Structured Approaches to Data Selection for Speaker Recognition

Howard Hao Lei

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2010-150
http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-150.html

December 8, 2010

Acknowledgement

I want to thank the following people, whose work contributed to my completion of this dissertation: Nikki Mirghafori, Eduardo Lopez-Gonzalo, and Andreas Stolcke. I also want to thank Prof. Nelson Morgan for being my advisor through my Ph.D dissertation. I also thank the following people for offering useful advice that contributed to my dissertation: Gerald Friedland, Lara Stoll, Mary Knox, Oriol Vinyals, Kofi Boakye, David Van Leeuwen, George Doddington, Christian Mueller, and David Imseng.

Structured Approaches to Data Selection for Speaker Recognition

by

Howard Hao Lei

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Engineering - Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Nelson Morgan, Chair
Professor Michael Jordan
Professor Keith Johnson

Fall 2010

Structured Approaches to Data Selection for Speaker Recognition

Howard Hao Lei

Abstract

Structured Approaches to Data Selection for Speaker Recognition

by

Howard Hao Lei

Doctor of Philosophy in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Nelson Morgan, Chair

In this work, I investigated structured approaches to data selection for speaker recognition, with an emphasis on information theoretic approaches, as well as approaches based on speaker-specific differences that arise from speech production. These approaches rely on the investigation of speaker discriminability measures that detect speech regions that result in high speaker differentiation. I also attempted to understand why certain data regions result in better speaker recognition system performance.

The knowledge gained from the speaker discriminability measures was used to implement an effective data selection procedure, that allows for the prediction of how well a speaker recognition system will behave without actually implementing the system. The use of speaker discriminability measures also leads to data reduction in speaker recognition training and testing, allowing for faster modeling and easier data storage, given that the latest speaker recognition corpora uses hundreds of gigabytes.

In particular, I focused primarily on Gaussian Mixture Model- (GMM) based speaker recognition systems, which comprise the majority of current state-of-the-art speaker recognition systems. Methods were investigated to make the speaker discriminability measures easily obtainable, such that the amount of computational resources required to extract these measures from the data would be significantly less in comparison to the computational resources required to run entire speaker recognition systems to determine what regions of speech are speaker discriminative.

Upon selecting the speech data using these measures, I created new speech units based on the data selected. The speaker recognition performances of the new speech units were compared to the existing units (mainly mono-phones and words) standalone and in combination. I found that in general, the new speech units are more speaker discriminative than the existing ones. Speaker recognition systems that use the new speech units as data in general outperformed systems using the existing speech units. This work, therefore, outlines an effective approach that is easy to implement for selecting speaker discriminative regions of data for speaker recognition.

This thesis is dedicated to my parents who offered me support and encouragement throughout my Ph.D studies, as well as members of the Gracepoint community who offered me a home away from home in Berkeley, CA.

Contents

List of Figures

List of Tables

Acknowledgements

I want to thank the following people, whose past and current efforts contributed directly to my completion of this thesis: Nikki Mirghafori, Eduardo Lopez-Gonzalo, and Andreas Stolcke.

I also want to thank Prof. Nelson Morgan for being a wonderful advisor for me through my Ph.D dissertation, and all other past and current members of the ICSI speaker recognition and diarization group for offering me advice during group meetings. These people include: Gerald Friedland, Lara Stoll, Mary Knox, Oriol Vinyals, Kofi Boakye, David Van Leeuwen, George Doddington, Christian Mueller, and David Imseng.

Curriculum Vitæ

Howard Hao Lei

Education

2005	University of Michigan at Ann Arbor B.S.E, Electrical Engineering
2007	University of California at Berkeley M.S., Electrical Engineering
2010	University of California, Berkeley Ph.D., Electrical Engineering and Computer Science

Personal

Born	November 7, 1982, Beijing, China

Howard Hao Lei focused on signal processing, machine learning, and statistics, with applications to speech processing, during his Ph.D studies at UC Berkeley. He specifically focused on improving speaker recognition, and on the use of speech-units for improved data selection for speaker recognition. Some of Howard's main work included development and implementations of various unit-constrained speaker recognition systems involving the use of Hidden Markov Models and Support Vector Machines. These systems provided complementary information to the mainstream systems using Gaussian Mixture Models. Howard also focused on developing measures to improve data selection for speaker recognition, and using these measures to select new speech units that are speaker discriminative. Howard published 7 first-authored papers in the field of speaker recognition, 6 of which are accepted to major speech conferences, and the 7th being his UC Berkeley master's report. Howard also actively participated in the NIST speaker recognition evaluations of 2006, 2008, and 2010.

Howard also did some automatic speech recognition work while interning at Google in Mountain View, CA, as he focused on improving acoustic modeling. He also had many teaching experiences at UC Berkeley. He was a Graduate Student Instructor for a lower-division undergraduate signals and systems class for 5 semesters, and was a primary course lecturer/instructor for a circuits class in the summer of 2010.

Chapter 1

Introduction

1.1 Speaker Recognition Overview

The goal of speaker recognition is to associate a speaker identity to a given speech utterance (spoken by a single speaker). While in general, a speech utterance refers to a sentence or sentence-like construct from a given speaker, in this work, the definition of an utterance is expanded to include multiple sentences or chunks of a conversation. In the typical speaker recognition scenario, speaker models are built using utterances from a given speaker, and stored. Next, an utterance spoken by an unknown speaker (i.e. test utterance) is evaluated against the speaker model to determine if the identity of the unknown speaker matches that of the speaker model. Figure 1.1 illustrates this problem:

Speaker recognition is a research problem that requires the application of various signal processing, statistical, and machine learning techniques. Since the early to mid 1990s, a standard and effective approach to speaker recognition uses acoustic features capturing characteristics of the spectral content of speech. Gaussian mixture models (GMMs) are built to model the acoustic tendencies of certain speakers [41], while test utterances are evaluated against the speaker models via log-likelihood ratio scoring.

Since 2001, there has been more attention paid to the use of high-level features, such as words [16] and phonetic sequences [4] [22], for speaker recognition. The purposes of using words and phones are to capture idiolect-based tendencies of speakers (i.e. the choice of words a speaker uses), and to use the inter-speaker variability of such tendencies for speaker discriminative purposes. Such high-level features, whose data is more sparse compared to low-level acoustic features, have been shown to provide good speaker discriminative power [16] [22].

The concept of using text-constraining, or unit-constraining, is another advancement in speaker recognition, where only portions of speech where certain texts, or units, occur are used to implement entire speaker recognition systems. The texts that have been used are typically linguistically-defined speech units, such as word N-grams [44] [31]. While discarding much of the overall speech data, the use of unit-

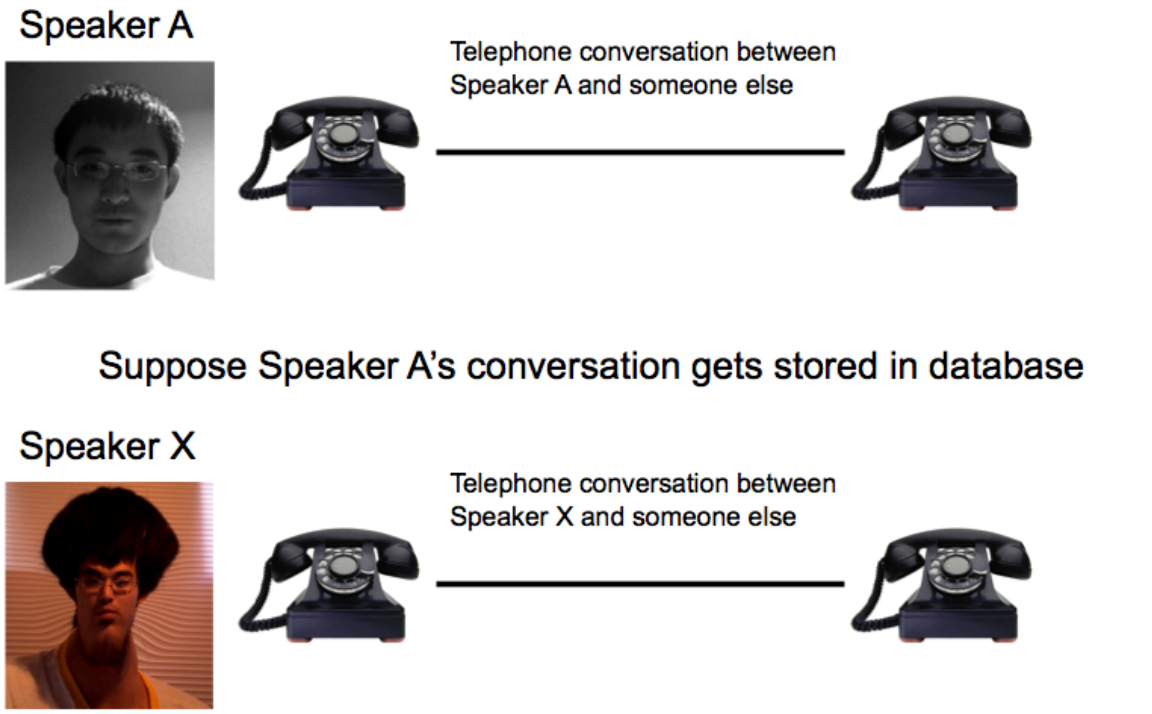

Figure 1.1. *An example of the speaker recognition question.*

constraining does introduce text-dependence in speaker recognition tasks where there are no constraints on the words that can be spoken. Figure 1.2 illustrates the concept of unit-constraining.

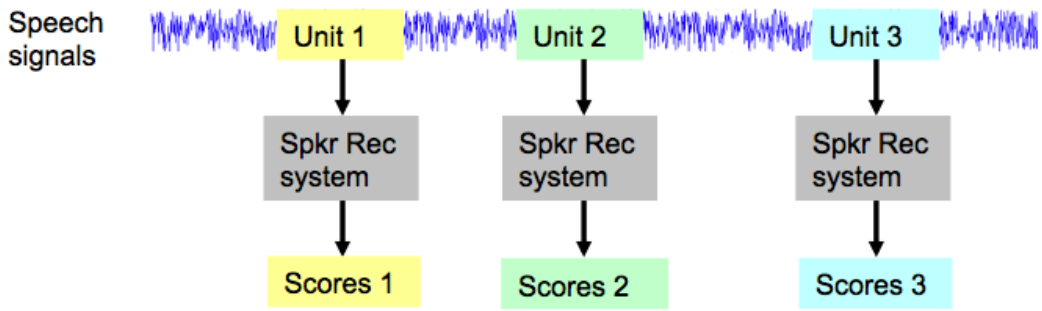

Figure 1.2. *Unit-constraining for speaker recognition.*

Such text-dependence compensates for the within-speaker variability resulting from the lexical variabilities of speakers at different times and in different utterances. For instance, if there are multiple utterances for a single speaker, and these utterances are all used in the speaker recognition system, then higher speaker recognition error

can occur for a particular speaker if his/her lexical content differ for the different utterances. Aside from compensating for within-speaker variability, unit-constraining also focuses speaker modeling power on more informative regions of speech. Thus, if certain speech units have high inter-speaker variability of pronunciation, a system can be built using only portions of speech containing those particular units.

Unit-constraining may be appropriate especially given the recent National Institute of Standards and Technology (NIST) speaker recognition evaluations where conversational speech data is used. Because there are no constraints to what words a speaker may choose to say, there can be a high degree of lexical variability for a single speaker across different recording sessions, or even within a single session. Unit-constraining thereby limits the speaker recognition analysis to only certain texts that are consistent across different recording sessions and within the same session. In addition, because the current NIST evaluation standards for speaker recognition systems have employed increasing levels of non-telephone data, the undesirable channel effects could potentially be mitigated if the units used to build the systems have higher signal-to-noise ratio.

Sturim et al., in 2002, introduced a system using GMM models trained on unit-constrained acoustic features, where the units are simple word unigrams [44]. The best unit-constrained result for this approach exceeds the result for the same approach without the use of word units. Various other speaker recognition systems - the keyword Hidden Markov Model (HMM) system and the keyword HMM supervector system [6] [31] - are extensions of the concept of unit-constraining. In these systems, HMMs, instead of GMMs, are used to model the acoustic feature distributions of speakers.

Unit-constraining via the use of linguistically defined units (i.e. words and phones) is an ad-hoc approach for selecting constraining data for speaker recognition, and depends on the existence of transcripts of such units. It does, however, illustrate the promise of data selection via unit-constraining for speaker recognition [31]. Certain other approaches to data selection have been proposed, such as a data-driven approach in [21], but these approaches are generally designed to re-create the phonetic units in the absence of an automatic speech recognizer.

In this work, I attempt to move beyond the typical linguistically-defined units for data selection via unit-constraining, and instead, select speech data based on a set of measures that can potentially determine the speaker discriminative power of various regions of speech. One of the measures investigated - mutual information - has been applied successfully to the problem of feature selection for automatic speech recognition, metal detection, and other classification tasks [28] [34] [17]. It's application to data selection in speaker recognition, however, has been minimal.

1.2 Overview of Popular Speaker Recognition Techniques

Because this work will heavily involve current speaker recognition approaches and techniques, it is worthwhile to briefly discuss the techniques.

1.2.1 Feature extraction

Two popular and effective acoustic features used for speaker recognition are Mel-Frequency Cepstral Coefficients (MFCCs) [14] and Perceptual Linear Predictive (PLP) coefficients [23]. These features are first developed for automatic speech recognition, and have subsequently been found to perform well in speaker recognition [39]. The MFCC and PLP features use information in log-spaced frequency bands of short-time speech spectra to match the log-spaced frequency responses of the human ear. The features are typically extracted on a frame-by-frame level, with 25 ms frames overlapping by 10 ms. Figure 1.3 illustrates the main steps involved in MFCC feature extraction. Note that the acronym STFT in the figure stands for Short-Time Fourier Transform. Refer to [14] for additional details and analysis of MFCC feature extraction.

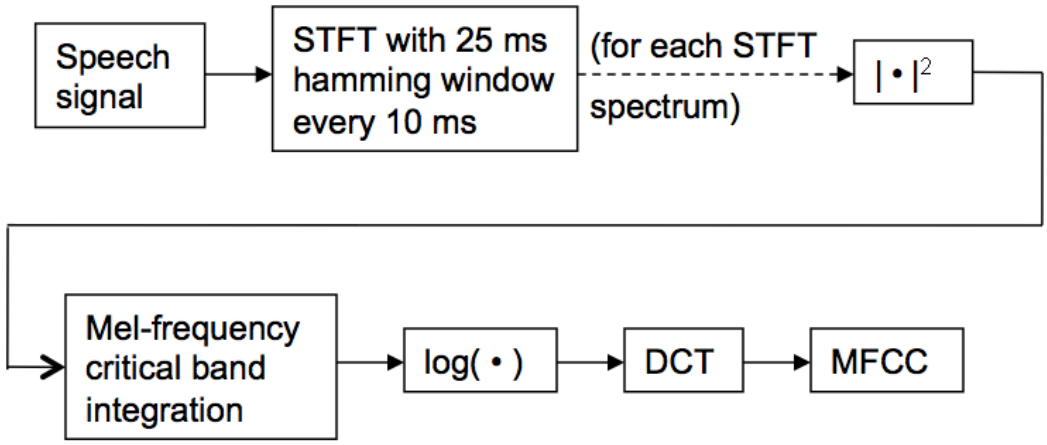

Figure 1.3. *Mel-Frequency Cepstral Coefficient feature extraction*

The mel-frequency scale is a log scale, and is given as follows:

$$f_{mel} = \frac{1000 \log_{10}(1 + f_{LIN}/1000)}{\log_{10} 2}$$

Often times, the temporal slope and accelerations of each acoustic feature vector component are used as well, and augment the basic feature vectors. These coefficients are generally referred to as "delta" and "double-delta" coefficients:

$$\text{Temporal slope parameters: } \Delta c_m = \frac{\sum_{k=-l}^{l} k * c_{m+k}}{\sum_{k=-l}^{l} |k|}$$

$$\text{Temporal acceleration parameters: } \Delta\Delta c_m = \frac{\sum_{k=-l}^{l} k^2 * c_{m+k}}{\sum_{k=-l}^{l} k^2}$$

In addition to acoustic features, prosodic features have been used as well, albeit less commonly and with different modeling approaches. The most effective prosodic features are pitch-related (f_0 mean and median), followed by energy and duration features [42]. Because prosodic features occur at longer time-scales than acoustic features, prosodic features offer complementary information to acoustic features, and the two types of features work well in combination [42].

Phonetic and lexical features have also been used successfully. Phone-lattice Viterbi decodings via automatic speech recognition provide many phone paths across speech utterances, and phone N-gram (in particular unigrams, bigrams, and trigrams) relative frequencies can be probabilistically extracted from the lattice for each utterance as features [22]. While less effective, systems based on phonetic features use complementary information from systems based on acoustic features, and the two types of systems have been shown to combine effectively [22].

1.2.2 Speaker model training via Gaussian mixture models

The distributions of features across utterances of a particular speaker are used to create speaker models for that speaker. Gaussian mixture models (GMMs) with 512-2048 mixture components have historically been used to model acoustic feature vector distributions [40] [24], because they allow for the modeling of a wide range of multi-dimensional distributions with no prior knowledge of the distribution. Given a set of acoustic feature vectors representing an utterance: $X = \{\vec{x}_1, \ldots, \vec{x}_T\}$, the likelihood of those feature vectors given a GMM model λ is the following:

$$p(\vec{x}|\lambda) = p(\vec{x}|w_i, \vec{\mu}_i, \Sigma_i) = \sum_{i=1}^{M} w_i p_i(\vec{x}) \tag{1.1}$$

where

$$p_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma_i|^{1/2}} e^{-(1/2)(\vec{x}-\vec{\mu}_i)^T \Sigma_i^{-1} (\vec{x}-\vec{\mu}_i)}$$

and

$$\sum_{i=1}^{M} w_i = 1$$

Here, there are M gaussians in the GMM and each mixture i is associated with a weight w_i, a mean $\vec{\mu}_i$, and a covariance Σ_i.

In order to obtain speaker-specific GMM parameters, a background GMM (called the Universal Background Model, or UBM) is first trained via the EM algorithm on feature vectors from a large set of utterances from many different speakers. The UBM is hence regarded as speaker-independent. The GMM parameters for each speaker are then adapted in a Bayesian context from the UBM parameters, where the UBM parameters act as priors. The speaker-specific GMM models are called target speaker models. The popular adaptation technique known as relevance MAP [41] performs EM adaptation utilizing UBM parameters as priors, and is done as follows:

E-step: Given the following statistic for mixture i of a GMM model:

$$P(i|\vec{x}_t) = \frac{w_i p_i(\vec{x}_t)}{\sum_{j=1}^{M} w_j p_j(\vec{x}_t)}$$

we have:

$$n_i = \sum_{t=1}^{T} P(i|\vec{x}_t)$$

$$E_i(\vec{x}) = \frac{1}{n_i} \sum_{t=1}^{T} P(i|\vec{x}_t)\vec{x}_t$$

$$E_i(\vec{x}^2) = \frac{1}{n_i} \sum_{t=1}^{T} P(i|\vec{x}_t)\vec{x}_t^2$$

M-step: New model parameters obtained using statistics computed during E-step as follows:

$$\hat{w}_i = [\alpha_i n_i/T + (1-\alpha_i)\hat{w}_i]\,\gamma$$

$$\hat{\vec{\mu}}_i = \alpha_i E_i(\vec{x}) + (1-\alpha_i)\vec{\mu}_i$$

$$\hat{\vec{\sigma}}_i^2 = \alpha_i E_i(\vec{x}^2) + (1-\alpha_i)(\vec{\sigma}_i^2 + \vec{\mu}_i^2) - \hat{\vec{\mu}}_i^2$$

where the scale factor γ ensures that the new weights $\hat{w}_i$ sum to unity. In addition, α is the relevance factor, controlling the balance between the UBM prior and new estimates obtained in the E-step.

Note that for unit-constrained systems, where only data from certain speech units (i.e. word N-grams, phone N-grams) are used, speaker-specific and background GMM models are trained using only portions of speech data constrained by the particular units.

1.2.3 Speaker recognition testing

This is known as the testing phase for speaker recognition. There are two variations of classification, or testing, for GMM-based systems. In one approach, the log-likelihood of feature vectors from a test utterance is computed using the target speaker GMM. The classification score is equal to the value of this likelihood, normalized by the likelihood of the feature vectors computed using the UBM. The score represents the likelihood that a test utterance is spoken by the target speaker. This is referred to as the GMM-UBM approach, and is illustrated in figure 1.4

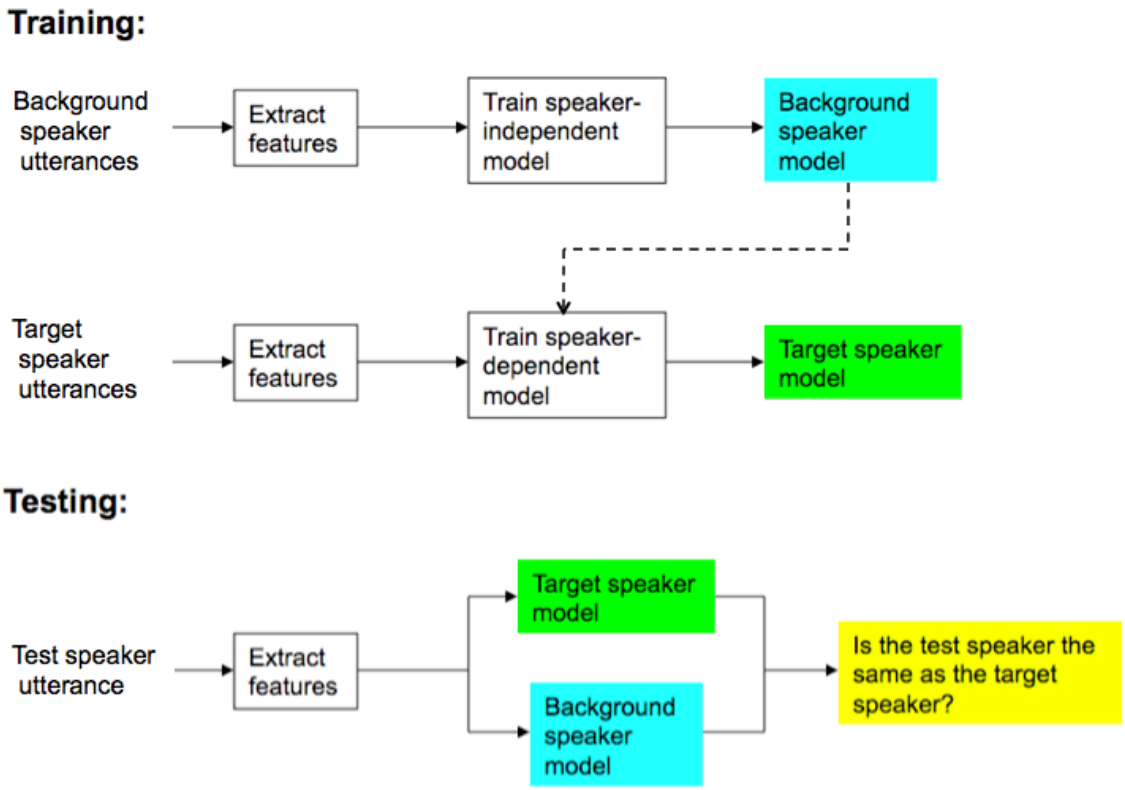

Figure 1.4. *GMM-UBM speaker recognition approach.*

In the second variation, the Gaussian mixture means of the GMM models are concatenated into a vector, known as a supervector, which is used in a Support Vector Machine (SVM) classifier [45]. This is referred to as the GMM-SVM approach. For each target speaker, its GMM means represent the positive training example(s), and the GMM means of an independent set of speakers (different from the target speaker) are the negative SVM training examples. The Gaussian mixture means of a test utterance are then given likelihoods of belonging to different speakers based on the SVM classification scores for the SVM models of those speakers. The GMM-SVM approach has been empirically determined to be superior to the GMM-UBM approach, perhaps because of the ability of SVMs to classify high-dimensional feature vectors [11].

