ACOUSTIC-FEATURE-BASED FREQUENCY WARPING FOR SPEAKER NORMALIZATION

Evandro B. Gouvêa

Department of Electrical
and Computer Engineering
Carnegie Mellon University

Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Electrical and Computer Engineering

Pittsburgh, Pennsylvania
December 1998

Table of Contents

List of Tables

List of Figures

Abstract

Speaker-dependent automatic speech recognition systems are known to outperform speaker-independent systems when enough data are available for training to overcome the variability of acoustical properties among speakers. Speaker normalization techniques modify the spectral representation of incoming speech waveforms in an attempt to reduce variability between speakers.

While a number of recent successful speaker normalization algorithms have incorporated speaker-specific frequency warping to the initial signal processing, these algorithms do not make extensive use of acoustic features contained in the incoming speech.

In this work we study the possible benefits of the use of acoustic features that are believed to be key to speech perception in speaker normalization algorithms using frequency warping. We study the extent to which the use of such features, including specifically the first three formant frequencies, can improve recognition accuracy and reduce computational complexity for speaker normalization compared to conventional techniques. We examine the characteristics and limitations of several types of feature sets and warping functions as we compare to their performance relative to that of existing algorithms.

We have found that the specific shape of the warping function appears to be irrelevant in terms of improvement in recognition accuracy. The use of a linear function, the simplest choice, allowed us to employ linear regression to define which features to use and how to weigh them. We present a method that finds the optimal set of weights for a set of speakers given the slope of the best warping function. Selection of a limited subset of features for use is a special case of this method where the weights are restricted to one or zero.

The application of our speaker normalization algorithm on the ARPA Resource Management task resulted in sizable improvements compared to previous techniques. Speaker normalization applied to the ARPA Wall Street Journal (WSJ) and Broadcast News (Hub 4) tasks resulted in more modest improvements. We have investigated the possible causes of this. Our experiments indicate that

normalization is less effective with a larger number of speakers presumably because in this case the output probability densities of HMMs tend to be broader and hence representative of a large class of speakers. In addition to this, increasing the vocabulary size tends to increase the search space, causing correct hypotheses to be replaced by errorful ones. The benefit brought about by normalization is thus diluted.

The amount of improvement provided by normalization also increases with increasing sentence duration in Hub 4. Since the actual Hub 4 contains a large number of short segments, the normalization provides a more limited improvement in performance.

Acknowledgments

I am grateful to the members of my Thesis committee, José Moura, Jordan Cohen, Raj Reddy, and Richard Stern. Dr. Jordan Cohen has especially provided me with valuable comments that highly improved my work.

I am especially indebted to my Thesis advisor, Dr. Richard Stern. During my almost five years at CMU, I often felt lost and unmotivated, and Rich has always been willing to help me come to terms with myself. He followed closely my work when I needed it, and gave me freedom to work when I wanted. His ethics, seriousness, and dedication to work have been inspirational. I owe a lot of my growth to Rich.

I am also grateful for the attention given me by the administrative staff at CMU during all these years, especially Elaine, Lynn, Carol, and Debbie.

From the robust group, I particularly owe a lot to Pedro Moreno and Bhiksha Raj. Pedro was the main creator and the soul behind Robust. His innate leadership and drive to work have made him a great role model. Bhiksha has always been willing to listen to my problems, professional or otherwise, and has always been a source of fruitful comments and discussions. Throw an idea at him, and he will get so excited that he will suggest zillions of experiments to test your idea. I owe most of what I know and of what I have become as a researcher in speech recognition to Pedro and Bhiksha. My work in speaker normalization started inspired by them.

Almost 5 years of life in Pittsburgh have given me a chance of building several friendships. I wish to thank the support given me during these years by Tom, Juan, Sam-Joo, Matt, Uday, Vipul, Daniel, Rita, Irina, Eric, and Ravi, and the rest of the speech group, and Greg, P. D., Miriam, Rosana, Emi, Ivan, Rosinha, Radu, Takumi, and the rest of the “Brazilian Mafia” for the friendship. Marcel, Clara, and Rafael have been especially important. My life in Pittsburgh has been made much more comfortable because they were here.

Thanks to the folks at the 61C Café for patiently coping with my thesis writing process. The many hours I spent there have always been pleasant thanks to them. The friendly environment made me always feel as if I were home.

A very special person I have to mention is Mirna Jusufbegovic. It has been extremely gratifying to have Mirna as a friend. It was Mirna who allowed me to see most of the good things I can think of myself today.

My family has also been an important support despite the distance. My parents, Oéliton and Maura, and my siblings, Patrícia and Maurício, have always been supportive about problems I have had. I am grateful for their love and affection and the great sense of family they have provided me.

I would also like to thank the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) for their financial support during my Ph.D. Thesis. I would not have a thesis today of I had not had their support.

Chapter 1
Introduction

We present a method of speaker normalization accomplished by a transformation of the frequency axis based on speaker specific acoustic features. Speaker normalization attempts to increase speech recognition accuracy by reducing speaker variability.

Speaker dependent systems are known to outperform speaker independent ones when enough training data are available. Speaker variability seems to account for much of this difference. Thus, reducing this variability can decrease this difference in performance.

Speaker variability appears for different reasons. We can point out extrinsic reasons, which have to do with external influences like a speaker's cultural background, emotional state, etc., and intrinsic reasons, which have to do with physiological differences between speakers such as differences in shapes and sizes of components of the vocal tract.

A clear effect of the differences in vocal tract components shows up in the locations of formants. Formants are the peaks of the envelope of the spectrum. Physically, they correspond to resonances of the vocal tract. Previous work [19] has shown that there is a correlation between the position of the formants and vocal tract length. Therefore, speaker normalization through a vocal tract length-dependent transformation of the frequency axis should yield formants which have less variability. Furthermore, for some phonemes this dependency between formants and vocal tract length seemed to be linear, suggesting a scaling of the frequency axis would be a good solution as a normalization technique. For some phonemes though the relation does not seem to be linear, suggesting that other transformations or "warping" of the frequency axis should be attempted.

Previous attempts in vocal tract length normalization [14][40][73] have mostly relied upon a "trial and error" approach: the algorithm tests several "solutions" for the warping function and picks the one producing the best results. Although these techniques have been motivated by the desire to compensate for differences in the vocal tract length, there is no clear relation between the warping

function chosen for a speaker and this speaker's acoustic characteristics.

Other work has relied upon speaker specific acoustic features, but with constraints placed on the shape of the warping function taking into account acoustic production arguments. The choice of warping function is limited by the assumed model. not by speaker specific features.

In this work we examine the feasibility of speaker normalization by a mapping directly between a new speaker's and a standard speaker's formants. In general, we separately consider two aspects of the warping function that maps one speaker to another: its shape and, given a shape, the parameters that define it. In our work, we select points which will define the warping function, and find a curve that will fit those points according to any chosen criterion. This framework allows us complete freedom in the choice of the shape of the warping function, as well as in the acoustic features which will define the warping function.

We study some of the choices for the selection of points that define the warping function, as well as the interpolation curve. We study the benefits brought about by each feature separately, and propose a method to determine their relative importance when used conjunctively. These methods are direct applications of extensions to the linear regression technique.

When fast adaptation is needed and the computational cost of the estimation of formants is unacceptable, it is worthwhile to categorize speakers into classes. We examine the simple case where speakers are classified according to their gender.

Our normalization technique is sensitive to errors in the estimation of formants. Ideally, these should be estimated from a large amount of speech recorded over close-talking microphones in quiet environments. Situations arising from deviations of this paradigm make the estimation of formants less reliable, and might bias features extracted from these estimates thus causing a degradation on the performance of the normalization technique. Estimation of formants is beyond the scope of this thesis. However we study some conditions where we know the formant estimation

will be less than ideal. We perform experiments where estimation of formants is performed on speech bandlimited by a telephone channel. In addition, we study cases where we limit the amount of speech from which we estimate formants and attempt to assess the effect of looser constraints in databases, namely databases with larger vocabularies and databases including spontaneous speech.

This thesis is organized as follows. In Chapter 2 we review basic concepts relevant to this work, including formants and their estimation, and we review speaker normalization techniques which employ warping functions. In Chapter 3 we present a brief overview of Sphinx-3, the recognition engine used throughout this thesis, and a description of the databases we employed. In Chapter 4, we present our speaker normalization algorithm, together with important topics related to it. Among these topics, we point out a study of possible shapes of warping function, a study of the criteria for selection and weighing of features that define the warping function, a comparison between gender generic normalized models and gender specific models, and a study of the effect in performance of telephone-like channel. In Chapter 5 we present comparisons of the normalization technique applied to different databases and attempt to assess reasons for differences in performance. Among the reasons, we study the influence of a limited amount of data and the influence of database size, both in terms of vocabulary size and number of speakers. Finally, in Chapter 6 we our present conclusions and suggestions for future work.

Chapter 2
Review of Basic Concepts and Speaker Normalization and Adaptation Techniques

2.1. Introduction

Speech from different speakers will sound differently due to factors that can be classified as both “extrinsic” and “intrinsic” [19]. Extrinsic factors are those that were learned by the individual, based on primarily cultural or emotional factors. They depend on cultural characteristics such as the environment where one grew up and the level of education attained, and emotional factors such as speech rate, and how angry or happy or nervous one is.

Intrinsic factors, on the other hand, are anatomically related. Those reasons would exist no matter what the particular individual circumstances would be. These factors are a consequence of variations in size and shape of the components of the vocal tract. These variations lead to changes in the resonance characteristics of the vocal tract. Formants, which are the peaks of the spectral envelop, are related to the resonances of the vocal tract, and will therefore be affected by the intrinsic factors.

Speech recognition is basically a pattern classification problem. Given a speech observation vector, we try to identify the phonemes occurring in the observation, or at least we try to identify the class of phonemes to which the observation belongs. This is an intrinsically difficult problem. The variability of spectral shapes, even for the same phoneme, makes the classification problem more difficult for two reasons. First, the variability for the same phoneme makes this class more spread out; second, since different phonemes uttered by different speakers typically have similar formants, the classes tend to overlap.

Speaker adaptation and normalization techniques attempt to reduce the effects of speaker variability on the performance of speech recognition systems. The techniques to cope with this variability, of course, depend on the technology being employed.

Most state-of-the-art speech recognition systems [12][28][70] make use of Hidden Markov Models (HMMs)[7] as a convenient statistical representation of speech. One way to account for the effects of speaker variability is to modify these models. “Model transformations” [22][41] are attempts to map output distributions of HMMs to a new set of distributions so as to make them a better statistical representation of a new speaker.

The other most common way to cope with the problem of speaker variability, which is the focus of this thesis, is to modify features of the incoming waveforms in either the spectral or cepstral domain. These changes aim at mapping spectra (or cepstra) so as to diminish interspeaker differences, so that a phoneme uttered by different speakers will have a characterization as close as possible to a standard.

Regardless of whether parameters of the internal HMMs or the incoming features are being modified, the mapping can be thought of as mapping within a space, for example, the space of Gaussian mixtures, in the case of HMMs. Speaker adaptation and normalization techniques can be characterized by what the space in which the mapping takes place, and how this mapping occurs. For example, approaches for mapping between distributions of HMMs or between spectra or cepstra differ.

The mapping of spectra is usually achieved by a warping of the frequency axis. As mentioned previously, differences in vocal tract dimensions will cause differences in spectra, even when the speakers are producing a sound that we perceive as the same phoneme. The warping of the frequency axis is an attempt to modify spectra so that the distance between the spectra of sounds perceived as the same phoneme is smaller. “Distance” might have several meanings. It could simply be the average squared difference, or it could be the average distance between formant frequencies. The definition of distance may yield different optimization criteria on the choice of the warping function.

A natural issue is the choice of the warping function, which specifies the relationship between the frequency axis used to represent spectra produced by the “standard” speaker and the frequency axis describing the spectral productions of a new speaker.

Methods that perform mapping of output distributions of HMMs are very flexible, and they can provide for compensation not only for speaker variability, but also for different environmental conditions. They are usually very computationally expensive, though, which renders them unsuitable for when there is a need for rapid adaptation.

Warping functions are not as likely to be useful in dealing with environmental effects, since the most common environmental degradations are typically modeled as a linear filter followed by additive noise applied to the original “clean” speech signal. In general, the types of compensation approaches that have been the most useful for environmental degradation are not likely to be helpful in counteracting the effects of frequency warping (and vice-versa) because of the very different nature of the types of distortions introduced.

In the following, we present a review of the concept of formants and of the difficulties involved in their estimation, followed by a review of techniques to perform speaker normalization of parameters derived from the incoming speech waveforms. We conclude the chapter with some remarks regarding the effects of these speaker normalization on formants.

2.2. Formant frequencies

A well accepted model of speech production characterizes the vocal tract as a tube or concatenation of tubes of varying cross sectional area [61]. The most energy will be transferred between the excitation and the output at the resonant frequencies of this resonance tube model. These resonant frequencies are the so called formant frequencies of the speech signal. As they correspond to high energy regions of the spectrum, the formant frequencies can be identified by the peaks of the spectral envelope.

Rabiner and Juang [61] point out that although vowels have very low relevance for recognition of written text, most practical speech recognition systems rely heavily on vowel recognition for high performance. If we remove vowels from a text, the average reader can easily figure out what the original text was. If we remove the consonants, the task is a lot more difficult to accomplish. On the other hand, when it comes to acoustic signals, vowels are more easily and reliably recognized because they are produced by exciting an essentially static vocal tract shape with a quasi-periodic excitation signal. They are usually long in duration and spectrally well defined.

From a pattern recognition point of view, it becomes relevant whether vowels' formants can be easily clustered. Peterson and Barney [58] carefully measured formants of vowel utterances from several speakers, and found out that measurements of formants of different vowels tend to cluster, although with some overlap between clusters. The centroids of these clusters can be represented in the so called "vowel triangle", which is depicted on Figure 1.

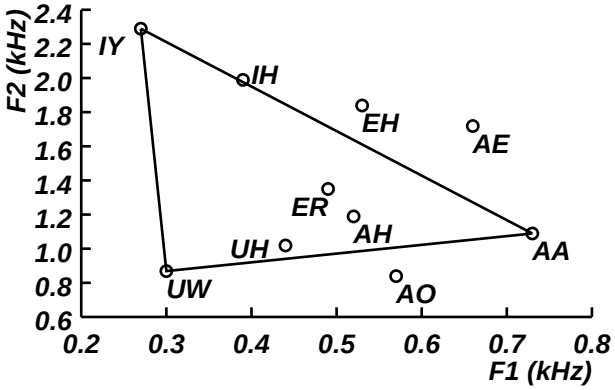

Figure 1. Centroids of clusters of common vowels defining the "vowel triangle".

2.3. Formant frequency estimation

Although the definition of formants is straightforward and the physical concept of formants is easy to understand, the estimation of formants given an utterance can be problematic. Estimation on a frame-by-frame basis is errorful. Moreover, formants are related to the position of the components of the vocal tract, which determine the characteristics of the vocal tract as a resonance cavity. Changes of position cannot occur suddenly because of mechanical limitations of the vocal tract. This imposes a continuity constraint on the formant track.

