations are known be be affected by several linguistic factors [13] such as "position in word", "position in sentence", "stress versus no stress" and also "speaking rate". Speaking rate has been addressed in chapter 4 and the partial compensation for rate shown to be valuable for speaker verification. In this chapter we implement some of the suggestions we made in chapter 4, section 4.4 by incorporating two of the other prominent factors known to influence phoneme durations; "position in word" and "position in sentence". We show how accounting for each of these factors improves speaker verification accuracy.

5.2 EXPERIMENTAL FRAMEWORK: DATABASE, PROTOCOL AND SYSTEMS

Experiments were conducted on the YOHO corpus (appendix 1). The test procedure described in section 3.3 was used, but in addition session 969 of speaker 240 was omitted from the testing procedure as she used a falsetto voice [15]. Other mistakes reported in [15] were corrected; empty headers for speaker 277 session 538 were replaced with correct headers and the label file for speaker 101 in enrollment session 2 was changed from the prompted phrase ($56 - 73 - 79$) to what was actually said ($53 - 73 - 79$).

Each experiment is briefly described in the sections that follow. Experiments 1 and 2 are subcategorized into sections a (not accounting for "position in sentence") and b (explicitly modeling "position in sentence"). "Position in sentence" was accounted for by creating 3 separate GMMs for sentence initial, sentence final and mid sentence contexts.

- *Experiment 1: Monophone duration models.* A complete SV test was conducted, characterizing the phonemes as monophones only. Single GMMs were trained for each of the 18 monophones. This test was also extended to test the effect of "position in sentence" on monophone durations (experiment 1b).

- *Experiment 2: Biphone duration models.* Biphones were used as a way of incorporating context into the duration models. As noted in [13], vowels preceding a voiced rather than an unvoiced plosive are generally longer. Taking the context of a phoneme into account should thus improve the modeling accuracy.

  - *Experiment 2.1: Biphone duration models, taking left context into account.* Biphones were constructed by taking the left context (or preceding phoneme) into account.

  - *Experiment 2.2: Biphone duration models, taking right context into account.* Biphones were constructed by taking the right context (or following phoneme) into account.

- *Experiment 3: Triphone duration models.* Triphones were modeled by taking both left and right context into account. "Position in sentence" was not taken into account.

- *Experiment 4: Analyzing the effect of position in sentence.* One-sided analysis of variance (ANOVA) was performed on each triphone to determine whether the different positions in a sentence could be viewed as groups with significantly different means. The hypothesis in each case was that the means of the 3 groups (initial, mid and final) are the same. In order to perform this test on a large enough dataset, the durations of each speaker was normalized by his/her sentence-independent triphone mean. The results of ANOVA indicated significant differences between group means for all triphones except some fricatives ("f+ih", "ih-f+t", "th+er", "k-s+t"), a nasal ("ay-n+t") and a voiced stop ("ao-t+iy"). Although this does not necessarily mean that every group mean for each phoneme is significantly different from both other group means, it is sufficient evidence for the claim that "position in sentence" is an important factor to be considered in duration modeling. The position in sentence factor was thus taken into account while modeling triphones.

- *Experiment 5: Analyzing the effect of position in word.* The "position in sentence" factor was taken into account as well as "position in word", while modeling triphones.

There are only a few triphones that are affected by “position in word”, since taking both left and right context into account limits the number of unique triphones occurring both in word initial as well as final positions. Triphones that occur in both contexts are shown in table 5.1 together with examples and the UBM means for the respective cases.

- *Experiment 6: Addressing automatic alignment errors.* The number of duration samples used for verification were filtered. If a duration was more than 3 standard deviations away from the triphone mean and the LL was positive, the duration was rejected.
- *Experiment 7: Sequential forward selection (SFS).* Sequential feature selection (SFS) was applied. Duration models were trained for each speaker using only enrollment sessions $1 - 3$. Enrollment session 4 was then used in a SV setup as described in section 3.3 to determine which phonemes are most useful. All phonemes which made a negative contribution to SV accuracy was then removed from the experiment on the test set.

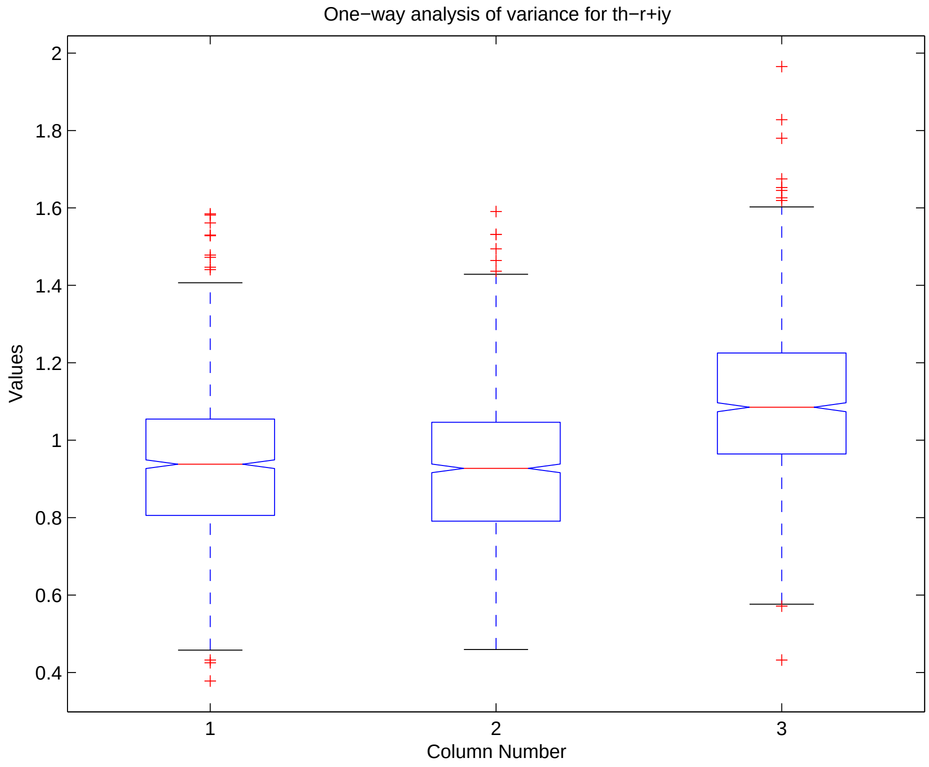

Figure 5.1: Results of ANOVA for th-r+iy, testing for significant differences between the normalized durations of this triphone when occurring in sentence initial (1), mid sentence (2) and sentence final (3) positions.

Table 5.1: Triphones occurring in both word initial and final positions.

Triphone	Word initial example	Word final example	UBM mean (ms)	
			Word initial	Word final
n-ay+n	ninety-seven	seventy-nine	105	225
s-eh+v	seventy-one	twenty-seven	79	113
s-ih+k	sixty-four	fourty-six	57	101
ih-k+s	sixty-four	fourty-six	42	75
eh-v+n	seventy-one	twenty-seven	60	86

5.3 RESULTS

The results of experiments $1 - 7$ can be seen in table 5.2. They compare well to results obtained in Chapters 3 and 4, but cannot be directly compared since the testing procedure was changed to exclude known errors in the YOHO corpus (see section 5.2). The best EER of 0.0949% obtained in experiment 6 can be seen to compare well with that obtained by others (table 5.3). Because of deviations from the original test protocol proposed in [20], the results cannot be directly compared. The result in [2] can be directly compared apart from the fact that two unspecified test sessions were omitted from the test protocol. An interesting observation in [52] was made in that there are two particular speakers who have a detrimental effect on the EER, since they have EERs of approximately 5% and 10% respectively. By completely removing these two (speakers 162 and 169), an EER of 0.13% was obtained. When we did the same, our EER dropped to 0.03527%.

When the phoneme duration results are fused with the GMM's, the overall results can be seen to improve significantly, whereas the addition of the HMM results makes much less of an impact. This can be attributed to among others the very low correlation between the phoneme durations and GMM results (0.0365). The correlation between the HMM and duration results is much higher (0.2158) but still relatively low while the correlation between the HMM and GMM results is significant (0.5049).

The effect of the different factors are clearly visible from the bar graphs depicting phoneme durations from randomly selected speakers. Figures 5.2,5.3,5.4,5.5,5.6 illustrate the effect of "position in word" on phoneme durations. It is clear that the same triphone occuring towards the end of the word is generally longer than when occuring in the onset. Figures 5.10,5.11,5.12 on the other hand illustrates the effect of "position in sentence" on some phonemes. The EERs obtained by incorporating these factors in the duration model are illustrated in the DET curves in figures 5.8,5.9,5.9,5.13,5.14.

Table 5.2: EER for different experiments.

EER (%)					
Exp. #	**Duration**	**Duration + GMM**	**Duration + GMM + HMM**	**# Triphones rejected**	**# Triphones used**
$1a$	15.25	0.2564	0.1922		
$1b$	14.33	0.2615	0.1922		
$2.1a$	8.34	0.2249	0.1501		
$2.1b$	6.91	0.1558	0.1456		
$2.2a$	8.82	0.1456	0.1353		
$2.2b$	7.75	0.1518	0.1456		
3	6.35	0.1450	0.11705		
4	5.71	0.1456	0.1313		
5	4.78	0.1290	0.1187		
6	4.57	0.1193	0.0949	$8.57 \cdot 10^4$	$1.66 \cdot 10^7$
7	4.40	0.1290	0.1159	$1.69 \cdot 10^6$	$1.50 \cdot 10^7$

Table 5.3: Equal error rates obtained on the YOHO database by other researchers. The superscripts distinguish between experiments that use impostors that have been seen^1, and those that use impostors not yet seen^2

Research group	Type of system	EER
ITT [4]	Continuous Speech Recognition (CSR)	1.7%
ITT [15]	Neural Network (NN)	0.5%
Rutgers [51]	Neural Tree Network (NTN)	0.65%
Reynolds [2]	Gaussian Mixture Model	0.58%
Wan & Campbell [52]	SVM, normalized polynomial kernel	$0.34\%^1$, $0.59\%^2$
Campbell & Assaleh [53]	Polynomial classifier	$0.18\%^1$, $0.31\%^2$
MIT/LL's [2]	Gaussian Mixture Model	0.51%
van Heerden & Barnard	GMM fused with phoneme durations	0.0949%

5.4 CONCLUSION AND FUTURE DIRECTIONS

This chapter investigated if and how much SV accuracy could be improved by incorporating factors known to influence phoneme durations in the duration model. Accounting for "position in sentence" was shown to be beneficial and "position in word" even more so. Furthermore, phoneme durations were shown to be a valuable attribute by itself and especially useful when fused with cepstral features.

While refining these models we constantly encountered data scarcity problems. Some triphones were not observed frequently enough to be reliably modeled and subsequently we had to back off to the UBM model in certain cases. We believe that although we have conclusively shown that phoneme durations are useful in a text-dependent speaker verification

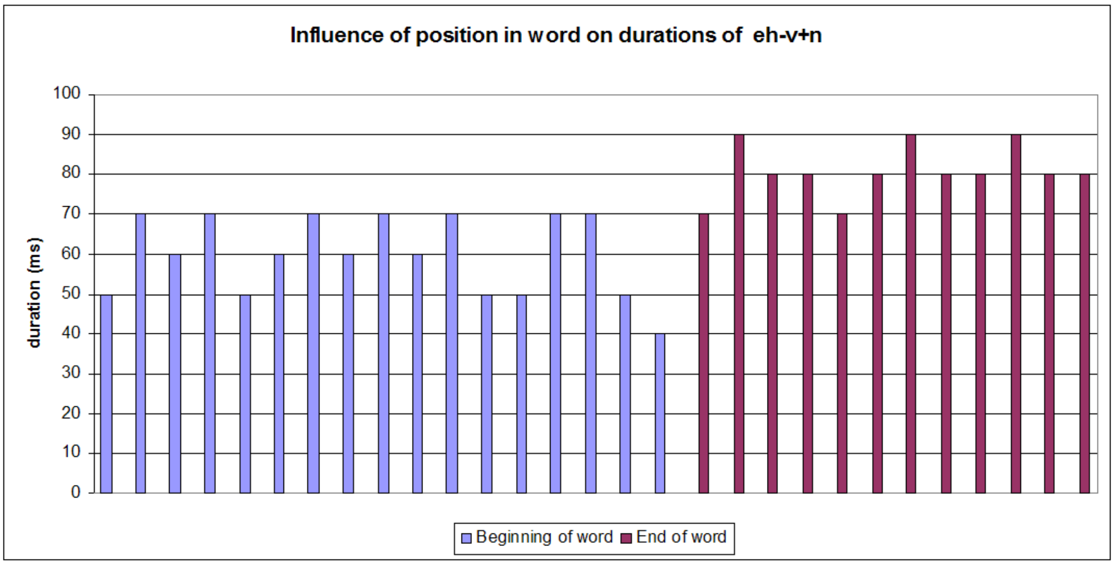

Figure 5.2: Bar graph illustrating the effect of “position in word” on eh-v+n.