1.2.4 Scoring

Many test utterances, containing speech from one speaker only, are scored against many speaker models, where each test utterance and speaker model comparison is referred to as a trial. The speaker in each test utterance can be thought of as having a claimed identity: the speaker of the speaker model. If the speaker in the test utterance matches its claimed identity, the trial is referred to as a true speaker trial; otherwise, the trial is referred to as an impostor trial [5].

For speaker recognition system scoring, a threshold needs to be set, such that scores above the threshold are classified as true speaker trials, and scores below the threshold are classified as impostor trials. True speaker trial scores are often higher than impostor trial scores because the test utterances in true speaker trials tend to be "better matched" to the corresponding speaker models. True speaker trials misclassified as impostor trials are known as misses; impostor trials misclassified as true speaker trials are known as false alarms.

The Equal Error Rate (EER) occurs at a scoring threshold where the number of false alarms equal the number of misses. Note that it is often preferable to have lower false alarms at the cost of higher misses (for a typical biometrics access control scenario, it is better to avoid impostor speakers gaining access as much as possible), meaning that it is better to examine speaker recognition system performances at higher scoring thresholds. In this work, however, only the EER will be dealt with. The benefits to using EER is that it gives a measure of speaker recognition performance involving no prior assumptions of task requirements and distribution of classification scores.

A typical plot of the false alarm probability versus miss probability (obtained from [5]) is shown in figure 1.5, and the EER point is denoted.

1.2.5 Speaker recognition system fusion

Note that for unit-constrained systems, where only data from certain speech units are used, classification scores are obtained only for particular units. Classification scores for multiple unit-constrained systems can be fused using a separate machine-learning algorithm [44], such as a multi-layer perceptron (MLP), an SVM, or logistic regression. In this sense, each speaker recognition system implemented on a particular unit can be thought of as a weak classifier, and the classifiers can be fused via a separate algorithm.

Multiple speaker recognition systems with scores over a common set of trials can be fused for those trials, whereby the EER of the fused system often outperforms the EERs of the individual systems. This is known as score-level fusion, where the only component of each individual system used in fusion are the scores. Figure 1.6 illustrates the fusion process for a set of speaker recognition system scores (perhaps one system for each speech unit), where a single set of scores is obtained at the output.

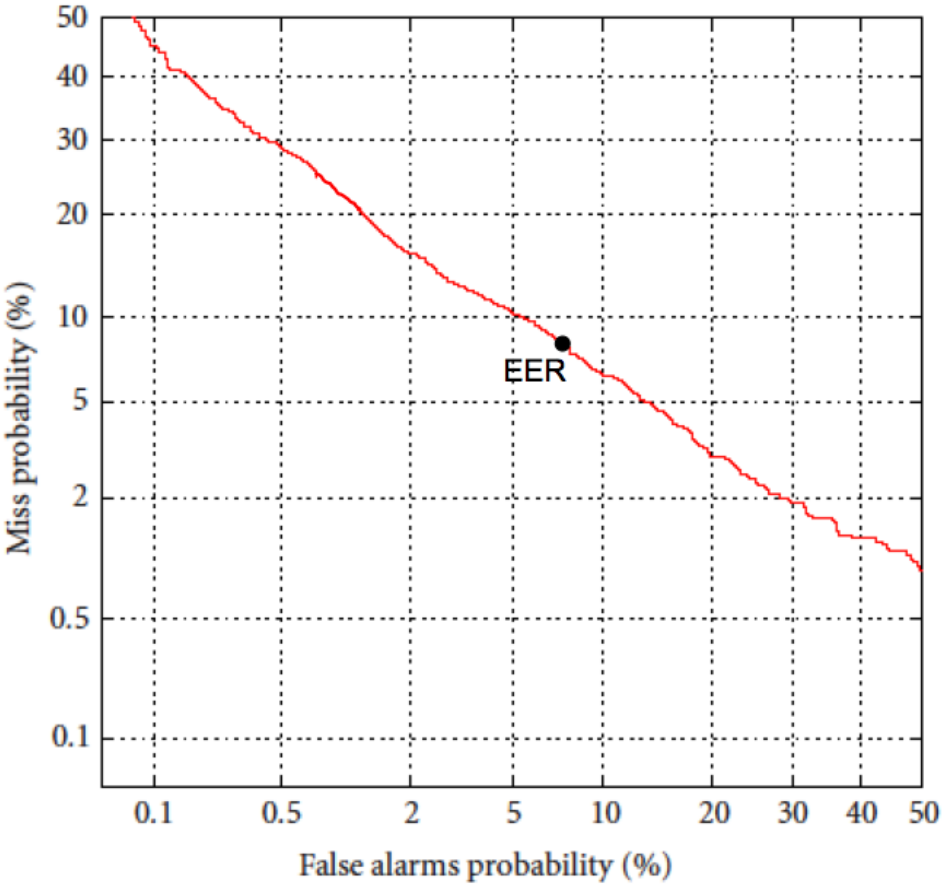

Figure 1.5. *Detection Error Tradeoff (DET) plot example*

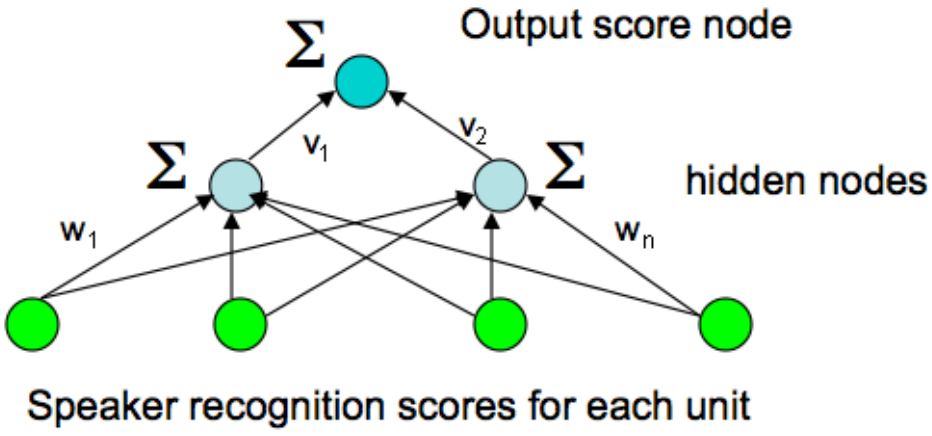

Figure 1.6. *Score-level fusion of multiple speaker recognition systems using the MLP*

Other types of fusion can also be performed for speaker recognition systems. For instance, if SVMs are used, fusion can occur by concatenating the supervectors of each system prior to SVM training and classification [31]. Feature-level fusion of-

ten outperforms score-level fusion when discriminative classifiers, such as SVMs, are involved.

1.3 Summary

This chapter gives a brief overview of the main traditional approaches to speaker recognition, and discusses unit-constraining. In particular, speaker recognition systems have traditionally relied on acoustic features, and uses GMM and SVM speaker models. For unit-constraining, only data where certain speech units (i.e. word N-grams, phone N-grams) exist are used to implement entire speaker recognition systems. The benefits of unit-constraining are the reduction of within-speaker lexical variability, as well as the focusing of speaker modeling power on more informative regions of speech. The EER is a standard metric by which speaker recognition systems are evaluated. Scores for individual speaker recognition systems implemented on particular speech units can be fused to create a more accurate system for speaker recognition.

Chapter 2

Speaker Recognition Approaches based on Unit-Constraining

As previously mentioned, the main emphasis of this work involves structured approaches to unit-based data selection, with emphasis on unit-constrained speaker recognition systems. In unit-constrained speaker recognition, only portions of speech where certain speech units (i.e. word N-grams, phone N-grams, syllables) occur are used to implement entire speaker recognition systems. Previous speaker recognition systems I've implemented for unit-constrained data selection revolve around the use of word N-grams as units [31]. These approaches and their results are described below.

2.1 Systems involving Word N-grams

For data selection using word N-grams, speech data corresponding only to a set of word N-gram units are used to construct entire speaker recognition systems. These N-grams are obtained using SRI's DECIPHER automatic speech recognizer [43]. The systems are denoted as the following: keyword-HMM system (HMM), supervector keyword-HMM system ($SVHMM$), and the keyword phone-lattice HMM system ($PLHMM$). These are systems that I previously implemented.

The HMM system uses HMM speaker modeling and MFCC feature sequences, with log-likelihood scoring [6]; the $SVHMM$ system is the supervector variant of the HMM approach, using SVM speaker modeling [11]; $PLHMM$ system is the same as the HMM system except that the HMMs are trained on phonetic versus acoustic feature sequences.

Figure 2.1 illustrates the key differences amongst the 3 systems. The main differences lie in the features, speaker models, and classifiers used to obtain the system scores [31].

Table 2.1 displays results on a portion of the SRE06 data set for a set of common word N-gram units for the HMM, $SVHMM$, and $PLHMM$ systems. Along with

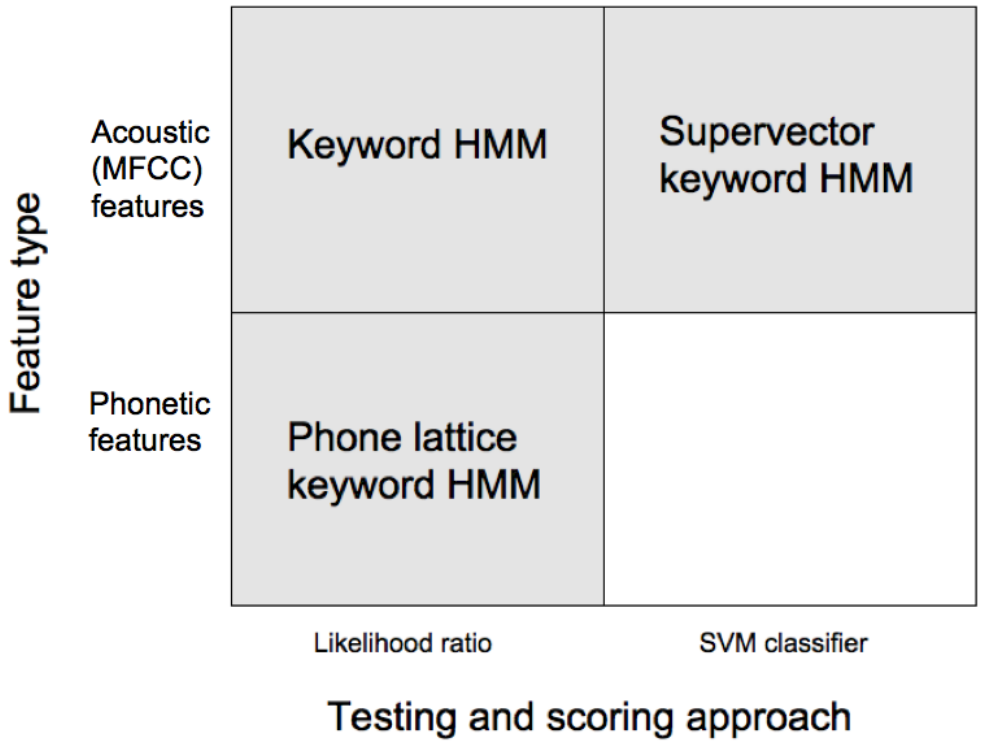

Figure 2.1. *Comparison of three existing speaker recognition approaches involving word N-gram .*

the results are the number of times they occur in a set of 1,553 background utterances from the Fisher [13] and Switchboard II [10] data sets. Refer to section 3.5.1 in chapter 3 for a description of the data sets. A total of ∼7,600 SRE06 conversation sides, ∼500 speakers, and ∼17,000 trials are used, including ∼2,000 true speaker trials.

According to table 2.1, word N-gram units that perform well for one system tend to perform well for others. This suggests that certain units have inherently superior speaker discriminative capabilities independent of the approach used for each system. While units that perform well occur more frequently in the background data, there are many exceptions such as the units I_THINK, RIGHT, and BECAUSE, which perform well according to table 2.1, but have relatively fewer occurrences compared to the units YEAH and YOU_KNOW. Figure 2.2 illustrates the performances of word N-gram units.

In addition to determining how individual units perform relative to one another, results involving combinations of units can also be obtained. The unit-combination results examine the effectiveness of the collective power of a set of units for each system. A set of 18 word N-gram units - ACTUALLY, ANYWAY, I_KNOW, I_MEAN, I_SEE, I_THINK, LIKE, NOW, OKAY, RIGHT, SEE, UH, UHHUH, UM, WELL, YEAH, YEP, YOU_KNOW, along with 20 high-frequency word unigrams – ABOUT, ALL, BECAUSE, BUT, HAVE, JUST, KNOW, MEAN, NO, NOT, ONE, PEOPLE, REALLY, SO, THAT, THERE, THINK, THIS, WAS, WHAT – are examined for

Unit	*EER* (%) results HMM	SVHMM	PLHMM	# occurrences in background data
YEAH	11.4	17.0	29.7	26,530
YOU_KNOW	11.9	17.5	26.0	17,349
I_THINK	14.7	23.5	34.0	6,288
RIGHT	14.7	22.7	30.2	8,021
UM	14.8	19.3	30.7	11,962
THAT	14.9	19.2	30.1	26,277
BECAUSE	15.2	24.1	32.3	5,164
LIKE	15.2	21.7	26.7	18,058
I_MEAN	15.8	26.8	34.0	5,470
BUT	16.6	22.9	32.6	12,766
PEOPLE	17.2	26.5	34.6	4,906
SO	17.4	24.7	34.5	14,291
HAVE	18.0	25.4	35.2	9,610
JUST	18.1	28.4	35.3	8,660
NOT	18.3	26.0	36.2	6,817
REALLY	18.6	28.5	32.2	6,674
UHHUH	20.1	26.4	37.0	8,371
THINK	20.3	30.2	39.3	3,179
OKAY	20.4	28.1	38.8	4,322
ABOUT	20.7	30.1	37.8	5,769
UH	21.3	23.6	37.7	18,065
NOW	22.5	34.2	40.9	2,851
ACTUALLY	23.1	31.5	37.9	2,240
THIS	24.8	33.9	43.7	5,408
WAS	25.0	34.4	38.3	9,888
WHAT	25.7	31.7	39.1	8,088
I_KNOW	25.8	33.1	42.4	2,142
ONE	26.2	33.3	41.4	4,559
NO	26.3	33.0	42.1	4,245
THERE	26.5	33.7	40.3	4,716
KNOW	27.8	34.8	39.9	4,767
SEE	28.0	35.7	43.1	2,006
ALL	28.4	35.5	41.7	4,681
WELL	29.7	34.3	39.7	7,590

Table 2.1. Individual word N-gram unit results for keyword HMM, supervector keyword HMM, and phone lattice keyword HMM systems. Results obtained on a subset of the SRE06 data set.

the HMM-based systems. Note that these 38 keyword units (the 18 word N-grams plus the 20 unigrams) represent 26% of the total data duration of the set of 1,553 background Fisher and Switchboard II speech utterances. The 18 word N-gram units

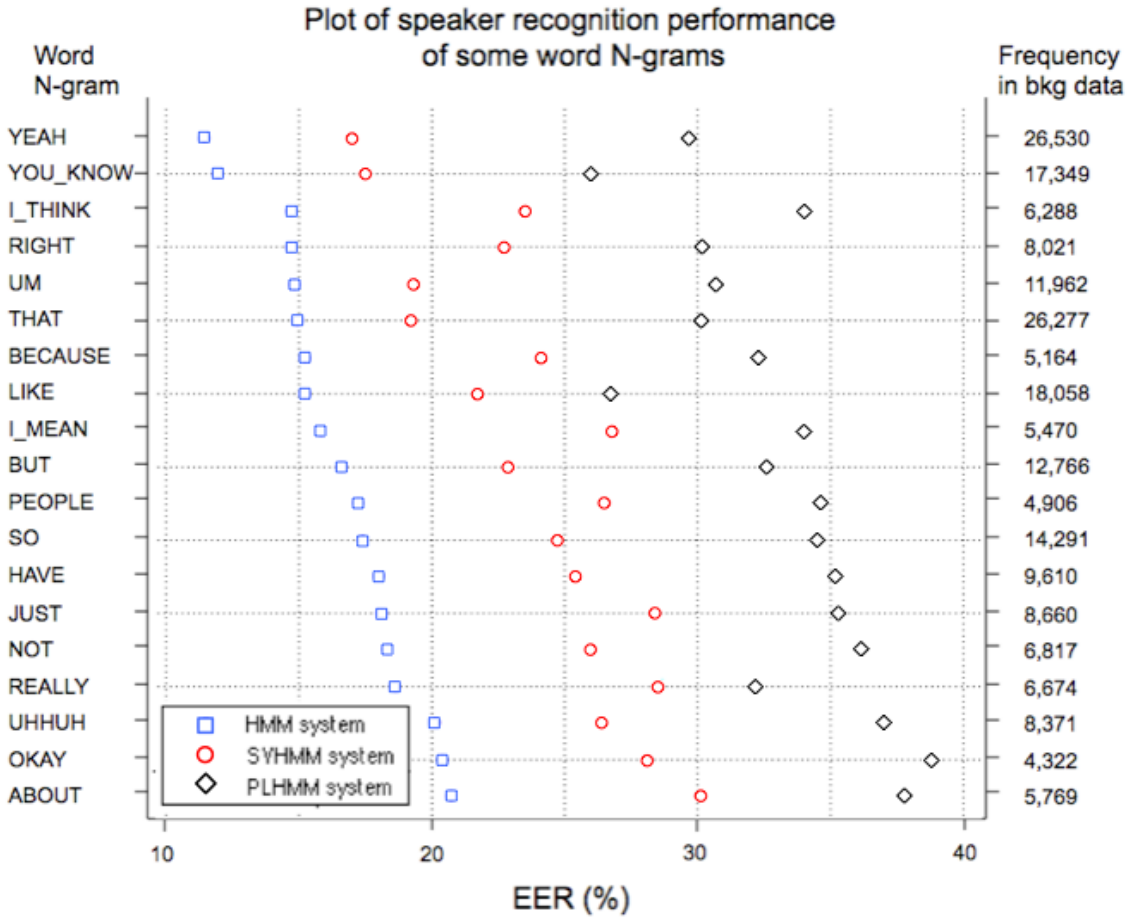

Figure 2.2. *EER of individual word N-grams and their frequencies for three systems.*

themselves represent 15% of the total duration of the background utterances. Refer to [31] for additional discussions of how the units are chosen.

Table 2.2 shows the unit-combination results for the HMM and $SVHMM$ systems, the two best HMM-based systems. Alongside the results for each system are the amounts of speech data used. For the HMM system, the speech units are combined at the log-likelihood scoring phase; for the $SVHMM$ system, speech units are combined at the SVM training and scoring phase (i.e. supervectors for each unit are concatenated to form longer supervectors comprising all units, prior to SVM training and scoring). Results for a baseline GMM-UBM system [24] is also shown. The baseline system is not a unit-based system, and hence use all of the speech data. A total of

The supervector keyword HMM ($SVHMM$) system out-performs all other systems. The 4.3% EER achieved by using 38 units in combination for the $SVHMM$ system is the best overall result, with a 6.5% relative EER improvement over the GMM baseline (4.6% EER). This is interesting, in that only 26% of speech data is being used to give improvements over a system using 100% of the data. The results also demonstrate the benefits of using more units. Increasing the number of units

System	# of units	EER (%)	bkg speech data used (%)
HMM	18	5.5	15
HMM	38	5.0	26
SVHMM	18	4.9	15
SVHMM	38	4.3	26
GMM-UBM baseline	–	4.6	100

Table 2.2. *Unit-combination results*

from 18 to 38 results in a 12.2% improvement for the supervector keyword HMM system (4.9% EER to 4.3% EER), and a 9.1% improvement for the keyword HMM system (5.5% EER to 5.0% EER).

However, as more and more units are used, one advantage of using unit-constraining, namely, reducing the amounts of speech data required, diminishes. The 38 units (26% of background data) represent a 73% increase in data over the 18 units (15% of background data). This increase in data usage greatly increases the need for memory and computation when implementing the systems.

2.2 Summary

The results demonstrate the benefits of using speech units for speaker recognition, and how good speaker recognition results can be achieved in spite of using less data than the entire amount of available speech. The speech units explored in these aforementioned systems involve word N-grams. A particular system (the $SVHMM$ system) using only portions of speech constrained by these word N-grams outperforms a system using all speech data. The word N-gram units are chosen simply because they are available to us, however, and no effort is made to determine whether or not they are the optimal speech units to use. The following chapters will provide a framework for selecting speech units that are more optimal for speaker recognition, so that a more informed choice of units can be had.

Chapter 3

Towards Structured Approaches to Data Selection

In this chapter, I attempt to gain an understanding of why certain speech units are more effective than others for speaker recognition. The primary approach consists of computing a set of measures on the feature vectors of speech data constrained by certain speech units, and seeing how indicative these measure values are of the speaker recognition system performance of the data constrained by the unit. Two information-theoretic measures are examined: Shannon's mutual information and KL-distance. In addition to these measures, a set of 11 nasality parameters that determine the nasal content of speech [38] at different temporal locations are investigated as well. Lastly, a set of miscellaneous measures, such as kurtosis, f-ratio, inter- and intra-speaker variance, unit duration, and unit frequency, are also investigated.

Once the measures are computed for each unit of interest, a relationship can be determined between the values of the measures and the performance of the units. For instance, the correlations between the measure values and the speaker recognition performances of the units (in EER) can be obtained to determine how well the measures predict the speaker discriminative ability of each unit. Figure 3.1 illustrates this process.

This method of unit selection is based on the filtering approach, as opposed to the wrapping approach. For the wrapping approach, units are selected by giving consideration to the classifier, and empirically verifying the classification result given a set of units. For the speaker recognition task, implementing the wrapping approach for unit selection would require the full training and testing of a speaker recognition system for every unit of interest. For the filtering approach, however, a set of values independent of the classifier is typically used to determine which units to use.