The estimation of formants on a frame-by-frame basis is usually accomplished by finding some smoothed approximation of the spectrum, for example, by LPC or cepstral analysis, and then locating the peaks on this envelope. Among the difficulties one might encounter when estimating formants on a frame-by-frame basis [56], we can point out:

- Spurious peaks: the process of finding the spectrum might lead to spurious peaks. For example, if one uses LPC analysis to find the spectrum envelope, the order of analysis is usually bigger than necessary to provide for flexibility. These extra poles might yield more peaks than the real ones.
- Blending: two peaks might be so close together that they actually get merged into one during the peak-finding process.
- High-pitched speech: Makhoul [46] has shown that in this kind of speech LPC envelope peaks tend to be drawn away from their true values towards the harmonic peaks, which are more widely spaced in this case.

For each frame, formant trackers find a smoothed approximation of the spectrum and locate peaks on this approximation. The first N peaks are assigned to the first N formants, thus producing N formant tracks. These tracks however are not necessarily smooth due to errors in the combination of the spectrum estimation and the peak finding process.

One possible way to cope with errors in the peak finding process is to use a tracking algorithm, in which one predicts a range of possible values for the estimates, and rejects values out of that range. This technique is prone to errors itself, because bad judgements will propagate. Besides, it assumes that all transitions will be smooth, and this is not always the case with speech.

A better approach would be to post-process the formant tracks, so that obvious outliers can be eliminated. The simplest idea is a moving average, where a weighted average is computed on the values of a sliding window. This has the effect of smoothing the sequence of estimates, maybe more than we want. The transition is important in some processes in speech, and smoothing is not desirable in these circumstances.

A possible alternative is the use of a median smoothing, where the median of values in a sliding window is computed instead of an average. This has the benefit of preserving most of the real discontinuities while being less affected by the presence of outliers.

A more sophisticated way to cope with outliers employs dynamic programming. The basic idea is that a cost is associated with every transition, and the path which leads to the minimum cost is chosen as the formant track. The solution is provided though a Viterbi search algorithm.

Formant tracking is beyond the scope of this thesis. In our experiments, we employ Waves+ [18], which finds candidates for formants by finding roots of the LPC polynomial and then performing formant track smoothing by dynamic programming. This state of the art formant tracker provides reliable estimates of formants.

2.4. Review of speaker normalization techniques

2.4.1 Normalization using warping functions

Warping functions are an attempt at mapping between spectra of two speakers so as to decrease systematic variations. Therefore, a warping function in the context of speaker normalization is a function mapping two spectra. If we intend to map the spectrum $X(\omega)$ to the spectrum $Y(\omega)$ we can use a function $f(\omega)$ so that

$$\hat{Y}(\omega) = X(f(\omega)) \quad (1)$$

The effect of the function $f(\omega)$ will be an expansion or compression of the spectrum $X(\omega)$ depending on whether the first derivative of f, $f'(\omega)$, is bigger or smaller than the unit. Figure 2 illustrates these effects. Notice that the general shape of the original spectrum does not

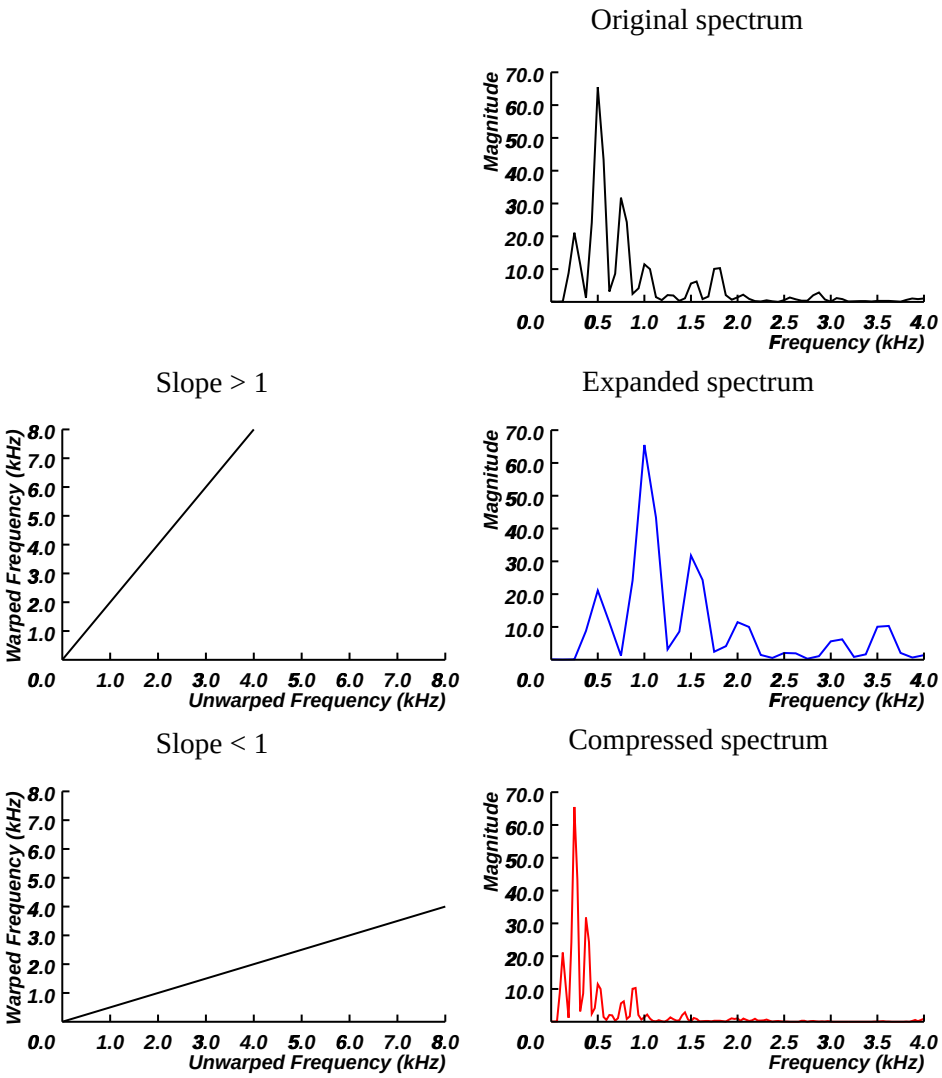

Figure 2. Effect of a warping function over the spectrum. The warping function causes an expansion or compression of the spectrum depending on whether its slope is bigger or smaller than unit. In this example, a linear warping function has been used.

change.

The important issues when examining normalization techniques are the choice of warping function and how it is selected. A very popular choice of warping function, appealing for its simplicity, is the linear function. Other common choices are curves based on speech perception studies, such as the bilinear transform or transforms based on the mel scale, along with curves based on speech production models. The selection of warping function is sometimes based on maximization of likelihood, and sometimes based directly on speaker specific parameters. In the following sections, we present some of the relevant normalization techniques that have been investigated.

2.4.2 Selection based on maximization of likelihood

In a seminal study, Cohen *et al*. [14] proposed a frequency-warping scheme in which the warping function for a given speaker is iteratively chosen as the one which maximizes the likelihood of the hypothesis transcription at the output of the decoder. This criterion takes into account everything that is involved in the process of recognizing speech, and guarantees that the selected warping function will be optimal for the decoder being employed.

In the training phase, the training set is split in two sets, training (T) and aligning (A). HMMs are created with T, and the best warping function is chosen for every element of A using these models. Then, with utterances in A warped according to this selection, T and A are interchanged, and new models are created. This process (training, selection, interchanging) is iterated until convergence is achieved, that is, until the warping function chosen for each speaker remains the same. In the testing phase, each sentence is warped using all warping functions, and decoded. The output with highest likelihood is chosen as the decoded output. While the warping function used by Cohen *et al.* was a linear function, *i.e.*, a linear scaling of the frequency axis, the general approach could in principle be applied to warping functions of any shape.

A significant disadvantage of this general approach is that running a decoder multiple times is extremely computationally costly. A number of similar studies followed that of Cohen *et al.*, and most of them proposed faster alternatives to selection based on decoder output. For example, Wegmann *et al*. [73] proposed an algorithm where the warping function for a particular speaker is chosen from a set of warping functions, based on maximization of likelihoods associated with a Gaussian mixture model that statistically represents the standard speaker. The shape of the warping functions proposed by Wegmann *et al.* corresponds to two straight lines whose point of intersection varies. Specifically, they define a straight line going through the origin and another straight line going through the point defined by the Nyquist frequency in both axes. (In their experiments the Nyquist frequency was 4 kHz.) The first line has a slope between 0.88 and 1.12, to account for variations in vocal tract length around 12%. The two lines intercept at 3500 Hz (see Figure 3).

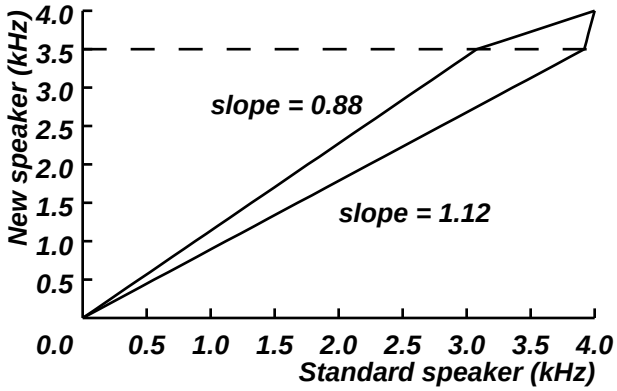

Figure 3. Examples of warping functions used by Wegmann *et al*. [73]. They are defined by two segments of straight lines that meet at 3.5 kHz The first one starts at the origin with slope between 0.88 and 1.12. The second connects the endpoint of the first with the point defined by the Nyquist frequency in both axes.

We implemented a version of the algorithm of Wegmann *et al*. [26], but we used a warping curve derived from the mel scale. The mel scale is motivated by psychophysical studies that show that human perception of frequency content does not follow a linear scale. A reference frequency is chosen, and the subjective pitch values are obtained as the ones perceived as half or twice the perceived pitch of the reference. The original Mel scale [70] was developed to represent the spectral resolution of the auditory system as a function of frequency. In our implementation, we approximated it by a curve that is linear up to a certain frequency and logarithmic afterwards. Varying this frequency, we obtain different warping functions (see Figure 4).

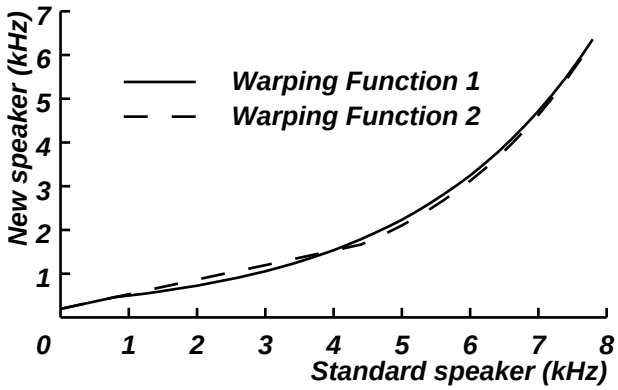

Figure 4. Examples of warping functions used in our implementation [26] of Wegmann *et al*.'s algorithm. They are variations of a curve approximating the mel scale. This approximation is linear up to a frequency, and logarithmic afterwards. Varying this frequency, we obtain different warping curves.

The first step in our implementation consists of an iterative derivation of the Gaussian mixture

model. At each step in this iteration, we first compute the Gaussian mixture model for the prototype speaker and then find the optimal warping function for each speaker. This optimal warping function is chosen to be the one that maximizes the a posteriori log-likelihood computed using this Gaussian mixture model. After choosing the optimal warping function for all speakers in the set, we re-compute the Gaussian mixture model, find new optimal warping functions based on the new Gaussian mixture model, and proceed in this manner until convergence is achieved. Convergence in this case means that for consecutive iterations the same optimal warping function is chosen for every speaker.

The second step of the implementation concerns the derivation of speaker-normalized HMMs from the optimally-warped training set. The training steps of our implementation consist of the iterative derivation of the Gaussian mixture model and the derivation of the HMMs. During recognition, the best warping function is obtained for each new speaker in the same manner that is derived for the utterances in the training set and the warped utterances are recognized using the previously-derived HMMs.

This scheme does not presume any particular shape of warping function. In fact, we also implemented this normalization method using linear warping functions, as opposed to the piecewise linear functions used by Wegmann *et al*. or the nonlinear function of our implementation.

Lee and Rose [40] propose a method similar to both methods mentioned above. The training phase is exactly the same as Cohen and his colleagues' work. A linear warping function is iteratively selected for every speaker in the training set, according to a maximization of likelihood at the output of the decoder. The test phase, on the other hand, resembles Wegmann's approach. Instead of using only one Gaussian mixture to represent the standard speaker, though, they create *N* Gaussian mixture models for a set of *N* warping functions. They create the Gaussian mixture model i using the unwarped utterances for which the best warping function is the i-th. This approach precludes the need to warp utterances in the test set in order to select the warping function.

An alternative to the previous selection criteria, which nonetheless involves a large computational load, was proposed by Fukada and Sagisaka [20]. They accomplish a time-varying warping of the frequency axis. Other algorithms define a speaker specific warping function which remains constant throughout the utterance. In their proposed method, the warping function is chosen not at front end level, but at training and decoding level. Successful state of the art speech recognition systems model speech use Hidden Markov Models. An utterance can be modeled as a path on a two-dimensional space defined by the time or frame index on one axis and by the HMM state on the other. Fukada and Sagisaka define a three dimensional space, where the third dimension varies with the warping function, or, more precisely, with a parameter defining the warping function. This parameter could be, for example, the slope of a linear curve, or, as in their paper, the α parameter of a bilinear transform.

The decoding, usually performed with the Viterbi algorithm, employs a modified Viterbi search, which takes into account transition probabilities on the warping function parameter dimension. This has the advantage over the previous methods that only one pass of recognition is enough to simultaneously find the best recognition hypothesis and to find the best warping function, or, more appropriately, the best sequence of warping functions.

Among the drawbacks of these methods is the requirement for a finite number of warping functions. Moreover, we need to compute features (such as cepstra) for all speakers, for all warping functions, and all these data need to be available at the time the warping functions for each speaker are selected. This means that if, for example, one out of 10 warping functions is to be chosen, we will need to compute the features 10 times.

2.4.3 Selection based on speaker-specific acoustic parameters

Eide and Gish [17] described a study in which measures of formants are used directly to perform speaker normalization. They devised an algorithm for normalization where the spectra of waveforms would undergo warping. The warping function chosen by Eide and Gish was:

$$f' = k_s^{\frac{3f}{8000}} f \tag{2}$$

where k_s is the median of the speaker's third formant (F3) over a set of frames, divided by the median of F3 taken across the set of training speakers. Figure 5 shows this function for two different values of k_s.

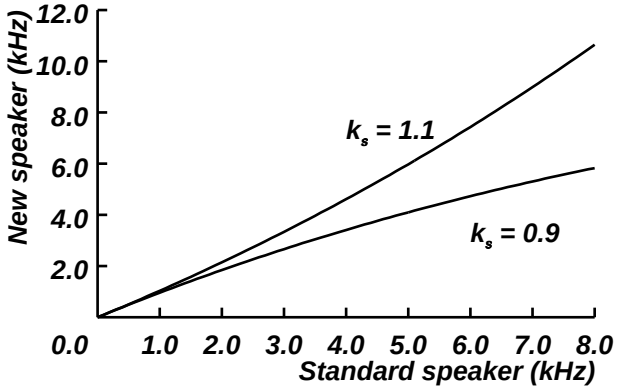

Figure 5. Warping functions used by Eide and Gish [17]. The constant k_s is defined as the ratio between the median of the speaker's third formant and the median of the third formant across all speakers.