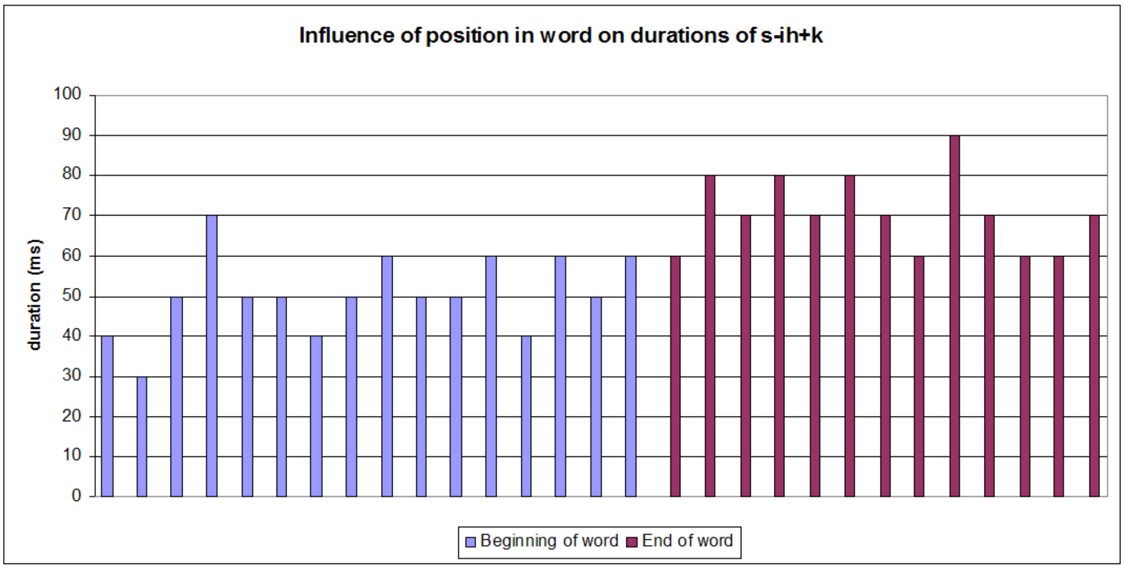

Figure 5.3: Bar graph illustrating the effect of “position in word” on s-ih+k.

system, the feature can be substantially improved with either more training data (which is an unlikely scenario in many speaker verification applications) or sophisticated duration prediction techniques that can maximally utilize the information available (such as correlations between phonemes and phonetic classes).

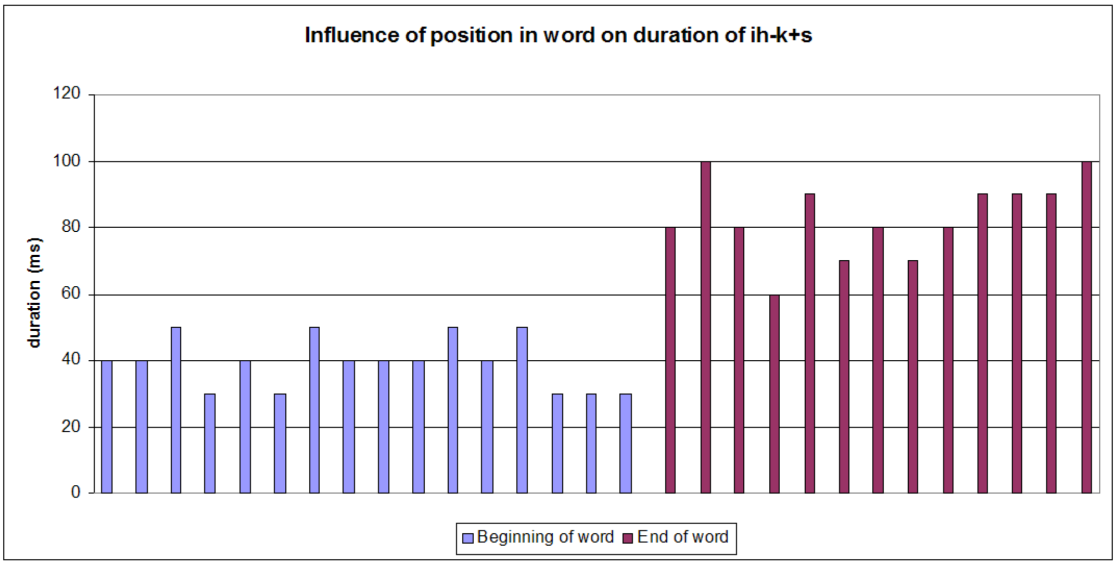

Figure 5.4: Bar graph illustrating the effect of “position in word” on ih-k+s.

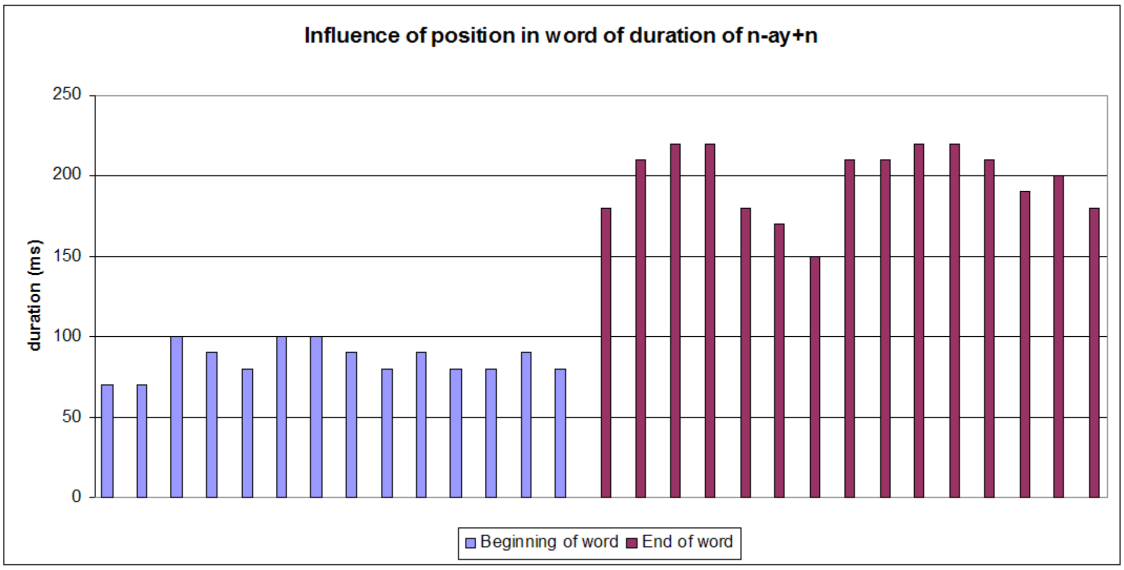

Figure 5.5: Bar graph illustrating the effect of “position in word” on n-ay+n.

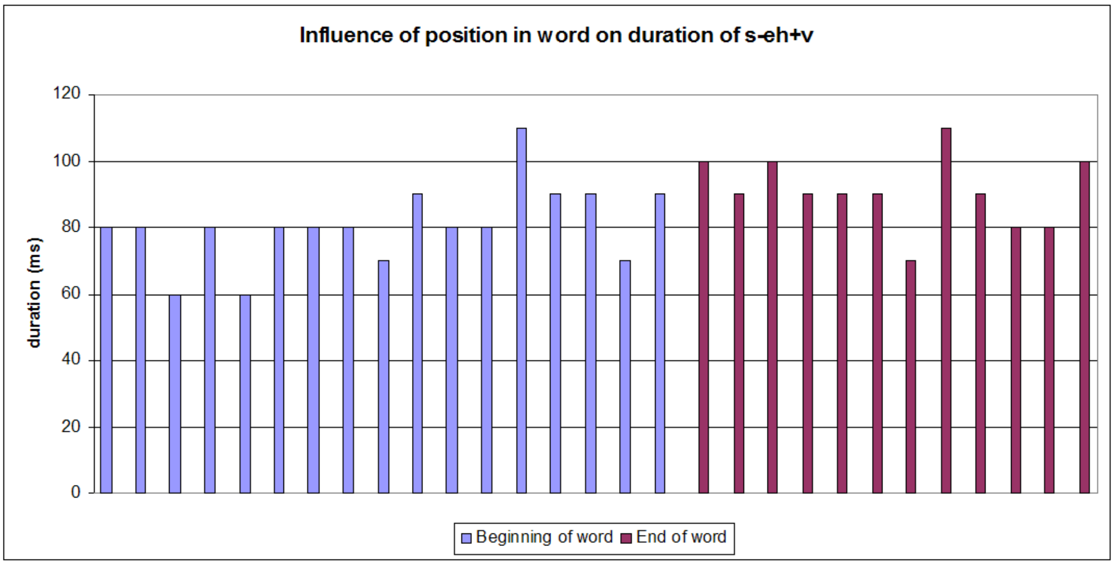

Figure 5.6: Bar graph illustrating the effect of "position in word" on s-eh+v.

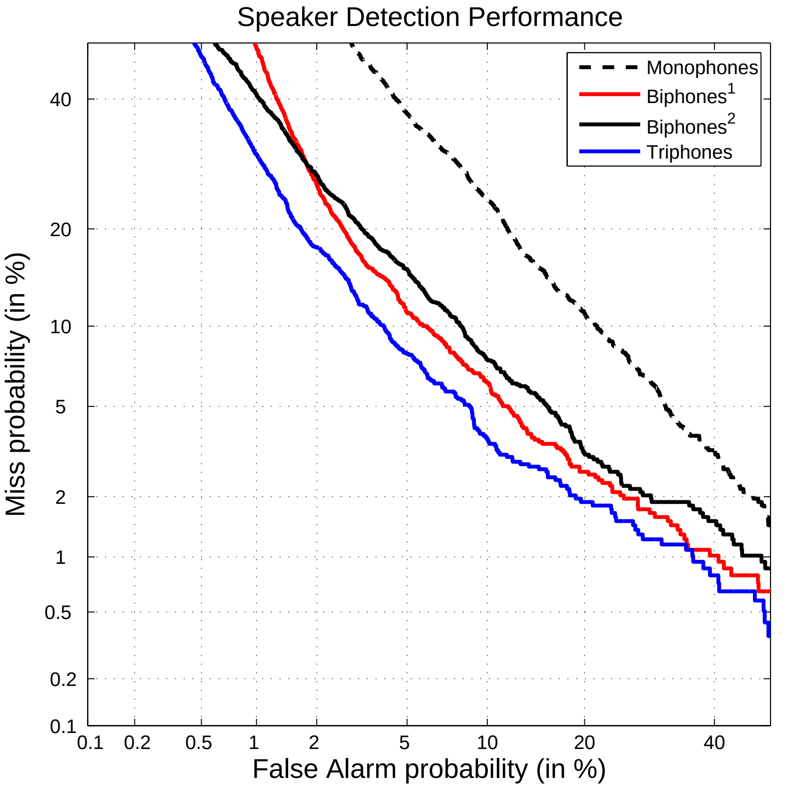

Figure 5.7: DET curve illustrating the improvement in SV accuracy when using triphones over biphones or monophones. [1] Constructing biphones taking preceding phoneme into account. [2] Constructing biphones taking following phoneme into account.