Hence, for the filtering approach, entire speaker recognition systems do not have to be run to determine which speech units are speaker discriminative. While it is okay to use the wrapping approach when there are a limited number of units of interest,

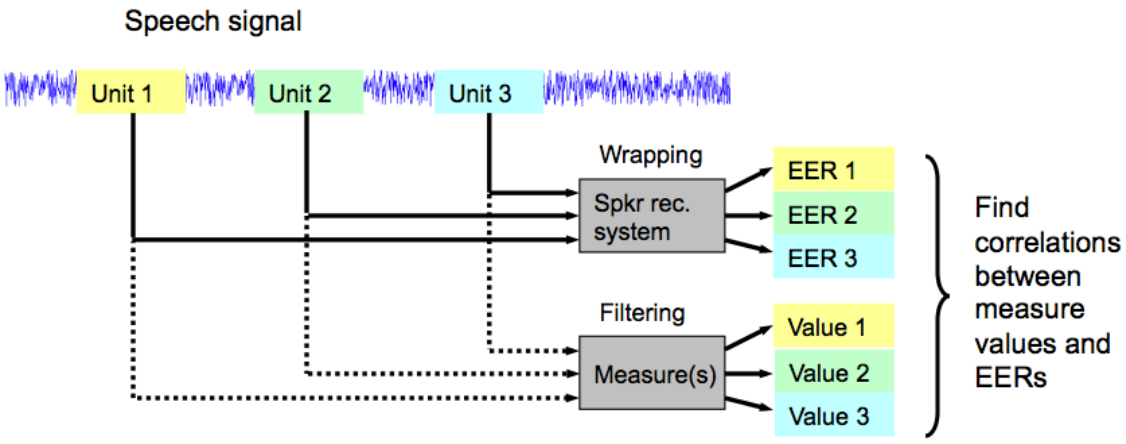

Figure 3.1. *Filtering versus wrapping for unit-selection.*

there can potentially be thousands of potentially speaker discriminative units, and running a speaker recognition system on each one is computationally infeasible.

Note that measures which are used to predict the speaker recognition performance of different speech units are regarded as relevance measures, as they measure the relevance of a particular speech unit to the speaker recognition task [37]. In the following chapter, I will also examine how the measures can be implemented on large numbers of frame sequences of speech to identify arbitrary frame sequences (and hence obtaining arbitrary speech units) that may be speaker discriminative. This results in significant computational cost savings over the brute force approach of implementing a speaker recognition system for each frame sequence to determine its speaker discriminative power.

3.1 Relevance Measures

I now introduce each of the relevance measures I've investigated to predict the speaker recognition performance of speech units. As previously mentioned, the measures investigated include the following: mutual information, KL-distance, nasality features, kurtosis, f-ratio, inter- and intra-speaker variance, unit duration, and unit frequency.

3.1.1 Mutual information

Mutual information, which measures the mutual dependence of two variables, has historically been used successfully in the related area of feature selection, such as in [28], [37], and [17]. In typical feature selection algorithms involving mutual information, the idea is to select features with high mutual information with respect to a classification label or class, such that the features are relevant to the classification

task (note that this is equivalent to selecting features based on information gain). The features can also be selected to have low mutual information with respect to one another, such that the selected features are not redundant [37]. In the feature selection approach of [37], a criterion based on mutual information is first used to filter a set of features to arrive at a set of candidate features for use with the classification task.

This work involves the selection of feature vector components. Using mutual information-based filtering to select feature vector sequences (i.e. each instance of a speech unit is comprised of a feature vector sequence) instead of feature vector components, however, is a much more complicated task. There are typically many more feature vectors in an utterance compared to its dimension (the data used typically consist of 2.5 minute utterances, which give approximately 30,000 feature vectors), and there are no set orderings of the feature vectors. For instance, with feature vector component selection, every feature vector contains features $f_1, f_2, \ldots, f_n$ in the same ordering, such that f_1 from one utterance corresponds to f_1 from another utterance. With feature vector sequence selection, however, feature vector f_n from utterance 1 probably does not correspond to feature vector f_n from utterance 2, due to the differences in utterance length and lexical variability.

Before describing the work involving the mutual information measure, essential definitions and implementation techniques are presented.

Mutual information definition

The mutual information between two continuous random variables X and Y (with distributions $p(x)$, $p(y)$, and $p(x, y)$), is given as follows:

$$I(X;Y) = \int \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{3.1}$$

It can be written equivalently as:

$$I(X;Y) = H(X) - H(X|Y) = \int_x p(x) \log p(x) dx - \int_x \int_y p(x|y)p(y) \log p(x|y) dxdy \tag{3.2}$$

where $H(X)$ and $H(Y)$ are the entropies of X and Y respectively, $H(Y|X)$ and $H(X|Y)$ are the conditional entropies, and $H(X,Y)$ is the joint entropy between X and Y. Note that equation 3.2 is also the information gain of X from knowing Y. Note that the mutual information $I(X;Y)$ can also be written as $I(X;Y) = H(Y) - H(Y|X)$.

In this work, X will typically be a time sequence of feature vectors, whereas Y is the speaker identity (a discrete random variable). If the set of speakers all have a roughly equal number of utterances (as will be the case in this work), then $p(Y) = \frac{1}{N}$, where N is the total number of speakers.

Mutual information computation

To implement mutual information for real data, the popular and effective Parzen Windowing technique is used [35]. Here, the probabilistic distribution of all feature vectors (for a given speaker class of a particular unit) are established by assigning each feature vector as the center of a Gaussian distribution. In particular, to model the distribution of a set of feature vectors $X = \vec{x}_1, \ldots, \vec{x}_n$, we have:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^{n} \phi(\vec{x} - \vec{x}_i, h)$$

where $\phi(\vec{x} - \vec{x}_i, h)$ is the Parzen window function of width h centered at x_i. In the case of the Gaussian function that we're dealing with, we have:

$$\phi(\vec{z}, h) = \frac{1}{(2\pi)^{\frac{d}{2}} h^d |\sigma|^{\frac{1}{2}}} e^{-\frac{\vec{z}^T \sigma^{-1} \vec{z}}{2h^2}}$$

where Σ is the covariance matrix, and $\vec{y}$ is a d-dimensional vector.

To compute the mutual information $I(X;Y) = H(Y) - H(Y|X)$ for n feature vectors and N speakers for a given unit, we use the technique for calculating the mutual information between continuous feature vectors and discrete classes by Kwak and Choi [28]. We first compute:

$$H(Y|X) = -\sum_{j=1}^{n} \frac{1}{n} \sum_{y=1}^{N} p(y|\vec{x}_j) \log p(y|\vec{x}_j)$$

where $\vec{x}_j$ is the j^{th} feature vector. The following computes an estimate of $p(y|\vec{x}_j)$, using Bayes' rule and the Gaussian Parzen Windowing estimation technique:

$$\begin{aligned}
p(y|\vec{x}) &= \frac{p(\vec{x}|y)p(y)}{p(\vec{x})} \\
&= \frac{p(\vec{x}|y)p(y)}{\sum_{k=1}^{N} p(\vec{x}|k)p(k)} \\
&= \frac{p(\vec{x}|y)}{\sum_{k=1}^{N} p(\vec{x}|k)} \\
&= \frac{\sum_{i \in y} \phi(\vec{x} - \vec{x}_i, h)}{\sum_{k=1}^{N} \sum_{i \in k} \phi(\vec{x} - \vec{x}_i, h)} \\
&= \frac{\sum_{i \in y} exp\left(-\frac{\vec{x}-\vec{x}_i)^T \Sigma^{-1} (\vec{x}-\vec{x}_i)}{2h^2}\right)}{\sum_{k=1}^{N} \sum_{i \in k} exp\left(-\frac{\vec{x}-\vec{x}_i)^T \Sigma^{-1} (\vec{x}-\vec{x}_i)}{2h^2}\right)}
\end{aligned}$$

This technique is used due to its effectiveness and ease of implementation. Note that in the mutual information implementation, only diagonal covariance matrices are used due to its computational cost savings, especially given the huge amounts of data we're dealing with.

KL-distance

Simple experiments are also performed using the KL-distance metric to see how well the metric models the performances of the units. For speaker recognition using GMM models, one quick and effective way to estimate the KL-distance between two models is to use its upper bound [12] [15]:

$$D(M_a||M_b) = \int p_{M_a}(x) \log \frac{p_{M_a}(x)}{p_{M_b}(x)} \leq \frac{1}{2} \sum_m w_m (\mu_m^a - \mu_m^b) \Sigma_m^{-1} (\mu_m^a - \mu_m^b) \qquad (3.3)$$

where M_i is model i, and the rest of the terms are defined in equation 1.1 in the introduction. Note that fixing the mixture weights and covariances, as is done in typical GMM-based speaker recognition systems - can still lead to effective modeling if the number of mixtures is increased. This upper bound provides a reliable distance metric estimate as it has been successfully used in deriving the popular distance-metric-based SVM kernel for the GMM-SVM system [12].

One way to use the KL-Distance as a measure is to compute the KL-distances between GMM speaker models of different speakers for a particular speech unit. Units that have high KL-distances amongst its speaker models would likely be better for speaker recognition, because its speaker models are more separate from one another, and hence become less confusable.

Nasality measures

Previous work suggests that nasal regions of speech are an effective speaker cue, because the nasal cavity is both speaker specific, and fixed in the sense that one cannot change its volume or shape [3]. Hence, different speakers should have distinct nasal sounds, and nasal regions of speech may hold high speaker discriminative power.

Various acoustic features have been proposed for detecting nasality. Glass used six features for detecting nasalized vowels in American English [20]. Pruthi extended Glass's work and selected a set of nine knowledge-based features for classifying vowel segments into oral and nasal categories automatically [38].

The goal, however, is to determine if the nasality features allow for the identification of regions of speech that have good speaker discriminative power. The fact that the features have been used to detect nasalization in vowels would possibly allow the features to better determine which speech units hold greater speaker discriminative power, since nasals themselves hold good speaker discriminative power [3]. The

means and variances of each nasality feature, computed over all data constrained by a speech unit, are used as relevance measures for that unit.

All nasality features described below are computed using 25 ms windows with 10 ms shifts. A total of 11 nasality features are implemented.

a1h1max800: This feature is the difference, measured in the log magnitude squared spectrum, between the amplitude of the first formant (a1) and the first harmonic (h1) [38]. *a1* is estimated using the amplitude of the maximum value in the band between 0 and 800 Hz. *h1* is obtained using the amplitude of the peak closest to 0Hz which had a height greater than 10dB and a width greater than 80Hz. This feature is found to be slightly smaller on average for nasals compared to non-nasal consonants and vowels.

Figure 3.2 shows the histogram of the *a1h1max800* feature values for nasal, non-nasal consonant, and vowel phonetic units. The histograms agree with the fact that this feature is slightly smaller on average for nasals.

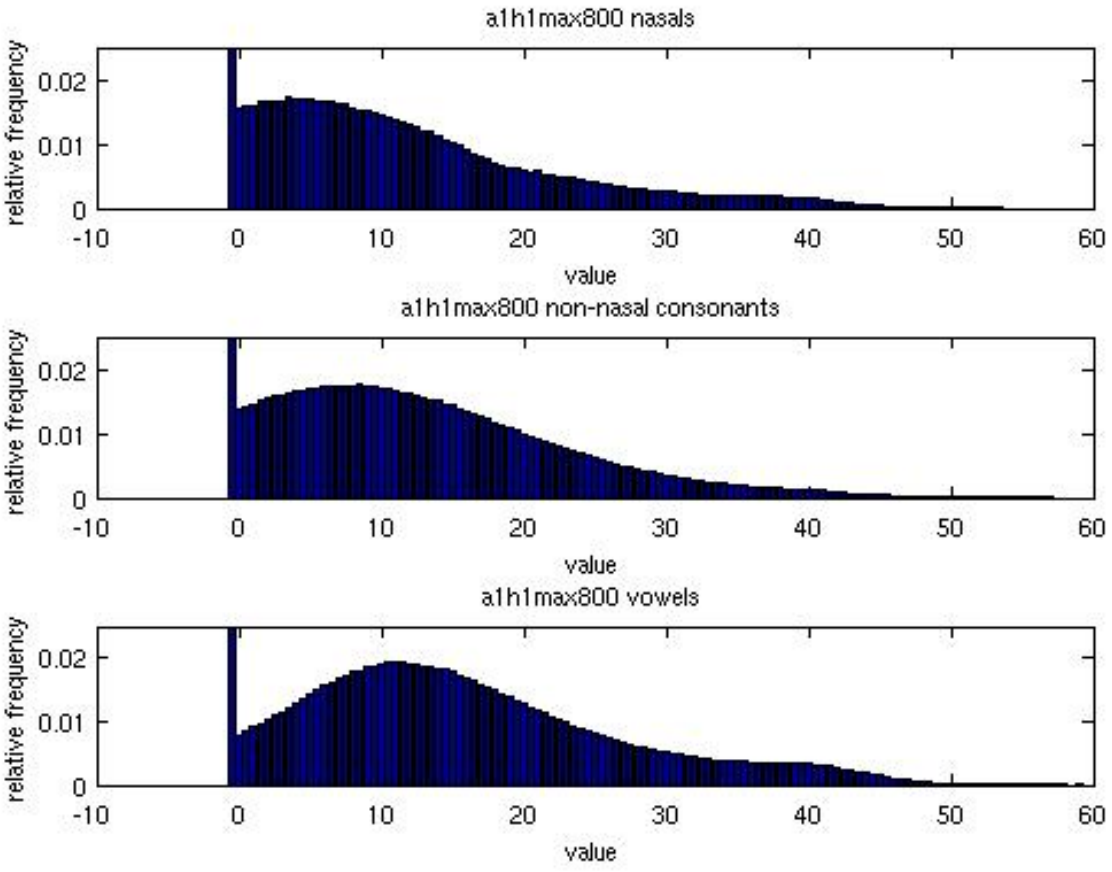

Figure 3.2. *Histogram of the a1h1max800 nasality feature for nasals, non-nasal consonants, and vowels within the set of 30 phones in e06tel1060female.*

std01k: This feature is the standard deviation of frequency around the center of mass of the frequency region below 1000Hz. Standard deviation is calculated using the spectral amplitudes 500 Hz on each side of the center of mass, but constrained to within 0 and 1000 Hz [20]. This feature is found to be smaller on average for nasals compared to non-nasal consonants and vowels, according to figure 3.3.

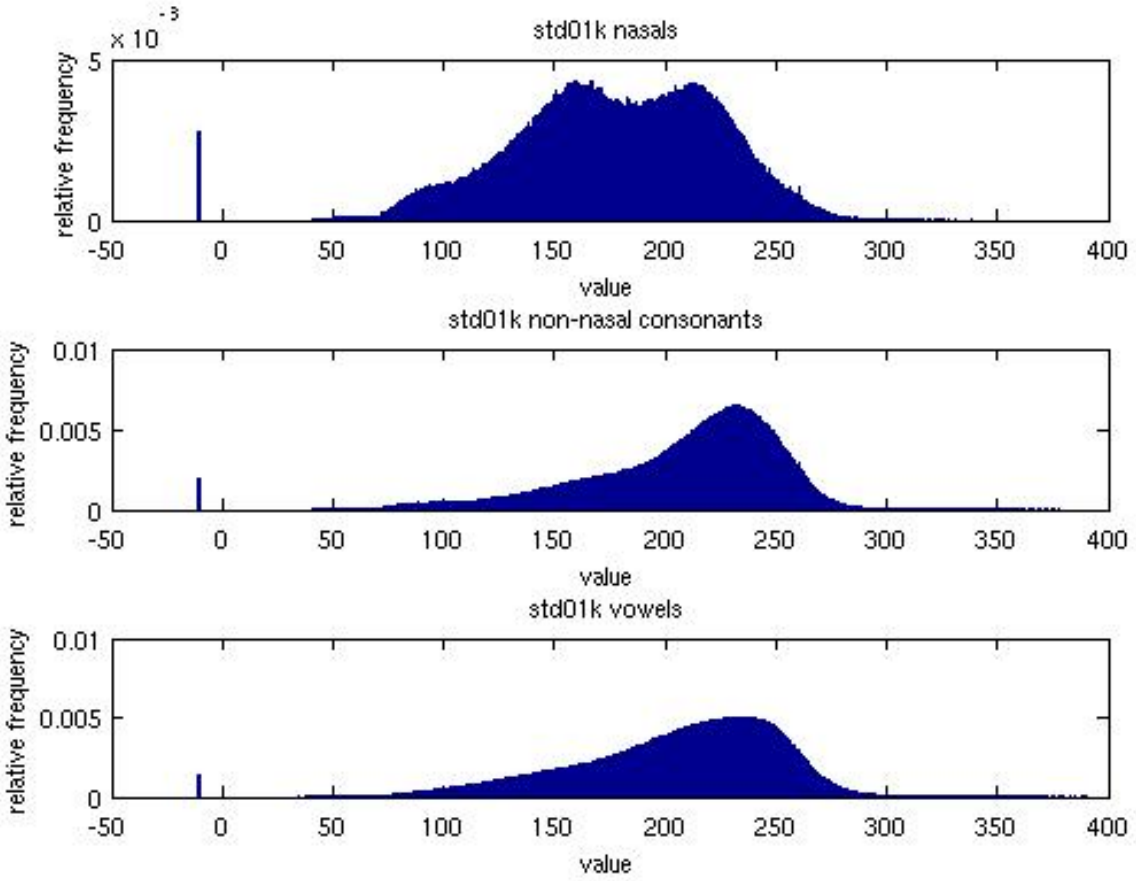

Figure 3.3. *Histogram of std01k nasality feature for nasals, non-nasal consonants, and vowels within the set of 30 phones in e06tel1060female.*

ctm01k: This feature is the center of mass of the short-term log magnitude squared (dB) spectrum amplitude in the frequency band between 0 and 1000 Hz. It is computed using a trapezoidal window with flatness between 100-900Hz. This feature is expected to be closer to 500 Hz for nasals.

a1max800: This feature is the amplitude of the first formant (A1) relative to the total spectral energy between 400 Hz and 800 Hz.

tef1: This feature is the teager energy operator for detection of hypernasality [9]. It finds the correlation between the teager energy profiles of narrow bandpass-filtered

speech and wide bandpass-filtered speech centered around the first formant. This feature is supposed to be closer to zero for nasals.

c0: The feature is the 0th cepstral coefficient representing the energy of the spectrum. The intuition is that this feature would be smaller on average for nasals because nasals appear to be softer in amplitude in general.

frat: This feature is the ratio of the spectral energies between 300 to 700 Hz and between 2,500 to 3,400 Hz. The ratio is observed to be higher on average for nasals, according to figure 3.4.

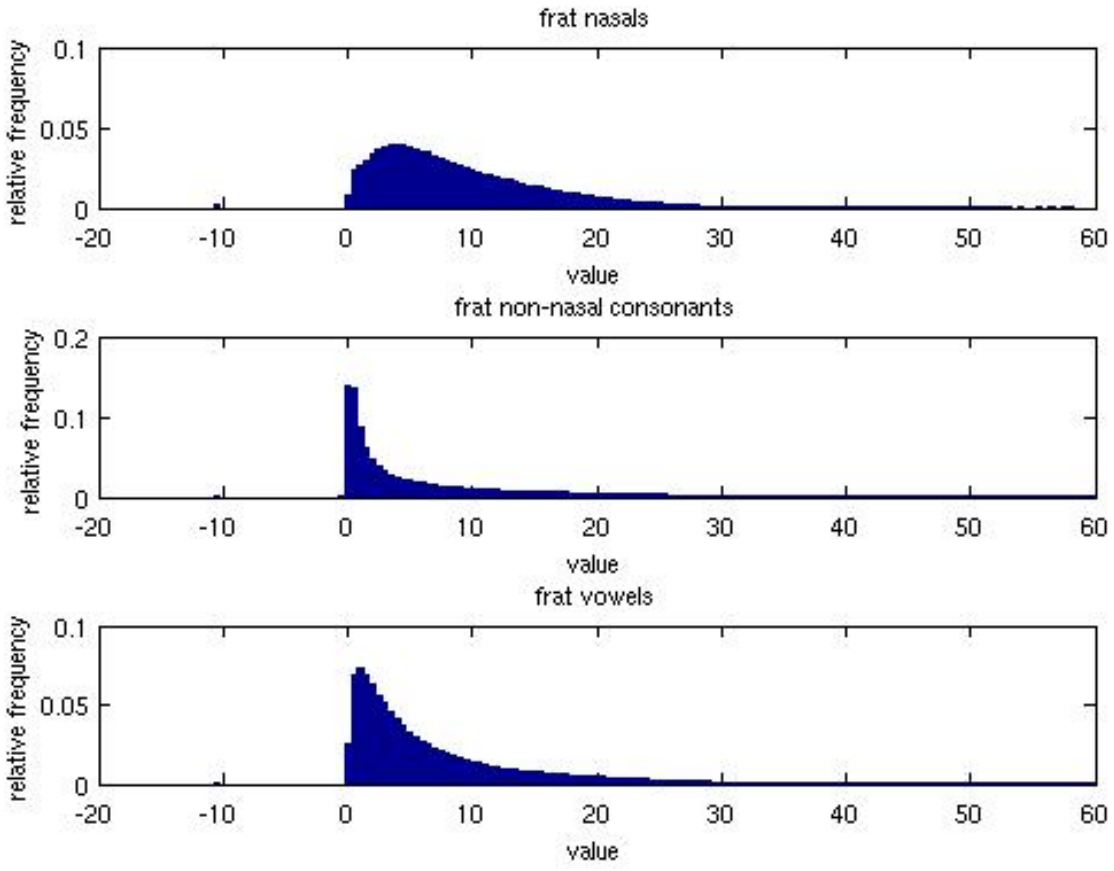

Figure 3.4. *Histogram of frat nasality feature for nasals, non-nasal consonants, and vowels within the set of 30 phones in e06tel1060female.*

Four additional features are extracted based on the detection of possible poles below and above the first formant. These poles are computed using a smoothed version of the FFT spectra. Denote *p0* and *fp0* as the amplitude and frequency of the pole below the first formant, *p1* and *fp1* as the amplitude and frequency of the pole above

the first formant, and *a1* and *f1* as the amplitude and frequency of the first format. The features are *a1-p0*, *a1-p1*, *f1-fp0* and *fp1-f1*.

a1-p0 is the difference in cepstrally smoothed FFT spectra between the amplitude of the first formant and the amplitude of the pole below f1, *a1-p1* is the difference in cepstrally smoothed FFT spectra between the amplitude of the first formant and the amplitude of the pole above f1, *f1-fp0* is the difference between the frequency of the first formant and the frequency of the pole below f1 and *fp1-f1* measures the difference between the frequency of the pole above the first formant and the frequency of the first formant. All four features are supposed to be lower for nasals according to observations.