This choice of warping function was motivated by resonance tube models of speech production [19]. Tube dimensions and configurations would differ for each phoneme being produced, and therefore the dependence of resonance frequencies on tube lengths would change. Imagine that all dimensions in a model are changed by a factor of k. So, for example, the length goes from l to kl. For some phonemes, this change would produce a change by a factor of $\frac{1}{k}$ in the resonance frequencies. For other phonemes, with different models, this change would cause a change by a factor of $\frac{1}{\sqrt{k}}$ in resonance frequencies. Eide and Gish chose a function that represents a compromise among all dependencies.

Eide and Gish do not elaborate on why they chose the third formant rather than some other feature. However, Claes *et al.* [13] point out that the third formant is the one with least variation across different phonemes. Therefore, its estimation is statistically more robust.

This same point is raised by Zhan and Westphal [77]. They extend Eide and Gish's work and perform normalization using the same kind of warping function but estimating the parameter k_s in equation (2) also from the first and second formants. Moreover, they compare this particular shape of function with linear functions obtained with the same parameter. Their results show that for a nonlinear warping function the first and third formant are equally helpful, although for the linear case the first formant seems to be most helpful. However, they conclude that a maximum likelihood approach, based on an exhaustive search over a grid of values, brings about the best performance.

Lin and Che [43] described a rather different approach. Using an articulatory speech synthesizer, they observed that changes in vocal tract length have the same effect on the computed values of cepstra as the computation of cepstra using just a limited range of the spectrum. They came to this conclusion by synthesizing the same utterance with different vocal tract lengths, computing cepstra, and then comparing these cepstra to the ones computed from "truncated" spectra. Lin and Che proposed a method in which, instead of computing cepstra from a DFT using the whole frequency range from 0 to the Nyquist frequency, they use just the range from 0 to a lower end frequency F_e. This corresponds to a linear scaling of the frequency axis. The end frequency assigned to a particular speaker is presumed to depend on the vocal tract length. Lin and Che suggest that if the vocal tract length is l, then $lF_e = \text{constant}$. If the vocal tract length is not available, they proposed that several different end frequencies be tried in parallel, and that the speaker-specific end frequency that yields the greatest likelihood be chosen.

Lincoln *et al.* [44] came up with a different approach that resembles the work developed in this thesis in some ways. They used the TIMIT database, which is segmented and labeled according to the beginning and ending of each phonemic segment. They used the time-labeled phonemes in the database to develop an algorithm that takes advantage of classes of phonemes. Use of a non-segmented database would require an expensive preprocessing to perform automatic segmentation.

In their method, the normalization of the training data works in two stages. In the first stage, estimates for the first and second formants (F1 and F2) are made for each frame of each vowel segment in the data. They assume a Gaussian distribution for each formant for each vowel class and estimate its parameters. In the second stage, which is also the only stage of the normalization of the test data, the estimates of the two formants for each frame, together with the estimated distributions, are used to compute a normalization factor a_i for each frame:

$$a_i = \arg\max_a \Pr(a f_i^1 \,|\, F^1) \Pr(a f_i^2 \,|\, F^2) \tag{3}$$

where f_i^1 and f_i^2 are the estimates of the first and second formant for frame i, and F^1 and F^2 are the Gaussian distributions for F1 and F2, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. The closed form solution for a_i is:

$$a_i = \frac{f_i^1 \mu_1 / \sigma_1^2 + f_i^2 \mu_2 / \sigma_2^2}{(f_i^1 / \sigma_1)^2 + (f_i^2 / \sigma_2)^2} \tag{4}$$

These frame based normalization factors are then combined into an overall normalization factor $a(I)$ for each speaker I:

$$a(I) = \frac{\sum_{i=1}^{N_v} a_i \Pr(a_i f_i^1 \,|\, F^1) \Pr(a_i f_i^2 \,|\, F^2)}{\sum_{i=1}^{N_v} \Pr(a_i f_i^1 \,|\, F^1) \Pr(a_i f_i^2 \,|\, F^2)} \tag{5}$$

where N_v is the number of vowel frames for speaker I.

It is interesting to notice that the solution provided by Equation 4 is exactly the solution obtained when fitting the points defined by the means of F1 and F2 with a linear function using the least squares formulation presented in Appendix A.

2.4.4 Warping functions in the cepstral domain

Motivated by previous works that used a bilinear transform to mimic the nonlinear characteristics

of human speech perception, Acero [1] proposed a method for warping in the cepstral domain using a bilinear transform. The bilinear transform is a mapping in the complex plane which maps the unit circle onto itself through the relation:

$$z_{new}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, -1 < \alpha < 1 \quad (6)$$

In the general case, taking the above equation into account, we can relate the unwarped and the warped cepstra by:

$$c_w = L(\alpha)c \quad (7)$$

where $L(\alpha)$ is a matrix which depends on α. It can be shown that if α is small and we neglect higher powers of α, the relation above can be expressed as:

$$c_w[n] = -(n-1)\alpha c[n-1] + c[n] + (n+1)\alpha c[n+1] \quad (8)$$

Acero selected the α by minimizing the overall VQ distortion of the cepstral vector.

Zahorian and Jagharghi [75] presented a technique for normalization in the cepstral domain which is an extension of the idea of warping function to this domain. They proposed the use of a polynomial of the form

$$c' = \alpha c^2 + \beta c + \gamma \quad (9)$$

as the warping function, where c corresponds to each cepstral coefficient before warping, and c', after warping. The coefficients were chosen to minimize the squared error between the normalized cepstra and a target cepstra over an utterance. The target was chosen as the speaker closest to the average of all speakers. The concept of the standard speaker as a single real speaker who is closest to the average is an interesting alternative to just averaging.

2.5. Effects of normalization on formants

In this section, we present some results showing the expected formant-related effects caused by normalization techniques. These results were obtained using a linear warping function, as detailed in [25]. Results presented in this section are supposed to be considered typical among normaliza-

tion techniques, not a special case of the cited implementation.

2.5.1 Effects related to gender

Females usually have vocal tracts with smaller dimension than males. Therefore, a female's formants are usually higher than a male's formants. We would expect that a reasonable normalization would cause a female's formants to get compressed in the process, while a male's formants are expected to be expanded, so that they come closer together. Compression, as mentioned previously, can be achieved by a warping function that is a straight line with a slope less than one, and expansion, with a slope greater than one.

Figure 6 shows histograms of slopes of a set of speaker-specific linear warping function separated by gender. We note that the clusters separate well by genders, and that the dependence of slope on gender is as expected.

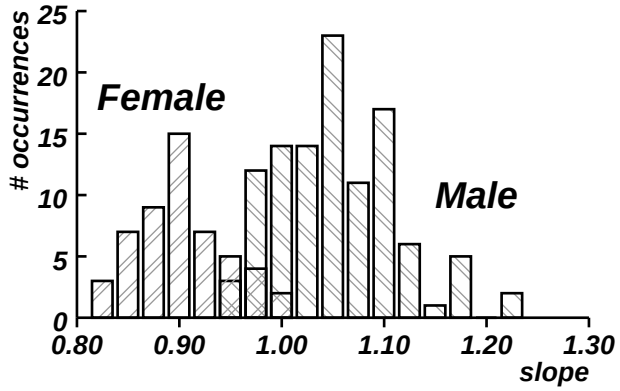

Figure 6. Distribution of slopes according to gender. These are the slopes of a linear warping function obtained according to Gouvêa and Stern [25].

2.5.2 Effects on clusters of vowels

The recognition process is ultimately a pattern recognition problem. Given an observation, we aim to label it as belonging to a particular class, which could be a phoneme, word, etc. We expect a normalization algorithm to modify the spatial distribution of observations so that clusters result more easily identifiable. Figure 7 represents a replotting of the data collected by Peterson and Barney [58], who collected vowels uttered by around sixty speakers, each speaker repeating each of ten vowels twice, and estimated formants from these data. Figure 8 represents the same data after

performing normalization. For clarity's sake, we plotted just three of those vowels in figures 7 and 8. These three vowels are the vowels that define the "vowel triangle" proposed by Peterson and Barney.

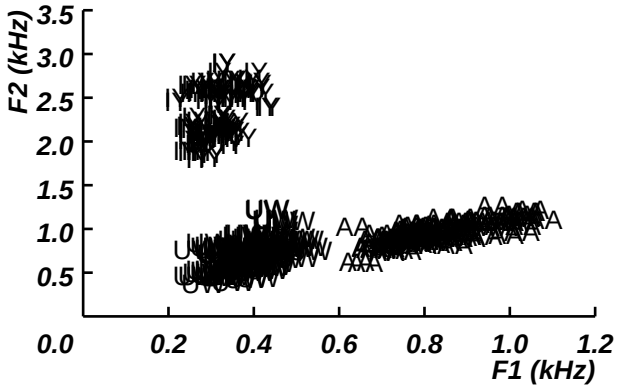

Figure 7. Formants of the phonemes /AA/, /IY/, and /UW/ before normalization. Data collected by Peterson and Barney.

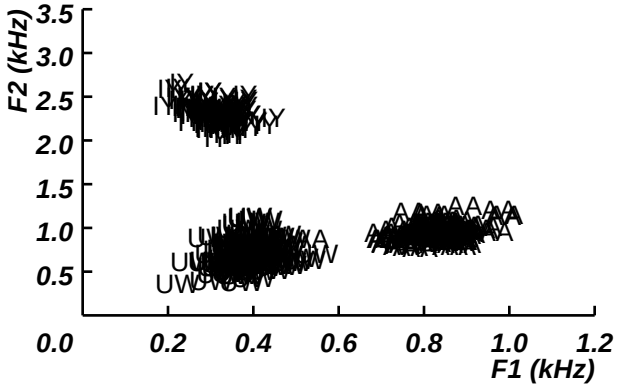

Figure 8. Formants of the phonemes /AA/, /IY/, and /UW/ after normalization. Data collected by Peterson and Barney.

The Fisher ratio [16], which is the ratio of the variance of the means divided by the means of the variances across several different clusters, is a commonly-used measure of cluster separation relative to intrinsic cluster variability. For the data replotted in Figures 7 and 8, the Fisher ratio of the clusters before normalization is 12; after normalization the Fisher ratio increases to 29. Therefore, from a pattern recognition perspective, the transformation is making the classes more compact and farther apart, which would provide greater classification accuracy.

2.6. Speaker adaptation with maximum likelihood linear regression

The focus of this work is on speaker normalization. However, we present here a brief review of one

common speaker adaptation technique as an example of an alternative approach to reducing speaker variability.

Speaker adaptation is usually understood as the modification of output distributions of HMMs so as to better match a speaker's characteristics. A very common technique is the Maximum Likelihood Linear Regression (MLLR) [41].

MLLR attempts to modify mean vectors of the output distributions so as to maximize the likelihood of the observed data from a new speaker, that is, the mean vector of the output probability distribution of the adapted models are given by:

$$\mu_{adapted} = A\mu_{base} + B \tag{10}$$

where A is an $n \times n$ matrix, B, $\mu_{adapted}$, and μ_{base} are n dimensional vectors, and A and B are found through a maximum likelihood approach.

No restriction is imposed in the matrix A or the vector B, thus rendering this technique very flexible. In fact, although originally designed as a speaker adaptation technique, MLLR has been successfully used to perform compensation to new environments, replacing many noise and channel compensation algorithms.

As MLLR involves reestimation of means of output pdfs of HMMs, the estimation of the entities A and B in Equation (10) requires estimation of the likelihoods of the adaptation sentences according to the current HMMs. Therefore, there is a need for transcription of the adaptation sentences. If we have the exact transcriptions for some sentences uttered by a speaker, we can use these sentences to perform supervised adaptation. A more realistic scenario, though, is one where transcriptions for the adaptation sentences are not available. In this case, transcriptions produced by a decoder are an alternative for the real perfect transcriptions.

2.7. Conclusions

We presented a review of basic concepts relevant to this work, as well as a review of major approaches to speaker normalization, with representative examples of their implementation. With the exception of the studies of Eide and Gish and of Lincoln *et al*., the methods reviewed do not make use of acoustic features that are relevant to speech perception such as formant frequencies. It is our belief that the use of acoustic features should be central in the selection of the warping function, for human perception of speech uses exactly these clues to perform normalization.

Chapter 3
The SPHINX-3 Recognition System

3.1. Introduction

Since the normalization algorithms to be developed will be evaluated in the context of continuous speech recognition, this chapter provides an overview of the basic structure of the recognition system used for the experiments described in this thesis. The algorithms developed in this thesis are independent of the recognition engine used, and in fact they are part of the acoustic preprocessing of speech. Hence, the results and conclusions of this thesis should be applicable to other recognition systems.

The most important topic of this chapter is a description of various aspects of the SPHINX-3 recognition system. We also summarize the databases used for evaluation in the thesis.

3.2. An overview of the SPHINX-3 system

SPHINX-3 is a large-vocabulary, speaker-independent, Hidden Markov Model (HMM)-based continuous speech recognition system like its predecessors, the original SPHINX system and SPHINX-II. SPHINX was developed at CMU in 1988 [39][37] and was one of the first systems to demonstrate the feasibility of accurate, speaker-independent, large-vocabulary continuous speech recognition. SPHINX-II [28] was one of the first systems to employ semi-continuous HMMs.

SPHINX-3 has a more flexible structure than SPHINX-II. The user can determine if continuous or semi-continuous models will be used, as well as how many streams of data the system will be using and how these streams are organized. In our work we always make use of continuous models. Therefore, the following description, especially regarding the front end, concentrates on SPHINX-3 running on continuous mode. In semi-continuous mode, SPHINX-3 is fully compatible with SPHINX-II.

Figure 9 shows the fundamental structure of the SPHINX-3 [60] system. We briefly describe the

functions of each block.