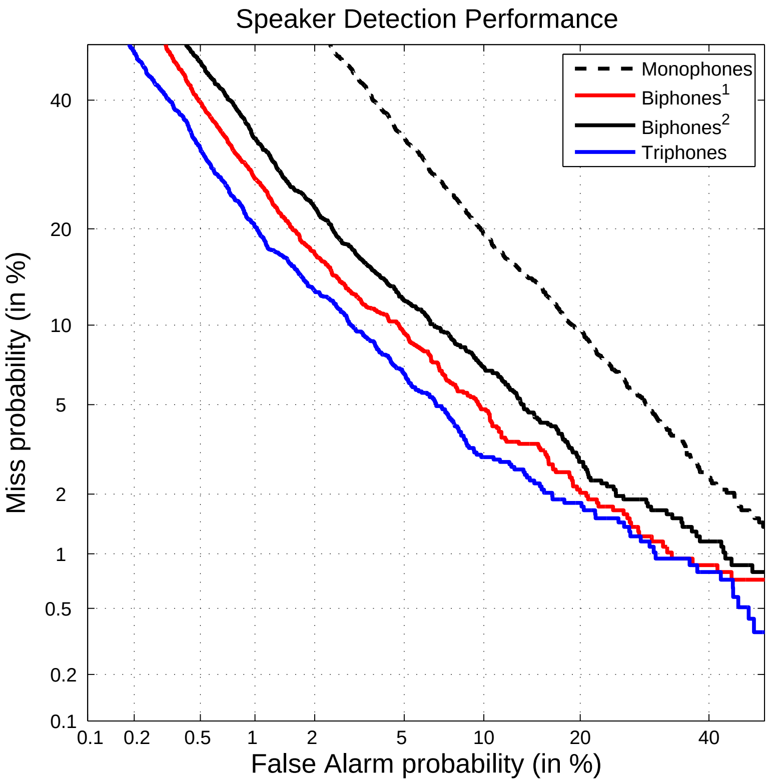

Figure 5.8: DET curve illustrating the improvement in SV accuracy when using triphones over biphones or monophones (modeling position in sentence as well). [1] Constructing biphones taking preceding phoneme into account. [2] Constructing biphones taking following phoneme into account.

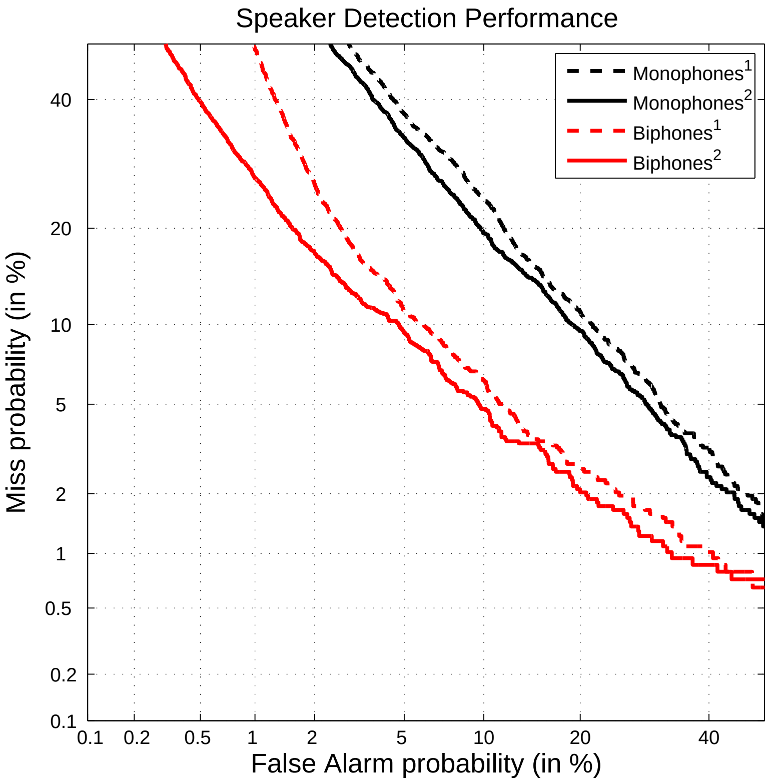

Figure 5.9: DET curve illustrating the improvement in SV accuracy when taking the "position in sentence" factor into account for monophones and biphones. [1] Ignoring "position in sentence" factor. [2] Explicitely modeling "position in sentence" factor.

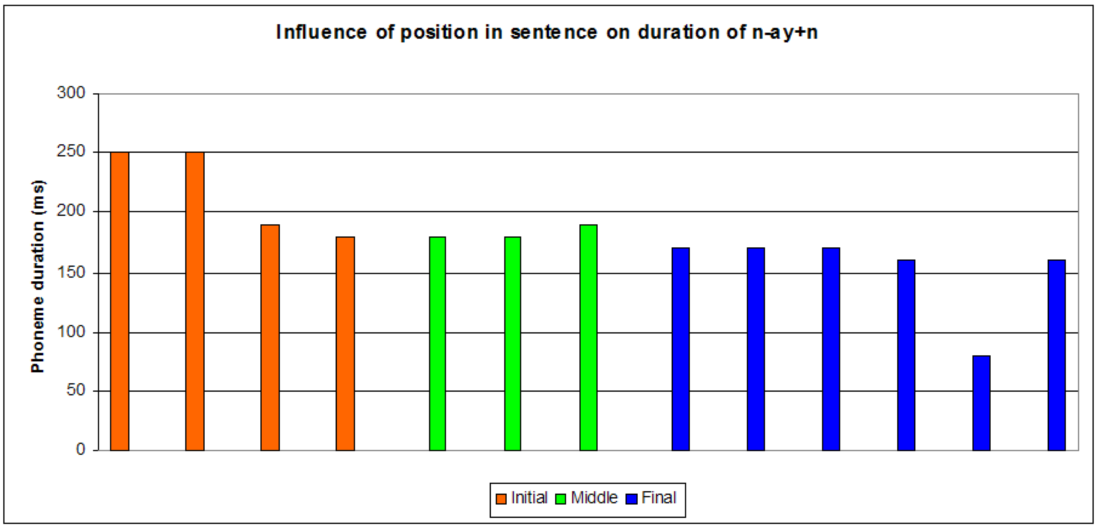

Figure 5.10: Bar graph illustrating the effect of "position in sentence" on n-ay+n.

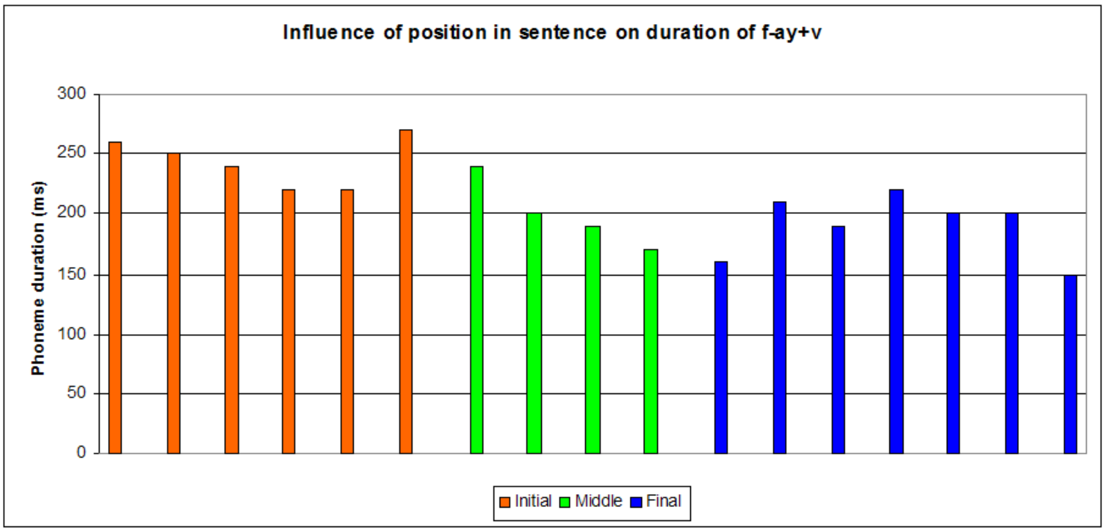

Figure 5.11: Bar graph illustrating the effect of "position in sentence" on f-ay+v.

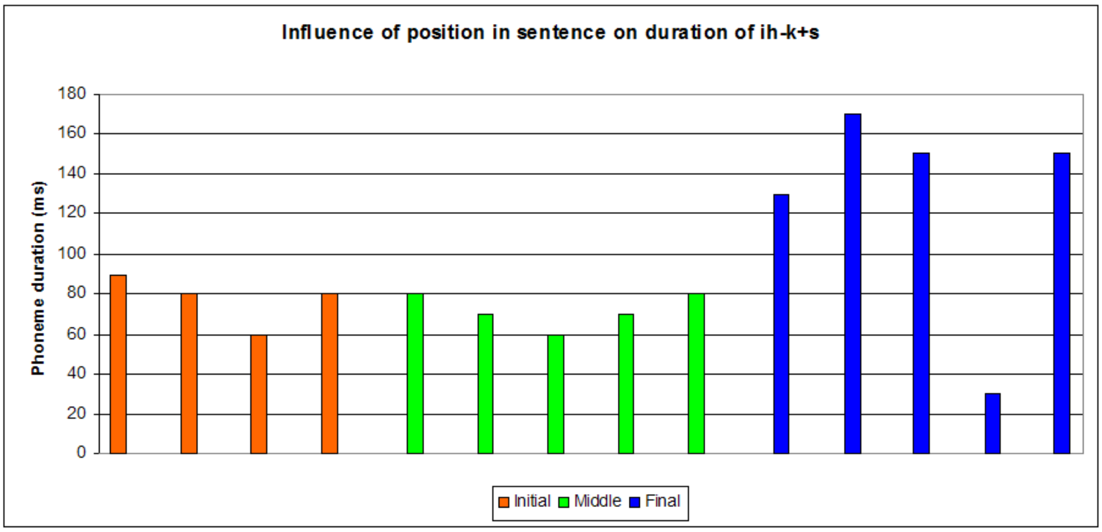

Figure 5.12: Bar graph illustrating the effect of "position in sentence" on ih-k+s.

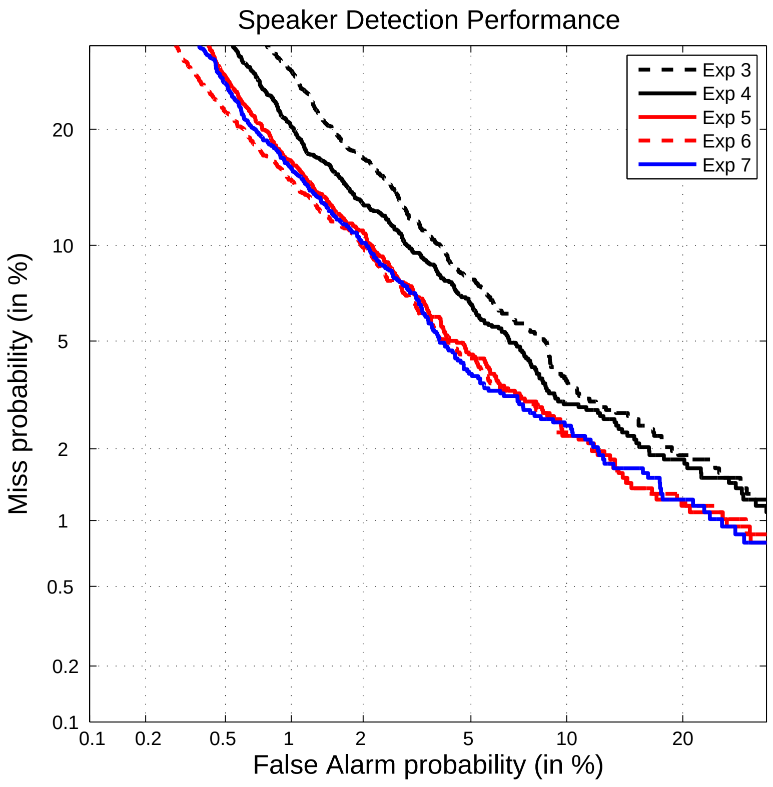

Figure 5.13: DET curve illustrating the improvement in SV accuracy for experiments $3 - 7$, using duration features only.

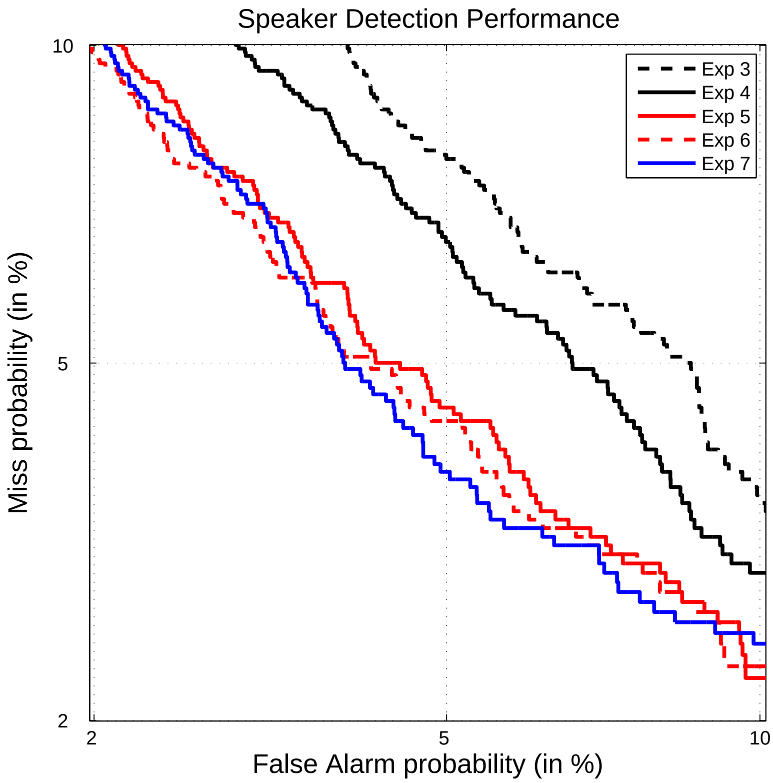

Figure 5.14: Closer look at the EER operating region in the DET curve illustrated in figure 5.13.