Kurtosis

Kurtosis can also be potentially effective in predicting the speaker recognition performance of speech units. Kurtosis is a measure of peakiness and/or non-Gaussianity of a random variable, and is defined for random variable X as:

$$Kurtosis(X) = \frac{E(x^4)}{E(x^2)^2} - 3 \tag{3.4}$$

Kurtosis mismatches between training and test utterances have been shown to adversely affect speaker recognition performance, and kurtosis normalization is an effective way to improve speaker recognition performance [48]. It has been shown that MFCC feature vectors have excess kurtosis, and removing the excess kurtosis improves speaker recognition [48]. Past work have also shown that the warping of feature vectors so that they conform to a Normal distribution improves speaker recognition performance [36]. Such feature-warping effectively removes excess kurtosis in the feature vector distribution.

F-Ratio, intra- and inter-speaker variances

F-ratio and intra- and inter-speaker variances all give measures of class-separability, whereby features/data with high f-ratio, high inter-speaker variances, and low intra-speaker variances have high relevance with respect to the classification task [47]. For this work, f-ratio is the ratio of the inter- to intra- speaker variances of the feature vectors of a unit, where the inter-speaker variance is estimated as follows:

$$\frac{1}{N} \sum_{speaker:s} (\vec{\mu}_s - \vec{\mu})^T (\vec{\mu}_s - \vec{\mu}). \tag{3.5}$$

and the intra-speaker variance as follows:

$$\frac{1}{N}\sum_{speaker:s}\frac{1}{N_s}\sum_{i\in s}(\vec{x}_i-\vec{\mu}_s)^T(\vec{x}_i-\vec{\mu}_s). \tag{3.6}$$

where N is the number of speakers, N_s and $\vec{\mu}_s$ are the number and average of feature vectors respectively for speaker s, $\vec{\mu}$ is the overall average of the feature vectors, and $\vec{x}_i$ is feature vector i.

Unit duration and unit frequency

Unit duration has been shown to affect systems in the areas of automatic speech recognition and speech synthesis [33] [27]. In the automatic speech recognition of Dutch digits, training with longer utterances has been shown to improve recognition results using longer test utterances, and vice versa [33]. The Dutch digit recognition may lead one to question whether high variance in the feature sequence lengths of the various instances of a particular unit may negatively affect the speaker discriminative capabilities of the unit.

As for unit frequency, prior work has shown that units with higher frequencies tend to perform better for speaker recognition, because more data is available to train its speaker models [31] [29].

3.2 Redundancy Measure

While it is nice to have measures that determine the relevance of a speech unit to the speaker recognition task, when combining units, it is necessary to determine which units combine well with others. This is done by determining the amount of speaker recognition redundancy that units have amongst one another [37]. Units with complete redundancy (i.e. their feature vectors offer the exact same information for the speaker recognition task) would likely not improve speaker recognition results when combined. Units with low redundancy offer complementary information to the speaker recognition task, and should combine well.

3.2.1 Pearson's correlation

For a pair of units, Pearson's correlation is computed using the average feature values of each unit for each utterance. Specifically, for each utterance, the average values of the MFCC feature vectors for each unit are computed. Pearson's correlation between the averaged values of each unit is computed across all utterances. Note that the correlation is computed separately for each dimension of the feature vectors, and an overall correlation is obtained by averaging the correlations of each dimension. Figure 3.5 illustrates this computation.

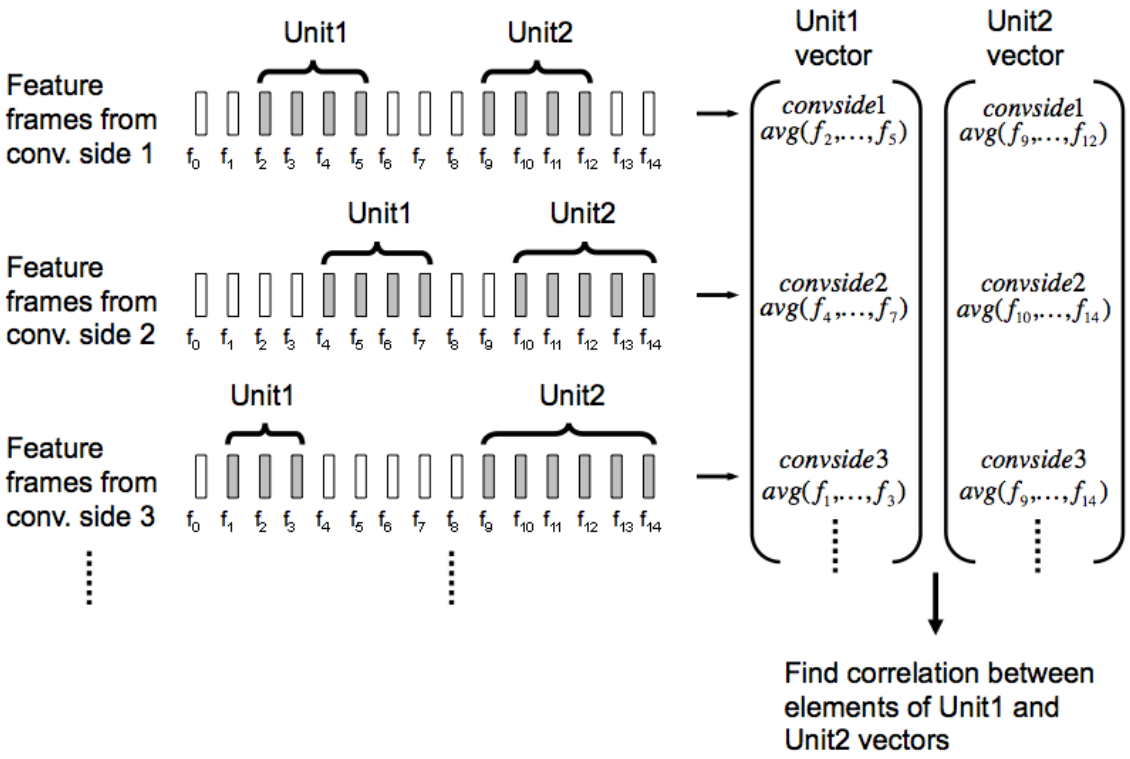

Figure 3.5. *Illustration of the procedure for computing Pearson's correlation as a redundancy measure.*

Hence, a Pearson's correlation value is associated with each pair of units. The correlation between this correlation and the relative MLP-based score level combination improvement of the unit pair is obtained to determine how well the measure predicts the redundancy of the unit pair. The relative MLP-based score level combination is determined by the relative score-level combination EER improvement over the average EER of the units standalone.

Note that I've also implemented mutual information as a redundancy measure, but found Pearson's correlation to be much more effective.

3.3 Data Selection Scheme Involving Relevance and Redundancy Measures

Given the set of relevance and redundancy measure values, an approach must be provided that integrates the values for selecting which units to use for speaker recognition. The data selection scheme involving the measures is based on the feature selection approach in [37]. Specifically, given a set of units, the task is to select N units that produce the best speaker recognition result in combination. Given the relevance measures for each unit and redundancy measures for unit pairs, the data

selection approach is the following: for a given set of pre-selected units P, determine if an additional unit Q should be selected by maximizing the following objective OBJ:

$$OBJ(Q) = Rel(Q) - \alpha \sum_{p \in P} Red(Q, p). \tag{3.7}$$

where $Rel(Q)$ is the value of the relevance measure for unit Q, $Red(Q, p)$ is the value of the redundancy measure between Q and p, and α is a weight between the relevance and redundancy factors. This objective allows one to select units that have good standalone speaker discriminative power (according to $Rel(Q)$) and are not redundant in their speaker discriminative characteristics with pre-selected units.

3.4 The Units

In order to determine how valuable the above measures are to relevance- and redundancy-based unit selection, a set of units for which measures and EERs can be computed must be arrived at. Because the word and phone transcripts from SRI's DECIPHER recognizer are available, a logical set of units to experiment with are the word and phones in the speech utterance transcripts. Note that there may be a 23% word error rate from the word transcriptions from which force-aligned phone transcripts are obtained [25]. This is okay, howeverk, because it is not necessary for the units to correspond to exact phone entities for the purposis of this work.

The following set of 30 mono-phones, as used by SRI's DECIPHER recognizer, is used in this work: AA, AE, AH, AO, AX, AY, B, D, DH, EH, ER, EY, F, HH, IH, IY, K, L, M, N, OW, P, PUH, R, S, T, UW, V, W, and Z. Note that PUH is the vowel in a filled pause, and the rest of the phonetic symbols are the ones used by SRI's DECIPHER recognizer. The table in figure 3.6 gives a mapping between the DECIPHER phonetic symbols and their rough IPA equivalents.

These phones are selected from the set of all phones because they occur most frequently in a set of 1,060 SRE06 utterances (refer to section 3.5.1 for a description of the data sets). Phones represent a good starting point for the evaluation of measures because they cover a wide range of possible speech entities, and hence provide a rich set of units which can be used to evaluate the measures.

In addition to the phones, the following set of 52 words are also experimented with: A, ABOUT, ALL, AND, ARE, BE, BECAUSE, BUT, DO, FOR, GET, HAVE, I, IF, IN, IS, IT, JUST, KNOW, LIKE, MEAN, MY, NO, NOT, OF, OH, OKAY, ON, ONE, OR, PEOPLE, REALLY, RIGHT, SO, THAT, THE, THERE, THEY, THINK, THIS, TO, UH, UHHUH, UM, WAS, WE, WELL, WHAT, WITH, WOULD, YEAH, YOU. These words have been chosen because of their high frequencies of appearance in the set of 1553 background utterances of the Fisher [13] and Switchboard II [10] data sets, which have been used as the background data for many of the previous systems. See section 3.5.1 below for a description of the data sets.

DECIPHER recognizer symbol	IPA symbol
AA	/ɑ/
AE	/æ/
AH	/ʌ/
AO	/ɔ/
AX	/ə/
AY	/ɑʸ/
B	/b/
D	/d/
DH	/ð/
EH	/ɛ/
ER	/ɝ/
EY	/e/
F	/f/
HH	/h/
IH	/I/
IY	/i/
K	/k/
L	/l/
M	/m/
N	/n/
OW	/o/
P	/p/
R	/r/
S	/s/
T	/t/
UW	/u/
V	/v/
W	/w/
Z	/z/

Figure 3.6. *Mapping between IPA phonetic symbols and the symbols used in SRI's DECIPHER recognizer.*

3.5 Experiments and Results

A set of experiments is performed illustrating the relevance and redundancy results for the above measures on the sets of units described. Before discussing the results, a description of the data sets used for obtaining the results is first provided.

3.5.1 Data, preprocessing, and speaker recognition details

The most popular data sets used by current state-of-the-art speaker recognition systems is the NIST Speaker Recognition Evaluation (SRE) data set, which is drawn from the MIXER corpus [32]. MIXER consists of conversational telephone speech between two speakers, with about 2-2.5 minutes of speech for each speaker. A conversation side refers to speech from one speaker only, and is used as the speech utterances. SRE06, SRE08, and SRE10, where the last two digits indicate the year of the NIST evaluation, are the most recent NIST SRE data sets. SRE08 and SRE10 contain not only conversational telephone speech, but also interview-style speech.

In this work, various data sets from SRE06 and SRE08 were used. Table 3.1 summarizes the main data sets used, excluding the background data sets. All data sets used contain telephone conversations between two unfamiliar speakers. For preliminary experiments, data sets consisting of 1,060 SRE06 utterances with 128 female speakers (denote this set of utterances as *e06tel1060female*), 666 SRE06 utterances with 84 speakers (denoted as *e06tel666male*), and 1,108 SRE08 utterances with 160 female speakers (denoted as *e08tel1108female*) are used. There are ∼55,000 total trials for *e06tel1060female* with ∼7,000 true speaker trials, ∼26,000 total trials for *e06tel666male* with ∼4,000 true speaker trials, ∼47,000 trials for *e08tel1108female* with ∼6,500 true speaker trials, and ∼33,000 trials for *e08tel710male* with ∼3,800 trus speaker trials.

Later experiments used a bigger set of 3,519 female and 2,533 male utterances from SRE06. Each set of female and male utterances are broken into two splits. For female utterances, split 1 has 2,001 utterances with 182 speakers (denoted as *e06tel2001female*), while split 2 has 1,518 utterances with 137 speakers (denoted as *e06tel1518female*). For male utterances, split 1 has 137 speakers with 1,508 utterances (denoted as *e06tel1508male*), while split 2 has 91 speakers with 1,025 utterances (denoted as *e06tel1025male*). Splits *e06tel2001female* and *e06tel1508male* are used for training and development, while splits *e06tel1518female* and *e06tel1025male* are used for testing. There are ∼45,000 total test trials with ∼8,000 true speaker trials for *e06tel1518female*, and ∼45,000 total trials with ∼11,000 true speaker trials for *e06tel1025male*. Note that only English language speech data is used.

In addition to the NIST SRE data sets, the Fisher and Switchboard II data sets are used as background data. Utterances in these data sets have the exact same format and stucture as those in the NIST SRE data sets, and portions of Switchboard II became part of the NIST SRE data set [10]. A total of 1,553 utterances are used in Fisher and Switchboard II.

Force-aligned phone Automatic Speech Recognition (ASR) decodings for all utterances, obtained via SRI's DECIPHER recognizer [43], are provided. A GMM-UBM system [41] with MAP adaptation and MFCC features C0-C19 (with 25 ms windows and 10 ms shifts) with deltas is used for computing the EERs of units. Various numbers of GMM mixtures, ranging from 32 to 512, are used. The ALIZE speaker

Data set	Gender	# Utterances	# Speakers
e06tel1060female	female	1,060	128
e06tel666male	male	666	84
e08tel1108female	female	1,108	160
e08tel710male	male	710	102
e06tel2001female	female	2,001	182
e06tel1508male	male	1,508	137
e06tel1518female	female	1,518	137
e06tel1025male	male	1,025	91

Table 3.1. *NIST SRE data sets used*

recognition system implementation is used [7], and the MFCC features are extracted using the HTK toolkit [1].

3.5.2 Unit-based speaker recognition results

Table 3.2 shows the EER results for each of the 30 phones on the data sets *e06tel1060female* and *e06tel666male*, along with the number of occurrences (counts) in each of the data sets. The results are sorted from the lowest to highest EERs in the data set *e06tel1060female*.

The results show that many of the non-nasal and nasal consonants performed well in speaker recognition. It is hypothesized that the use of delta features, which capture transitions into and out of the consonants, may have improved the speaker discriminative abilities of the consonants.

Results also demonstrate a -0.489 correlation between EER and the number of counts (i.e. unit frequency) in *e06tel1060female* and a -0.370 correlation for *e06tel666male*. A 0.381 correlation of the average phone duration (over each of the instances) with EER for *e06tel1060female* and a 0.399 correlation with EER for *e06tel666male*, are also observed. Correlations of phone duration variances with EERs are also computed; a 0.445 correlation is observed for females, and a 0.437 correlation is observed for males.

Overall, these results suggest that there is a significant correlation between unit frequency and EER, as well as unit duration variance and EER. The greater the unit frequency, the lower the EER, and the greater the unit duration variance, the higher the EER. This latter result suggests that units whose instances are consistent in terms of duration have lower EER and better speaker recognition performance. Results for unit duration variance resembles those in [33], which suggest that duration mismatches between training and test utterances harm the speech recognition accuracy of Dutch digits. Results for average phone duration do not have as significant a correlation, but does suggest that shorter units perform better in EER.

Unit	*e06tel1060female*		*e06tel666male*	
	EER (%)	# occurrences	EER (%)	# occurrences
T	21.2	7864	21.9	5383
DH	21.5	10222	22.7	6622
D	21.6	10627	21.5	7006
Z	21.7	8409	21.3	5212
K	22.0	7374	22.0	5011
S	22.1	6256	22.0	4030
B	22.4	10370	20.8	6932
P	23.0	9162	22.3	6308
F	23.7	8159	25.2	5412
N	23.7	6351	19.5	4422
M	23.8	7780	20.8	5230
V	24.9	11678	23.4	7843
AE	25.7	3966	23.9	2763
HH	26.0	8170	26.0	5760
PUH	26.1	3079	29.5	1796
W	26.2	7666	23.1	5037
IH	26.9	6850	25.8	4516
R	27.0	6501	24.6	4810
AA	27.2	0.152	23.6	5543
AY	27.3	3989	25.2	2772
ER	27.4	6603	26.0	5034
UW	27.8	5916	24.1	3972
IY	28.4	4834	24.7	3285
AH	28.8	7836	25.3	5809
EY	28.4	5241	25.4	3634
AO	29.3	6753	27.5	4473
EH	30.0	7677	26.1	5445
OW	30.6	3417	26.9	2592
L	30.8	6340	29.1	4455
AX	31.0	7953	31.6	5522

Table 3.2. *EER results for each of the 30 phones on the data sets e06tel1060female and e06tel666male, along with the number of occurrences (counts) in each of the data sets.*

3.5.3 Mutual information as relevance measure

For preliminary experiments, the mutual information is implemented as a relevance measure for each of the 30 phones on *e06tel1060female* and *e06tel666male*. Mutual information is the most effective measure, giving a -0.8352 correlation between the mutual information values and EERs of the phones for *e06tel1060female* and a -0.587 correlation for *e06tel666male*. This correlation implies that in general,

phones with good speaker recognition performance (low EER) also have high mutual information, and that mutual information is an effective measure for speaker recognition performance prediction.

Because the EERs of female speakers seem to be more easily predictable using the mutual information measure (i.e. higher correlation magnitude between EER and mutual information), giving results where the effects of the measures can be more easily characterized, only female speakers are used in determining the effectiveness of the measures. The differences in correlation with respect to gender could be a future topic for investigation. Note that for the female speakers, the phones with the lowest EER and highest mutual information involve the nasals and some consonants: T, K, S, P, F, V, D, DH, Z, B, M, N. The male speakers, however, have a 5.68% average phone EER improvement over the female speakers across the 30 phones.

Figure 3.7 plots the EER vs. mutual information for the 30 phones and 128 speakers on *e06tel1060female*.

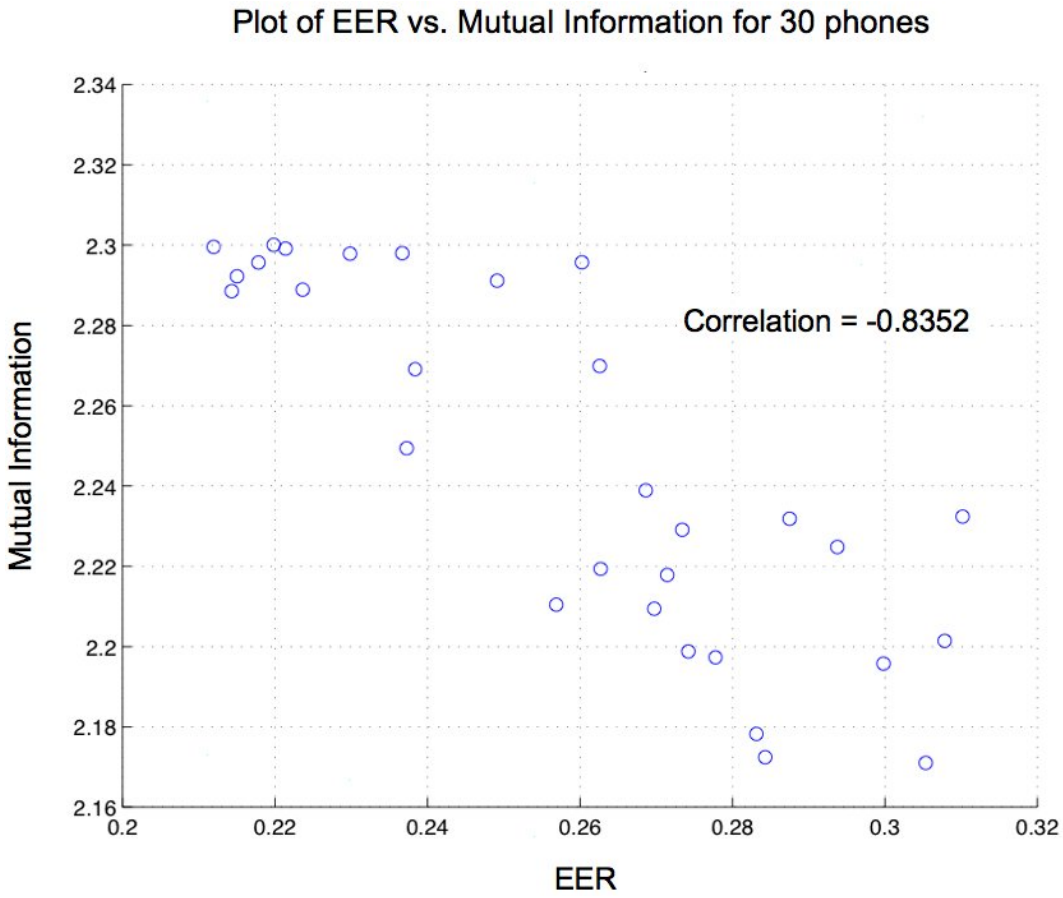

Figure 3.7. *Plot of speaker recognition EER vs. mutual information for 30 phones using 1,060 utterances and 128 female speakers on SRE06.*

Note that the following 6 phones - T, D, B, M, IH, and EH - resulted in a -0.9969 correlation between mutual information and EER for *e06tel1060female*. The same

phones show a -0.9819 correlation for *e08tel1108female*, suggesting that if all speech data are comprised of the 6 phones, a good indication of which phones are speaker discriminative based on their individual mutual information values can be achieved.

The mutual information measure performed far less well for the 52 words (-0.0871 correlation) compared to the phones. This leads me to suspect that the words may not be ideal units to begin with in terms of evaluating the measures. One reason for this may be that the words are far less stationary than the phones in the temporal context, where the feature vectors may be wildly distributed, affecting the mutual information value in unpredictable ways. Also, intra-speaker differences in pronouncing words can have a negative effect as well, where the feature vector distribution from a certain word instance pronounced by a speaker may differ significantly from the feature vector distribution of a different word instance.

Lastly, there is a greater range of frequencies for the different words (refer to 3.2). Because it has been shown that the EER of a certain unit depends on the frequency of the unit [29] [31], it is possible that the EERs of the words are heavily affected by the word frequencies (where more accurate models are obtained from words with higher frequencies). Hence, the word units are ignored in evaluating the measures in terms of their usefulness in predicting the speaker recognition performance of regions of speech.