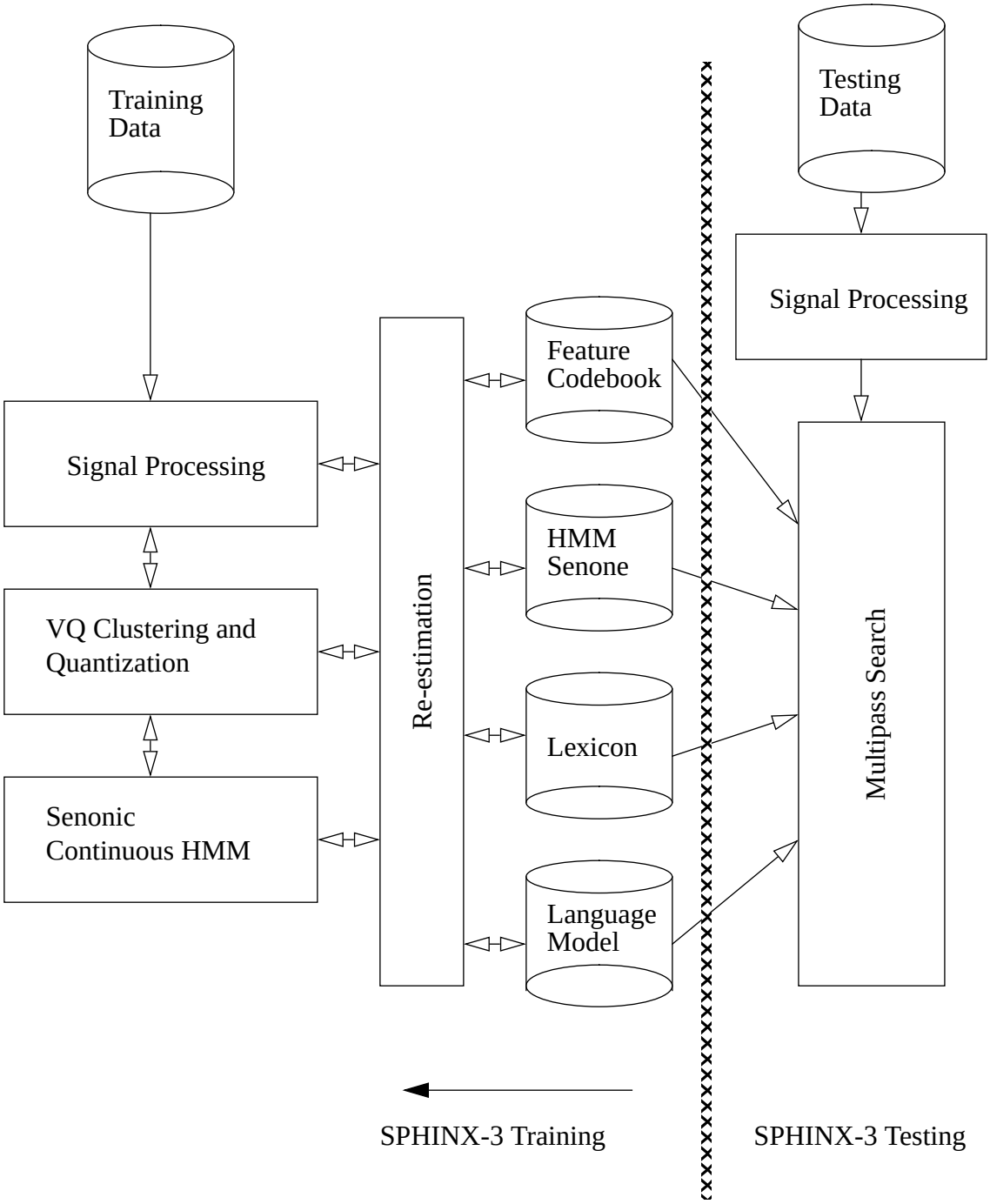

Figure 9. Block diagram of SPHINX-3.

3.2.1 Signal processing

Almost all speech recognition systems use a parametric representation of speech rather than the waveform itself as the basis for pattern recognition. The parameters usually carry the information about the short-time spectrum of the signal. SPHINX-3 uses mel-frequency cepstral coefficients (MFCC) as static features for speech recognition [15]. First-order and second-order time derivatives of the cepstral coefficients are then obtained, and power information is included as a fourth feature.

The front end of SPHINX-3 is illustrated in Figure 10. We summarize this feature extraction procedure as follows:

1) The input speech signal is digitized at a sampling rate of 16 kHz.
2) A pre-emphasis filter $H(z) = 1 - 0.97z^{-1}$ is applied to the speech samples. The pre-emphasis filter is used to reduce the effects of the glottal pulses and radiation impedance [47] and to focus on the spectral properties of the vocal tract.
3) Hamming windows of 25.6-ms duration are applied to the pre-emphasized speech samples at an analysis rate (frame rate) of 100 windows/sec.
4) The power spectrum of the windowed signal in each frame is computed using a 512-point DFT.
5) 40 mel-frequency spectral coefficients (MFSC) [15] are derived based on mel-frequency bandpass filters using 13 constant-bandwidth filters from 100 Hz to 1 kHz and 27 constant-Q filters from 1 kHz to 7 kHz.
6) For each 10-ms time frame, 13 mel-frequency cepstral coefficients (MFCCs) are computed using the cosine transform, as shown in Equation (11)

$$\boldsymbol{x}_t[k] = \sum_{i=0}^{39} X_t[i] \cos[k(i + 1/2)(\pi/40)] \qquad 0 \le k \le 12 \tag{11}$$

where $X_t[i]$ represents the log-energy output of the i^{th} mel-frequency bandpass filter at time frame t and $\boldsymbol{x}_t[k]$ represents the k^{th} cepstral vector component at time frame t. Note that unlike other speech recognition systems (*e.g.* [74]), the $\boldsymbol{x}_t[0]$ cepstrum coefficient here

is the sum of the log spectral band energies as opposed to the logarithm of the sum of the spectral band energies. The relationship between the cepstrum vector and the log spectrum vector can be expressed in matrix form as

$$\begin{bmatrix} \boldsymbol{x}_t[0] \\ \dots \\ \boldsymbol{x}_t[k] \\ \dots \\ \boldsymbol{x}_t[12] \end{bmatrix} = \left[d_{k,\,i} \right] \begin{bmatrix} X_t[0] \\ \dots \\ X_t[i] \\ \dots \\ X_t[39] \end{bmatrix} \qquad (12)$$

$$d_{k,\,i} = \cos[k(i + 1/2)(\pi/40)]$$

where $\left[d_{k,\,i} \right]$ is a 13x40 dimensional matrix.

7) The derivative features are computed from the static MFCCs as follows,

(a) Differenced cepstral vectors consist of 40-ms differences with 12 coefficients

$$\Delta \boldsymbol{x}_t[k] = \boldsymbol{x}_{t+2}[k] - \boldsymbol{x}_{t-2}[k], 1 \le k \le 12 \qquad (13)$$

(b) Second-order differenced MFCCs are then derived in similar fashion, with 12 dimensions.

$$\Delta\Delta \boldsymbol{x}_t[k] = \Delta \boldsymbol{x}_{t+1}[k] - \Delta \boldsymbol{x}_{t-1}[k], 1 \le k \le 12 \qquad (14)$$

(c) Power features consist of normalized power, differenced power and second-order differenced power.

$$\begin{aligned} \bar{\boldsymbol{x}}_t[0] &= \boldsymbol{x}_t[0] - \max\{\boldsymbol{x}_i[0]\} \\ \Delta \boldsymbol{x}_t[0] &= \boldsymbol{x}_{t+2}[0] - \boldsymbol{x}_{t-2}[0] \\ \Delta\Delta \boldsymbol{x}_t[0] &= \Delta \boldsymbol{x}_{t+1}[0] - \Delta \boldsymbol{x}_{t-1}[0] \end{aligned} \qquad (15)$$

In SPHINX-3, the speech representation uses only one stream of features including: (1) 12 Mel-frequency cepstral coefficients (MFCC); (2) 12 40-ms differenced MFCC; (3) 12 second-order differenced cepstral vectors; and (4) power, 40-ms differenced power, and second-order differenced power. These features are all assumed to be statistically independent for mathematical and implementational simplicity.

The normalization technique presented in this Thesis adds one step to the algorithm described above. Warping of the frequency axis is performed between steps 4 and 5.

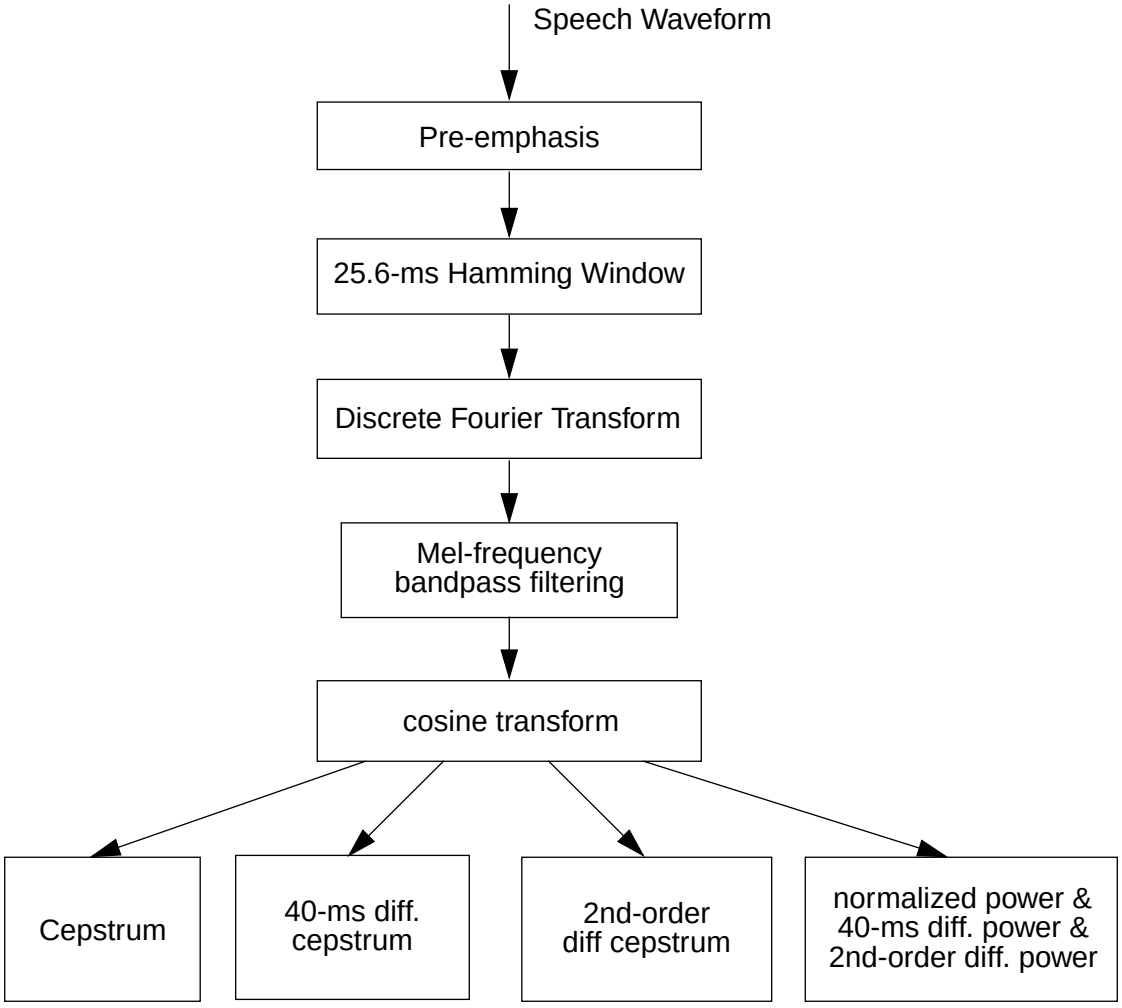

Figure 10. Block diagram of SPHINX-3's front end.

3.2.2 Hidden Markov Models

In the context of statistical methods for speech recognition, hidden Markov models (HMM) have become a well known and widely used statistical approach to characterizing the spectral properties of frames of speech. As a stochastic modeling tool, HMMs have an advantage of providing a natural and highly reliable way of recognizing speech for a wide variety of applications. Since the HMM also integrates well into systems incorporating information about both acoustics and syntax, it is currently the predominant approach for speech recognition. We present here a brief summary of the fundamentals of HMMs. More details about the fundamentals of HMMs can be found in [7][30][37][42][61].

Hidden Markov models are a “doubly stochastic process” in which the observed data are viewed as the result of having passed the true (hidden) process through a function that produces the second process (observed). The hidden process consists of a collection of states (which are presumed abstractly to correspond to states of the speech production process) connected by transitions. Each transition is described by two sets of probabilities:

- A **transition probability**, which provides the probability of making a transition from one state to another.
- An **output probability** density function, which defines the conditional probability of observing a set of speech features when a particular transition takes place. For semicontinuous HMM systems (such as SPHINX-II) or fully continuous HMMs [31], pre-defined continuous distribution functions are used for observations that are multi-dimensional vectors. The continuous density function most frequently used for this purpose is the multivariate Gaussian mixture density function.

The goal of the decoding (or recognition) process in HMMs is to determine a sequence of (hidden) states (or transitions) that the observed signal has gone through. The second goal is to define the likelihood of observing that particular event given a state determined in the first process. Given the definition of hidden Markov models, there are three problems of interest:

- **The Evaluation Problem**: Given a model and a sequence of observations, what is the probability that the model generated the observations? This solution can be found using the forward-backward algorithm [9][61].
- **The Decoding Problem**: Given a model and a sequence of observations, what is the most likely state sequence in the model that produced the observation? This solution can be found using the Viterbi algorithm [72].
- **The Learning Problem**: Given a model and a sequence of observations, what should the model’s parameters be so that it has the maximum probability of generating the observations? This solution can be found using the Baum-Welch algorithm (or the forward-backward algorithm) [5][9].

3.2.3 Recognition unit

An HMM can be used to model a specific unit of speech. The specific unit of speech can be a word, a subword, or a complete sentence or paragraph. In large-vocabulary systems, HMMs are usually used to model subword units [7][12][36][38] such as phonemes, while in small-vocabulary systems HMMs tend to be used to model the words themselves.

SPHINX-3 is based on phonetic models because the amount of training data and storage required for word models is enormous. In addition, phonetic models are easily trainable. However, the phone model is inadequate to capture the variability of acoustical behavior for a given phoneme in different contexts. In order to enable detailed modeling of these co-articulation effects, triphone models were proposed [67] to account for the influence by the neighboring contexts.

Because the number of triphones to model can be too large and because triphone modeling does not take into account the similarity of certain phones in their effect on neighboring phones, a parameter-sharing technique called distribution sharing [29] is used to describe the context-dependent characteristics for the same phones.

3.2.4 Training

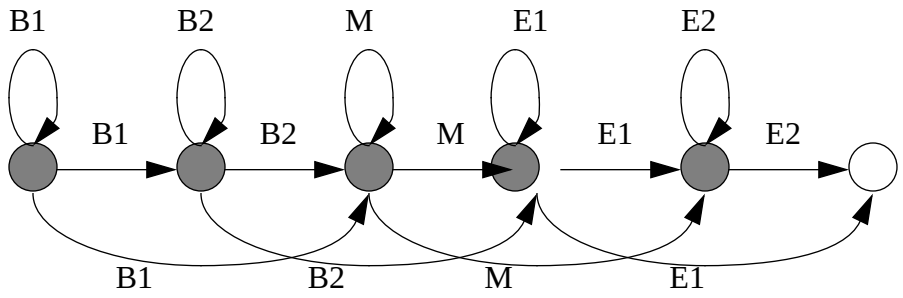

Figure 11. The topology of the phonetic HMM used in the SPHINX-3 system.

SPHINX-3 is a triphone-based HMM speech recognition system. Figure 11 shows the basic structure of the phonetic model for HMMs used in SPHINX-3. Each phonetic model is a left-to-right Bakis HMM [8] with 5 distinct output distributions.

SPHINX-3 [28] uses a subphonetic clustering approach to share parameters among models. The output of clustering is a pre-specified number of shared distributions, which are called senones [29]. The senone, then, is a state-related modeling unit. By using subphonetic units for clustering, the distribution-level clustering provides more flexibility in parameter reduction and more accurate acoustic representation than the model-level clustering based on triphones.

The training procedure involves optimizing HMM parameters given an ensemble of training data. An iterative procedure, the Baum-Welch or forward-backward algorithm [9][61], is employed to estimate transition probabilities, output distributions, and codebook means and variances under a unified probabilistic framework.

The optimal number of senones varies from application to application. It depends on the amount of available training data and the number of triphones present in the task. For the training corpus and experiments in this thesis which will be described in Section 3.3, we use 2500 senones for the Resource Management task with 5000 training sentences, whereas we use 7000 senones for the ARPA Wall Street Journal task with 7200 training sentences and 6000 for the Broadcast News task.

3.2.5 Recognition

For continuous speech recognition applied to large-vocabulary tasks, the search algorithm needs to apply all available acoustic and linguistic knowledge to maximize recognition accuracy. In order to integrate the use of all the lexical, linguistic, and acoustic sources of knowledge, SPHINX-3 uses a multi-pass search approach [2]. This approach uses the Viterbi algorithm [72] as a fast-match algorithm, and a detailed re-scoring approach to the N-best hypotheses [66] to produce the final recognition output.