CHAPTER SIX

SPEAKER SPECIFIC VARIABILITY OF PHONEME DURATIONS

6.1 INTRODUCTION

Developing accurate phoneme duration models has been a topic of discussion for several years, especially with regard to the potential benefits for automatic speech recognition (ASR) [13]. In [36],[37] as well as chapters 3,4,5 we showed that accurate phoneme duration models can significantly improve state of the art speaker recognition (SR) systems in a text-dependent environment. Data scarcity has been identified as a pitfall and needs to be addressed by either more training data, or models that predict missing phoneme durations. Also, for practical applications of both ASR and speaker recognition, duration models have to be developed for text-independent speech. This is not a trivial problem as there are many factors influencing the duration of phonemes in text-independent speech, such as position in word, position in sentence, stress, preceding and following phonemes, speech rate etc. Although the work done in [36] was in a text-dependent environment, it did confirm earlier findings by [7] that phoneme durations are also speaker-specific to a large extent, which adds another dimension to the model estimation. All these factors contribute to making data scarcity a significant obstacle to characterizing phoneme durations accurately. This obstacle, which was identified in 1988 already by Crystal and House [12], remains arguably the most significant one to the more general use of phoneme durations.

An attempt to estimate the individual contributions of the abovementioned factors to the total variance was made in [13]. A hierarchical analysis of variance was performed and it was found that much of the variance can indeed be explained by these factors. Because of the type of ANOVA performed, it was not possible to examine interactions among the factors, which may omit important information. Duration patterns were also modelled in [8],[9],[10],[11] in order to improve speaker recognition performance. It was observed that significant improvements in accuracy can be achieved by separately modelling word durations, single phoneme durations and state durations using 3-state hidden Markov models (HMMs). Data sparseness was addressed in all cases by a back-off technique, through which word-models would be backed off to triphone models and the latter to single phoneme models. This ignores the effect of the specific factor being addressed on the particular phoneme. Rao Gadde [10] also performed a simple speech rate normalization. The speech rate was calculated as the number of phonemes per second. By applying this simple normalization technique, a consistent improvement in word recognition was observed over several databases.

Taken together, these studies are strong evidence that accurate phoneme duration models can greatly benefit both ASR and speaker recognition. However, no sophisticated model exists yet because of data scarcity (which limits the number of factors that can be modeled), the many different factors which have an influence on the duration of phonemes and the fact that interaction effects between the different factors are not incorporated into the models.

In this concluding chapter we present some introductory work towards the goal of building alternative models that more efficiently incorporate the abovementioned factors and their interactions. In particular, we have focused on two of the factors that have been found to be important, namely “speaker”, and “phoneme type/class”. Our objective was to see if it is possible to make better duration predictions of unknown speakers than the back-off approach, given a model that was trained from other speakers’ data. We also believe that certain phonemes are more predictable than others and that certain classes of phonemes tend to have greater predictive value for the durations of others. All of these experiments were conducted on the TIMIT corpus because of the availability of accurate manual phoneme segmentations.

6.2 TIMIT CORPUS

The TIMIT corpus is a speech corpus of 630 speakers from eight major dialect regions in the United States. Each speaker spoke 10 utterances resulting in 6300 utterances in TIMIT. The training set consists of 462 speakers, which comprise 326 males and 136 females. Three types of sentences were read: sx, si and sa. The sx sentences were read from a list of 450 phonetically balanced sentences that were designed at MIT, the si sentences from 1890

phonetically diverse sentences designed at TI and two dialect sentences designed at SRI. The test set consists of 168 speakers, which were selected so that no sentence text appears in both the training and test set. The whole of the TIMIT corpus was hand-labeled and segmented.

6.3 MODELLING APPROACH

The modelling approach we took was influenced by two questions. From a theoretical viewpoint we wanted to know if and how different phonemes' durations co-vary under different conditions such as those mentioned above. In particular we decided to investigate this question under the "speaker" condition. In practical terms, we wanted to see whether it is possible to reduce the data requirements of phoneme duration predictions by using inter-phoneme information. This knowledge would be useful for scenarios where data scarcity is an issue.

As already mentioned, there are several factors that act together in a complex and as yet unknown fashion in influencing the durations of the phonemes. A good understanding of each factor is necessary before attempting to model them together. In answering the questions we posed, we wanted to isolate the "speaker" factor. For that reason, we decided to conduct independent experiments where we worked with mean phoneme durations per speaker in an attempt to smooth out the other factors. Our first set of measurements therefore consisted of computing the correlations between mean phoneme durations across speakers.

In order to get a perspective of the extent of the influence of this factor on the durations, an eigenvector analysis was done. The directions and magnitudes of the principal contributions to variance were obtained by calculating the eigenvalues and eigenvectors. By then projecting speaker-specific data onto the eigenvectors a good indication is obtained of the speaker differences for the specific factor. The directions of the eigenvectors explain how each of the input factors contributes to the specific dimension.

The eigenvectors and eigenvalues are obtained from the covariance matrix of the $n \times m$ data where n is the number of levels of a factor (number of speakers in our case) and m the number of phonemes. Before calculating the covariance matrix, the data matrix is normalized by subtracting the mean values from the column vectors and then dividing by the standard deviation. This ensures that phonemes with a high variance do not dominate the analysis.

We decided to use a maximum likelihood (ML) approach for cross-phoneme duration estimation, because this enabled us to utilize the information provided by the eigenvectors in a practical model. It was assumed that the data can be approximated by a normal distribution with the covariance matrix calculated as described above. Suppose one has a vector $\overline{x} = \{x_0 ... x_m\}$ representing normalized phoneme durations with x_{m-p} unknown, $p < m$. The ML approach will find x_{m-p} such that the probability $P(x_0 ... x_m)$ is maximized. If one defines x to be $\overline{x} = \overline{d} - \overline{mu}$ with d the original duration, the ML solution given $\Sigma^{-1}\overline{x}$ can be found from

$$\frac{\partial \Sigma}{\partial x_{m-p}} = 0 \tag{6.1}$$

The solution to (6.1) is simply

$$\Sigma^{-1}\overline{x} = 0 \tag{6.2}$$

on condition that $\overline{x} = \overline{k}$ with the exception of x_{m-p}, with $\overline{k}$ being the given data vector. (6.2) can easily be solved by simple linear algebra. This method was then extended by allowing several unknown durations to be estimated simultaneously using exactly the same approach as described above.

6.4 EXPERIMENTAL SETUP

The proposed models were tested using the TIMIT database as described in section 6.2. TIMIT contains 52 different phone symbols. This set was reduced to the well-known ARPABET owing to data scarcity, which consists of 48 symbols, by combining em, en and eng into en, hh and hv to h and zh and z to z. ARPABET was then reduced by one symbol to 47 symbols by combining [ʊ] and [Λ].

The training set of 462 speakers was used to estimate the covariance matrix, as well as the eigenvectors and eigenvalues. The test set of 168 speakers was then used to test the models by trying to predict phoneme durations. For every speaker, sample means of all present phonemes were calculated.

Our initial measurements of the correlations in phoneme durations across speakers focused on the relationships between phonemes that co-vary in duration. Thereafter, we conducted a number of experiments to investigate duration prediction using the ML method described above.

- *Experiment 1: Estimation of phoneme durations given examples of all phonetic classes.* An iterative approach was used to estimate every single phoneme duration, given all the other phoneme durations. The objective of this experiment was twofold: to determine if the ML approach can be used to make better predictions than simply predicting the global mean of each phoneme and also to determine the relative predictability of individual phonemes.

- *Experiment 2: Estimation of phoneme durations given examples of same phonetic class.* The same iterative approach was used to estimate the durations of all phonemes, but this time only phonemes of the same class as the phoneme in question was given as input. The hypothesis that was to be tested was that phonemes tend to vary in classes. There were five phoneme classes: stops, fricatives & affricates, nasals, semivowels & glides and vowels.

- *Experiment 3: Estimation of phoneme durations given examples of all but same phonetic class.* Experiment 2 was repeated, but instead of using phoneme durations of the same class, all durations *except* the durations of phones of the same class were given as input. An interesting observation from the covariance matrix was tested here in that there seems to be a "cross-class" correlation between certain phonemes.
- *Experiment 4: Simultaneous estimation of several phoneme durations.* For every speaker 50 duration estimates were done. Every estimation entailed three vowels and three consonants to be estimated simultaneously, with the rest of the observed phonemes given as data to the ML model estimator.
- *Experiment 5: Estimation of phoneme durations, conditioning on correlation.* For every speaker, each phoneme was estimated iteratively, each time adding phonemes in descending order according to their correlation with the phoneme to be estimated.
- *Experiment 6: Determining the theoretical minimum of experiment 5.* The theoretical minimum of experiment 5 was calculated by adding phonemes until just before the error started to increase again.

6.5 RESULTS

6.5.1 THE CORRELATION OF PHONEME DURATIONS

The correlations across speakers between the durations of all phonemes were computed; because of the normalization employed, these are equivalent to Pearson correlation coefficients. Some typical results are shown in figures 6.11 to 6.14, which represent the largest and smallest measured correlation values between four different phonemes and all other phonemes in our set. We see that some groups of phonemes (including most vowels) have high correlations with all other phonemes in the same group. Other phonemes have a more diverse set of correlations - for example, the duration of "p" correlates highly not only with other plosives ("k" and "t" in Figure 6.12), but also with the fricatives "s", "z" and "sh", the nasal "n", etc. Similarly, the duration of "r" in Figure 6.13 correlates highly with the expected "l" and "w", but also with several vowels. Finally, some phonemes have few strong correlations - for example, "dx" in Figure 6.14, which has reasonably weak correlation with the other flap ("nx"), and no other notable correlations.

6.5.2 EIGENVECTOR ANALYSIS

We now describe the results obtained in our analysis of the eigenstructure of the correlation matrix. Firstly, the magnitudes of the eigenvalues indicate how much weight or value a particular eigenvector carries. From Fig. 6.1 it can be seen that the first eigenvector contains more than 22% of the total information and that approximately 65% of the information is contained within the first 10 eigenvectors. This is a strong indication that a significant amount of information is contained in a relatively small number of factors. As can be seen from Fig. 6.2, the first eigenvector corresponds to a simultaneous stretching of *all* phonemes - this can therefore be seen as an indication of speaking rate. The vowels and fricatives are seen to be the most consistent participants in this change. The second eigenvector, shown in Fig. 6.3, corresponds to a differential lengthening of vowels in comparison with consonants, whereas the third eigenvector (Fig. 6.4) seems to indicate a distinction between the relative lengths of liquids, glides and nasals, on the one hand, in comparison with plosives, fricatives and certain vowels, on the other.