3.5.4 Kurtosis, f-ratio, intra- and inter-speaker variances, and KL-distance as relevance measures

Kurtosis, f-ratio, intra- and inter-speaker variances, and KL-distances are also computed, using equations 3.4, 3.6, 3.5, and 3.3, on each of the 30 phones. The KL-distance measure is computed as follows: for each speech unit, a KL-distance is computed for the GMM models of each pair of speakers in the *e06tel1060female* data set, and the KL-distances are averaged to produce the overall KL-distance for the speech unit. The mean and variances of the pairwise KL-distances are used as measures. The values of all measures are then compared to the EERs of the phones, and a correlation between measure value and EER is obtained for each measure. Results on *e06tel1060female* for the correlations of kurtosis, f-ratio, intra- and inter-speaker variances, and KL-distance for each phone with respect to the EERs are shown in table 3.3. The results for mutual information, unit duration, and unit frequency measures are shown as well, and the overall set of measures are sorted from top to bottom by their correlation magnitudes. Results on *e08tel1108female* are shown for the top two overall measures: mutual information and kurtosis.

According to table 3.3, mutual information and kurtosis have the most significant correlations (-0.835 and 0.715 for *e06tel1060female*, and -0.814 and 0.709 for *e08tel1108female*) with the EERs of the 30 phones. Note that the correlation be-

Measure	Correlation w/ phone EERs for	
	e06tel1060female	*e08tel1108female*
Mutual information	-0.835	-0.814
Kurtosis	0.715	0.709
Intra-speaker variance	0.580	—
Inter-speaker variance	0.539	—
Unit frequency	-0.489	—
Unit duration variance	0.445	—
Unit duration average	0.381	—
F-ratio	0.363	—
KL-distance variance	0.114	—
KL-distance mean	0.102	—

Table 3.3. Correlations of the values of 10 measures for each phone with the respective EERs of the phones. Results obtained for *e06tel1060female* and *e08tel1108female*.

tween inter-speaker variance and EER is positive, which is counter-intuitive, since the inter-speaker variance should be high for phones with good speaker discriminative ability (and hence low EER). While this is rather strange, past results on Nuisance Attribute Projection (NAP) have suggested that minimizing inter-speaker variance helps speaker recognition performances [46][30].

One possible explanation for this is that features with high inter-speaker variance also have high intra-speaker variance in general (this has been shown by examining plots of the feature vectors along the top 2 PCA dimensions for speaker pairs). Nevertheless, these results demonstrate a significance in the correlations between a majority of the measures and EER of the phones. Thus, many of the measures are useful for speaker recognition performance prediction.

3.5.5 Nasality measures as relevance measure

As discussed in 3.1.1, the mean and variance of each of the 11 nasality features constrained by a unit are used as relevance measures for that unit. Each unit is thus associated with 11 nasality means, and 11 nasality variances, such that 22 total nasality-based relevance measures are obtained. Table 3.4 shows the correlations of each nasality relevance measure with EER for each of the phones for *e06tel1060female* and *e08tel1108female*. The correlations for the mutual information measure (the best standalone measure) are also shown.

According to table 3.4, the *a1h1max800* mean (0.796 and 0.809 correlations for *e06tel1060female* and *e08tel1108female*) and *tef1* variance (-0.764 and -0.757 correlations for *e06tel1060female* and *e08tel1108female*) are nasality measures able to most strongly predict the EER. Note that these correlations are significant at the 1% level. However, the standalone nasality measures themselves do not outperform mutual information (-0.835 and -0.814 correlations for *e06tel1060female* and *e08tel1108female*).

Measure	Mean or Var	Correlation w/ phone EERs for	
		e06tel1060female	*e08tel1108female*
a1max800	Mean	-0.346	-0.316
a1max800	Var	-0.314	-0.465
a1h1max800	Mean	0.796	0.809
a1h1max800	Var	0.748	0.699
c0	Mean	0.174	0.252
c0	Var	0.682	0.640
ctm01k	Mean	0.470	0.471
ctm01k	Var	-0.565	-0.502
frat	Mean	0.452	0.394
frat	Var	0.367	0.340
std01k	Mean	0.030	-0.041
std01k	Var	-0.474	-0.510
tef1	Mean	0.169	0.198
tef1	Var	-0.764	-0.757
a1-p0	Mean	0.397	0.373
a1-p0	Var	0.496	0.486
a1-p1	Mean	0.079	0.086
a1-p1	Var	-0.024	-0.182
f1-fp0	Mean	0.135	0.067
f1-fp0	Var	0.127	0.055
fp1-f1	Mean	-0.217	-0.238
fp1-f1	Var	0.382	0.344
Mutual information		-0.835	-0.814

Table 3.4. *Correlations of the means and variances of each nasality feature with the EERs of each phone. Results obtained on e06tel1060female and e08tel1108female.*

Note that the nasality measure results for the two data sets have a 0.991 correlation with each other, indicating that the results are largely independent of data set.

According to figure 3.2, which shows a histogram of the *a1h1max800* nasality feature for non-nasal consonants, nasals, and vowels within the set of 30 phones in *e06tel1060female*, the feature values are lower for the nasals than for the other two classes. This agrees with the fact that the *a1h1max800* mean is positively correlated with EER, because lower *a1h1max800* values correspond to units that are more nasal in nature, and such units produce lower EERs.

In order to determine the effectiveness of the measures in combination, a subset of the nasality measures are first selected via leave-one-out (LOO) selection. Starting with the full set of measures, leave-one-out selection removes one measure each iteration, selecting the measure whose removal produces the highest correlation between the combination of the remaining measures and the EER. The measures are combined via linear regression and stronger correlations between the EER and combined mea-

sures are obtained. Because of the necessity to use a training set for LOO selection, along with training the correlation coefficients, the data sets *e06tel2001female* and *e06tel1518female* are used to train LOO selection and correlation coefficients, while results will be obtained on *e08tel1108female*. Table 3.5 shows the correlations obtained from the linear regression of the remaining nasality measures after removing one measure each iteration of LOO selection.

Iteration	Nasality feature eliminated	Correlation w/ phone EERs for *e06tel1518female*
1	*a1max800* Var	0.915
2	*a1-p1* Var	0.917
3	*frat* Mean	0.918
4	*a1max800* Mean	0.919
5	*tef1* Var	0.922
6	*f1-fp0* Mean	0.923
7	*fp1-f1* Var	0.921
8	*a1-p1* Mean	0.922
9	*a1h1max800* Mean	0.927
10	*ctm01k* Var	0.924
11	*a1-p0* Mean	0.925
12	*tef1* Mean	0.924
13	*c0* Var	0.923
14	*a1-p0* Var	0.911
15	*c0* Mean	0.903
16	*frat* Var	0.895
17	*std01k* Var	0.886
18	*f1-fp0* Var	0.886
19	*std01k* Mean	0.885
20	*a1h1max800* Var	0.768
21	*fp1-f1* Mean	0.518
22	*ctm01k* Mean	0.000

Table 3.5. *All iterations of leave-one-out selection. Results show correlations obtained via linear regression of the remaining nasality measures after the specified measure is removed each iteration. Results are for e06tel1518female.*

Interestingly, the measures that perform best individually (*a1h1max800* Mean and *tef1* variance) are dropped within the first 5 iterations of LOO selection. It is good to not select too many measures to combine via linear regression so that regression weights would be over-trained (there are only 30 phones), while not too few such that the set of nasality measures combined would not be enough to produce strong correlations with EER for the 30 phones. Hence, the final 7 nasality measures are kept via LOO selection - *ctm01k* Mean, *std01k* Mean, *tef1* Var, *f1-fp0* Var, *fp1-f1* Mean, *a1-p0* Mean, and *f1fp0* Mean - whereby the correlation exceeds 0.9 (iteration 15).

Using *e06tel2001female* and *e06tel1518female* to train the regression weights for these 7 measures, and applying the weights to the nasality measures and EERs for the same set of phones on the data set *e08tel1108female*, a 0.886 correlation is obtained between the combination of the set of 7 nasality measures and EER. Recall that the correlation obtained from the mutual information measure on *e08tel1108female* is -0.814. Hence, the combination of the 7 nasality measures give an 8.8% relative improvement in correlation between measure values and EER for the set of 30 phones on *e08tel1108female*. I've experimented with keeping different subsets of measures obtained via LOO selection, as well as selecting the measures based on their individual correlation magnitudes with EER (according to table 3.5). Table 3.6 summarizes these results.

Measure(s)	Correlation w/ phone EERs for *e08tel1108female*
LOO last 2 nasality measures	0.696
LOO last 3 nasality measures	0.866
LOO last 4 nasality measures	0.879
LOO last 5 nasality measures	0.878
LOO last 6 nasality measures	0.890
LOO last 7 nasality measures	0.886
LOO last 8 nasality measures	0.892
LOO last 9 nasality measures	0.886
LOO last 10 nasality measures	0.684
LOO last 11 nasality measures	0.734
LOO last 12 nasality measures	0.662
Measures with 2 highest correlation	0.847
Measures with 3 highest correlation	0.862
Measures with 4 highest correlation	0.780
Measures with 5 highest correlation	0.826
Measures with 6 highest correlation	0.570
Measures with 7 highest correlation	0.573
Measures with 8 highest correlation	0.539
Mutual information	0.814
Mutual information + LOO last 5 nasality measures	0.878
Mutual information + LOO last 6 nasality measures	0.883
Mutual information + LOO last 7 nasality measures	0.881
Mutual information + LOO last 8 nasality measures	0.892

Table 3.6. *Results on e08tel1108female showing correlations between mutual information, and combinations of various nasality measures, and EER on 30 phones.*

These results indicate that while simply combining measures with high individual correlations with EER may produce a higher correlation than LOO selection when only a couple measures are to be combined (in this case, only the first two), LOO

selection selects measures that combine better overall, especially as greater numbers of measures are selected to be combined.

This also shows that different nasality measures offer complementary information, because the nasality measures selected via LOO do not have the strongest individual correlations, yet combine to give higher correlations with EER than those selected for their strong individual correlations. Note that the correlation with nasality and mutual information measures is roughly equivalent to the correlations with nasality measures alone, indicating that mutual information does not contribute to correlation improvements when combined with nasality measures.

3.5.6 Pearson's correlation as redundancy measure

I also implemented the Pearson's correlation (described in section 3.2) as a measure of the redundancy between pairs of units on *e06tel1060female*. Correlations are obtained between the feature vectors of all distinct pairs of the 30 phones, along with the relative improvement in the MLP-based score-level combinations of the pairs. The latter is obtained by computing the EER improvements of phones in combination over the average of the standalone phone EERs.

The optimal correlation between the correlation of feature vectors and the EER relative improvements of phone pairs is -0.486, which is obtained by considering only C1 and C2 of the MFCC feature vectors without their deltas (a -0.409 correlation is obtained when considering all MFCC coefficients). This result suggests that if the correlation between feature vectors of two phones is high, then the relative improvement of their score-level combination is low, and vice versa. Hence, Pearson's correlation is a suitable measure of unit redundancy. Table 3.7 shows the relative improvements of the MLP-combined EERs over the average of the individual EERs, and Pearson's correlation coefficient averages of C1 and C2, of the top 40 phone pairs on *e06tel1060female*. Note that the MLP weights are also trained on *e06tel1060female*, so a bit of overtuning occurs. However, these results still indicate that Pearson's correlation coefficient is a valid indicator of the potential for EER improvements of phones in combination.

According to table 3.7, 24 of the 30 least redundant phone pairs are vowel-consonant pairings. This makes intuitive sense, as vowels and consonants probably have feature vector distributions with the least in common.

3.5.7 Preliminary data selection investigation and discussion

The measures are then applied to the selection of phones in combination, using the data selection scheme in section 3.3, whereby a set of final units are arrived at for combination based on their relevance and redundancy measure values. Obtaining the relevance and redundancy measures requires a small fraction of the computational costs of running the speaker recognition system for all phones, and combining based on

Phone pair	Combined EER Improvement	Pearson's corr. coef.
AE, T	0.262	-0.083
DH, O	0.260	-0.076
AY, Z	0.257	-0.112
AE, D	0.254	-0.118
O, Z	0.253	-0.245
R, T	0.253	-0.047
K, O	0.248	-0.109
AE, M	0.248	0.005
N, UW	0.248	0.107
OW, T	0.245	-0.113
AX, T	0.244	-0.125
AE, K	0.244	-0.079
M, UW	0.244	0.056
HH, T	0.244	0.058
AY, T	0.243	-0.058
AE, Z	0.243	-0.118
R, Z	0.242	-0.106
AE, S	0.241	-0.050
B, OW	0.241	-0.051
AX, DH	0.241	-0.028
IY, K	0.240	-0.046
UW, Z	0.240	-0.036
AA, DH	0.240	-0.098
DH, L	0.240	0.006
PUH, T	0.239	-0.142
R, S	0.239	-0.152
AY, K	0.239	-0.067
L, T	0.239	-0.146
AX, D	0.239	-0.047
EH, T	0.239	-0.077

Table 3.7. Individual EERs, combined EERs and Pearson's correlation coefficients (averaged over C1 and C2) of the top 40 phone pairs with the best EER improvement over their averaged individual EERs in combination on *e06tel1060female*.

their EERs. The measures are computed using roughly the same numbers of CPUs (∼30 CPUs, with ∼100 percent usage per CPU), hardware, and coding language (C/C++). While I do not have the exact CPU clock time or some other concrete measure from which the comparisons are based, my sense of computation time comes from my general experiences with our compute cluster.

The standalone EERs of the individual phones are used as the baseline relevance measure. Only C1 and C2 are used for Pearson's correlation measure, which pro-

duces the optimal correlation according to section 3.2. All measures (including the standalone EERs) are obtained on *e06tel1060female*.

The top 5, 10, and 15 phones are selected for MLP-based score-level combination on *e08tel1108female* (with MLP weights trained on *e06tel1060female*). Relevance measures that are used include combinations of nasality measures after various leave-one-out iterations (i.e. *NAS LOO4* = nasality measures after 4th leave-one-out iteration), combinations of nasality measures and mutual information (*NAS+MI*), standalone mutual information (*MI*), standalone kurtosis (*KURT*), standalone phone EERs (*EER*), and standalone EERs without the use of the redundancy measure (*EERonly*). α for equation 3.7 is trained using *e06tel2001female* and *e06tel1518female*.

EER results are obtained on *e08tel1108female*. Tables 3.8, 3.9, and 3.10 show the results for α equal to its optimal value, for combinations of 5, 10, and 15 phones respectively. Note that all results are for female speakers only, as female speakers are better able to illustrate the efficacy of the measures. Results for both male and female speakers will be shown in the actual arbitrary data selection experiments in section 4.4 in the next chapter.

Relevance measure	5 phones selected	EER (%)
MI	HH W N K V	13.14
NAS LOO4	S HH M N B	12.83
NAS LOO5	S HH AE N B	12.42
NAS LOO6	S HH AE N B	12.42
NAS LOO7	S HH AE N B	12.42
NAS LOO8	S HH AE N B	12.42
NAS+MI LOO4	S HH M N B	12.83
NAS+MI LOO5	S HH AE N B	12.42
NAS+MI LOO6	S HH M N B	12.83
NAS+MI LOO7	**Z S AE N B**	**12.23**
NAS+MI LOO8	S HH AE N B	12.42
KURT	Z HH AY P DH	13.36
EER	Z M T AE N	12.38
EERonly	Z M T D N	12.39

Table 3.8. *MLP score-level combination of top 5 phones selected according to relevance and redundancy measures with optimal α. Results obtained on e08tel1108female.*

According to tables 3.8, 3.9, and 3.10, unit-selection based on the relevance and redundancy criterion using the various relevance measures previously discussed produces performance gains compared to selecting units based on standalone EERs. For the selection of 5 phones, the lowest EER (12.23%) is produced using the combination of nasality measures for LOO7 and mutual information (the measure *NAS+MI*

Relevance measure	10 phones selected	EER (%)
MI	AH S T W N K V HH AY DH	10.16
NAS LOO4	F S N P B Z HH M AE PUH	9.68
NAS LOO5	**F S N B Z HH M D AE PUH**	**9.48**
NAS LOO6	**S N P B Z HH M AE PUH AY**	**9.48**
NAS LOO7	**F S N B Z HH M D AE PUH**	**9.48**
NAS LOO8	S N B V Z HH M AE PUH AY	9.99
NAS+MI LOO4	F S N P B Z HH M AE PUH	9.68
NAS+MI LOO5	F S P B Z HH M AE AY PUH	9.94
NAS+MI LOO6	**S N P B Z HH M AE PUH AY**	**9.48**
NAS+MI LOO7	**F S N B Z HH M AE D PUH**	**9.48**
NAS+MI LOO8	S N B V Z HH M AE PUH AY	9.88
KURT	T K P EY B Z HH AE AY DH	10.35
EER	T N K V Z M AE D AY PUH	10.04
EERonly	F S T N P B Z M D DH	10.35

Table 3.9. *MLP score-level combination of top 10 phones selected according to relevance and redundancy measures with optimal α. Results obtained on e08tel1108female.*

Relevance measure	15 Phones selected	EER (%)
MI	AH F S W T N P K B V Z HH M D DH	9.22
NAS LOO4	F S N P EY B Z AA HH M AE R AY PUH IY	9.00
NAS LOO5	F S N P EY B Z AA HH M AE R AY PUH IY	9.00
NAS LOO6	F S UW N P EY B Z HH AA M AE R PUH AY	9.02
NAS LOO7	F S N P EY B Z AA HH M AE R AY PUH IH	9.00
NAS LOO8	F S UW N P B V Z HH M D AE AY PUH DH	8.89
NAS+MI LOO4	**F S N P EY B Z HH M AE R PUH AY DH IY**	**8.70**
NAS+MI LOO5	F S N P EY B Z AA HH M AE R AY PUH IY	9.00
NAS+MI LOO6	**F S N P EY B Z HH M AE R PUH AY DH IY**	**8.70**
NAS+MI LOO7	F S N P EY B V AA HH M AE AY PUH DH IY	8.96
NAS+MI LOO8	F S UW N P B V Z HH M D AE AY PUH DH	8.89
KURT	S F T K P EY B Z HH AO D AE AY DH EY	9.55
EER	S T N K EY V Z AA HH M D AE R AY PUH	9.02
EERonly	F S T N K P B V Z HH M AE D AY DH	9.33

Table 3.10. *MLP score-level combination of top 15 phones selected according to relevance and redundancy measures with optimal α. Results obtained on e08tel1108female.*

LOO7). Using standalone EERs with no redundancy information for the selection of 5 phones produces an EER of 12.38%. Although the EER obtained from *NAS+MI LOO7* in the combination of 5 phones via relevance and redundancy selection is not significantly better than from simply selecting phones with the best EERs, the former does not require a speaker recognition system to be run, producing computational cost savings.

Using redundancy measures in unit selection slightly improves the EER of the combined phones (10.04% versus 10.35% using standalone EERs as relevance measures). However, the results did not change much with respect to the number of LOO iterations used for selecting the nasality measures. Using mutual information and kurtosis standalone are both worse in general than incorporating nasality measures as relevance measures - the worst result using nasality measures is 12.83% EER (obtained using *NAS LOO4*, *NAS+MI LOO4*, and *NAS+MI LOO6*), which is still better than the best mutual information/kurtosis result (13.14% EER). The standalone mutual information measure performed slightly better than the standalone kurtosis measure (13.14% EER versus 13.36% EER), albeit insignificantly.

For the selection of 10 phones, the lowest EER (9.48%) is produced using nasality measures standalone for LOO5, LOO6, LOO7, and nasality measures in combination with mutual information for LOO6 and LOO7. This is a 9.18% relative improvement over the EER obtained via selecting phones with best individual EERs (10.35%), and a 5.91% relative improvement over selecting phones using the individual EERs as the relevance measure and Pearson's correlation as the redundancy measure. Once again, incorporating nasality measures into the relevance measure improves the EER of the combination of phones over standalone mutual information and kurtosis. The lowest EER obtained using nasality measures (9.48%) is 6.69% better relatively than the EER obtained using standalone mutual information (10.16%), while standalone mutual information is slightly better than standalone kurtosis (10.16% EER versus 10.35% EER).

For the selection of 15 phones, the use of nasality measures for LOO4 or LOO6, in combination with the mutual information measure, produces the best overall EER (8.70%). Similar to the selection of 5 and 10 phones, the best overall EER obtained via relevance and redundancy phone selection outperforms phone selection via best individual EERs (9.33%). Likewise, the combinations of phones selected by the incorporation of nasality measures into the relevance measure produce EERs that outperform the combinations of phones selected using standalone mutual information and kurtosis. Consistent with the results for the selection of 5 and 10 phones, standalone mutual information produces a slightly lower overall EER than standalone kurtosis (9.22% EER versus 9.55% EER).

Note that in all cases, the incorporation of the redundancy measure allows for the selection of phones that combine to produce slightly lower EERs. Comparing the *EER* and *EERonly* results for the selection of 5, 10, and 15 phones shows that incorporating the redundancy measure slightly improves the EER from 10.35% to

10.04% for 10 phones, and from 9.33% to 9.02% for 15 phones. For the selection of 5 phones, the EER stays the same (12.39% vs 12.38%).

According to the results, I have demonstrated that it is possible to select effective units for speaker recognition without running the actual speaker recognition system, by obtaining relevance and redundancy measures from acoustic feature vectors. These measures are able to predict effectively the speaker recognition performances of the regions of data they are computed on. Note that using the standalone phone EERs as a baseline relevance measure requires running the speaker recognition system, but selects phones that combine to produce higher EERs than selecting phones using a combination of nasality measures and mutual information as relevance measures, and Pearson's correlation as the redundancy measure. The results indicate that taking both relevance and redundancy (as opposed to just relevance) into consideration leads to better unit selection. Interestingly, only the MFCC C1 and C2 coefficients are sufficient for computing the redundancy measure.

3.6 Summary

The main elements of this part of the work examines various measures, which include mutual information, kurtosis, KL-distance, f-ratio, intra- and inter-speaker variance, unit duration, unit frequency, and 22 nasality measures as they relate to the speaker recognition performances of units standalone and in combination. The measures are computed using MFCC acoustic feature vectors constrained by speech units, which consists of a set of 30 phones and 52 words. Results showed that mutual information and kurtosis are good standalone measures in predicting the performances of the 30 phones, while nasality measures (when combined together) are effective standalone and in combination.

Data selection for sets of 5, 10, and 15 phones shows that relevance and redundancy-based data selection (with Pearson's correlation between feature vectors of different units used as the redundancy measure) involving the measures allow for the selection of phones with combined EER better than the EER obtained by simply combining phones with top standalone EERs. Hence, the approach involving the measures allows for the selection of phones without having to determine how well they perform individually by running entire speaker recognition systems, resulting in computational cost reductions. The following chapter will discuss the selection of arbitrary sequences of feature vectors as units, allowing one to move beyond the use of linguistically-defined units such as phones and words.

Chapter 4

Measure-Based Selection of Arbitrary Speech Segments

In this chapter, I create and implement a new data-selection approach for speaker recognition, which utilizes some of the measures I've investigated in the previous chapter. In particular, the data selection approach allows for the selection of arbitrary frame sequences as speech units that are speaker discriminative according to the measures. While it is infeasible to examine the set of all possible frame sequences in a utterance (which may consist of over 20,000 speech frames from a single speaker), the approach does allow for the examination of potentially interesting regions of speech, and to select the most speaker discriminative segments in such regions of speech.