The SPHINX-3 decoder [62] is designed to exploit all available acoustic and linguistic knowledge in several search phases. Initially, a Viterbi beam search produces a single recognition hypothesis

as well as a word lattice that includes word segmentations and acoustic scores. The word lattice is transformed into a directed acyclic graph (DAG). With DAGs, it is possible to use other possibly larger language models and perform a much quicker search for the best hypothesis. DAGs are also used to generate N-best lists for rescoring with parameters that are empirically optimized, like language weight and insertion penalty.

3.3. Experimental tasks and corpora

To develop the algorithm reported in this thesis, we have mainly used the DARPA Naval Resource Management Continuous Speech Database (RM1). This database is small enough to allow for a bearably short train-test cycle and yet large enough to make comparisons significant. The main limitation of this database is the fact its language model is too restrictive. If we allow the recognition engine to take the language model into account, performance differences of different implementations of the vocal tract length normalization algorithms considered are not discernible. In order to observe more clearly the effect on recognition of the various normalization algorithms considered, we disable the language model for the RM1 task in our experiments. This is accomplished by setting the language weight to a very small value.

Other foci of experiments, although to a much more limited extent, were the Wall Street Journal (WSJ) and Broadcast News (Hub-4 or H4) tasks. These databases cover a much broader range of speakers and speaking styles, as described later. For each database, we have created a single set of gender-independent HMMs.

3.3.1 Resource Management (RM1)

The Resource Management (RM1) corpus is a collection of recordings of spoken sentences pertaining to a naval resource management task. Subjects read the sentences from written prompts in low background noise. The material was recorded using a Sennheiser HMD-414 headset microphone and digitized at a 20-kHz sampling frequency with 16-bit quantization. The digitized speech data were downsampled to 16 kHz and segmented into files corresponding to individual sentence-utterances. Most of these sentence utterances are approximately 3 to 5 seconds in duration.

The corpus was originally partitioned into training, development testing, and evaluation testing data. In the experiments reported on this thesis, we have used the training and development testing data as training material, totalling 4797 sentences uttered by 120 speakers, 85 male and 35 female speaker. The test portion consists of 1600 sentences uttered by 40 speakers, out of which 23 are male and 17 are female. It contains 14968 words from a 1000-words vocabulary. The language weight is 0.001 unless otherwise noted.

3.3.2 Wall Street Journal (WSJ)

The WSJ database [57] consists of several subsets encompassing different vocabulary sizes, environmental conditions, foreign accents, etc.

In our experiments with the Wall Street Journal corpus, we use the official speaker-independent training corpus, referred to as “WSJ0+1-si_trn”, or “SI-284”, supplied by the National Institute of Standards and Technology (NIST) containing 35776 utterances of read WSJ text. These sentences were recorded using a Sennheiser close-talking noise-cancelling headset. The training utterances are collected from 284 speakers. All these data are used to build a single set of gender independent HMMs to train the SPHINX-3 system.

For testing purposes, we have used ARPA 1994 CSR Evaluation H1 Hub test data [33]. This corpus was similarly collected with a high quality close-talking, noise-cancelling microphone. Each of 20 subjects produced about 15 sentences to a total of 316 sentences containing 8184 words. The sentences were read from articles published in a variety of North American Business News Services: Reuters News Service, New York Times, Washington Post and Los Angeles Times as well as WSJ; all texts were drawn from financial news articles. The testing domain is therefore consistent with the training data. We have employed a dictionary with around 20000 words and a language weight of 9.5. These settings are more realistic than the ones used for RM1 in the sense that a real life system for this task would have similar settings.

3.3.3 Broadcast News (Hub-4)

The Hub-4 corpus [27] has been recorded from broadcast news shows, both from TV and radio broadcasts. It consists entirely of “found speech”, that is, speech that has been observed and captured in actual day-to-day usage, contrasting completely with the other corpora, which were based on speech elicited solely for purposes of speech recognition research.

Television broadcasts were received via cable TV network and recorded simultaneously to both Super-VHS video tape and digital audio tape (DAT). Radio broadcasts were received by means of a common high-fidelity stereo receiver with a digital FM tuner. An amplifying FM antenna was attached to the receiver to maximize received signal strength. Distance to the broadcasting antennas was approximately 10 miles, well within the broadcasting range of the local NPR affiliate station (which reaches a radius of at least 60 miles).

The DAT recordings were played through a Townshend DATLink digital audio converter for downsampling from the 32-kHz DAT sample rate to 16 KHz and storage of the left channel only to 16-bit PCM-encoded waveform sample files. In most cases where a single broadcast episode lasted for an hour or more, the waveform data were split into consecutive 30-minute segments for transcription and distribution.

The variety of conditions in this kind of task led DARPA to classify different segments of each show into a number of focus conditions [21], summarized below.

We followed two approaches for training models. In one of them we use all the available data, which is around 109 hours of recording. In the second approach, we use just a portion of the F0 focus condition containing about 8 hours of speech. The best performance is achieved when models are trained using all available data. The motivation to limit training to a portion of one of the conditions stems from the dependency of parts of the algorithm developed in this thesis on the availability of speech with clean background and high fidelity. For development purposes, we also limited the amount of speech to expedite training, keeping in mind that performance would deteri-

Condition	Dialect	Mode	Fidelity	Background
Baseline Broadcast (F0)	native	Planned	High	Clean
Spontaneous Speech (F1)	native	Spontaneous	High	Clean
Reduced Bandwidth (F2)	native	(any Mode)	Med/Low	Clean
Background Music (F3)	native	(any Mode)	High	Music
Degraded Acoustics (F4)	native	(any Mode)	High	Speech or Noise
Nonnative Speakers (F5)	nonnative	Planned	High	Clean
Other Combinations (FX)	—	—	—	—

Table 1. Hub-4 focus condition definitions

orate.

The testing data were recorded under the same conditions, although at a different epoch from the training data. Development and Evaluation test data are delivered as half an hour long blocks. No information is provided as to where utterances start or end or who the speakers are. Even if this piece of information were provided, there is no control over the amount of speech from each speaker. We employed a one hour subset of the Hub-4 1997 Development Test data. This subset was designed so as to reflect the amount of speech under each condition contained in the whole Development test data. Results obtained with this subset would then be scalable.

3.4. Statistical significance of differences in recognition accuracy

The algorithms we propose in this dissertation are evaluated in terms of recognition accuracy observed using a common standardized corpus of speech material for testing and training. Recognition accuracy is obtained by comparing the word-string output produced from the recognizer (hereafter referred to as the hypothesis) to the word string that had been actually uttered (hereafter referred to as the reference). Based on a standard nonlinear string-matching program [53], word error rate is computed as the percentage of errors including insertion, deletion and substitution of words.

It is important to know whether any apparent difference in performance of algorithms is statistically significant in order to interpret experimental results in an objective manner. Gillick and Cox [23] proposed the use of McNemar's test and a matched-pairs test for determining the statistical significance of recognition results. Recognition errors are assumed to be independent in the McNemar's test or independent across different sentence segments in the matched-pairs test, respectively. This assumption implies that the McNemar's test is meaningful only if the errors are sufficiently far from each other. Picone *et al.* [59] also advocated a phone-mediated alternative to the conventional alignment of reference and hypothesis word strings for the purpose of analyzing word errors. NIST has implemented several automated benchmark scoring programs to evaluate statistical significance of performance differences between systems.

Many results produced by different algorithms do not differ from each other by a very substantial margin, and it is to our interest to know whether these performance differences are statistically significant. A straightforward solution is to apply the NIST "standard" benchmark scoring program to compare a pair of results.

In general, the statistical significance of a particular performance improvement is closely related to the differences in error rates, and it also depends on the number of testing utterances, the task vocabulary size, the positions of errors, the grammar, and the range of overall accuracy. Nevertheless, for the RM1 task with the SPHINX-3 system, a rule of thumb we have observed is that performance improvement is usually considered to be significant if the absolute difference in accuracy between two results is greater than 3%. For the other databases, the rule of thumb we observed is that an absolute difference greater then 1% is significant. Whenever there was any doubt about significance of differences in results, we have used the scoring program provided by NIST.

3.5. Conclusions

In this chapter, we reviewed the overall structure of SPHINX-3 that will be used as the recognition system in our study. We also described the training and evaluation speech corpora that we employ

to evaluate the performance of our algorithm in the following chapters. The primary vehicle for research of this thesis will be the RM1 1,125-word 4797 sentences training corpora from which a single set of gender independent 2500-senonic HMMs will be constructed, together with the 1600 sentences evaluation test data. The other databases we will employ, although to a lesser extent, are the WSJ0+1 and Hub-4 corpora.

Chapter 4
Vocal Tract Length Speaker Normalization

4.1. Introduction

Speaker normalization through a warping function can be thought of as a mapping between two spectra. Speaker normalization techniques deal with two problems that are different in nature. One of them is how to model this mapping, i.e., how to choose a generic mathematical description relating two spectra. The second problem is, given the function, how to choose the parameters that uniquely define the mapping between two particular spectra.

Mathematically, we can see the problem of defining a speaker specific warping function as the definition of a curve on a plane whose axes are the frequency scales before and after warping. We approach this problem by defining points on this plane and fitting a curve to these points as in Figure 12. The problem is therefore restricted to choosing the points and what kind of curve to use to fit them.

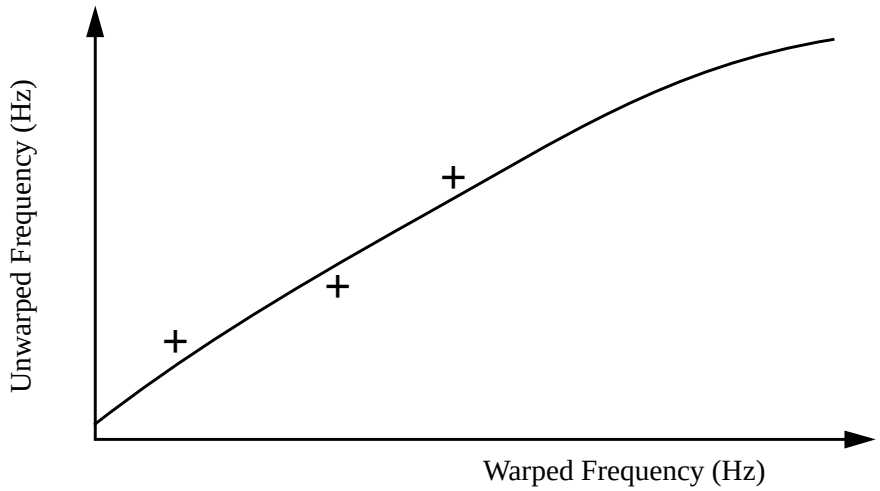

Figure 12. Warping function is chosen by defining points on a plane and fitting a curve to the points.

The speaker dependency obviously lies in the choice of points. We define these points based on acoustically-related features, like formants and high frequency parameters. We compute histograms of distributions of these features and employ statistics computed from these histograms in

the definition of the previously mentioned points. Initially we employ each parameter separately to analyze their possible individual benefits to speech recognition performance.

Thereafter, we study some aspects involved in the combination of features. These aspects are the choice of what features to combine and how to weigh them so as to achieve the best combination. The problem of choosing what features to use can be interpreted as a special case where the weights are 1 or 0 depending on whether the feature is utilized or not respectively. Nonetheless, we treat the choice of features and their weighing separately. We start with a fairly large number of possible features, select a possibly small number of them, and determine weights for this final set.

Given a set of points with respective weights, we face the problem of how to fit a curve to them. Linear regression is a powerful tool for this purpose. Linear or, more generically, affine functions can easily be found to fit the mentioned points. Nonlinear functions can be mapped to another space where points along the nonlinear curve can be fitted to a straight line. We study some possibilities and analyze the influence of different choices of curves on the performance of the recognition system.

A crucial point in the estimation of the warping function is the definition of points based on acoustic features, which are extracted from formant distributions. There is an inherent dependency on a formant tracker. This dependency affects the overall performance of the algorithm in terms of computational cost and reliability.

Computational cost can be reduced by defining a small number of classes of speakers and performing normalization based on these classes instead of on a speaker-specific basis. We examine the extreme case where these classes are the two genders. The cost of classifying an utterance as having been spoken by a male or female speaker is much less than that of estimating a speaker's formants.

The reliability of a formant tracker can be affected by the quality of the recorded speech. The pres-

ence of disturbances, such as noise, music, or concurrent speech, can undermine the estimates of formants. Limited bandwidth can easily mislead the formant tracker into picking the wrong peaks of the spectral envelope. If, however, we have high-quality speech recorded by a speaker, we can estimate formants taking advantage of these segments of speech alone. We study the extent to which our algorithm can still be employed for normalization in this scenario where both clean and noisy speech are available from a speaker.

In this chapter we explore the issues outlined in this introduction. We examine the effects on recognition accuracy of the shape of the warping function, the features used to define the warping function, and various computational simplifications. All experiments reported in this chapter make use of the Resource Management (RM1) database described in Section 3.3.1.

4.2. Speaker normalization methods proposed by other authors

In this section we present recognition results obtained with speaker normalization methods proposed by other researchers. These methods were described in Chapter 2. We present recognition results on the Resource Management (RM1) database on Table 2. Notice that these implementations are our implementations of algorithms developed by other authors. Our implementations differ from Wegmann *et al*. in the shape of the warping function and they differ from Eide and Gish in the parameter used to define the warping function. Specifically, Eide and Gish use the median of the third formant F3, whereas we ran experiments using the medians of the first three formants. In fact, our work differs from the one carried on by Zhan and Westphal in that they use means instead of medians.

These results seem to confirm results obtained by Zhan and Westphal in the sense that the best result is achieved when selecting warping functions based on maximization of likelihood. Moreover, these experiments seem to show that the linear warping function, besides the appeal of simplicity, which brings about lesser cost, also provides the best benefit. A linear warping function seems to capture most of the information necessary to achieve speaker normalization.

Our implementation of method originally by	WER
Baseline	38.0
Wegmann *et al*. with nonlinear function	35.1
Wegmann *et al*. with linear function	31.2
Eide and Gish using F1	37.3
Eide and Gish using F2	33.0
Eide and Gish using F3	34.5

Table 2. Word Error Rate (WER) on Resource Management (RM1) for our implementation of some normalization techniques presented on the review.

Results based on Eide and Gish's work also seem to indicate that we can gain by exploring different choices of parameters. It has the advantage of not requiring an iterative process in the definition of the warping function, thus being a simpler process.

In the remainder of this work, we will explore these two dimensions: different shapes and different parameters defining a warping function.

4.3. Issues related to shape of the warping function

As mentioned in the introduction to this chapter, we define points on a plane whose axes are the frequency axes before and after warping. The warping function is a curve fitting these points. The coordinates of these points are defined according to speaker-specific acoustic features, such as statistics extracted from distributions of formants. We search for a best combination of warping function shapes and acoustic features by choosing a set of features, testing for different shapes of warping functions, and choosing the best set of features according to the resulting best shape of function. Iterations of these steps might be performed as needed. Therefore, the first step in this iteration is the choice of a set of features.

We have chosen to use the medians of the three first formants as the features defining points on the plane. No assumption is made of the optimality of this choice. Some other authors [17][77] have

used medians of formants as the guiding acoustic features in their works, suggesting that medians are a good choice at least for an initial approach to the experiments.