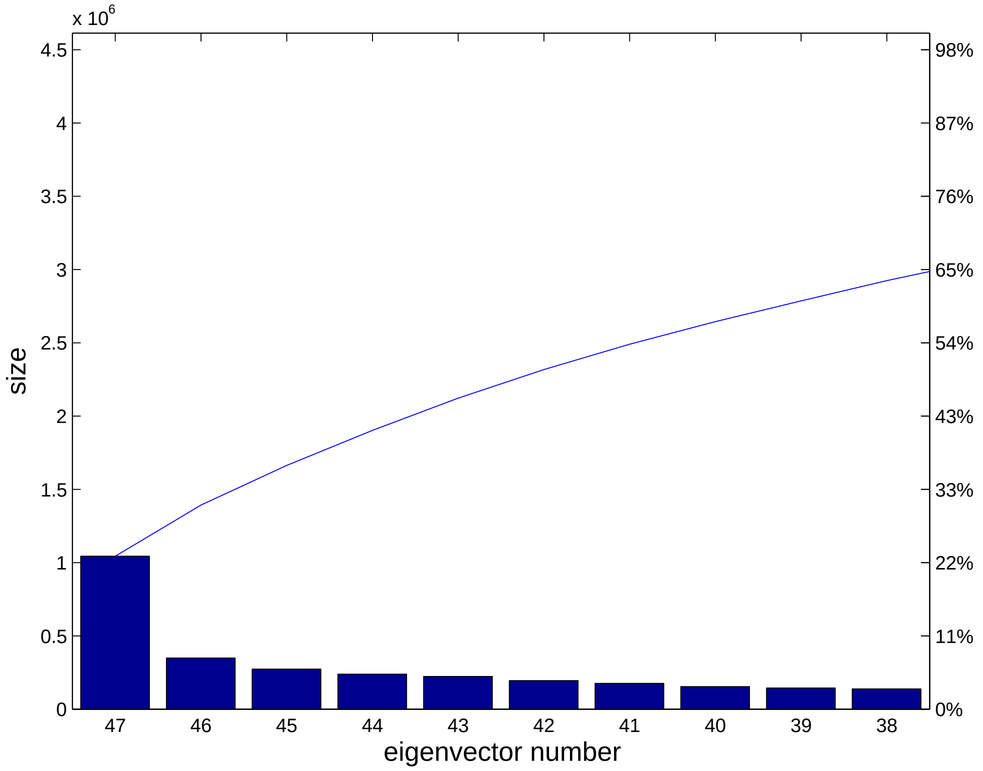

Figure 6.1: *Pareto chart of the eigenvalues obtained from the speaker/phoneme covariance matrix.*

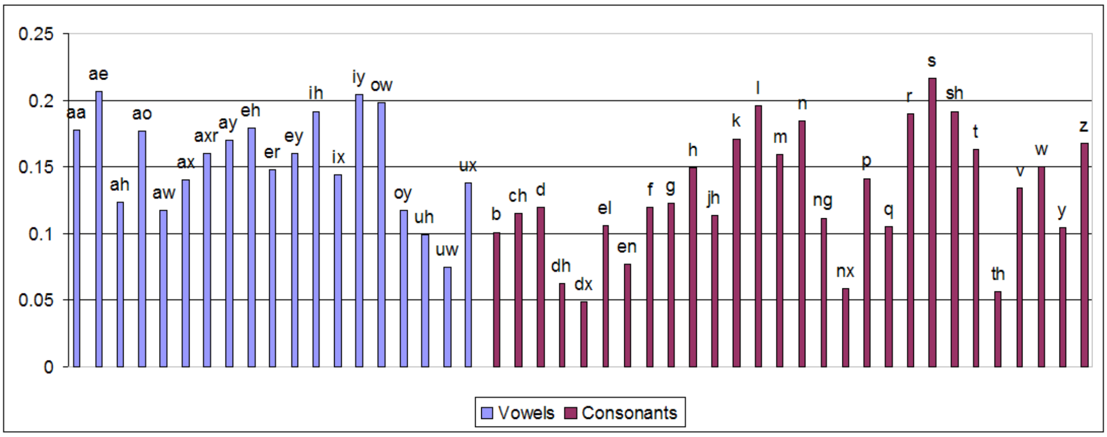

Figure 6.2: *Components of first eigenvector.*

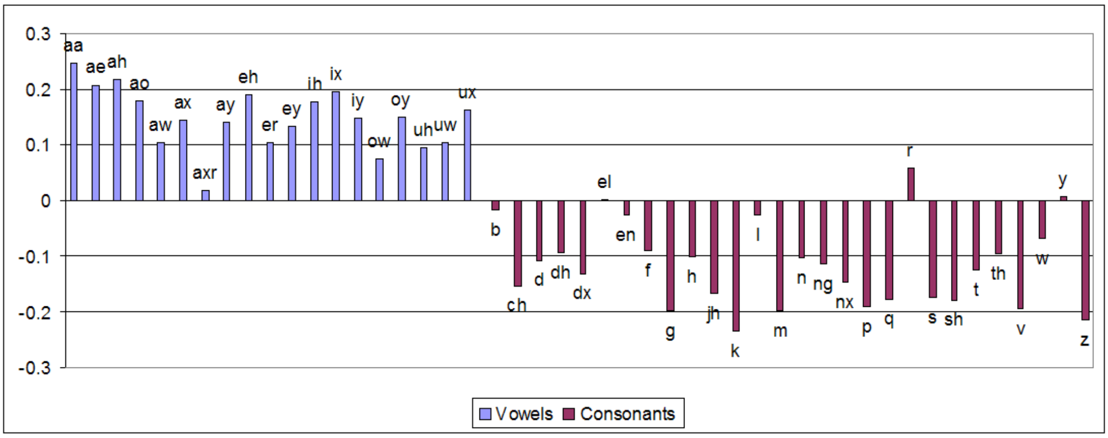

Figure 6.3: *Components of second eigenvector.*

6.5.3 ML ANALYSIS

The performance of the ML model in the five experiments described in Section 6.4 was calculated in terms of the variance normalized mean squared error (MSE) between the correct duration and the estimated duration. A total of 7522 estimations were done for the first three and also the last experiment and 50400 for the fourth. The latter will be normalized to the other four experiments in order to give comparable results. A baseline against which the results can be tested must also be established. Two baselines were selected: a nearest neighbor approach (where the closest training speaker based on all the known phoneme durations is calculated, using the Euclidean distance) and simply using the global mean for the specific phoneme. The results can be seen in Table 6.1. Experiment 5 was conducted to evaluate individual phoneme errors and is thus not presented in the table. Several interesting observations can be made from Table 6.1. Firstly, we note that the ML approach consistently outperforms the global mean approach. In experiment 1 the percentage improvement is approximately

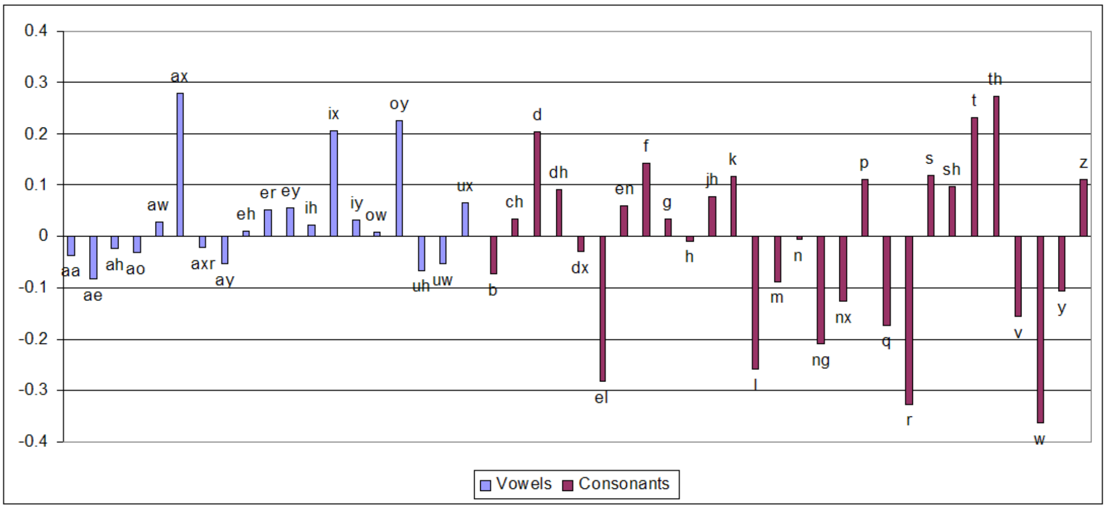

Figure 6.4: *Components of third eigenvector.*

Table 6.1: *Variance normalized MSE of the ML model, nearest neighbor and mean model from the four experiments.*

Exp.	ML	Global Mean	Eucl. dist
1	0.874	1.070	1.601
2	0.871	1.070	1.658
3	1.007	1.070	1.730
4	0.874	1.087	1.606
6	0.815		

18.3% and this increases to 18.6% for experiment 2. As could be expected, this percentage drops significantly for experiment 3 (to only 5.9%). The interesting observation here is that this approach still performs better than the global mean approach. This phenomenon confirms that the many non-zero correlations between different classes of phonemes can be employed usefully. Surprisingly, the average improvement jumps to 19.6% for experiment 4 where six unknown phoneme durations were estimated simultaneously. This is promising, since this experiment is a better reflection of a practical application than the other experiments, as one will rarely have all phoneme examples. It must be noted that the error values for experiment 4 have a much larger variance, since phonemes to be estimated were chosen randomly every time. The ratios between vowel and consonant occurrences are also equal, whereas the other experiments have more consonants that are estimated than vowels. The mean normalized MSE for vowels is also slightly lower than that of consonants and thus the error value in experiment 4 will tend to be slightly lower than if the conditions had been exactly the same as in the other experiments.

Although there is a slight decrease in the overall MSE when only within-class information is used for duration estimation (Exp. 2), this improvement is not uniform. A combined analysis of experiments 2 and 3 and the correlation coefficients indicates that the cross-class correlations are the reason for this behaviour. Examples include vowels such as "ae" and semivowels such as "r" and "l".

Note that the nearest neighbor approach performs significantly worse than the other two methods. This may seem counterintuitive, but if there is limited (or even negative) correlation between the estimated and nearest neighbor estimates, one can easily see that larger error values will be observed in this case. If we let x_1 be the duration we want to estimate and x_2 the predictor, the expected value of the MSE can be expressed as

$$< (x_1 - x_2)^2 > \tag{6.3}$$

Multiplying out gives

$$< x_1^2 - 2x_1x_2 + x_2^2 > \tag{6.4}$$

Subtracting the mean value from x_1 and x_2 respectively will not change the expected value. It then follows that

$$< x_i^2 >= \mu_i^2 + \sigma_i^2 \tag{6.5}$$

but $\mu_i^2 = 0$ since the means are subtracted, giving $< x^2 >= \sigma^2$. (6.5) can be rewritten as

$$\sigma_1^2 - 2 < x_1x_2 > +\sigma_2^2 \tag{6.6}$$

For the global mean case this is equivalent to

$$2\sigma_1^2 - 2 < x_1x_2 >, \tag{6.7}$$

where we have assumed that the training speakers and testing speakers all have roughly the same variance per phoneme ($\sigma_1 \approx \sigma_2$).

From the above analysis it can be seen that for large correlations the error will tend to zero, but for small correlations the global mean approach will tend to have double the error of the nearest neighbor approach.

The variance-normalized MSE values for all phonemes in experiment 1 are shown in Fig. 6.5 and Fig. 6.6. Fig. 6.5 shows that there are significant differences in the relative predictabilities of the different phonemes, with the vowels "iy", "eh", "ae", the fricatives "s", "sh" and the liquid "l" being most predictable. These are also the phonemes whose durations correlate most strongly with those of other phonemes. The least predictable phonemes are characterized by factors such as data scarcity ("oy"), phonemic ambiguity ("uw") and weak correlation with other phoneme durations (“nx”). It is interesting that the plosives “t”, “k”, and “d” are fairly predictable, whereas the other three plosives are less so.