The measures I've investigated allow for a quick determination of the speaker recognition performance of each unit without having to run a speaker recognition system, giving indications of the relevance of a unit with respect to the speaker recognition task. Measures that have high correlation (in magnitude) with the speaker recognition EERs of the units have good predictive value for speaker recognition, and are good measures for the task of arbitrary data selection.

My data selection approach involves computing the most relevant measures on subsequences within phone N-grams. The subsequences that are most relevant for the speaker recognition task (indicated by the measures) are selected as new units. The ANN/HMM approach for speech recognition [8] is used to train models for the new units, and to decode new data to determine the locations of the units in all utterances. For each new unit, a speaker recognition system is run to evaluate the speaker recognition performances of the units. Note that while it is possible to run a speaker recognition system separately for each phone N-gram subsequence to determine what the most speaker-discriminative units are, its computational cost would be high due to the large numbers of feature sequences requiring consideration.

The data sets used for training and testing are *e06tel2001female*, *e06tel1518female*, *e06tel1518male*, *e06tel1025male*, *e06tel1108female*, and *e06tel710male*, described in the previous chapter.

4.1 Data Selection

I begin by using the two most reliable standalone measures discussed in the previous chapter - mutual information and kurtosis. Because my goal is to compute the measures on corresponding feature vector sequences across utterances for all speakers, I need to be able to identify corresponding regions of speech across the utterances so that regions used for measure computation across all utterances are consistent with one another.

The data selection approach involves computing the mutual information and kurtosis measures on every feature vector sequence within phone N-grams (4-, 5-, 6-grams are used) across utterances for all speakers. The reason phone N-grams are used is that they represent relatively lengthy segments of speech containing an abundance of arbitrary frame sequences that are unconstrained by the boundaries of individual phonetic units. In addition, because phonetic transcripts for all speech data are readily available via the DECIPHER recognizer [43], it is easy to spot multiple instances of phone N-grams such that sufficient data can be acquired for measure computation.

A more complicated approach would be to use an MLP to determine portions of all utterances where the same word or phone are spoken. Note that this approach is similar to the MLP phonetic matching approach described in [19]. However, the MLP-based data selection approach is computationally expensive even when trying to extract similar segments across only two utterances [19], and multiple utterances produce exponential increases in computational complexity. Hence, this approach is ignored in favor of the simpler approach of simply utilizing the phonetic transcripts.

4.1.1 Finding similar segments across utterances

The first step involves finding long phone N-grams that frequently occur across many utterances, so that there's enough data for measure computation. Table 4.1 shows all the high-frequency phone N-grams of length 4 or more that have been considered (15 in all), and their counts across all utterances in the data set *e06tel1060female*. These phone N-grams are considered due to their optimal combinations of length and frequencies of appearance.

4.1.2 Unit length normalization and measure computation

Because different phone N-gram instances differ in length, dynamic time-warping (DTW), computed using acoustic feature vectors (MFCC C0-C19 and deltas), is used to normalize all phone N-gram instances in all utterances, as shown in figure 4.1.

For each length-normalized phone N-gram, mutual information and kurtosis are dynamically computed for all relevant frame sequences within the phone N-gram using instances across all speakers. Each frame sequence is hence associated with a mutual information and kurtosis values. Only the SRE06 training portions of the

Phone N-gram	Length (# phones)	Frequency of occurrence in *e06tel1060female*
Y_UW_N_OW	4	8,364
TH_IH_NG_K	4	4,474
D_OW_N_T	4	3,751
DH_AE_T_S	4	3,171
JH_AX_S_T	4	2,538
P_IY_P_AX_L	5	2,357
AY_M_IY_N	4	2,063
B_AX_K_AH_Z	5	2,043
AX_B_AW_T	4	2,021
HH_AE_V_T_AX	5	1,262
S_AH_M_TH_AX_NG	6	966
K_AY_N_D_AX_V	6	879
AY_TH_IH_NG_K_DH	6	844
Y_UW_Z_AX_K	5	699
G_OW_AX_NG_T_AX	6	678

Table 4.1. 15 Phone N-grams considered, based on individual length and frequencies of occurrence.

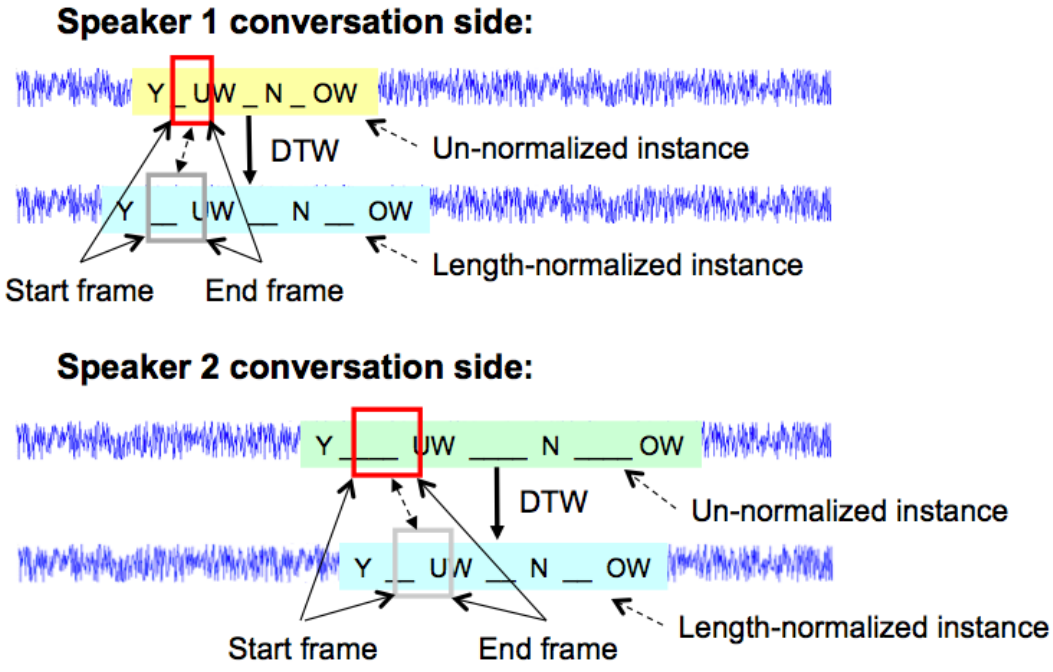

Figure 4.1. *Length normalization of phone N-gram sequences across all speakers.*

data set (*e06tel2001female* and *e06tel1508male*) are used for this purpose. Figure 4.2 illustrates this process.

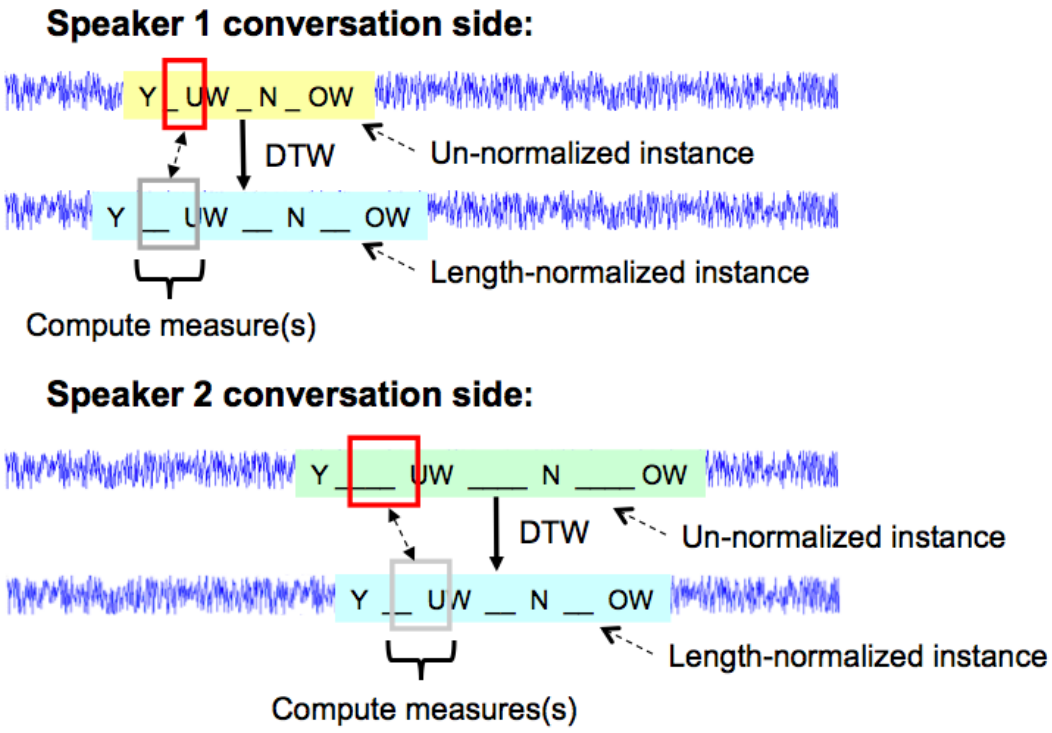

Figure 4.2. *Computing measures on length-normalized phone N-gram sequences.*

4.1.3 Frame sequence selection

Upon computing the mutual information and kurtosis measures on the length-normalized phone N-gram frame sequences, sequences within each phone N-gram are ranked according to the values of its mutual information and kurtosis measure. Desired frame sequences from the length-normalized phone N-grams are selected based on a simple linear combination of the mutual information and kurtosis measure values. The new units consist of the highest-ranking frame sequences within a particular phone N-gram, according to their measure values. For each new unit, the length-normalized start and end frames are mapped to corresponding start and end frames of the un-normalized phone N-gram instances (see figure 4.1), and used as training instances for decoding the new units. This procedure is referred to as inverse-DTW mapping.

4.2 Decoding New Units

The ANN/HMM speech recognition paradigm [8] is used as a keyword-spotting mechanism for decoding the new units. The reason this approach is used is that it is a fast and effective approach readily available to us, and outperforms various other approaches at our disposal. Specifically, an Artificial Neural Network (ANN) with 1 hidden layer, 500 hidden nodes, and a context window of 9 frames is trained using

utterances containing the training instances of the new units. Each frame consists of MFCC C0-C19 and their deltas. The ANN output consists of classes corresponding to each new unit, as well as a junk class, consisting of all other data. Note that silences have been removed, such that a silence output class is unnecessary. Once the ANN is trained, a forward-pass is run on all training and test utterances, such that each frame is mapped to a vector of class posteriors.

An HMM-based Viterbi decoding of the ANN class posterior vectors is then run on the training and test utterances. One or two HMM states is typically used per class. Note that for the training utterances, the labels corresponding to the output classes generated by the HMM, as opposed to the original labels obtained via inverse-DTW mapping, is used in subsequent speaker recognition training and testing. This is to ensure a level of consistency between the training and testing utterances for speaker recognition. Figure 4.3 summarizes the unit selection procedure.

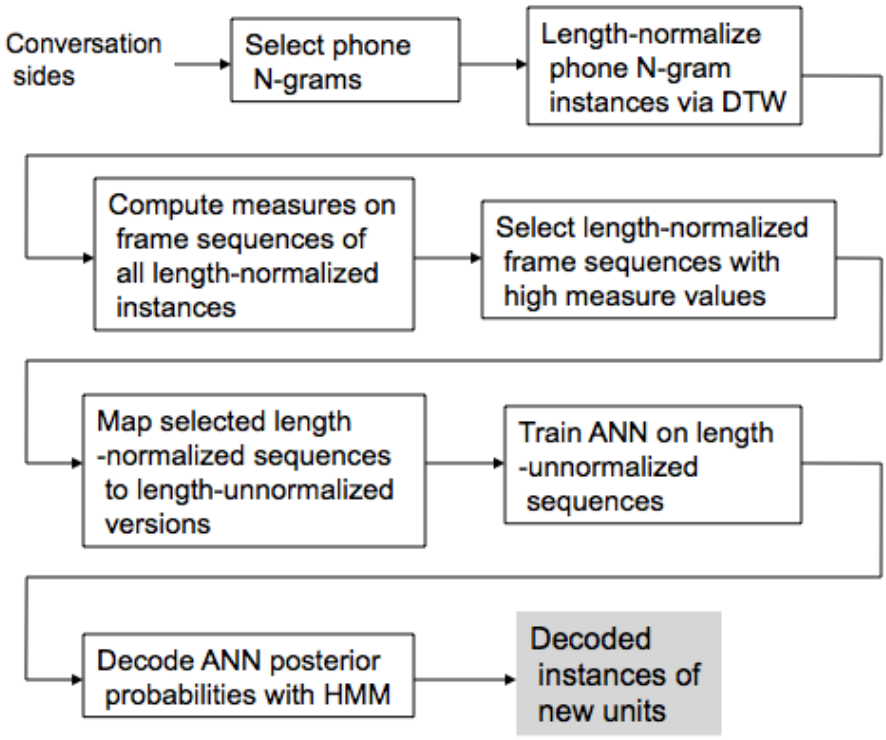

Figure 4.3. *Summary of unit selection procedure*

4.3 The Overall Set of Units Used

Along with the new units, the same set of 30 mono-phones introduced in the previous chapter are used as standalone units: AA, AE, AH, AO, AX, AY, B, D, DH, EH, ER, EY, F, HH, IH, IY, K, L, M, N, OW, P, PUH, R, S, T, UW, V, W, Z, where

PUH is the vowel in a filled pause. These phones are selected so that their speaker recognition performances can be compared to those of the new units.

Recall that a set of 15 phone N-grams from which frame sequences are used for the new units is selected according to their length and high frequencies of occurrence in a subset of the SRE06 utterances. The phone N-grams chosen based on these criterion are different for male and female speakers, because the phonetic content of speech for the genders slightly differ.

Recall also that once the phone N-grams are chosen, their instances are length-normalized via DTW to a length of 40 frames. For each length-normalized phone N-gram, one of the frame sequences with the highest linear combination of mutual information and kurtosis measure values is chosen, so that the frame sequence may have high speaker discriminative power. Each frame sequence and its corresponding phone N-gram comprises a new unit, and are shown in table 4.2. Frame sequences are shown in brackets, with the start frame followed by the end frame, after each phone N-gram. Because the frame sequences are selected from phone N-grams length-normalized to 40 frames, they can range from 0 through 39.

Phone N-grams and frame sequences for new units for male speakers	
Y_UW_N_OW [16,23]	P_IY_P_AX_L [22,30]
AY_M_IY_N [10,18]	S_AH_M_TH_AX_NG [19,27]
JH_AX_S_T [27,35]	K_AY_N_D_AX_V [4,11]
D_OW_N_T [20,28]	Y_UW_Z_AX_K [31,38]
DH_AE_T_S [26,34]	AY_TH_IH_NG_K_DH [11,17]
TH_IH_NG_K [11,18]	G_OW_AX_NG_T_AX [0,10]
AX_B_AW_T [5,12]	B_AX_K_AH_Z [11,18]
Phone N-grams and frame sequences for new units for female speakers	
Y_UW_N_OW [13,19]	P_IY_P_AX_L [6,13]
AY_M_IY_N [11,18]	S_AH_M_TH_AX_NG [17,24]
JH_AX_S_T [25,33]	K_AY_N_D_AX_V [2,11]
DH_AE_T_S [29,37]	AY_TH_IH_NG_K_DH [9,16]
TH_IH_NG_K [2,10]	G_OW_AX_NG_T_AX [2,11]
AX_B_AW_T [7,15]	B_AX_K_AH_Z [11,21]
HH_AE_V_T_AX [24,31]	–

Table 4.2. *High-ranking frame sequences and corresponding phone N-grams of new units for male and female speakers. The start and end frames of the length-normalized phone N-grams from which the new units are selected are denoted in the brackets.*

I computed the frame-level precision and recall values for each of the new units with respect to their actual instances (obtained via inverse-DTW mapping) in the data sets for which the ANN/HMM decoder is trained (*e06tel2001female* and *e06tel1508male*). Table 4.3 shows the results.

Note that for the training utterances, the HMM-labeled classes corresponding to

For *e06tel2001female*		
Phone N-gram	Precision (%)	Recall (%)
AX_B_AW_T [5,12]	10.5	28.6
AY_M_IY_N [10,18]	9.4	43.7
JH_AX_S_T [27,35]	4.8	71.2
P_IY_P_AX_L [22,30]	12.1	80.0
Y_UW_N_OW [16,23]	21.7	67.5
AY_TH_IH_NG_K_DH [11,17]	4.0	22.5
G_OW_AX_NG_T_AX [0,10]	6.2	30.3
K_AY_N_D_AX_V [4,11]	9.8	56.5
S_AH_M_TH_AX_NG [19,27]	10.4	56.7
Y_UW_Z_AX_K [31,38]	30.9	30.8
B_AX_K_AH_Z [11,18]	4.4	23.6
DH_AE_T_S [26,34]	8.4	57.9
D_OW_N_T [20,28]	9.0	32.3
TH_IH_NG_K [11,18]	9.7	61.5
For *e06tel1508male*		
Phone N-gram	Precision (%)	Recall (%)
AY_M_IY_N [11,18]	5.3	21.8
B_AX_K_AH_Z [11,21]	13.3	57.3
DH_AE_T_S [29,37]	8.5	19.3
P_IY_P_AX_L [6,13]	10.8	80.1
Y_UW_N_OW [13,19]	12.3	67.7
AY_TH_IH_NG_K_DH [9,16]	4.4	15.7
G_OW_AX_NG_T_AX [2,11]	4.8	39.5
HH_AE_V_T_AX [24,31]	3.8	75.9
K_AY_N_D_AX_V [2,11]	7.5	74.6
S_AH_M_TH_AX_NG [17,24]	5.1	75.8
AX_B_AW_T [7,15]	4.0	32.7
JH_AX_S_T [25,33]	5.1	45.1
TH_IH_NG_K [2,10]	4.8	42.3

Table 4.3. *Frame-based precision and recall for ANN/HMM decoder of new units for e06tel1508male and e06tel2001female.*

instances of the new units have precision rates ranging from 4.0% to 30.9%, and recall rates ranging from 15.7% to 80.1%. While it is true that the frame-level precision and recall rates are not particularly high, for my purposes, I am only looking for segments of speech that are similar to the units of interest, of which there could be much more than simply the ground-truth segments.

In addition, due to the relatively higher recall rates compared to the precision rates, more of the speech is labeled as one of the new units compared to the originally labeled data. However, the fact that a significant portion of the ground truth is

recalled suggests that even those HMM-labeled classes that do not match the ground truth are in some way similar to the ground truth labels, and may behave similarly in a speaker recognition system. In general, given that only a small fraction of the training data contains instances of the new units, while most of the data contains speech corresponding to the Junk class, it is satisfying to see that significant portions of the ground-truth instances of the new units have been recalled.

4.4 Experiments and Results

To test my data selection framework, I've used the ALIZE [7] implementation of a 512-mixture GMM-UBM system, with MFCC C0-C19 and deltas (with cepstral mean subtraction), and relevance MAP adaptation for the speaker-specific GMM models [41]. While more advanced MAP adaptation approaches, such as factor analysis-based MAP [26], have been developed, I have observed using past systems that such techniques do not significantly improve results on unit-constrained data. This is because factor analysis-based MAP attempts to remove intra-speaker variability across multiple utterances, and much of this variability is due to lexical variability [18]. Hence, I believe that a simple GMM-UBM system using relevance MAP is a valid starting point for testing the data selection framework.

One gender-dependent speaker recognition system is implemented using data from each of the standalone phones, along with the new units. The UBMs are trained using the SRE06 training utterances, while speaker model training and testing are performed using the SRE08 test utterances (refer to section 3.5.1). A common set of roughly 36,000 trials with 6,000 true speaker trials are used. The trial counts are determined by the amounts of data available for all units.

The new speech units are split into 3 groups of 5, 5, and 4 units for males, and 3 groups of 5, 5, and 3 units for females (there is one fewer new unit for female speakers due to the availability of data). The EERs of new units within each group are separately compared to the EERs of the 30 mono-phone units, also belonging to the corresponding group. Hence, the 3 male groups comprise 35, 35, and 34 total units, and the 3 female groups comprise 35, 35, and 33 total units (with the 30 mono-phone units belonging to each of the groups). The groups are created so that the ANN/HMM decoder does not have to decode between too many classes of new units, to ensure better decoding accuracy; it only decodes between the 3, 4, or 5 new units within a particular group, and is run multiple times to decode units from all groups.

Within each group, I've ensured that roughly similar amounts of speech data is used for the new units and the 30 mono-phone units, to minimize the bias in speaker recognition performances due to differing amounts of data usage. This means, however, that different groups will have different average EERs based on the amount of data selected for each group. Typically, more data used leads to lower EER due to better speaker models being trained [29].

EER results for all units and phones in each group are obtained on speaker recog-

nition systems giving the lowest average EER results for all phones and units within a particular group. Due to computational cost and time, not all systems have been run for all groups of units. The GMM-UBM and GMM-SVM systems are implemented on each of the phones, and the number of mixtures for each system are varied from 16, 32, 128, and 512.

I've found that, overall, for these experiments, systems with 16 and 32 mixtures produce the lowest EERs for the unit-constrained data. While typical GMM-based speaker recognition systems may use mixtures of 512, 1024, or 2048 [11], the fact that unit-constrained systems are used (where the data for each system is more homogenous compared to the total amounts of speech data) means that the speaker models do not have to be as expressive in their modeling power. I've also found that, for these experiments, the GMM-UBM system (with the log-likelihood ratio scoring) produces significantly lower EERs on average (∼4% absolute) compared to the GMM-SVM system.

Tables 4.4, 4.5, and 4.6 show the speaker recognition EER results of each new unit, along with its rank amongst all units (new and old) for each group. Recall that there are 33 to 35 units in each group. Results are obtained on *e08tel710male* and *e08tel1108female*, using the GMM-UBM system with varying numbers of mixtures.

Table 4.7 compares results for the units in group 1 for the GMM-UBM and GMM-SVM systems (the configurations of which are set so that the units have similar EER values) on *e06tel1025male* and *e06tel1508female*, and table 4.8 shows the EER results for the units in group 1 for the GMM-SVM system on *e06tel1025male* and *e06tel1508female*. The lower the rank of a unit within a group, the lower its EER amongst all other units in the group. The lower the EER of a unit, the more speaker-discriminative it is.