After we define points on a plane, we attempt to find a function that uses those points as constraints or guidelines. We make a distinction between interpolation and fitting of points. When we interpolate, we end up with a curve that necessarily contains the points, as if we utilized a data-driven method to find a function. When we fit a curve to the points, we use a model-based approach: we have a model for what the warping curve should be in the form of a mathematical description, and we find the parameters of this model so that the difference between the model, *i.e.*, the warping function, and the data is minimal.

When interpolating points, we are completely driven by data, so it makes no sense to impose constraints on the shape the final function will have. A simple piecewise-linear curve connecting points has been used in our experiments. The number of points changes only the coarseness of the interpolating curve.

As a matter of fact, we can take the idea of interpolation to an extreme and define a large amount of points on the plane, so the interpolation can be as refined as we want. This large amount of points can be chosen based on an extension of how we previously defined medians and extremal points: mapping points that define regions of equal areas in the histograms of formant frequencies. Therefore, we can slice the histograms into, say, 100 pieces, corresponding to the different percentiles. These percentiles will define 101 points on the plane which will be interpolated defining the warping function. The resulting warping function maps regions of equal area of both distributions and is referred to as “Equal area” warping function in following sections.

When using a model to fit to the data, though, the restrictions on the shape become important. Vocal tract length studies [19] have shown that for some phonemes, the formants are inversely proportional to the vocal tract length, that is, formant frequencies for some phonemes depend on the inverse of the vocal tract length l. Some phonemes though have a dependency with powers of l.

This suggests that a linear warping function might be a good choice for speaker normalization, but the use of nonlinear functions might lead to gains.

Regarding the linear function, there are still some extensions that can be applied. The linear function $y = ax$ can be used, but the affine function $y = ax + b$ can also be employed. This choice would imply we do not impose constraints regarding whether the warping function maps the frequency 0 Hz to the frequency 0 Hz. Intuitively this constraint sounds appropriate, but we do not have experimental evidence to impose it. An alternative would be to use the affine function with the additional condition that the origin will be one of the points in the collection of points we use to fit the line. This additional condition represents and intermediate solution between the linear function and the affine function using merely the acoustic features.

We have not imposed constraints on the choice of nonlinear function except that it should be monotonically increasing. It does not make sense to map low frequencies to higher frequencies while mapping the high frequencies to lower ones. The monotonicity can be justified on the argument that we do not want to map two different frequencies to the same one.

In speech recognition, we traditionally employ warping curves based on studies from speech perception. These studies, especially the concept of critical band, led to the use of the mel and Bark scales. We reasoned that this warping can be thought of as a mapping from a generic speaker to a standard speaker. We selected warping functions that approximate or resemble these functions but which can be completely defined with a few parameters.

One such a function is the normal-deviate obtained in the context of “receiver operating characteristic” (ROC) or “isosensitivity curve”. In this context, if we are given two sources which generate signal with a Gaussian distribution, the ROC curve, which relates the probability of detection P_D to probability of false alarm P_F, has a shape that fits the criteria described previously. Moreover, when this curve is transformed from the linear coordinate system to a normal-deviate coordinate system, the resulting ROC curve is a straight line. The transformation from the linear coordinates

(P_F, P_D) to the normal-deviate coordinates (Z_F, Z_D) is defined by:

$$P_F = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_F} \exp\left(-\frac{X^2}{2}\right) dX \tag{16}$$

$$P_D = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_D} \exp\left(-\frac{X^2}{2}\right) dX \tag{17}$$

The straight line defined by the coordinates (Z_F, Z_D) can be defined in terms of a parameter C and the means and variances of the two Gaussians as:

$$Z_F = \frac{(\mu_1 - C)}{\sigma_1} \tag{18}$$

$$Z_D = \frac{(\mu_2 - C)}{\sigma_2} \tag{19}$$

In the context of speaker normalization, we are given speaker specific features and we fit a curve to these points. Considering the normal-deviate transform, we transform the features to the normal-deviate space, fit a straight line to the points in that space, and map points along this line back to the linear system to find the warping function. The same idea of mapping a nonlinear curve to another space where this curve is linear can be extended to other curves.

We have benchmarked the shapes described in this section using the medians of the three first formants as the speaker specific features. Results are presented on Table 3. Notice that as mentioned in Section 3.3.1 we have used a very small language weight which essentially disables the language model. The reason behind this choice lies in the fact that the language model for this task is so constrained that differences between implementations cannot be assessed.

All different implementations of warping function result in improvement compared to no normalization. Moreover, differences between implementations are not statistically significant. Even if the shape of the warping function implies a counter-intuitive transformation, like the mapping of 0 Hz to nonzero, the specifics of the shape seem to be irrelevant. As differences are not significant,

Warping function	Datapoints	WER
Baseline	—	38.0
Affine ($y = ax + b$)	Medians	31.1
Affine ($y = ax + b$)	Medians + Origin	30.7
Linear ($y = ax$)	Medians	31.1
Normal-deviate	Medians	30.7
Equal Area	Percentile	30.4

Table 3. Word error rate (WER) for normalization using the medians of the three first formants and several different shapes of warping function. Normalization is effective, but differences in implementation do not seem to be relevant.

we can make no claim as to whether this constraint would bring degradation or improvement. If the particular shape is not important, we obviously choose the simplest function, the linear function, in most of the remaining experiments in this thesis.

4.4. Formant-based features

As pointed out in Section 2.2, formants are a very important cue in the identification of vowels. Motivated by this fact, we sought to achieve normalization by the use of features extracted from formants. Statistical measures were employed for mathematical simplicity and practical flexibility.

Means are the simplest and most straightforward measure we could make of a distribution of points. In the context of speaker normalization through a warping function, though, they might not make much sense. If we attempt to map distributions of formants using a warping function, the mean is most likely not going to be mapped correctly, unless the warping function is an affine function [54], due to the linearity of the expectation operator.

Medians represent a robust statistic of a set of data. Its estimate is less affected by outliers than means, for example. Moreover, we intend to map between two distributions of formant frequencies. The use of a median point is a good first approximation for this mapping. Any mapping, pro-

vided it is monotonically increasing, will map the medians of the distributions onto each other based on the definition of medians.

Extending this idea of medians mapping to medians, we also chose the extremal points of the distribution as features to be mapped. In the estimation of extremal points, the presence of outliers would be devastating. To cope with this fact, we made use not of the minimum or maximum points themselves, but of points defining the 5% and 95% percentiles under the assumption that the measurement of areas is less susceptible to be affected by mavericks.

4.5. Non-formant-based features

Formants that can be consistently estimated, that is, the first three or four formants, provide enough information to define the warping function in the frequency range from 0 to around 3.5 Hz. Most of the speech used in the experiments reported on this thesis was sampled at 16 kHz, providing us with signals whose spectrum ranges from 0 to 8 kHz. The range from 3.5 to 8 kHz goes unattended. As we intend to use speaker-specific features to constrain and define the warping function, we face the problem of how to handle frequencies in this range.

Formants are essentially a vowel-like-phoneme's characteristic. Fricatives have spectra that resemble white or colored noise in a range that usually goes beyond 4 kHz. They could therefore provide us with features in the range not covered by formants.

The problem of what measures to use arises again. With fricatives, there is no clear peak in the spectrogram as in the case of formants of vowel-like phonemes. However, the spectra of fricative sounds resemble a Gaussian-mixture distribution. We make use of the mathematical description of the spectrum as a distribution for modeling convenience, without rigorously imposing all constraints the concept of distribution would require.

The fitting of a mixture of Gaussians to the spectrum allows us to extend the idea of formants to fricative sounds. Instead of the peak of the spectral envelope, we collect at each frame the mean of

each Gaussian component, or at least the means located in a high frequency range. If we utilize only one Gaussian, we effectively compute the center of mass of the spectrum of a frame of signal. If we make use of a mixture of two Gaussian components, we discard the smaller one, and so on. We then estimate the distribution of these peaks, or of the means of Gaussian components. Statistics from these distributions are the high frequency features.

4.6. Features in isolation

Although we have complete freedom to choose features, we should not do so blindly. Andrews *et al.* [3] report on an attempt to select features mechanically. The authors devised a method to automatically select features from a large number of randomly chosen ones. They report this approach was a failure, and advise to use intuition as a guide on the search of features.

In the previous sections, we mentioned the reasons for choices of the parameters we experimented with. In a first attempt to try to understand their importance for speaker normalization, we studied the separate benefits of using each feature separately. In an attempt to once again assess the relevance of the choice of the shape of the warping function, we have used both a linear function and a normal-deviate curve. Notice that if we use only one feature at a time, we define only one point on the plane where the warping function is to be defined. Therefore we need to impose further constraints at least for the normal-deviate curve. For this curve, we further require the normal-deviate curve to be symmetric around the diagonal $y = 1 - x$. Recognition results with individual features for both the linear and the normal-deviate curve are presented on Table 4.

Some of the numbers presented in the table are very close. Particularly for these numbers, it is important to consider statistical significance. As a rule of thumb, we have observed that differences in word error rates around 3 % can be considered significant. This empirical rule is based on our observation of results of the matched pairs test applied to experiments on RM1 in the same conditions as the experiments reported in this section. Matched pairs test, which was used whenever there was any doubt about significance in improvements, is a more precise way to make pairwise comparisons that take statistical variations into account. Similarly to other statistical significance

Feature	Formant	Warping function	
		Linear	Normal-deviate
Baseline	—	38.0	
Maxima	f1	45.1	43.7
	f2	43.4	45.8
	f3	43.8	46.9
Median	f1	36.3	35.3
	f2	29.5	30.7
	f3	32.5	33.2
Minima	f1	40.5	38.7
	f2	35.2	34.2
	f3	34.1	34.9
Mean	f1	36.1	35.3
	f2	30.1	30.9
	f3	33.7	34.0
High Frequency	#1	38.3	37.9
	#2	37.2	36.6
	#3	36.9	36.4

Table 4. Word error rate (WER) for normalization employing linear and normal-deviate-based warping functions defined by features used separately. Experiments on Resource Management (RM1). Baseline implies no speaker normalization performed at any stage. Otherwise, normalization is applied to training and testing data.

tests, it requires analysis of where the errors occur and whether both systems produced the same kinds of errors, as mentioned in Section 3.4.

Firstly, we can reorganize the table above in order of relative benefit to performance in terms of word error rates. Without regard for the absolute numbers, we sorted the results from Table 4 to produce Table 5. Features which are not statistically significant according to the matched pairs test

[23] are grouped together.

Order	Linear	Normal-deviate
Best	Med F2	Med F2 Mean F2
	Mean F2	
	Med F3	Med F3 Mean F3 Min F2
	Mean F3 Min F3	
	Min F2	Min F3 Mean F1 Med F1
	Mean F1 Med F1 HF 3 HF 2	
		HF 3 HF 2
	HF 1	HF 1 Min F1
	Min F1	
	Max F2 Max F3	Max F1
		Max F2 Max F3
Worst	Max F1	

Table 5. Isolated features sorted by WER. Features in the same cell do not present statistically significant differences.

The relative position of features is essentially the same for both the linear and normal-deviate-based warping functions, although these functions have fairly dissimilar characteristics. One is linear and unconstrained on the upper end of the frequency range. The other is nonlinear and bounded both at the origin and at the Nyquist frequency. The largest dissimilarity appears when we use maximal points to define the warping function. But even in this case it might well be the case that differences arise because the maximum of F1, which is “out of order”, just should not be used at all.

We have attempted to relate the order of features found on Table 5 with the distributions of slopes produced by these features. We have measured distances between distributions of slopes grouped

by gender. We have assumed these distributions to be Gaussians. Distance between the distributions with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 is given by

$$D = \frac{\mu_1 - \mu_2}{\left(\frac{\sigma_1 + \sigma_2}{2}\right)} \tag{20}$$

We have computed the distance D for slopes produced with each of the features presented on Table 5 and plotted the results on Figure 13.

Figure 13 clearly displays a relation between the distance between the distributions and the relative importance of the feature. This relation is to be expected. A smaller distance implies either that the means of the distributions are closer or that the variances are larger. In either case, the distributions will have a larger overlap, meaning that there are more speakers being warped in the wrong "direction".

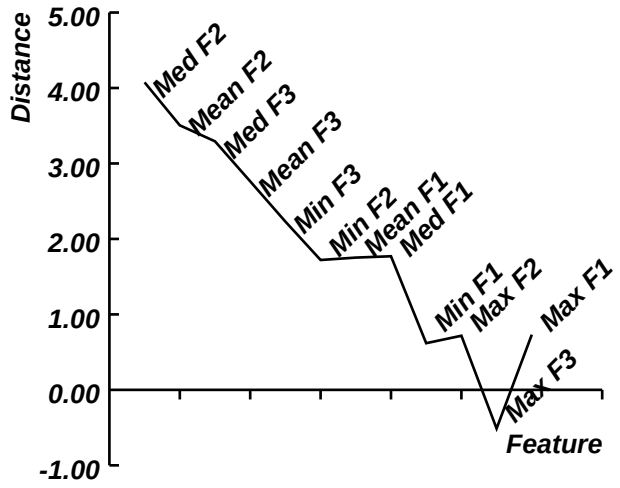

Figure 13. Distance between distributions of slopes according to gender for several features examined in this section. We can see a clear relation between the best features and a greater distance between distributions.

The fact that medians and means tend to be better than the minimal points, together with the fact that statistics of F2 and F3 tend to be better than those of F1, also deserve attention. Extraction of formants is an errorful process. Mapping of areas, as is the case with medians as well with extre-

mal points, according to the definitions we used, alleviates this drawback. Nonetheless, the definition of the warping function might be much affected by bad estimates. However, the farther the feature is from the origin, the less influence errors in its estimation will have on the warping function, provided the warping function is forced to include the origin. If we err by the same amount in the estimation, say, of the minimum of F1 and the median of F3, this error causes a greater effect on the warping function estimated from the former.

Our findings that medians of F2 and F3 are among the best features confirm the choice made by Eide and Gish [17], who used the median of the third formant as the guiding parameter in their definition of warping function. However, we should point out that their selection of feature was based on other criteria, and the warping function they use also has a different shape. Eide and Gish chose the median of F3 because this formant has the least variation across different phonemes uttered by a speaker, thus making its estimation more reliable. Their choice of warping function was described in Section 2.4.3.

Zhan and Westphal [77], on the other hand, found slightly differing results. They have attempted to use both linear and nonlinear warping functions, as described in Section 2.4.3. Their choice of nonlinear function was based on Eide and Gish's work. For both linear and nonlinear shapes they reached the conclusion that the mean of the first formant was the most effective in performance improvement. They in fact report that the second formant brings about a deterioration in performance. In our experiments, we found that the mean of the second formant works better than the third, and both work better than the first. We still found that, when used in isolation, the median seems to provide a better feature than the mean. In either case, using medians or means, the improvement achieved with the second formant is the best, and with the first, it is the worst.

4.7. Combination of features

In this section, we study issues related to the combination of features. If we combine the best features from the previous section, that is, the best features considered separately, we are not guaranteed to achieve the best possible result. Other combinations of features could result in a better

performance.

We explore two issues related to combination of features: what features will be used and what weight we will assign to each of them. We explore the two issues with the same approach. We compare estimates of slopes based on uni-dimensional and multi-dimensional linear regression and derive weights from this comparison. Obviously, we need some extra information regarding the slopes so we can estimate the weights. We assume we know the optimal slope for a set of speakers in terms of recognition performance.