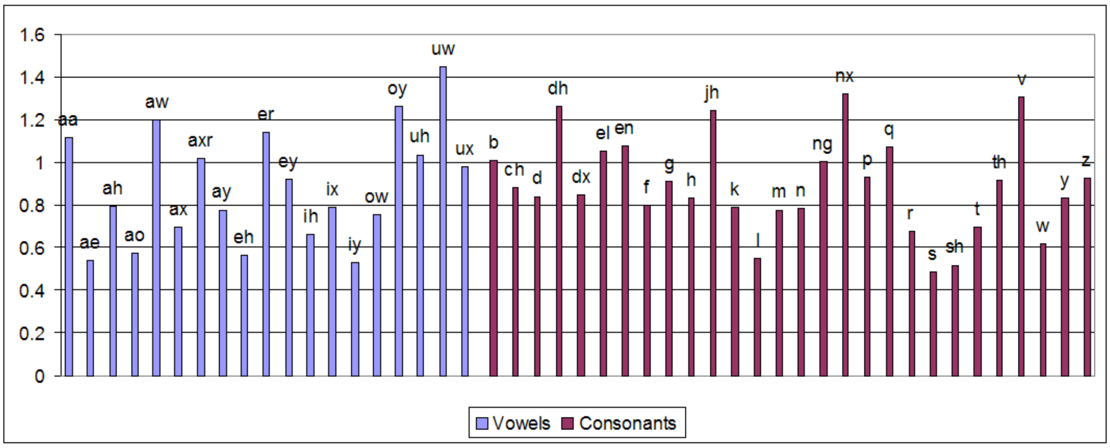

Figure 6.5: *Variance normalized MSE for the different phonemes using the ML approach.*

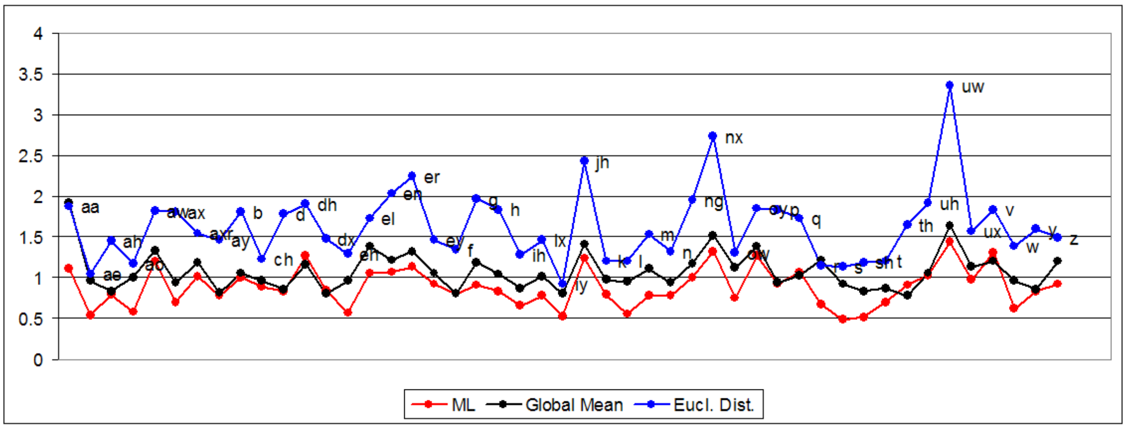

Figure 6.6: *Variance normalized MSE for the different phonemes using the ML, global mean and Euclidean distance approaches.*

The results of experiment 5 are summarized in Figures 6.7 to 6.10. As expected the error value decreases rapidly when the phonemes with the highest correlation are given as examples. An unexpected phenomenon is that even the highly predictable phonemes’ errors start to increase after a moderate number of phonemes have been added as examples. This is probably a result of the Gaussian distribution, which is assumed during our ML estimation, and deserves further attention.

6.6 DISCUSSION AND CONCLUSIONS

The Pareto chart in Fig. 6.1 is a confirmation of the claim that much of the variation observed in the duration of phonemes, as caused by the variable “speaker”, can be explained by a

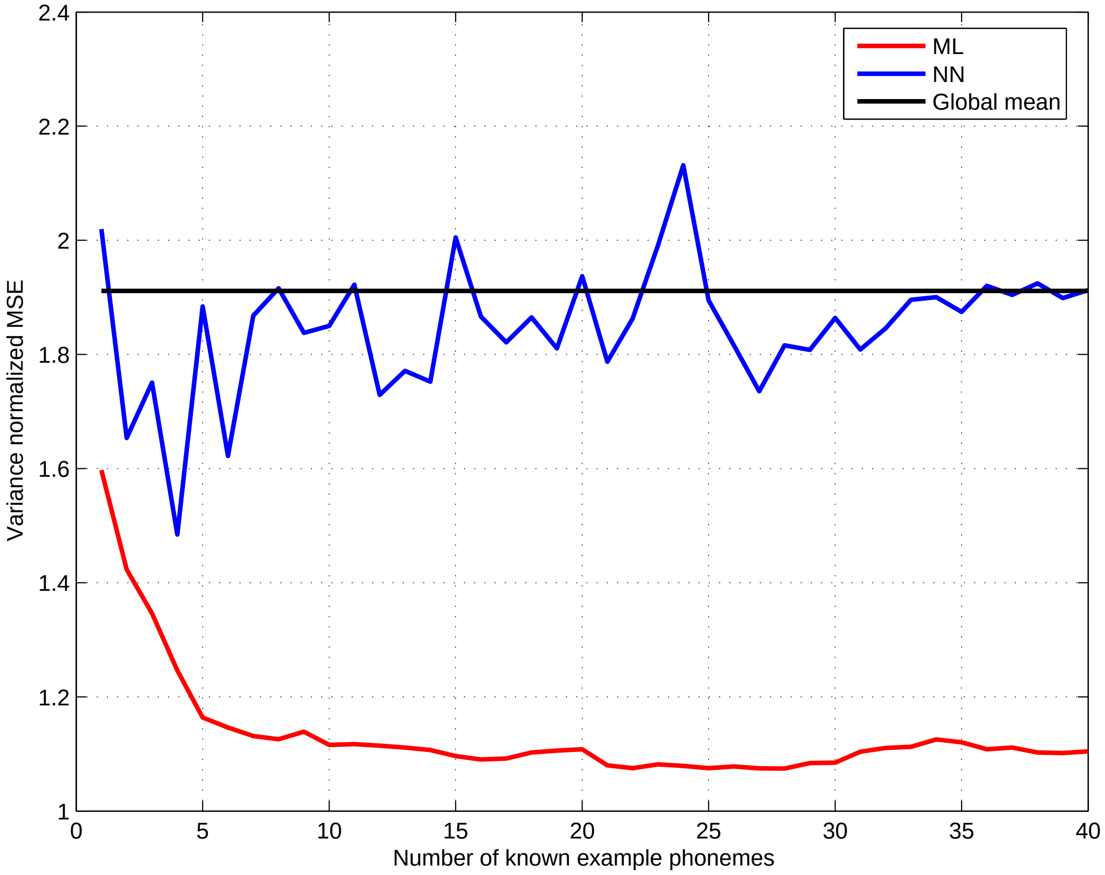

Figure 6.7: *Variance normalized MSE for aa vs the number of known phonemes during estimation, added in descending order of correlation with aa.*

relatively small number of factors. Figures 6.2, 6.3 and 6.4 show that a common lengthening or shortening of all phonemes is the strongest single effect, but that differential stretches between and within phoneme classes also play a significant role.

This knowledge was then applied by estimating an ML model from the training data in the TIMIT corpus. The model was tested using the testing data, also from the TIMIT corpus. From Table 6.1 and Figure 6.6 it can be seen that the ML approach performs significantly better than the mean phone duration approach. Thus, the observed intra-speaker correlations between phoneme durations are practically usable.

High correlations between phonemes in the same class, but also across classes were observed. It was found that most phonemes correlate well with only a few other phonemes (on the order of 10), and that accurate duration estimation is achieved using only those phonemes. As can be seen in Table 6.1, the lowest achievable error rate when selecting input phonemes in this fashion is 0.815, approximately 6.5% better than the result from experiment 2.

Our results also emphasize the importance of combining the various effects that influence the durations of phonemes. We found that about 15% to 20% of the intra-speaker variability

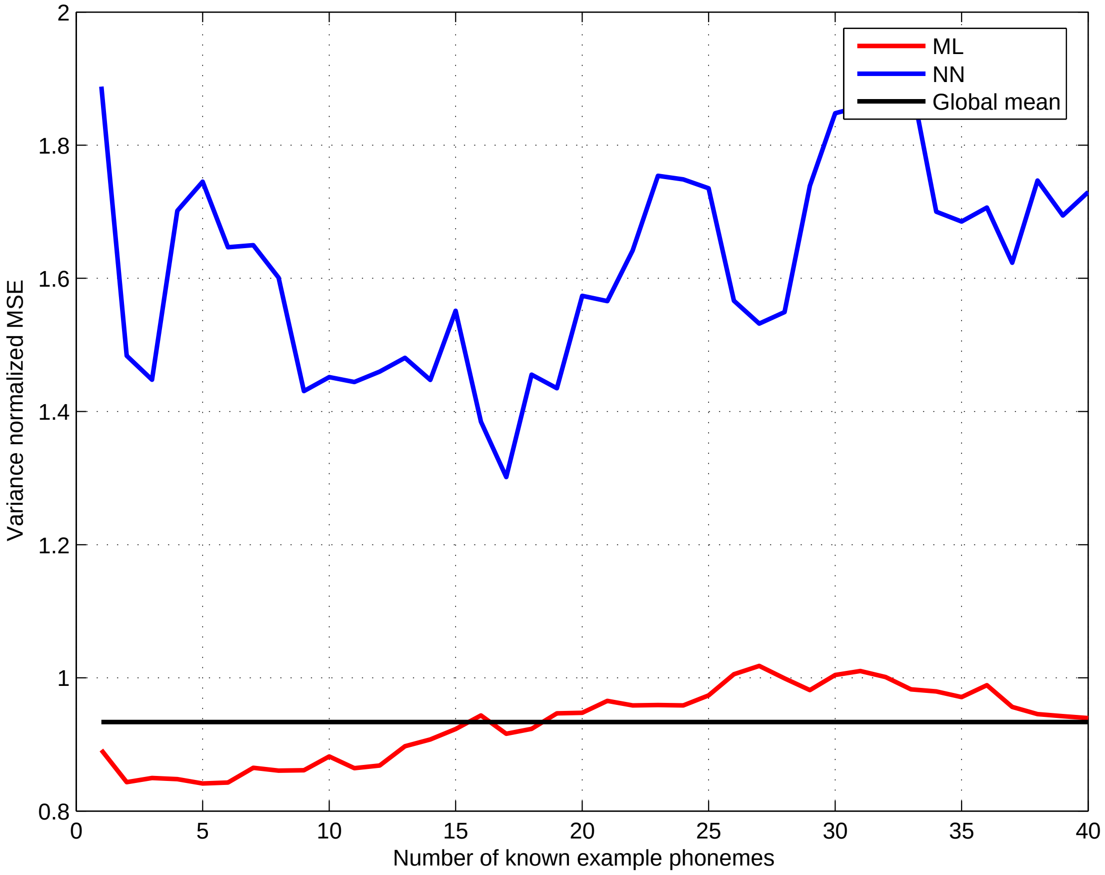

Figure 6.8: *Variance normalized MSE for p vs the number of known phonemes during estimation, added in descending order of correlation with p.*

in phoneme durations can be explained without reference to other factors, which indicates a significant role for those factors.

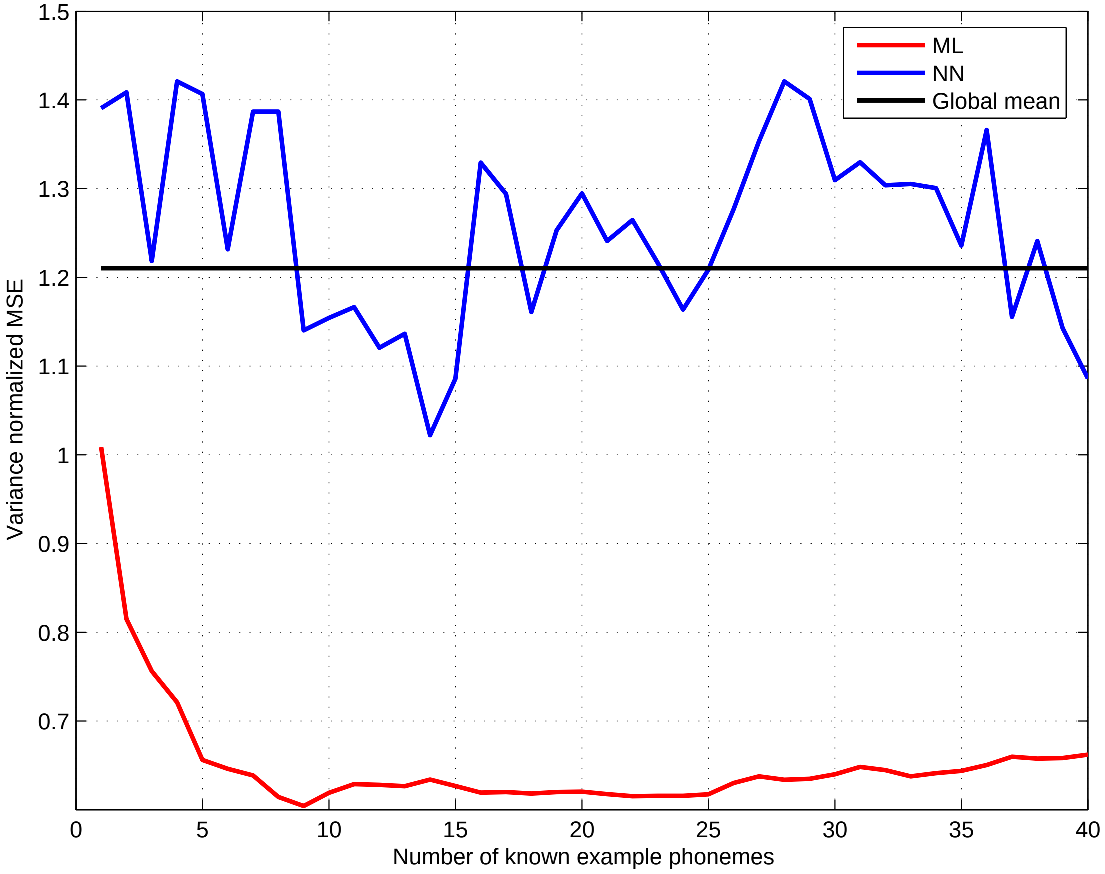

Figure 6.9: *Variance normalized MSE for r vs the number of known phonemes during estimation, added in descending order of correlation with r.*