32-mixture GMM-UBM System for *e08tel710male*		
New units - group 1	EER rank	EER (%)
JH_AX_S_T [27,35]	**1 of 35**	31.4%
Y_UW_N_OW [16,23]	**2 of 35**	31.7%
AX_B_AW_T [5,12]	4 of 35	33.1%
P_IY_P_AX_L [22,30]	6 of 35	33.3%
AY_M_IY_N [10,18]	12 of 35	34.9%
32-mixture GMM-UBM System for *e08tel1108female*		
New units - group 1	EER rank	EER (%)
B_AX_K_AH_Z [11,21]	6 of 35	31.8%
DH_AE_T_S [29,37]	7 of 35	32.1%
P_IY_P_AX_L [6,13]	10 of 35	33.5%
AY_M_IY_N [11,18]	12 of 35	34.5%
Y_UW_N_OW [13,19]	14 of 35	37.7%

Table 4.4. *Speaker recognition results and rankings of new units in group 1. Results obtained on e08tel710male and e08tel1108female.*

512-mixture GMM-UBM System for *e08tel710male*		
New units - group 1	EER rank	EER (%)
S_AH_M_TH_AX_NG [19,27]	**2 of 35**	38.6%
K_AY_N_D_AX_V [4,11] [16,23]	8 of 35	40.4%
Y_UW_Z_AX_K [31,38]	9 of 35	40.5%
AY_TH_IH_NG_K_DH [11,17]	14 of 35	41.6%
G_OW_AX_NG_T_AX [0,10]	29 of 35	44.3%
512-mixture GMM-UBM System for *e08tel1108female*		
New units - group 1	EER rank	EER (%)
AY_TH_IH_NG_K_DH [9,16]	**3 of 35**	31.8%
S_AH_M_TH_AX_NG [17,24]	10 of 35	33.4%
K_AY_N_D_AX_V [2,11]	13 of 35	34.8%
HH_AE_V_T_AX [24,31]	14 of 35	35.2%
G_OW_AX_NG_T_AX [2,11]	18 of 35	36.5%

Table 4.5. *Speaker recognition results and rankings of new units in group 2. Results obtained on e08tel710male and e08tel1108female.*

16-mixture GMM-UBM System for *e08tel710male*		
New units - group 1	EER rank	EER (%)
D_OW_N_T [20,28]	**3 of 34**	30.2%
DH_AE_T_S [26,34]	5 of 34	30.5%
TH_IH_NG_K [11,18]	6 of 34	30.6%
B_AX_K_AH_Z [11,18]	7 of 34	30.6%
16-mixture GMM-UBM System for *e08tel1108female*		
New units - group 1	EER rank	EER (%)
JH_AX_S_T [25,33]	10 of 33	21.5%
AX_B_AW_T [7,15]	11 of 33	23.4%
TH_IH_NG_K [2,10]	14 of 33	25.7%

Table 4.6. *Speaker recognition results and rankings of new units in group 3. Results obtained on e08tel710male and e08tel1108female.*

Results suggest that the data selection approach yields speech units that perform well compared to the existing mono-phone units, especially for male speakers, where several of the new units outperform all existing mono-phone units. The fact that the new units perform well compared to the existing mono-phone units is consistent across the two systems (according to the left and right columns of table 4.7), as well as different configurations of the same system (according to tables 4.7 and 4.8).

Analyzing the results for the new units, consider first the results in tables 4.4, 4.5, and 4.6. For males, 11 of the 14 new units perform amongst the top 26 percent (in relative EER) in their respective unit-groups, while 12 of the 13 new units perform amongst the top 40 percent. 4 of the new

512-mixture GMM-UBM System for *e06tel1025male*		
New units - group 1	EER rank	EER (%)
Y_UW_N_OW [16,23]	**1 of 35**	31.8%
AY_M_IY_N [10,18]	6 of 35	35.6%
JH_AX_S_T [27,35]	15 of 35	36.2%
P_IY_P_AX_L [22,30]	16 of 35	36.6%
AX_B_AW_T [5,12]	21 of 35	38.0%
32-mixture GMM-SVM System for *e06tel1025male*		
New units - group 1	EER rank	EER (%)
Y_UW_N_OW [16,23]	**1 of 35**	33.4%
P_IY_P_AX_L [22,30]	8 of 35	36.5%
AY_M_IY_N [10,18]	11 of 35	37.2%
JH_AX_S_T [27,35]	14 of 35	37.5%
AX_B_AW_T [5,12]	15 of 35	38.6%

Table 4.7. *Speaker recognition results and rankings of new units in group 1. Results obtained on e06tel1025male and e06tel1518female, comparing GMM-UBM and GMM-SVM systems.*

512-mixture GMM-SVM System for *e06tel1025male*		
New units - group 1	EER rank	EER (%)
Y_UW_N_OW [16,23]	**1 of 35**	36.5%
JH_AX_S_T [27,35]	**2 of 35**	40.5%
AX_B_AW_T [5,12]	**3 of 35**	41.1%
AY_M_IY_N [10,18]	6 of 35	41.2%
P_IY_P_AX_L [22,30]	11 of 35	42.2%
512-mixture GMM-SVM System for *e06tel1518female*		
New units - group 1	EER rank	EER (%)
B_AX_K_AH_Z [11,21]	5 of 35	40.5%
DH_AE_T_S [29,37]	8 of 35	41.5%
Y_UW_N_OW [13,19]	9 of 35	41.5%
P_IY_P_AX_L [6,13]	13 of 35	42.4%
AY_M_IY_N [11,18]	14 of 35	42.5%

Table 4.8. *Speaker recognition results and rankings of new units in group 1. Results obtained on e06tel1025male and e06tel1518female, using GMM-SVM system.*

units - D_OW_N_T [20,28], Y_UW_N_OW [16,23], JH_AX_S_T [27,35], and S_AH_M_TH_AX_NG [19,27] - are in the top 3 in terms of having the lowest EER amongst all units in their respective groups. Results for females are slightly worse than for males, but still demonstrate that the new units are in general more speaker discriminative than the existing mono-phone units. In general, no correlations have

been found between the locations of the frame sequences within the phone N-grams, and the EER performances of the frame sequences comprising the new units.

The fact that the absolute EER values themselves are rather high is not surprising, given that each units uses only a small fraction of the overall data, and the amount of data for each unit is further reduced to ensure that all units have similar amounts of data. In certain groups (group 3 in particular), I've been forced to reduce the amount of per-utterance data for each unit to less than half a second, to achieve a balance of data for all units across all utterances.

Overall, however, the results suggest that it is possible to design speech units - comprised of arbitrary feature frame sequences - with good speaker-discriminative power. In fact, several of the new units have the lowest relative EER compared to the existing mono-phone units. The data selection scheme allows one to step beyond the use of traditional speech units - such as phones, words, and syllables - that may or may not be highly speaker-discriminative for unit-based text-dependent speaker recognition. The results also demonstrate the effectiveness of the mutual information and kurtosis measures in predicting the speaker-discriminative power of units.

The new method for selecting units is also faster and more computationally efficient than simply implementing and running a speaker recognition system for each arbitrary frame sequence and determining what the most speaker-discriminative sequences are. Running the GMM-UBM speaker recognition system described above (including UBM training) takes several CPU-hours on our cluster using the same SRE06 and SRE08 data sets. Roughly 2,500 frame sequences have been considered, and running a separate speaker recognition system for each frame sequence would require thousands of CPU hours. In contrast, the data selection approach to unit selection, including the speaker recognition system runs, requires a small fraction of the computation time.

Upon determining the relative standalone-performance of the new units in comparison to the existing units, the new units have been combined according to the relevance and redundancy criterion described in section 3.3 in the previous chapter. Combination is again performed at the score-level using an MLP (with MLP weights trained using *e06tel1518female* and *e06tel1025male*), and results for *e08tel1108female* and *e08tel710male* are obtained.

Because the nasality measures, when combined together, produce high correlations with EER for female speakers (and produce insignificant correlations with EER for male speakers according to some preliminary experiments), I've decided to apply the nasality measures for unit selection in their combination for females but not for males. The nasality measures and mutual information are combined via linear regression using the same regression weights for the *NAS+MI LOO6* measure in section 3.5.7 of the previous chapter. This particular measure - the combination of mutual information and 6 nasality measures obtained via leave-one-out selection - is used because it has been able to select units that produce among the best overall combination results according to tables 3.8, 3.9, and 3.10 in the previous chapter.

The 6 nasality measures used are the following: *ctm01k* Mean, *std01k* Mean, *tef1* Var, *f1-fp0* Var, *fp1-f1* Mean, and *a1-p0* Mean.

However, because the weights are not trained by taking into account the new units, they are slightly ill-suited for the combination of new units, and hence combination results involving both the new units and nasality measures are omitted. To train weights for both the old and the new units would require many more speaker recognition systems to be run on the new units. The standalone mutual information measure is used for both males and females, as it has been determined in the previous chapter to be the single most reliable measure for data selection (and hence unit selection).

All results are obtained using the same optimal α values for the relevance-redundancy selection scheme as used in tables 3.8, 3.9, and 3.10 in the previous chapter. The α values used are the optimal α values corresponding to the combination conditions involving the same set of parameters as in the previous chapter. For males, only the mutual information measure is used, as it illustrates well the main results of the unit-selection approach.

Results are shown in tables 4.9, 4.10, and 4.11 for female data set *e08tel1108female*, and tables 4.12, 4.13, and 4.14 for the male data set *e08tel710male*. For females, table 4.9 is only for units in group 1, table 4.10 is only for units in group 2, and table 4.11 is only for units in group 3. Likewise for males, table 4.12 is only for units in group 1, table 4.13 is only for units in group 2, and table 4.14 is only for units in group 3. For each table, there are unit combination results that include the new units, and results that do not include the new units (these latter results include only the 30 mono-phones). Results that do not take the new units into consideration are marked with an asterisk (*) next to the group number in the first column of each table. Each of the new units that are selected are marked in bold.

Results for female speakers in tables 4.9, 4.10, and 4.11 show that overall, selecting units for combination using the nasality measures in combination with the mutual information measure leads to lower EERs in general. Consider the results using mutual information standalone, along with mutual information in combination with nasal measures, applied to the original mono-phones. For female unit group 1 results in table 4.9, the 5-unit combination result using the nasality measures gives a 8.90% relative EER improvement over the result using only the mutual information measure. A 7.01% relative EER improvement is observed for the 10-unit combinations results, and a 1.13% relative EER improvement is observed for the 15-unit combination results.

For female unit group 2 results in table 4.10, the 5-unit combination result using the nasality measures gives a 10.39% relative EER improvement over the result using only the mutual information measure. A 1.79% relative EER improvement is observed for 10-unit combination results, while for 15-unit combinations, the EERs are equal.

For female unit group 3 results in table 4.11, the 5-unit combination result using the nasality measures gives a 6.78% relative EER improvement over the result

Results for *e08tel1108female* , Unit group 1			
Measure	# units	Units selected	EER(%)
MI	5	**B_AX_K_AH_Z [11,21]**, **DH_AE_T_S [29,37]**, **Y_UW_N_OW [13,19]**, B, HH	***25.6***
MI	5	HH, M, AY, K, V	28.1
NAS+MI LOO6	5	S, HH, D, N, B	25.6
MI	10	**Y_UW_N_OW [13,19]**, **B_AX_K_AH_Z [11,21]**, **AY_M_IY_N [11,18]**, **DH_AE_T_S [29,37]**, **P_IY_P_AX_L [6,13]**, W, B, V, HH, M	***21.2***
MI	10	AH, S, W, K, V, Z, HH, M, D, AY	21.4
NAS+MI LOO6	10	S, N, B, V, Z, HH, M, D, AE, PUH	19.9
MI	15	**Y_UW_N_OW [13,19]**, **AY_M_IY_N [11,18]**, **B_AX_K_AH_Z [11,21]**, **P_IY_P_AX_L [6,13]**, **DH_AE_T_S [29,37]**, S, T, W, P, V, Z, HH, M, DH, D	17.9
MI	15	F, S, W, T, N, P, K, B, V, Z, HH, M, D, AY, DH	***17.7***
NAS+MI LOO6	15	F, S, W, N, B, V, Z, AA, HH, M, D, AE, AY, PUH, DH	**17.5**

Table 4.9. *EER results for combinations of 5, 10, and 15 mono-phones and units in group 1 for e06tel1108female, using the GMM-UBM system with optimal numbers of mixtures for each group of units.*

using only the mutual information measure. A 15.31% relative EER improvement is observed for 10-unit combinations results, and a relative 9.20% EER improvement is observed for the 15-unit combinations. Hence, in general, using nasality measures in conjunction with the mutual information measure allows for the selection of units which combine at the score-level to produce lower EERs than using only the mutual information measure standalone. Note that a 6.6% improvement is obtained for the average of all female EERs using the nasality measures, over the average of all female EERs using mutual information standalone (for only the results involving the 30 mono-phone units).

Results for *e08tel1108female* , Unit group 2			
Measure	# units	Units selected	EER(%)
MI	5	**S_AH_M_TH_AX_NG [17,24]**, **K_AY_N_D_AX_V [2,11]**, AH, Z, HH	***22.1***
MI	5	HH, M, AY, K, V	23.1
NAS+MI LOO6	5	S, HH, N, D, B	20.7
MI	10	**HH_AE_V_T_AX [24,31]**, **K_AY_N_D_AX_V [2,11]**, **AY_TH_IH_NG_K_DH [9,16]**, S, N, V, HH, M, D, AE	17.0
MI	10	S, N, V, Z, HH, M, D, AE, PUH, DH	***16.8***
NAS+MI LOO6	10	S, N, B, V, Z, HH, M, D, AE, DH	16.5
MI	15	**G_OW_AX_NG_T_AX [2,11]**, **HH_AE_V_T_AX [24,31]**, **S_AH_M_TH_AX_NG [17,24]**, **K_AY_N_D_AX_V [2,11]**, **AY_TH_IH_NG_K_DH [9,16]**, D, DH, S, W, T, N, V, HH, M, Z	***13.7***
MI	15	AH, F, S, W, T, N, P, K, V, Z, HH, M, D, AY, DH	***13.7***
NAS+MI LOO6	15	F, S, W, N, B, V, Z, AA, HH, M, D, AE, AY, PUH, DH	**13.7**

Table 4.10. *EER results for combinations of 5, 10, and 15 mono-phones and units in group 2 for e06tel1108female, using the GMM-UBM system with optimal numbers of mixtures for each group of units.*

As mentioned previously, because the nasality measure weights have been trained without the new units, they are better suited for the combination of only the 30 mono-phone units, and hence the results for the combinations involving the new units are omitted here. However, examining the combination results involving only the mutual information measure, it is apparent that the combinations involving the new units are in general either better than the combinations involving only the old units, or not significantly worse.

Consider the results involving the standalone mutual information measure for female speakers in tables 4.9, 4.10, and 4.11. There are 9 comparisons of results involving the old and new units (one for each of the 3 unit groups, and each of the 5, 10, and 15 unit combinations) for female speakers, and a 1.9% relative improvement

Results for *e08tel1108female* , Unit group 3			
Measure	# units	Units selected	EER(%)
MI	5	**JH_AX_S_T [25,33]**, AA, HH, M, B	12.2
MI	5	S, AA, HH, M, V	***11.8***
NAS+MI LOO6	5	S, HH, D, N, B	11.0
MI	10	**JH_AX_S_T [25,33]**, V, Z, HH, AA, M, AY, DH, W, T	***9.5***
MI	10	S, W, P, V, Z, HH, AA, M, D, AY	9.8
NAS+MI LOO6	10	S, N, B, V, Z, AA, HH, M, D, AE	**8.3**
MI	15	**TH_IH_NG_K [2,10]**, **JH_AX_S_T [25,33]**, **AX_B_AW_T [7,15]**, DH, S, W, T, N, V, Z, AA, HH, M, D, P	9.1
MI	15	F, S, W, T, N, K, B, V, Z, Z, AA, HH, M, D, AY, DH	***8.7***
NAS+MI LOO6	15	F, S, N, B, V, Z, AA, HH, M, D, AE, AY, PUH, IY, DH	***7.9***

Table 4.11. *EER results for combinations of 5, 10, and 15 mono-phones and units in group 3 for e06tel1108female, using the GMM-UBM system with optimal numbers of mixtures for each group of units.*

of the EER average of the new units compared to the EER average of the old units is observed. For the 9 such comparisons for male speakers in tables 4.12, 4.13 and 4.14, there is a 4.3% relative improvement of the EER average of the new units compared to the EER average of the old units. In 7 of the 9 comparisons for male speakers, the EERs involving the new units are lower than the EERs involving the old units, and in 5 of the 7, the EERs involving the new units are significantly lower ($>$ 10% relative improvement).

These results demonstrate the value of the new units as well as the nasality and mutual information measures in unit combination. One thing to note is that in combinations involving the new units, several of the new units are always selected for combination via the relevance-redundancy unit selection scheme, most likely because their measure values indicate that they have high relevance to speaker discrimination. This is consistent with the fact that the new units have come to exist because they represent regions of speech with high speaker discriminative power as determined by the measures. One thing that is less clear, however, is the redundancy these units have with other units.

Results for *e08tel710male*, Unit group 1			
Measure	# units	Units selected	EER(%)
MI	5	**JH_AX_S_T [27,35]**, **Y_UW_N_OW [16,23]**, **AX_B_AW_T [5,12]**, HH, K	**23.8**
MI	5	S, D, AY, L, V	27.9
MI	10	**JH_AX_S_T [27,35]**, **Y_UW_N_OW [16,23]**, **AX_B_AW_T [5,12]**, S, K, V, HH, D, L, AY	**19.7**
MI	10	S, UW, N, P, V, HH, D, AY, EH, L	22.5
MI	15	**AY_M_IY_N [10,18]**, **JH_AX_S_T [27,35]**, **P_IY_P_AX_L [22,30]**, **Y_UW_N_OW [16,23]**, **AX_B_AW_T [5,12]**, F, S, T, K, V, HH, AO, D, AY, DH	**17.5**
MI	15	F, S, N, P, K, V, Z, AA, HH, AO, D, PUH, AY, DH, L	18.3

Table 4.12. *EER results for combinations of 5, 10, and 15 mono-phones and units in group 1 for e06tel710male, using the GMM-UBM system with optimal numbers of mixtures for each group of units.*

It may be the case that even though each new unit performs well individually in speaker recognition, they are less effective when combined together, and hence their combination results do not stand out as much from combination results involving only the old units. While Pearson's correlation does provide a measure of redundancy, its correlation with the EER improvement of phone pairs in combination over the EER average of the phones standalone is less than 0.5 in magnitude (discussed in section 3.5.6 of the previous chapter). Until now, I have been unable to find a more effective redundancy measure for the speech units, and this is something that can be explored in the future.

4.5 Summary

The results in this chapter demonstrate the effectiveness of the measures in selecting units with good speaker discriminative power. The approach used to select the units is fast and effective, producing units that perform well standalone compared to existing mono-phone units (recall that for males, 11 of the 14 new units perform

Results for *e08tel710male*, Unit group 2			
Measure	# units	Units selected	EER(%)
MI	5	**AY_TH_IH_NG_K_DH [11,17]**, **S_AH_M_TH_AX_NG [19,27]**, **K_AY_N_D_AX_V [4,11]**, AY, V	32.4
MI	5	S, AY, PUH, DH, V	**29.8**
MI	10	**AY_TH_IH_NG_K_DH [11,17]**, **S_AH_M_TH_AX_NG [19,27]**, **K_AY_N_D_AX_V [4,11]**, **Y_UW_Z_AX_K [31,38]**, EY, V, Z, HH, AY, DH	28.2
MI	10	S, N, K, P, V, HH, D, PUH, AY, L	**25.2**
MI	15	**AY_TH_IH_NG_K_DH [11,17]**, **S_AH_M_TH_AX_NG [19,27]**, **K_AY_N_D_AX_V [4,11]**, **G_OW_AX_NG_T_AX [0,10]**, **Y_UW_Z_AX_K [31,38]**, F, T, K, V, Z, HH, AO, PUH, AY, DH	**23.9**
MI	15	F, S, T, N, K, P, V, Z, HH, AO, M, D, PUH, AY, DH	24.0

Table 4.13. *EER results for combinations of 5, 10, and 15 mono-phones and units in groups 2 for e06tel710male, using the GMM-UBM system with optimal numbers of mixtures for each group of units.*

amongst the top 26 percent in relative EER in their respective unit-groups, while for females, 13 of the 14 new units perform amongst the top 40 percent), with decent performance in combination with existing units. The approach for unit selection is based on the most reliable standalone measures mutual information and kurtosis, which relate feature vector distributions to speaker recognition performances. Not only do the units that have been discovered perform well when used for speaker recognition, but a greater understanding as to why certain units are more speaker-discriminative (i.e. due to their high mutual information and low kurtosis values) is obtained. Overall, this work offers a new approach in determining regions of speech with good speaker discriminative power, and offers unique insights into the problem of speaker recognition.

Future work may involve testing additional relevance and redundancy measures as a basis for selecting new units. Other automatic speech recognition classifiers besides the ANN/HMM can be tested as well to see if precision and recall rates of the new units with respect to the training data can be improved. In addition, other types of units, such as syllables, can be experimented with as well. Because I do not have

Results for *e08tel710male*, Unit group 3			
Measure	# units	Units selected	EER(%)
MI	5	**D_OW_N_T [20,28]**, S, D, AY, V	**20.5**
MI	5	S, D, AY, L, V	23.2
MI	10	**D_OW_N_T [20,28]**, **TH_IH_NG_K [11,18]**, **B_AX_K_AH_Z [11,18]**, S, N, V, HH, D, AY, L	**15.4**
MI	10	F, S, UW, N, V, HH, D, AY, EH, L	17.7
MI	15	**D_OW_N_T [20,28]**, **TH_IH_NG_K [11,18]**, **B_AX_K_AH_Z [11,18]**, S, N, P, V, Z, AA, HH, D, PUH, AY, DH, L	**14.2**
MI	15	F, S, UW, N, K, EY, V, Z, AA, HH, AO, D, AY, DH, L	15.8

Table 4.14. *EER results for combinations of 5, 10, and 15 mono-phones and units in group 3 for e06tel710male, using the GMM-UBM system with optimal numbers of mixtures for each group of units.*

access to an automatic syllabifier, I have not been able to examine the effectiveness of using syllables instead of phones as speech units. While currently, my approach relies on the existence of an automatic speech recognizer to provide phonetic transcripts so that common segments of speech across multiple utterances can be found, future approaches may attempt to locate the common segments automatically. Overall, there are many things that can be explored using the framework I have created.

Chapter 5

Analysis and Interpretation of the New Units

Upon obtaining a set of new units based on phone N-grams, and demonstrating that they perform well in speaker recognition standalone and reasonably well in combination, it is worth analyzing the phonetic composition of the new units to determine characteristics that allow them to perform well. Through this, an even greater understanding of the phonetic composition of speaker discriminative regions of speech can be obtained.

The previous chapter describes the main algorithm for unit selection. Within the algorithm, the inverse-DTW mapping step maps the newly obtained speech unit instances with DTW-normalized start and end frames, to their original, non DTW-normalized locations found in the actual speech data. Refer to figure 4.3 in the previous chapter for an overview of this procedure. Recall that each new unit is a subsequence of a phone N-gram. Hence, the locations of the new unit instances found in the actual speech data can be used to determine the phonetic composition underlying the units (i.e. which phones of the phone N-gram does the subsequence span across?).