Uni-dimensional linear regression gives an estimate of the slope of a linear function that best fits a set of points on a plane given the coordinates of these points and corresponding weights. All these pieces of information refer to one particular speaker. However, the weights and the features of the standard speaker are common to all speakers, which suggests we can estimate the slope of the speaker's warping function from the speaker specific features alone.

Multi-dimensional linear regression gives an estimate of the parameters of a linear function where the independent variable is multi-dimensional. In this context, the set of speaker specific features is the independent variable from which we estimate the slope of the warping function. Each speaker defines a point in this space. A set of speakers defines a "cloud" of points that we model as a straight line whose parameters we find with multi-dimensional linear regression. We assume knowledge of the best slope of the warping function for each speaker in this set. A comparison between the slope estimate from this multi-dimensional model and the uni-dimensional estimate allows us to find the optimal set of weights.

The slope of the best warping function for each speaker is determined by exhaustive search. From the results of this procedure, we find the weights, common to all speakers. New speakers will use the same set of weights, thus precluding the need for any further search. Moreover, search over a coarser grid of values, which reflects in quantization errors in the multi-dimensional linear regression space, will not have such a dramatic effect in the weights.

In this section we present a detailed study of the issues raised in this introduction.

4.7.1 Selection of features

In this section, we present a method for selection of features given a large set of features. This method takes into account recognition results as a criterion, although indirectly. We constrain the set of weights of the linear regression to be either one or zero for each feature. Each combination gives an estimate for the slope. We sort the combinations according to how close they are to the best slope, thus obtaining an optimal combination for the set of speakers.

In general, given a set of features chosen with arbitrary criteria, we need to select some of them so that their combination will be optimal for recognition purposes. If we just consider recognition results of features taken separately, we might be misled into believing the combination of the best features will bring about the best combination. It might be the case, for example, that the separate features lead to very similar warping functions for all speakers, even though this warping function might not be optimal for any speaker. In this case, we gain nothing from their combination.

The two main techniques to select features from a set of arbitrary ones are essentially complementary ideas. In one of them, the so called knock-out method [64], performance is evaluated for a system using all N features together. Then, performance is evaluated for each combination of $N-1$ features, that is, each feature is discarded at a time. The feature which causes less degradation in performance is eliminated. The process is repeated until there is one feature left. The last feature left is the most influential and the first one removed is the least important. Besides, combinations have been examined with different numbers of features.

The other method, named add on [24], starts with separate features. Each feature is combined with the best one in turn, and the best combination is selected. The process continues, each time growing the set containing the best combination by one. Again, combinations with all numbers of features have been examined.

A weak point of these techniques is that both of them presume that the optimal set of feature is a subset or a superset of the current optimal set. So, for example, if the best set of features with three elements contains the features A, B, and D, but the best set with two elements contains the features A and C, neither of the methods would find these solutions.

If we consider performance of a system regarding speech recognition, we are of course considering word error rate. As we need to examine several combinations of features, it implies we have to run several passes of recognition. This approach would be extremely expensive. Instead, we adopt a simplified approach.

According to our experiments reported in Section 4.3, different shapes of the warping function result in statistically insignificant differences in recognition performances. Therefore, we make use of the simplest choice, the linear function $y = ax$. The use of this function implies that we only need to find one parameter for each speaker, the slope a.

In order to find the best combination of features, we will assume we know the slope of the best warping function for a set of speakers. We will then minimize the difference between this best slope and the ones obtained from a linear regression estimate when we vary the weights.

Given a set of features $\{(x_k, y_k)\}$ and corresponding weights W_k, the slope of the warping function can be estimated by linear regression as:

$$\hat{a} = \frac{\sum_k x_k y_k W_k}{\sum_k x_k^2 W_k} \tag{21}$$

If we pool together all of the chosen features, we can set each W_k successively to 0 or 1 and compute the estimate $\hat{a}^i$ for each speaker i. We know the best slope a^i for each speaker i, so we can easily compute the difference:

$$\sum_i (a^i - \hat{a}^i)^2 \tag{22}$$

for all combinations of the finite set $\{W_k | W_k = 0, 1, k = 0, 1, 2, \ldots\}$. Notice that the criterion imposed by Equation 22 is exactly the same imposed by the linear regression formulation. Therefore, this method can be interpreted as a special case of the method presented on Section 4.7.2. However, the set of features might be as large as the set of speakers. An analytical solution would be compromised, since we would be attempting to estimate a large number of variables from a small number of data points.

Setting the weight W_k successively to 0 or 1, we computed the squared error and found the best combination for the RM1 task was achieved using the minimum of F1 and the medians of F1 and F2. We tested for the optimal combination in two partitions of the test data (no overlap of speakers), and this combination and some variations of it (adding or not the mean of F1 for example) were among the top 5 (out of 2^{18}) for both partitions. As a comparison, we ran recognition results for the combination of the 2 and 3 best features according to Table 4. The results are summarized on Table 6.

Combined Features	WER(%)
Baseline	38.0
Min F1, Med F1, F2	29.4
Mean F2, Med F2	30.1
Mean F2, Med F2, F3	30.7
Med F1, F2, F3	31.1

Table 6. Recognition results for combinations of features. The first set (combination of minimum of F1 and medians of F1 and F2) was selected based on least squared error; the second and third sets (mean and median of F2; mean and median of F2 and median of F3) are combinations of the two best and three best features when used in isolation; the final set is a combination of features selected intuitively (medians of F1, F2, F3).

Results from Table 4 on page 47 confirm results from Eide and Gish in that F3 is a good parameters for the definition of the warping function. However, our current results show that when we seek a combination of features, the best choice might lie elsewhere.

4.7.2 Weighing of features

The use of variances as weighing factors seems intuitive because if the feature has a bigger variance, its estimate is less reliable, and therefore its weight should be smaller. But the fact that a feature has a smaller variance does not necessarily mean it is more important from a speech recognition standpoint.

To estimate proper weights, we introduce a method utilizing Multi-Dimensional Linear Regression (MDLR). As mentioned in the introduction to Section 4.7, we will compare the estimates of slope using uni-dimensional and multi-dimensional linear regression. In the uni-dimensional perspective, let's assume we have chosen the warping function to be a linear function $y = ax$, and we want to find the slope a^i for speaker i based on the features F_k^i, $k = 1, 2, 3, \ldots$, and corresponding weights W_k, $k = 1, 2, 3, \ldots$, that is:

$$a^i = f(F_1^i, F_2^i, \ldots, W_1, W_2, \ldots) \tag{23}$$

where the superscript i indicates a measurement for speaker i. Notice that the set of weights W_k, $k = 1, 2, 3, \ldots$, is the same for the entire set of speakers.

Given the best slope a^i for speaker i (for example, found because we performed an exhaustive search of parameters a), we can find the weights W_k which would yield the best estimates $\hat{a}^i$ of each slope a^i by minimizing the difference between $\hat{a}^i$ and a^i in the least squares sense. We can find the weights W_k based on a subset of speakers, and use these weights for all speakers afterwards.

In the multi-dimensional linear regression perspective, we model the slope of the warping function as a linear combination of speaker specific features. We thus find a model that provides an estimate of the slope so that the difference between this estimate $\hat{a}^i$ and the best slope a^i is minimal in the least squares sense for a set of speakers.

Mathematically, let x_k be the k-th feature for the standard speaker and y_k be the corresponding k-th feature for one particular speaker with corresponding weights W_k. The dependency between x_k and y_k can be modeled by the function $y = ax$ on the plane (x, y), where the slope a is given by:

$$a = \frac{\sum_k x_k y_k W_k}{\sum_k x_k^2 W_k} \tag{24}$$

Applying the normalization constraint for the $\{W_k\}$:

$$\sum_k x_k^2 W_k = \sum_k x_k^2 \tag{25}$$

we can rewrite Equation 24 as:

$$a = \frac{x_1 W_1}{\sum_k x_k^2} y_1 + \frac{x_2 W_2}{\sum_k x_k^2} y_2 + \frac{x_3 W_3}{\sum_k x_k^2} y_3 + \ldots \tag{26}$$

On the other hand, if we know the slope of the best warping function for a set of speakers but we do not know the weights, the slope can be modeled by multi-dimensional linear regression as a function of the multidimensional variable $y^i = (y_1^i, y_2^i, y_3^i, \ldots)$ for each speaker i of a set of speakers. MDLR finds the coefficients γ_k of the model:

$$a = \gamma_1 y_1 + \gamma_2 y_2 + \gamma_3 y_3 + \ldots \tag{27}$$

If we compare equations (26) and (27) and equate, we get:

$$\gamma_j = \frac{x_j W_j}{\sum_k x_k^2} \tag{28}$$

which is easily solvable for W_j. We can re-normalize the weights at will.

This technique usually yields weights that are seemingly inconsistent. We believe this is the consequence of finite precision in the computer's representation of real numbers. Figure 14 presents the

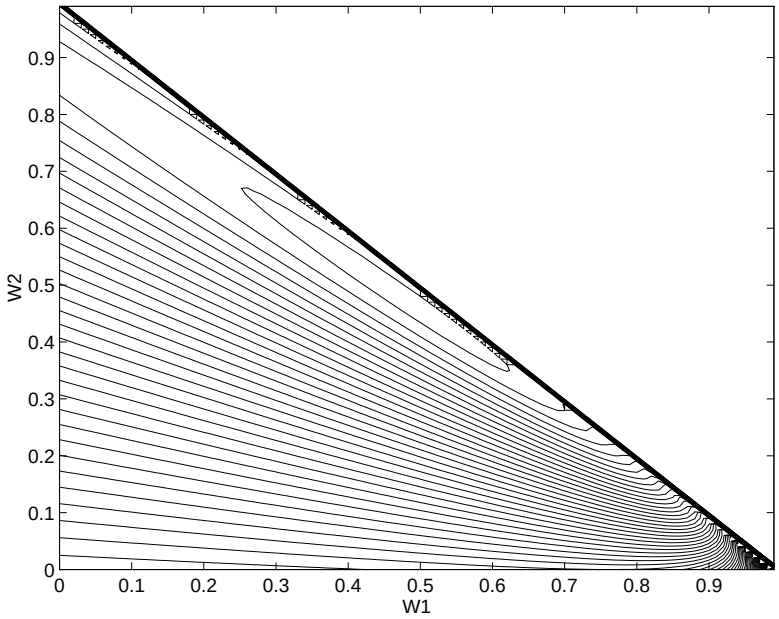

Figure 14. Contour plot of the error between estimated and measured slopes as function of combinations of W_1, W_2, W_3 (which add up to 1). Points in the central ellipse might be indistinguishable due to limited precision.

contour plot of the error, that is, of the summation of squared differences between the slopes obtained by exhaustive enumeration of a and the ones estimated using the set of weights for a range of weights. Each point in the $W_1 \times W_2$ plane represents a combination of W_1, W_2, W_3. The values of W_1 and W_2 are taken directly from the coordinates. The value of W_3 is so that the weights add up to 1.

The error in Figure 14 is minimum at the central ellipse and increases towards the edges of the plot. The error at the vertices have values which are consistent with the findings on Table 5, that is,

this error is largest at $W_1 = 1$, decreasing for $W_3 = 1$ and decreasing more for $W_2 = 1$. Table 5 indicates that the best feature taken in isolation is the median of F2 followed by the medians of F3 and F1. A feature in isolation corresponds to setting its weight to one and the other weights to zero.

The central ellipse in Figure 14 represents points that are very close to the absolute minimum for which the error function does not vary much. Limited precision in the representation of floating point numbers prevents us from finding this minimum. On the other hand, points in this region produce squared errors close to 0.02 accumulated over 40 speakers. The error, as mentioned previously, is the difference between slopes measured in an exhaustive search and slopes estimated from a set of weights. If we compare this difference between slopes with a typical slope, around 1.0, we can see the error at any point in this region as well as the difference between errors at different points in this region are both very small. If the difference in error is so small and if the error itself is so small, the estimated slopes, no matter which point we use in this region, will be very close to the best slope for each speaker. This closeness reflects in no difference in word error rate for points in a region around the absolute minimum.

4.7.3 Experiments involving optimal selection and weighing of features

We have run experiments investigating the efficacy of MDLR in finding weights as opposed to other more “intuitive” choices of weight for different combinations of features. We used the combination of minimum of F1 and medians of F1 and F2, which according to our findings in Section 4.7.1 provides the best performance. We also used a combination of the medians of F1, F2, and F3, because median has become a popular choice of measurement because of its statistical robustness to outliers. In Table 7 we present results for some choices of selection of weights. The warping function is a linear function. Experiments were performed on Resource Management, as described in Section 3.3.1.

MDLR involves the computation of weights given the best slope of a linear warping function for a set of speakers. In the experiments reported in this section we used disjoint sets of speakers for

estimation of weights and for recognition tests.

Weighing method	Combination of features	
	Med F1, Med F2, Med F3	Min F1, Med F1, Med F2
Baseline	38.0	
Equal weights	31.1	29.4
Inverse of variance	31.1	33.6
MDLR method	29.7	29.5

Table 7. WER for normalization using different sets of weights. The weights were chosen to be equal for all features, the inverse of the variance of each feature, and based on MDLR.

According to the matched pairs test, the difference between MDLR and the other choices is statistically significant for the combination of medians of F1, F2, and F3, but the difference between using MDLR and using equal weights is not statistically significant for the combination of minimum of F1 and medians of F1 and F2.

Although the result for the combination of minimum of F1 and medians of F1 and F2 might seem disappointing, we have to consider these features were already selected as an optimal combination using exactly the same criterion as the one used to find optimal weights. These results indicate that we can achieve improvements in performance if we select a combination of features based on a quantitative optimization criterion but might as well achieve a comparable improvement if we select a combination based on qualitative reasons and optimize the weights so as to minimize errors in the estimation of slopes.

4.7.4 Exhaustive search

In this section, we describe the method used to find the best warping function for a speaker, exhaustive search of slopes, as well as report on experiments showing how this method can be affected by limiting the search space.

When performing an exhaustive search, we select a number of warping functions, warp utterances by a speaker according to each of these functions, and select the warping yielding best recognition performance as the best warping function for the speaker. The precision of the set of “best” slopes of course depends on how many points we initially chose to search. It is obviously a costly operation, since we need to perform recognition for as many slopes as we have selected.

As an attempt to decrease this cost, we can choose a coarser grid of values. The search results in a set of “best” slopes subject to much greater quantization errors, because we search over a sparser grid. A finer grid certainly results in better performance, but we expect the weights we find with MDLR not to be much affected by a coarser search space.

In order to verify this claim, we have run experiments comparing the recognition results we obtain with the exhaustive search over a grid of 21 points and another sparser one of 5 points. We then employed MDLR to find weights from sets of “best” slopes estimated with each grid. Results are reported on Table 8.

Points in the grid	Exhaustive search	MDLR
21	28.3	29.7
5	29.7	29.9

Table 8. Comparison of the influence of the coarseness of initial values on the estimation of weights. The exhaustive search presumes availability of transcriptions. The estimation of weights was performed using a set of speakers not used for test. A coarser grid results in no deterioration in performance for MDLR.

As expected, the recognition result for a sparser grid resulted in degradation in performance for the exhaustive search. However, the use of a coarser grid with MDLR results in word error rates which are not statistically significantly different. A finer grid implies a higher computational cost. With MDLR, one can search a sparser grid without degradation in performance. Therefore we do not need to incur in more computational cost.

Moreover, MDLR allows us a complete independence between the set of speakers where the initial exhaustive search is performed and the set where recognition is to be performed. Exhaustive search by itself requires several runs of recognition on the test set, or at least some sort of maximization of likelihood over several warped instances of each utterance in the test set to be performed. MDLR allows for a much simpler scheme of normalization with slight degradation.