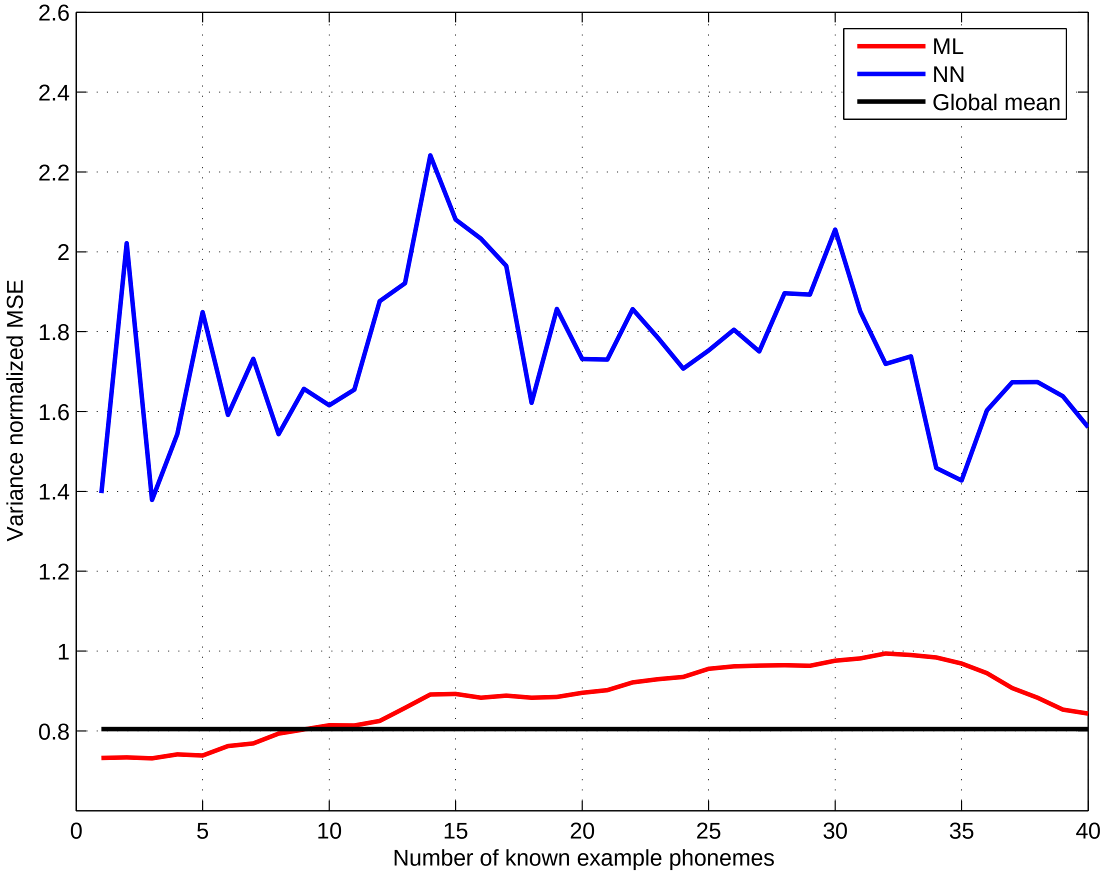

Figure 6.10: *Variance normalized MSE for dx vs the number of known phonemes during estimation, added in descending order of correlation with dx.*

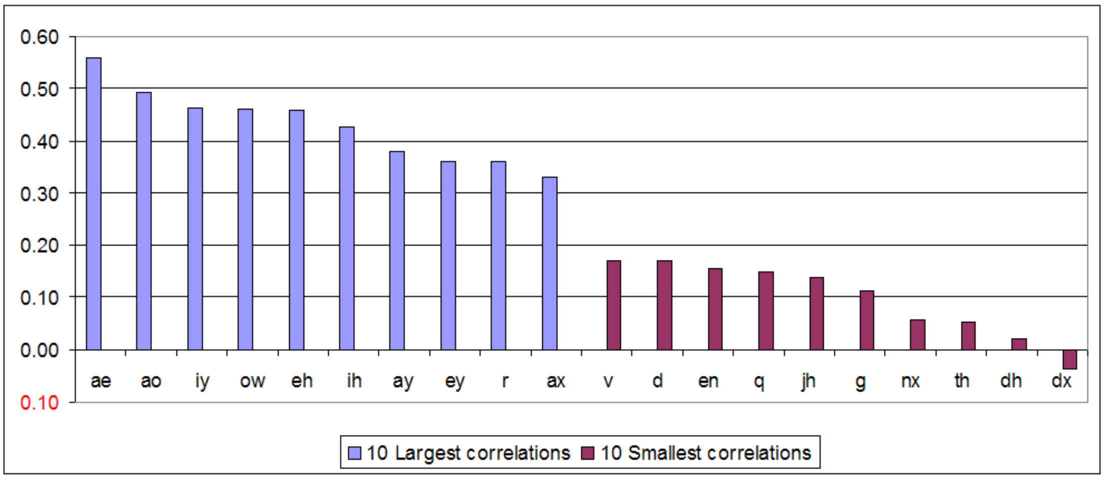

Figure 6.11: 10 *phonemes with the highest and* 10 *with the lowest correlation with aa.*

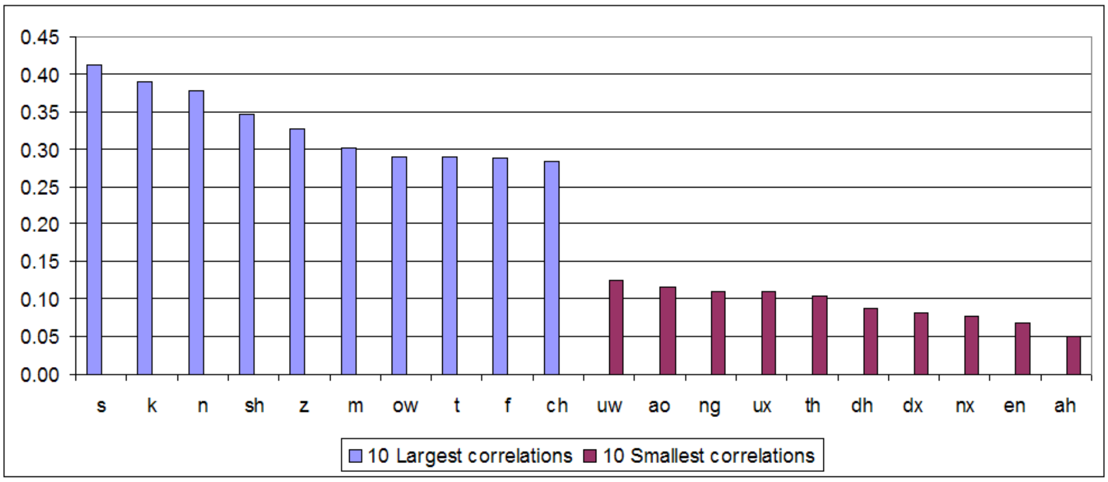

Figure 6.12: 10 *phonemes with the highest and* 10 *with the lowest correlation with p.*

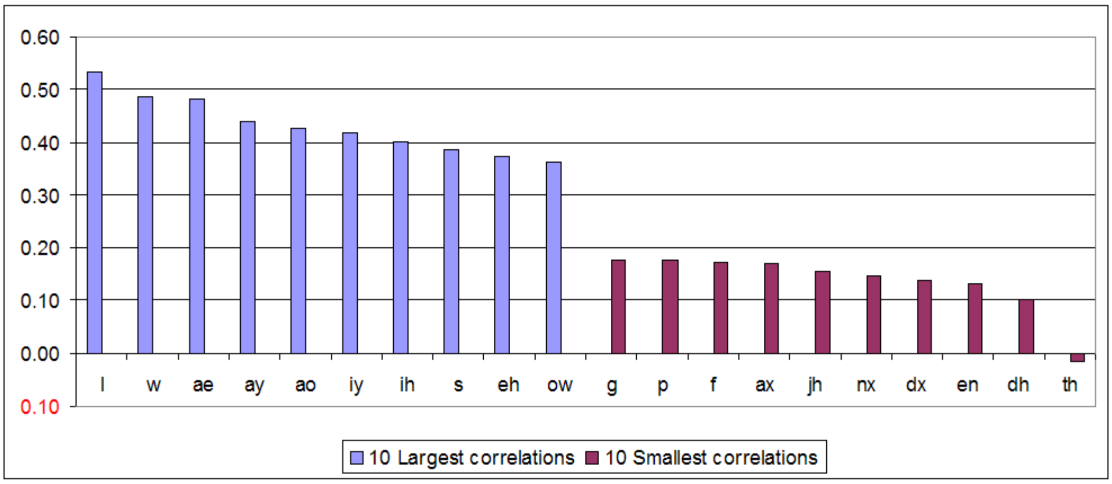

Figure 6.13: 10 *phonemes with the highest and* 10 *with the lowest correlation with r.*

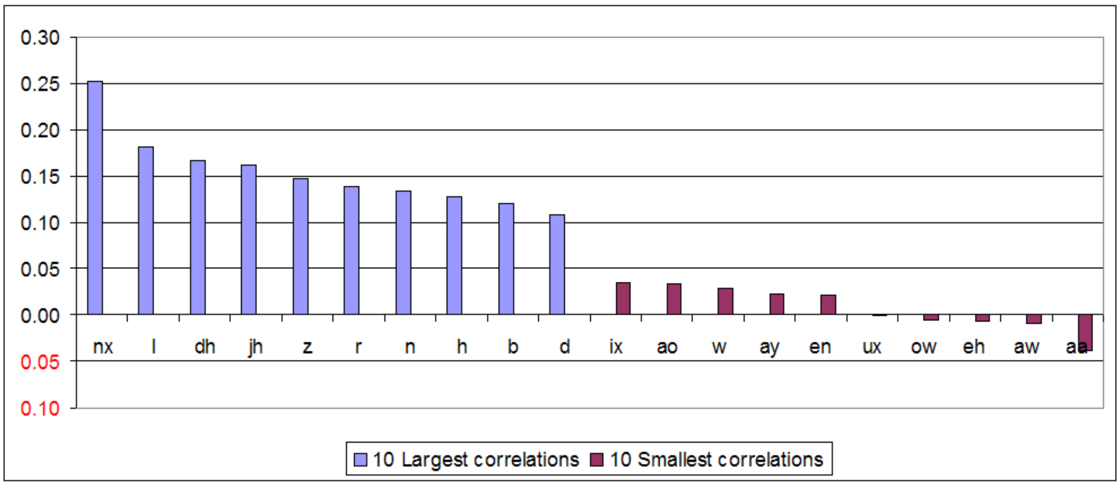

Figure 6.14: 10 *phonemes with the highest and* 10 *with the lowest correlation with dx.*

CHAPTER SEVEN

CONCLUSION

7.1 INTRODUCTION

This dissertation discussed research conducted in the pursuit of answering three questions related to phoneme durations: 1) If and how phoneme durations could be successfully applied to speaker verification; 2) how SRN would impact SV accuracy and 3) if unseen phoneme durations could be predicted by utilizing information from other phonemes, such that the prediction is better than traditional back-off approaches. The extent to which these questions have been answered will be summarized in this chapter, as well as future work.

7.2 SUMMARY OF CONTRIBUTION

The research presented in this dissertation has shown that phoneme durations is a valuable attribute to a text-dependent speaker verification system. In particular we have shown that:

- Durations of context-dependent triphones constitute a feature set that can improve the accuracy of speaker verification systems to a significant degree. Acoustic and temporal features seem largely uncorrelated and the improvements obtained should thus be possible in a text-independent setup as well.

- Even a novel approach to speech rate modeling could be effectively used to partially remove the degrading effects thereof. Differing speech rates is a prominent constituent of intra-speaker variability, which is known to be detrimental to SV accuracy. By removing speech rate, the robustness of phoneme durations as a discriminating speaker trait is improved.

- By refining the duration model, a significant increase in SV accuracy can be achieved. A simple refinement of the existing model entailed accounting for “position in sentence”, “position in word” and occasional failures of the automatic alignment process.
- A ML model could be constructed to predict the durations of unseen phonemes such that the predictions are more accurate than traditional backoff approaches. In particular, we observed that intra-speaker correlations between phoneme durations are practically usable.