For each of the new units, each unit instance in the original training data is analyzed for the durations of each phone that the instance spans across. Next, the durations of phones found in all instances of each new unit are accumulated, and a normalized distribution of the total durations of the individual phones is obtained. The normalized distribution is such that the sum of all components in the distribution is unity. Tables 5.1 and 5.2 show the distributions of the phone durations for each new unit for females and males respectively. Also shown in the tables are the phones with the longest durations in the normalized distributions.

From the tables, it appears that many of the phones with high durations spanned by the new unit frame sequences are also phones that perform well in EER according to table 3.2 in chapter 2, which shows the EER results of each phone on the *e06tel1060female* data set. Such phones include the following: N, P, M, S, K, T, V,

Female speech unit	Mono-phone distribution	Highest duration phone
Y_UW_N_OW [16,23]	[Y, UW, N, OW] = [0.06, 0.27, 0.36, 0.31]	N
P_IY_P_AX_L [22,30]	[P, IY, P, AX, L] = [0.00, 0.02, 0.44, 0.40, 0.14]	P
AY_M_IY_N [10,18]	[AY, M, IY, N] = [0.35, 0.44, 0.19, 0.02]	M
JH_AX_S_T [27,35]	[JH, AX, S, T] = [0.01, 0.12, 0.54, 0.33]	S
AX_B_AW_T [5,12]	[AX, B, AW, T] = [0.54, 0.43, 0.03, 0.00]	AX
S_AH_M_TH_AX_NG [19,27]	[S, AH, M, TH, AX, NG] = [0.01, 0.02, 0.47, 0.33, 0.16, 0.01]	M
K_AY_N_D_AX_V [4,11]	[K, AY, N, D, AX, V] = [0.56, 0.41, 0.03, 0.00, 0.00, 0.00]	K
AY_TH_IH_NG_K_DH [11,17]	[AY, TH, IH, NG, K, DH] = [0.45, 0.50, 0.04, 0.01, 0.00, 0.00]	TH
G_OW_AX_NG_T_AX [0,10]	[G, OW, AX, NG, T, AX] = [0.31, 0.46, 0.18, 0.05, 0.00, 0.00]	OW
Y_UW_Z_AX_K [31,38]	[Y, UW, A, AX, K] = [0.00, 0.00, 0.01, 0.37, 0.62]	K
B_AX_K_AH_Z [11,18]	[B, AX, K, AH, Z] = [0.04, 0.36, 0.40, 0.16, 0.04]	K
D_OW_N_T [20,28]	[D, OW, N, T] = [0.05, 0.50, 0.39, 0.06]	OW
TH_IH_NG_K [11,18]	[TH, IH, NG, K] = [0.56, 0.36, 0.07, 0.01]	TH
DH_AE_T_S [26,34]	[DH, AE, T, S] = [0.00, 0.09, 0.51, 0.40]	T

Table 5.1. *Distribution of mono-phone counts spanned by new unit sequences in new unit training data for females.*

B. Certain phones that did not perform well according to table 3.2 are also amongst the high duration phones: AX, OW, AY. What's interesting about these results is that phones that do not perform well individually are frequently found amongst speech units with good overall speaker discriminative ability.

Hence, the transitions between phones must have a strong influence in the speaker recognition performances of certain new units, and the use of such transitions as speech units would not have been evident simply from the phonetic transcriptions themselves. Adjacent phones with highest durations within a speech unit (according to tables 5.1 and 5.2) are shown in table 5.3.

Male speech unit	Mono-phone distribution	Highest duration phone
Y_UW_N_OW [13,19]	[Y, UW, N, OW] = [0.18 0.38 0.34 0.10]	UW
P_IY_P_AX_L [6,13]	[P, IY, P, AX, L] = [0.55, 0.43, 0.02, 0.00, 0.00]	P
AY_M_IY_N [11,18]	[AY, M, IY, N] = [0.36 0.47 0.15 0.02]	M
B_AX_K_AH_Z [11,21]	[B, AX, K, AH, Z] = [0.06, 0.38, 0.32, 0.22, 0.02]	AX
DH_AE_T_S [29,37]	[DH, AE, T, S] = [0.00, 0.04, 0.47, 0.49]	S
S_AH_M_TH_AX_NG [17,24]	[S, AH, M, TH, AX, NG] = [0.01, 0.06, 0.49, 0.41, 0.02, 0.01]	M
K_AY_N_D_AX_V [2,11]	[K, AY, N, D, AX, V] = [0.65, 0.34, 0.01, 0.00, 0.00, 0.00]	K
AY_TH_IH_NG_K_DH [9,16]	[AY, TH, IH, NG, K, DH] = [0.56, 0.42, 0.02, 0.00, 0.00, 0.00]	AY
G_OW_AX_NG_T_AX [2,11]	[G, OW, AX, NG, T, AX] = [0.46, 0.39, 0.13, 0.02, 0.00, 0.00]	G
HH_AE_V_T_AX [24,31]	[HH, AE, V, T, AX] = [0.00, 0.01, 0.54, 0.41, 0.04]	V
JH_AX_S_T [25,33]	[JH, AX, S, T] = [0.01, 0.14, 0.55, 0.30]	S
TH_IH_NG_K [2,10]	[TH, IH, NG, K] = [0.88, 0.11, 0.01, 0.00]	TH
AX_B_AW_T [7,15]	[AX, B, AW, T] = [0.23, 0.52, 0.25, 0.00]	B

Table 5.2. *Distribution of mono-phone counts spanned by new unit sequences in new unit training data for males.*

According to table 5.3, 10 of the 27 phone pairs consist of a consonant followed by a vowel, 10 of the 27 phone pairs consist of a vowel followed by a consonant, while 7 of the 27 phones pair consist of consonants only. 7 of the 27 phone pairs contain nasals, suggesting that nasals are a significant component of units that are speaker discriminative. While the standalone phone results in table 3.2 of chapter 2 suggest that consonants may perform better in speaker recognition, the frame sequences underlying the new speech units contain mixtures of vowels, nasal-, and non-nasal consonants. These results further suggest that it may be difficult to intuitively determine which phones or combinations of phones can produce good speaker recognition results. Hence, the framework developed and demonstrated for selecting speaker dis-

Female speech unit	Highest duration phone pairs
Y_UW_N_OW [16,23]	[N, OW]
P_IY_P_AX_L [22,30]	[P, AX]
AY_M_IY_N [10,18]	[AY, M]
JH_AX_S_T [27,35]	[S, T]
AX_B_AW_T [5,12]	[AX, B]
S_AH_M_TH_AX_NG [19,27]	[M, TH]
K_AY_N_D_AX_V [4,11]	[K, AY]
AY_TH_IH_NG_K_DH [11,17]	[AY, TH]
G_OW_AX_NG_T_AX [0,10]	[G, OW]
Y_UW_Z_AX_K [31,38]	[AX, K]
B_AX_K_AH_Z [11,18]	[AX, K]
D_OW_N_T [20,28]	[OW, N]
TH_IH_NG_K [11,18]	[TH, IH]
DH_AE_T_S [26,34]	[T, S]
Male speech unit	**Highest duration phone pairs**
Y_UW_N_OW [13,19]	[UW, N]
P_IY_P_AX_L [6,13]	[P, IY]
AY_M_IY_N [11,18]	[AY, M]
B_AX_K_AH_Z [11,21]	[AX, K]
DH_AE_T_S [29,37]	[T, S]
S_AH_M_TH_AX_NG [17,24]	[M, TH]
K_AY_N_D_AX_V [2,11]	[K, AY]
AY_TH_IH_NG_K_DH [9,16]	[AY, TH]
G_OW_AX_NG_T_AX [2,11]	[G, OW]
HH_AE_V_T_AX [24,31]	[V, T]
JH_AX_S_T [25,33]	[S, T]
TH_IH_NG_K [2,10]	[TH, IH]
AX_B_AW_T [7,15]	[B, AW]

Table 5.3. *Phone pairs with highest duration counts within new unit sequences with*

criminative regions of speech is all the more valuable in determining what units have good speaker discriminative power.

Chapter 6

Applications of the Developed Techniques

One main application of unit-based speaker recognition is that of biometric access control, where a user must speak a certain password to gain access to something. It may be the case that certain things a speaker says may be able to better distinguish that speaker from other speakers, and the measure-based techniques that select speaker discriminative regions of speech can be used to determine what a speaker should say so as to not be confusable with other speakers.

While the speaker recognition task is one application of the techniques that I have developed, the techniques can be used to improve other speech-processing tasks as well. In a sense, there are always benefits to being able to determine useful regions of speech data in any speech-processing task that involves training of models and classification. For instance, the measure-based approachs can be applied to the closely-related task of language recognition, where the task is to recognize the languages of speech utterances. It is conceivable that certain regions of speech may have more language-distinctive features, and the data-selection approaches can be used to identify such regions. For the task of automatic speech recognition (ASR), where different speakers exist in a given speech utterance, it may be the case that certain speakers may utter words that are easier to recognize. The measures can be used to identify such speakers. For instance, different words that a speaker speaks can be used as the class labels, and the mutual information between the speech feature vectors and class labels can be computed. It is hence possible to obtain measure values identifying which speakers are easier to perform speech recognition for, and perhaps use this information to improve the acoustic modeling for ASR.

A more direct application of the measure-based approaches is the identification of usable speech for acoustic model training - whether speaker, language, or ASR models. For instance, the measures can be applied in combination with a speech/non-speech detector to detect not only the non-silence regions, but also regions that are useful to the task. This is especially useful for data that are recorded in the presence of real

environmental noise, where some of the data may be too corrupted by the noise to be useful for speech processing.

In the speaker recognition experiments below, data from the Robust Open-Set Speaker Identification (ROSSI) data set, which consists of landline and cell-phone speech data recorded in various real environmental conditions, is used. The conditions include office, public place, vehicle, and roadside. This data set consists of 5 noisy conditions, each with 100 training speakers (with one utterance per speaker) 200 testing speakers, 100 of which are within the set of training speakers. Matching every training speaker against every test speaker produces a set of 20,000 trials, with 100 true speaker trials, for each condition. The 5 conditions are shown in table 6.1.

	Training condition	Testing condition
1	Cell-phone public place	Cell-phone public place
2	Cell-phone public place	Cell-phone vehicle
3	Landline office	Cell-phone office
4	Landline office	Cell-phone vehicle
5	Cell-phone roadside	Landline office

Table 6.1. ROSSI data set conditions.

The speaker recognition system used is a GMM-UBM system with 128 Gaussian mixtures, and the feature vectors are MFCC C0-C12 with deltas and double deltas. The MFCCs are computed on 25 ms speech frames, advancing every 10 ms. To detect usable speech for acoustic model training and testing, the kurtosis measure is implemented across every speech utterance to get a sense of which regions in the utterance may hold good speaker discriminative power. The reason kurtosis is used is that it is one of the easiest measures to implement, and nicely illustrates the usefulness of the measures.

To apply the kurtosis measure for data selection, a set of kurtosis values at different points along each utterance that indicate regions of speech to keep and regions of speech to discard must first be obtained. The kurtosis measure is first computed on feature vectors from 30 feature frames, advancing every five frames. Hence, each kurtosis value is computed across a 300 ms time span, shifting every 50 ms. Note that the kurtosis measure is computed for each feature vector dimension separately, and the final kurtosis value is averaged across all feature dimensions.

Having labeled the entire utterance with kurtosis values, parts of the utterance whose kurtosis values lie in the top 25% among all kurtosis values in that utterance are discarded. The 25% is rather arbitrarily chosen, but is based on the fact that speech data is in general leptokurtic (having a kurtosis value in excess of 0), which degrades speaker recognition performance [48]. Hence, lower kurtosis values are preferred. Regions of speech selected using the kurtosis measure are combined with those selected by the Shout speech/non-speech detector [2], which discriminates between regions of speech and silence. Table 6.2 shows the EER results on the ROSSI data set before and after applying the kurtosis measure for data selection.

Condition	EER (%) w/o kurtosis	EER w/ kurtosis (%)
1	10.0	9.0
2	11.0	10.0
3	10.1	9.1
4	12.1	12.1
5	10.1	10.1

Table 6.2. Results on ROSSI data set with and without kurtosis-based data selection.

Results show that applying the kurtosis measure for data selection on ROSSI data with real environmental conditions with the intent of keeping only the more speaker discriminative regions of speech improves speaker recognition performance in 3 of the 5 conditions for the ROSSI data set. The relative EER improvements for conditions 1, 2, and 3 are 10.0%, 9.1%, and 9.9%. Conditions for which the kurtosis-based data selection approach did not improve speaker recognition are conditions using landline office data for training and cell-phone vehicle data for testing, as well as cell-phone roadside data for training and landline office data for testing. Nevertheless, data selection using the kurtosis measure did not make the results worse in these last two conditions.

It is interesting that in general, results involving cell-phone speech data in environmental conditions such as the office, public place, and vehicle, can all be improved by a simple data selection procedure using the kurtosis measure. This is my first crack at implementing this data selection procedure in terms of selecting the parameters (i.e. discarding the data with kurtosis values amongst top 25% per utterance), and by choosing better parameters, the results can possibly be improved even further. Moreover, kurtosis is only one potential measure that can be implemented to select speech data. Implementing the nasality features for data selection can conceivably improve speaker recognition results even more.

6.1 Summary

Overall, various possible applications of the measure-based approaches have been suggested, and the successfulness of applying one of the measures for simple data selection on cell-phone speech data with real environmental noise has been demonstrated. There are potentially other applications that have not been conceived of, and future work can be geared towards improving various other speech processing tasks using the measure-based approaches.

Chapter 7

Conclusion and Future Work

This work provides a new framework for data selection for speaker recognition, based on a set of measures that indicate regions of speech of high relevance to the speaker recognition task. The first component of this work tests a set of measures, which include mutual information, kurtosis, KL-distance, f-ratio, intra- and inter-speaker variability, 11 nasality features, unit duration, and unit frequency. Each of these measures is computed on speech data constrained by a particular speech unit, and a correlation between the values of each measure and the speaker recognition EER of each unit is obtained. The measures are then ranked according to the magnitude of their correlations with the EERs, showing that the best standalone measures are mutual information and kurtosis. In addition, the combination of a set of nasality measures also produces high correlations between their values and EERs of speech units, and these correlations are higher than those obtained from the mutual information and kurtosis measures standalone.

The best standalone measures - mutual information and kurtosis - are then applied to select regions of speech comprising of arbitrary frame sequences to form new speech units that may be highly speaker discriminative. The measures need to be computed on matching regions of speech across all conversation sides in order to determine which frame sequences should be selected, however, and transcriptions of phone N-grams are used for this purpose. Specifically, long phone N-gram sequences that are commonly found amongst all speakers and conversation sides are chosen, and the measures are computed on frame sequences within each (length-normalized) phone N-gram across all speakers and conversation sides. Frame sequences with high-scoring measure values are chosen as new speech units.

Once the identities of the new speech units are determined, the ANN/HMM automatic speech recognition approach is applied to train models for each of the new units, and to decode new conversation sides for these new units. Once all conversation sides are decoded for these new units, a speaker recognition system is run on each new unit to determine how well the new units perform relative to the existing mono-phone speech units. Results show that most of the new units perform well above average compared to the existing units in terms of speaker recognition accuracy.

In order to combine the units to improve the overall unit-based speaker recognition performance, a unit-selection scheme that incorporates relevance and redundancy measures to select units with high relevance to the speaker recognition task, but with low redundancy with one another, is used. The top relevance measures consist of mutual information, kurtosis, and the nasality measures, while the redundancy measure consists of Pearson's correlation. These results show that the use of the nasality measures in combination with the mutual information measure as the overall relevance measure allows one to select speech units that combine together to produce the overall best combination results. In addition, the incorporation of the new speech units in unit combination improves upon results using only the existing mono-phone units.

Overall, the new data selection approach, in conjunction with the unit selection scheme, allows for the saving of much computation time over that of actually running a speaker recognition system on each frame sequence to determine which sequences are more speaker discriminative, and which units to use in combination. The data selection approach also allows one to discover new units that perform well in speaker recognition; the previously existing speech units are obtained simply by choosing a set of speech units from automatic speech transcriptions, with little regard to the actual speaker discriminative capabilities of the chosen units. The new approach, however, takes into account the speaker discriminative power of the units when selecting the new units.

This work also sheds new light on what makes speaker recognition systems perform well, starting with the speech feature vectors themselves. Having developed tools and techniques for selecting regions of speech with high speaker discriminative power, these techniques can then be applied to all kinds of speech data for various other speech tasks, such as automatic speech recognition, language recognition, and speaker diarization. Future work can involve actually applying the techniques for these tasks, and also attempting to improve upon the techniques by investigating other measures, testing other keyword-spotting techniques, as well as other techniques that improve upon the general data selection framework.

Bibliography

[1] "Hmm toolkit (htk)," *http://htk.eng.cam.ac.uk*.

[2] "Shout," *http://wwwhome.ewi.utwente.nl/ huijbreg/shout/index.html*.

[3] K. Amino, T. Sugawara, and T. Arai, "Idiosyncrasy of nasal sounds in human speaker identification and their acoustic properties," *Acoustic Science and Technology*, vol. 27, no. 4, 2006.

[4] W. Andrews, M. Kohler, and J. Campbell, "Phonetic speaker recognition," *Proceedings of Eurospeech*, pp. 149–153, 2001.

[5] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[6] K. Boakye, "Speaker recognition in the text-independent domain using keyword hidden markov models," *Masters Report, University of California at Berkeley*, 2005.

[7] J. F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," *Proceedings of ICASSP*, 2005.

[8] H. Bourlard and N. Morgan, "Connectionist speech recognition - a hybrid approach," *Kluwer Academic Press*, 1994.

[9] D. A. Cairns, J. H. Hansen, and J. F. Kaiser, "Recent advances in hypernasal speech detection using the nonlinear teager energy operator," *Proceedings of ICSLP*, pp. 780–783, 1996.

[10] J. Campbell and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," *Proceedings of ICASSP*, 1999.

[11] W. D. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using gmm supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, pp. 308–311, 2006.

[12] W. D. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "Svm based speaker verification using a gmm supervector kernel and nap variability compensation," *Proceedings of ICASSP*, 2006.

[13] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: A resource for the next generations of speech-to-text," *Proceedings of 4th International Conference on Language Resources and Evaluation*, pp. 69–71, 2004.

[14] S. Davis and P. Mermelstein, "Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences," *Proceedings of ICASSP*, 1980.

[15] M. Do, "Fast approximation of kullback-leibler distance for dependence trees and hidden markov models," *IEEE Signal Processing Letters*, vol. 10, no. 4, 2003.

[16] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," *Proceedings of Eurospeech*, pp. 2521–2524, 2001.

[17] D. Ellis and J. Bilmes, "Using mutual information to design feature combinations," *Proceedings of ICSLP*, 2000.

[18] D. Garcia-Romero and C. Y. Espy-Wilson, "Intersession variability in speaker recognition: A behind the scene analysis," *Proceedings of Interspeech*, 2008.

[19] M. Gerber, R. Beutler, and B. Pfisher, "Quasi text-independent speaker-verification based on pattern matching," *Proceedings of Interspeech*, 2007.

[20] J. R. Glass and V. W. Zue, "Detection of nasalized vowels in american english," *Proceedings of ICASSP*, 1985.

[21] A. Hannani, D. Toledano, D. Petrovska-Delacrétaz, A. Montero-Asenjo, and J. Hennebert, "Using data-driven and phonetic units for speaker verification," *Proceedings of IEEE Speaker Odyssey*, 2006.

[22] A. Hatch, B. Peskin, and A. Stolcke, "Improved phonetic speaker recognition using lattice decoding," *Proceedings of ICASSP*, vol. 1, pp. 169–172, 2005.

[23] H. Hermansky, "An efficient speaker-independent automatic speech recognition by simulation of some properties of human auditory processing," *Proceedings of ICASSP*, 1987.

[24] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, and R. R. Gadde, "Speaker recognition using prosodic and lexical features," *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 19–24, 2003.

[25] S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, and A. Stolcke, "2008 nist speaker recognition evaluation: Sri system description," 2008.

[26] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," *Proceedings of Speaker Odyssey - Speaker and Language Recognition Workshop*, 2004.

[27] S. Kishore and A. Black, "Unit size in unit selection speech synthesis," *Proceedings of Eurospeech*, 2003.

[28] N. Kwak and C. Choi, "Input feature selection by mutual information based on parzen window," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, 2002.

[29] H. Lei, "Applications of keyword-constraining in speaker recognition," *Master's report, UC Berkeley*, 2007.

[30] ——, "Nap, wccn, a new linear kernel, and keyword weighting for the hmm supervector speaker recognition system," *Technical report, International Computer Sciences Institute*, 2008.

[31] H. Lei and N. Mirghafori, "Comparisons of recent speaker recognition approaches based on word-conditioning," *Proceedings of Speaker Odyssey*, 2008.

[32] A. Martin, D. Miller, M. Przybocki, J. Campbell, and H. Nakasone, "Conversational telephone speech corpus collection for the nist speaker recognition evaluation 2004," *Proceedings of 4th International Conference on Language Resources and Evaluation*, pp. 587–590, 2004.

[33] A. Nagórski, L. Boves, and H. Steeneken, "In search of optimal data selection for training of automatic speech recognition systems," *Proceedings of ASRU*, 2003.

[34] M. Omar and M. Hasegawa-Johnson, "Maximum mutual information based acoustic-features representation of phonological features for speech recognition," *Proceedings of ICASSP*, 2002.

[35] E. Parzen, "On estimation of a probability density function and mode," *Annals of Math. Statistics*, vol. 33, 1962.

[36] J. Pelacanos and S. Sridharan, "Feature warping for robust speaker verification," *Proceedings of Speaker Odyssey*, 2001.

[37] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, 2005.

[38] T. Pruthi and C. Y. Espy-Wilson, "Acoustic parameters for the automatic detection of vowel nasalization," *Proceedings of Interspeech*, 2007.

[39] D. Reynolds, "Experimental evaluations of features for robust speaker identification," *Proceedings of IEEE Trans. Speech and Audio Processing*, vol. 2, 1994.

[40] D. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[41] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, pp. 19–41, 2000.

[42] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Proceedings of Speech Communication*, 2005.

[43] A. Stolcke, H. Bratth, J. Butzberger, H. Franco, V. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, "The sri march 2000 hub-5 conversational speech transcription system," *NIST Speech Transcription Workshop*, 2000.

[44] D. Sturim, D. Reynolds, R. Dunn, and T. Quatieri, "Speaker verification using text-constrained gaussian mixture models," *Proceedings of ICASSP*, 2002.

[45] V. Vapnik, "The nature of statistical learning theory," *Springer*, 1999.

[46] R. Vogt, S. Kajarekar, and S. Sridharan, "Discriminant nap for svm speaker recognition," *Proceedings of IEEE Speaker Odyssey*, 2008.

[47] M. Welling, "Fisher linear discriminant analysis," *Department of Computer Science, University of Toronto.*

[48] Y. Xie, B. Dai, Z. Yao, and M. Liu, "Kurtosis normalization in feature space for robust speaker verification," *Proceedings of ICASSP*, vol. 1, 2006.