7.3 CONCLUSION AND FUTURE DIRECTIONS

As mentioned in chapter 1, traditional cepstral features are susceptible to transmission line and cross-channel degradations. Higher-level features such as phoneme durations seem to be a viable supplement to increase the robustness of SV under these conditions. Data scarcity remains a significant obstacle though and needs to be addressed for phoneme durations to be of practical use in SV. The research reported in this dissertation showed that phoneme durations is a valuable attribute to acoustic SV systems and that data scarcity can be addressed to a degree by utilizing phoneme correlations in predicting unknown phoneme durations.

The duration models we used are still rather crude. All triphones still carry equal weight and are modelled by independent Gaussian distributions. We believe that the duration feature can be improved even more by incorporating other factors into the model, such as the frequency of observation of triphones, the acoustic reliability of the observation and giving greater weight to more discriminative triphones.

We addressed differing speech rates by performing speech rate normalization. Differing speech rates were not part of the YOHO protocol though and this technique may be even more useful when applied to more natural speech.

Although we have successfully incorporated phoneme durations into a text-dependent speaker verification system, the text-independent case still poses an interesting challenge. A text-independent solution to duration modeling would also be beneficial to other speech applications, such as ASR [43],[13] and TTS [47].

Furthermore our research indicated that about 15% to 20% of the intra-speaker variability in phoneme durations can be explained without reference to other factors, which indicates a significant role for those factors. Their exact contributions and interaction is an interesting field for further research.

Many more research questions were raised by the answers we found. The questions we did answer conclusively showed that phoneme durations is a valuable addition to text-dependent speaker verification systems.

REFERENCES

[1] P. Gutkowski, "Algorithm for retrieval and verification of personal identity using bimodal biometrics," *Information Fusion*, vol. 5, no. 1, pp. 65–71, March 2004.

[2] D.A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1, pp. 91–108, March 1995.

[3] A. Solomonoff, W.M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005, vol. 1, pp. 629–632.

[4] A. Higgens, L. Bahler, and J. Porter, "Speaker verification using randomized phrase prompting," *Digital Signal Processing*, vol. 1, no. 2, pp. 89–106, 1991.

[5] G.R. Doddington, M.A. Przybocki, A.F. Martin, and D.A. Reynolds, "The NIST speaker recognition evaluation - overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 3, pp. 225–254, June 2000.

[6] V.R.R. Gadde, "Modeling word durations," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, October 2000, vol. 1, pp. 601–604.

[7] H.R. Pfitzinger, "Intrinsic phone durations are speaker-specific," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, USA, September 2002, vol. 1, pp. 1113–1116.

[8] L. Ferrer, H. Bratt, V.R.R. Gadde, S. Kajarekar, E. Shriberg, K. Sönmez, A. Stolcke, and A. Venkataraman, "Modeling duration patterns for speaker recognition," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, September 2003, pp. 2017–2020.

[9] E. Shriberg, L. Ferrer, S. Kajarekar, A. Stolcke, and A. Venkataraman, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46, no. 3, pp. 455–472, February 2005.

[10] V. R. Rao Gadde, "Modeling word duration for better speech recognition," in *Proceedings of the NIST Speech Transcription Workshop*, College Park, USA, May 2000.

[11] E. Shriberg and L. Ferrer, "A text-constrained prosodic system for speaker verification," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Antwerp, Belgium, August 2007, vol. 1, pp. 1226–1229.

[12] T.H. Crystal and A.S. House, "Segmental durations in connected-speech signals: syllabic stress," *The Journal of the Acoustical Society of America*, vol. 83, no. 4, pp. 1574–1585, 1988.

[13] L.C.W. Pols, X Wang, and L.F.M. ten Bosch, "Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR," *Speech Communication*, vol. 19, no. 2, pp. 161–176, 1996.

[14] J.P. Campbell, "Speaker recognition: A tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, September 1997.

[15] J. Campbell, "Testing with the YOHO CD-ROM voice verification corpus," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, USA, May 1995, vol. 1, pp. 341–344.

[16] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1, pp. 19–41, January 2000.

[17] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, and D.A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 4, pp. 430–451, 2004.

[18] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Veltchev, and P. Woodland, *The HTK Book*, Cambridge University Engineering Department, Cambridge, 2005.

[19] Stuart J. Russell and Peter Norvig, *Artificial intelligence: a modern approach*, Prentice-Hall, Inc., Upper Saddle River, 1995.

[20] J.P. Campbell and D.A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Phoenix, USA, March 1999, vol. 2, pp. 829–832.

[21] B.S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, June 1974.

[22] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–32, October 1994.

[23] D.A. Reynolds, "Large population speaker identification using clean and telephone speech," *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46–48, March 1995.

[24] A.O. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for svm-based speaker recognition," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, USA, September 2006, vol. 3, pp. 1471–1474.

[25] S.S. Kajarekar and A Stolcke, "NAP and WCCN: Comparison of approaches using MLLR-SVM speaker verification system," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hawaii, USA, April 2007, vol. 4, pp. 249–252.

[26] R. Auckanthaler, "Score normalization for text-independent speaker verification," *Digital Signal Processing*, vol. 10, no. 1, pp. 42–54, January 2000.

[27] D.E. Sturim and D.A. Reynolds, "Speaker adaptive cohort selection for T-Norm in text-independent speaker verification," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, USA, March 2005, vol. 1, pp. 741–744.

[28] H. Melin, J.W. Koolwaaij, J. Lindberg, and F. Bimbot, "A comparitive evaluation of variance flooring techniques in hmm-based speaker verification," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Sydney, Australia, November 1998, vol. 1, pp. 2379–2382.

[29] D.A. Reynolds, W. Andrews, J.P. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and X. Bing, "The SuperSID project: exploiting high-level information for high-accuracy speaker recogni-

tion," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, April 2003, vol. 4, pp. 784–787.

[30] N.B. Yoma and T.F. Pegoraro, "Robust speaker verification with state duration modeling," *Speech Communication*, vol. 38, no. 1, pp. 77–88, September 2002.

[31] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews, and J. Abrahamson, "Combining cross-stream and time dimensions in phonetic speaker recognition," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, April 2003, vol. 4, pp. 800–803.

[32] J.P. Campbell, D.A. Reynolds, and R.B. Dunn, "Fusing high- and low-level features for speaker recognition," in *Proceedings of Eurospeech*, Geneva, Switzerland, September 2003, vol. 1, pp. 2665–2668.

[33] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2, pp. 210–229, April 2006.

[34] E. Shriberg, "Speaker classification i, fundamentals, features and methods," vol. 4343 of *Lecture Notes in Computer Science*, chapter Higher-Level Features in Speaker Recognition, pp. 241–259. Springer, Berlin, February 2007.

[35] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Prentice Hall, Upper Saddle River, 2001.

[36] C.J. van Heerden and E. Barnard, "Speech rate normalization used to improve speaker verification," in *Proceedings of the Symposium of the Pattern Recognition Association of South Africa*, Parys, South Africa, November 2006, pp. 2–7.

[37] C.J. van Heerden and E. Barnard, "Durations of context-dependent phonemes: A new feature in speaker verification," vol. 4441 of *Lecture Notes in Computer Science*, chapter Speaker Classification II, Selected Projects, pp. 93–103. Springer, Berlin, February 2007.

[38] M. Ostendorf and I. Bulyko, "The impact of speech recognition on speech synthesis," in *Proceedings of the 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, USA, September 2002, vol. 1, pp. 99–106.

[39] Izhak Shafran, Mari Ostendorf, and Richard Wright, "Prosody and phonetic variability: Lessons learned from acoustic model clustering," in *Proceedings of the ISCA workshop*

on Prosody in Automatic Speech Recognition, Red Banks, USA, October 2001, vol. 1, pp. 127–131.

[40] Diego H. Mielone and Antonio J. Rubio, "Prosodic and accentual information for automatic speech recognition," *IEEE Transactions of Speech and Audio Processing*, vol. 11, no. 4, pp. 321–333, July 2003.

[41] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," in *Proceedings of Eurospeech*, Rhodes, Greece, September 1997, vol. 4, pp. 2079–2082.

[42] G. Chung and S. Seneff, "A hierarchical duration model for speech recognition based on the ANGIE framework," *Speech Communication*, vol. 27, no. 2, pp. 113–134, March 1999.

[43] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Munich, Germany, December 1999, vol. 1, pp. 79–83.

[44] Bernd Möbius and Jan van Santen, "Modeling segmental duration in German text-to-speech synthesis," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, October 1996, vol. 4, pp. 2395–2398.

[45] R. Batusek, "A duration model for Czech text-to-speech synthesis," in *Proceedings of the International Conference on Speech Prosody*, Aix-en-Provence, France, April 2002, vol. 1, pp. 167–170.

[46] J.P.H. van Santen, "Assignment of segmental duration in text-to-speech synthesis," *Computer Speech and Language*, vol. 8, no. 22, pp. 95–128, April 1994.

[47] O.V. Goubanova, "Bayesian modelling of vowel segment duration for text-to-speech synthesis using distinctive features," in *Proceedings of the 15th International Conference on Phonetic Sciences(ICPhS)*, Barcelone, Spain, April 2003, vol. 3, pp. 2349–2352.

[48] J.L. Hieronymus, "Automatic sentenial vowel stress labeling," in *Proceedings of the 1st European Conference on Speech Communication and Technology (Eurospeech)*, Paris, France, September 1989, vol. 1, pp. 1226–1229.

[49] J.L. Hieronymus and B.J. Williams, "An investigation of the relation between perceived pitch accent and automatically-located accent in British english," in *Proceedings of the*

2nd European Conference on Speech Communication and Technology (Eurospeech), Genova, Italy, September 1991, vol. 1, pp. 1157–1160.

[50] C.J. van Heerden and E. Barnard, "Using timing information in speaker verification," in *Proceedings of the Symposium of the Pattern Recognition Association of South Africa*, Langebaan, South Africa, November 2005, pp. 53–57.

[51] H-S. Liou and R.J. Mammone, "A subword neural tree network approach to text-dependent speaker verification," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Detroit, USA, 1995, pp. 357–360.

[52] V. Wan and W. Campbell, "Support vector machines for speaker verification and identification," in *Proceedings of the IEEE Signal Processing Society Workshop*, Sydney, Australia, December 2000, vol. 2, pp. 775–784.

[53] W.M. Campbell and K.T. Assaleh, "Polynomial classifier techniques for speaker verfication," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Phoenix, USA, March 1999, vol. 1, pp. 321–324.

[54] N. T. Kleynhans and E Barnard, "Language dependence in multilingual speaker verification," in *Proceedings of the Symposium of the Pattern Recognition Association of South Africa*, Langebaan, South Africa, November 2005, pp. 117–121.

[55] A.E. Rosenberg, J. DeLong, C-H. Lee, B-H. Juang, and F.K. Soong, "The use of cohort normalized scores for speaker verification," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Banff, Canada, October 1992, vol. 1, pp. 599–602.

APPENDIX A

THE YOHO CORPUS

The YOHO corpus is a large supervised speaker verification database [15]. It consists of 138 speakers (106 males and 32 females) who spoke a number of prompted utterances (Table A.1) from a restricted grammar set of 56 two-digit numbers ranging from $21 - 97$ [4]. All decades ($30, 40$ etc.), numbers ending in 8 and double digits were omitted for various reasons. Although numbers are known to be difficult for verification purposes, they are attractive because of the randomization capability they introduce to the prompting process. The utterances thus comprised combination-lock phrases as proposed by [4], which makes it extremely difficult for a potential impostor to know beforehand which combination of phrases will be prompted. An example is $21 - 36 - 43$, which would be pronounced as "twenty-one, thirty-six, fourty-three". This randomization results in 56^3 possible combination-lock phrases to be prompted. Four such combination-lock phrases were prompted during a single verification session and 24 such phrases for a training/enrollment session. The YOHO corpus has 4 enrollment sessions per speaker and 10 verification sessions. The data was recorded with a microphone using a 3.8 kHz bandwidth in an office environment with normal background noise. The recording period stretched over approximately 3 months to incorporate the factor of slight changes to a speaker's voice. These changes are also known as intersession variability and can be detrimental to a speaker verification system if not accounted for. This corpus is widely used for the evaluation of text-dependent systems for speaker verification, and is therefore suitable for the comparative evaluation of such systems.

Table A.1: YOHO corpus summary: number of phrases per speaker

Session	**Nr of phrases**
Enrollment	96
Verification	40
Total	136