Regarding the results on Table 8, the word error rate relative to exhaustive search has the additional constraint that the best slope for each speaker was selected based on word error rate, which implies knowledge of the transcriptions. Of course this is not a realistic assumption, because if we know the transcriptions there is no need for recognition. However, these numbers can be considered as the best case, that is, the numbers we would obtain if the selection of slope for each speaker was perfect. The numbers for MDLR on the other hand follow a realistic assumption that transcriptions are not available for the test set, whereas they need to be available for some development data where the “best” slope for each speaker is to be found.

Yet exhaustive search seems to provide gain over MDLR, at least when a finer grid is used for search of the optimum slope. The question arises as to whether MDLR could realize the same potential of the exhaustive search. We have examined the slopes generated by linear regression when we vary the set of weights W_1, W_2, etc. for some speakers. The solution we find with MDLR corresponds to one point in this space. Varying the weights, we can see how much variability we can obtain for the slopes and therefore how potentially different the solutions found with MDLR will be. We plotted the slopes across weights for a typical speaker on Figure 15.

We can see from this plot that the slope does not vary much in this space. There is some variation which provides the improvement produced by MDLR, but not as much as to potentially realize the improvement brought about by exhaustive search, provided the exhaustive search is performed on a detailed enough grid.

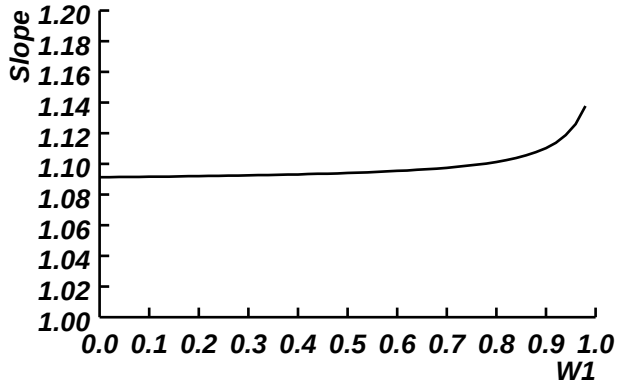

Figure 15. Slopes obtained for speaker vmh0 when varying W1 keeping W2 fixed. Curve shows little variation of possible slopes, thus showing limitation of MDLR.

4.8. Revisiting shapes of warping functions

In the process of optimizing for the features and shapes of warping function, we started by choosing a set of features, guided by intuition more than any scientific criterion. With this choice of features, we explored different shapes of warping functions, and with the optimal shape, we searched for the optimal combination of features. It might be the case that this optimal combination of features results in a different choice of optimal shape of warping function. Therefore, we need to benchmark the previous choices of warping function so as to account for the changes in features.

Initially, we used the combination of medians of F1, F2, and F3. For this new set of experiments, we have use the combination of the minimum of F1 and medians of F1 and F2 based on the results presented on Section 4.7.2. The shapes are the same utilized in Section 4.3, except of course that we do not need to benchmark the warping function defined on mapping of equal areas of histograms. Results are presented on Table 9.

Even though our experiments have led us to use a different set of features from the set used initially, we still find no statistically significant difference among most of the different warping functions we tried. In fact, we find a slight advantage to the specific warping function for which we optimized the set of features. This slight advantage, however, does not guarantee a much better performance with any particular warping function. Therefore, our initial conclusion that the shape of the warping function is not relevant still holds.

Warping function	Datapoints	WER
Baseline	—	38.0
Affine ($y = ax + b$)	Medians	30.4
Affine ($y = ax + b$)	Medians + Origin	29.6
Linear ($y = ax$)	Medians	29.4
Normal-deviate	Medians	31.2

Table 9. Word error rate (WER) for normalization using the optimal combination of minimum of F1 and the medians of F1 and F2 and several different shapes of warping function. Complexity of implementation and benefit to performance again point to the linear function as the optimal choice for the shape of a warping function.

4.9. Gender normalization

The need for estimation of formants for each speaker might still impose a computational cost beyond the acceptable for some applications. On the other hand, automatic labelling of an utterance according to the gender of the speaker who produced it is fairly cheap and accurate.

Moreover, it has been noted in Section 2.5.1 that the distribution of speaker specific parameters typically follows a bimodal distribution with the modes roughly corresponding to different genders. Speaker specific parameters, when grouped together by gender, clearly define distributions with different modes. In that section, the speaker specific parameter was the slope of a linear warping function. Acero [1] presents a similar result for the α parameter when the warping function is based on the bilinear transform.

The issue of computational cost and the observation of the bimodal distribution of speaker-specific parameters have led to the idea of using only two warping functions, one for each gender. Gender normalization precludes the need for formant estimation, thus greatly reducing cost. Gender normalization uses one of two different warping functions defined in advance rather than a speaker specific one computed from a speaker's acoustic characteristics. Linear warping functions require that two slopes be estimated, one for each gender. In order to estimate them, we first computed the

slopes from formants of speakers in the training set. We then considered speakers of each gender separately, and computed the medians of slopes of speakers of each gender. These medians were used as the slopes of the gender specific warping functions.

Computational cost is a more important concern during the testing stage, and the computational cost of the training stage is usually not of utmost importance. This flexibility allows for the use of speaker-specific normalization of the training data rather than gender normalization, even if the computational cost is an issue. A question that might be raised is whether different normalization schemes during the training and testing stages might lead to degradation in performance, or whether any normalization during training is necessary at all.

A related issue concerns the use of gender-specific models. In all experiments reported so far, we have used a single set of generic HMMs, created from speech from speakers of both genders. An alternative would be to use gender-specific models, that is, decoding speech from a male speaker using a set of HMMs created from speech from male speakers only, and similarly for female speakers. There is less variability among male speakers and among female speakers than if we consider all of them together. However, we have more data if we pool speech from all speakers when we create models. We need to assess if this trade-off between variability and amount of data is beneficial and how it compares to gender-specific normalization.

We have run experiments investigating these issues. Experiments were performed on Resource Management (RM1). Originally, this database contains roughly twice as many males as females. In the experiments reported in this section, we have used the same number of male and female speakers so that differences in results could be directly attributed to differences in gender rather than in the amount of data used to create models. Linear function was employed as the warping curve in all experiments. Results are reported on Table 10.

Regarding gender normalization, results show that measurable improvement in performance is achieved with gender normalization even if no normalization is employed with the training data.

Training conditions	Testing conditions	WER(%)
Unnormalized	Unnormalized	40.0
Unnormalized, male speakers	Unnormalized, male speakers	36.0
Unnormalized, female speakers	Unnormalized, female speakers	35.5
Unnormalized	Normalized	36.4
Normalized	Normalized	33.5

Table 10. Recognition results for gender normalization on RM1. The same amount of data was used for male and female speakers in all cases considered. This effects the results for no normalization to be slightly different from other results presented on this Thesis.

Normalization also of the training data results in improvement which is statistically significantly better. Gender normalization therefore provides improvement, and its potential can be fully realized only when normalization is performed on all the data.

These results also show that gender generic models with gender normalization outperform gender specific models. This result shows that normalization has been effective in bringing closer the speakers from both genders. This effectiveness allows us to use all available date to produce one single and arguably better set of HMMs.

4.10. Normalization on narrowband speech

In this section, we report on experiments using narrowband speech. Formants were extracted from narrowband speech. Models were also created from narrowband speech both when we normalize speech or not.

Formant estimation is always prone to errors. The algorithm presented on this chapter has an intrinsic dependency on the reliability of these estimates. Therefore, conditions that increase the chances of errors in the estimation of formants will lead to a poorer performance of the algorithm.

Formant frequency estimation is usually achieved by estimating the spectrum of a frame of speech, finding the peaks, and selecting some of them. Two potential problems are the presence of noise and the bandlimiting of speech, for example. These two conditions affect the process of formant frequency estimation differently.

Additive noise signals such as music or competing speech will add peaks to the spectrum. The estimation of spectrum as well as the location of peaks might be perfectly achieved, and yet the formant frequency estimates would not be reliable because the peak picking stage will likely select peaks introduced by the disturbance.

Bandlimited speech, particularly telephone speech, will cause the spectral estimation stage to be affected. Frequency components in the higher range of the spectrum will be attenuated, so peaks at the higher end might not be located accurately. Moreover, the peak picking stage might locate a peak below the first formant or between two formants, therefore selecting the spurious peak as a formant and shifting all formants upwards. Consequently, bandlimited speech, while in theory affecting only the extremes of the spectrum, can actually cause estimates of all formant frequencies to be wrong.

In the experiments reported in this section, we have artificially bandlimited speech by filtering it through a simulated telephone channel. Formants were extracted from the filtered bandlimited speech and normalization performed with statistics collected from these formant estimates.

Models for narrowband speech were created from narrowband speech. The issue of cross conditions, that is, of differences in channel and environment in the training and testing data, was examined by Acero [1] and Moreno [49]. They have shown that the best performance is achieved when the environments of the training and testing data match. We have run preliminary experiments wherein we recognized narrowband speech with models created from wideband speech, and recognition results are so much worse than when training and test data are similar that comparisons are meaningless.

The issue of influence of reduced bandwidth on the performance of the normalization algorithm has been examined, and results are presented on Table 11.

Model	Test data	Baseline	Normalized
Wideband	Wideband	38.0	29.4
Narrowband	Narrowband	46.8	41.9

Table 11. WER in narrowband speech. Speech was filtered through a simulated telephone channel. Normalization was applied to both wideband and narrowband speech. In both cases, a linear function fitting the minimum of F1 and the medians of F1 and F2 was used as warping function. Features have been computed from narrowband speech for experiments with narrowband speech and accordingly with wideband speech.

The performance of the recognition engine degrades when we use narrowband speech rather than wideband speech, as expected. However, normalization provides improvement in both cases. The improvement achieved through normalization is not as high as the one with wideband speech though. This is not a surprise since extraction of formants is affected by bandlimiting. This effects changes in the distributions of formants thus causing deviations to the features.

4.11. Conclusions

In this chapter, we presented a technique for speaker normalization which employs speaker specific features to define a warping function. Features determine points on a plane, which we fit to a curve thus defining the warping function. We studied several shapes of functions, concluding that a linear warping function captures most of the information necessary to accomplish speaker normalization. We examined several choices of features and their influence on the performance of the recognition system. This influence was examined taking each feature separately, as well as in combination. We presented techniques to select a subset of optimal features from a larger set, as well as a technique to find optimal weights given a set of features. These techniques use a similar approach based on comparisons between estimates provided by linear regression and slopes found through exhaustive search. We have found that both an optimal set of features and an optimal set of

weights for arbitrarily selected features bring improvement in recognition performance. A combination of these approaches, that is, optimal set of weights for an optimally selected set of features, however, does not bring improvement as compared to either method applied separately, since they both use the same criterion of optimality. Gender normalization was also examined as a simplification of speaker normalization. Gender-generic HMMs created from gender normalized speech outperforms gender specific models, thus showing the effectiveness of the presented technique for gender normalization. The effectiveness of the algorithm was also examined with bandlimited speech. Normalization was shown to perform well even under this constrained condition.

Chapter 5
Normalization Applied to Large Vocabulary Tasks

5.1. Introduction

In this chapter, we explore issues related to differences in performance when the speaker normalization is applied to different tasks. These issues have to do with implementation details or with peculiarities of the task.

So far, we have been using the Resource Management (RM1) database, which has been recorded under controlled conditions. Enough data from each speaker are available, and these data have been recorded under excellent audio conditions. Beside these facts, the vocabulary size is small, thus rendering the search during recognition much smaller.

The tasks we use for comparison are the Wall Street Journal (WSJ) and Broadcast News (Hub 4) databases, described in Sections 3.3.2 and 3.3.3 respectively. In this chapter, we explore differences by imposing artificial constraints in the databases. For example, we make use of a small amount of speech from each speaker on RM1 so as to simulate this limitation from Hub 4. With these simulations, we attempt to assess the effects on performance of these less than ideal conditions on different databases.

WSJ has also been recorded under good audio conditions, although the number of speakers and the vocabulary size as well as the total amount of data are much larger. These differences make WSJ a more realistic database than RM1, although a more difficult one under the point of view of the recognition system.

Hub 4 has been recorded from broadcast radio and TV. Therefore, there is no control over who the speakers are or how much data come from each speaker. Also, recording conditions vary both in channel and noise level or quality. All this variability makes Hub 4 a much more difficult task.

The first issue we examine regards the joint use of speaker normalization and adaptation. Normalization attempts to reduce variability between speakers by modifying speakers towards a common standard one. Adaptation attempts to reduce the differences by modifying models to make them more closely resemble the individual speaker. We study the possible benefits of joining these two complementary approaches.

WSJ and Hub 4 are much larger tasks than RM1. We studied how this difference affects the performance of the normalization algorithm. We have concentrated these studies on WSJ, since we have control over how much data is available from each speaker on this database.

Regarding the amount of data, WSJ has a much larger number of speakers than RM1 and a much larger amount of data per speaker. These differences influence the relative improvement in performance achieved by applying speaker normalization as well as whether normalization of the training data is essential or only beneficial. A larger number of speakers produces broader output distributions in the HMMs, thus increasing the probability of match between a new speaker and the ensemble of speakers in the training set. Reduction of variability achieved by speaker normalization has a less dramatic effect in this situation. On the other hand, more data per speaker has the opposite effect of producing narrower output distributions. Normalization of the training data reduces variability between speakers, thus resulting in less well tuned models.

On WSJ tasks, we employ a dictionary much larger than that used on RM1 tasks. The use of a smaller dictionary implies a smaller search space for the decoder. In this new search space, efficient normalization transforms the original data so that the correct path has higher likelihood than wrong paths for a larger number of sentences.

However, our objective is to reduce word error rate in general and not simply to make recognition results with normalization better than without normalization. Limiting the vocabulary, we might make the relative improvement larger. Yet, the absolute performance will probably be smaller if we use a larger dictionary. On a related matter, Schwartz *et al*. [68] report that for databases

recorded over more general conditions, the use of as much speech as possible, regardless of the quality, results in better models. We need to assess how normalization can benefit our system if we make use of an otherwise optimal system.

On Hub 4, there is no control over how much data we can use from each speaker. If we consider it on a sentence-by-sentence basis, some of these sentences are very short. We have attempted to assess the relevance of having less speech from a speaker by running a controlled experiment on RM1 where we varied the average amount of speech from each speaker. We also studied possible relations between sentence duration and performance improvement for RM1 and Hub 4.

A related matter regards clustering of speech data that have arguably been uttered by a single speaker. On Hub 4, we do not have information about where utterances begin or end, neither do we have knowledge of who the speakers are. We attempt to assess the effects of this lack of knowledge by restricting normalization, on RM1, to be performed on a sentence by sentence basis rather than on all utterances by a speaker. We further compare results obtained when we cluster segments on Hub 4 in an attempt to use as much data with similar acoustic characteristics as possible.

5.2. Experimental conditions for different tasks

5.2.1 Warping function

The normalization algorithm makes use of speaker specific features to define points on a plane. These points define a warping function, which is used to accomplish speaker normalization. In experiments reported in this chapter, we use an affine warping function $y = ax + b$ to fit the points defined by the minima and the medians of the three first formants F1, F2 and F3. As concluded in Section 4.3, the specifics of the warping function are not relevant. We have chosen this function for compatibility with experiments performed in other databases. The combination of minimal points and medians was among the top combinations optimized according to Section 4.7.1. One of the conclusions of Section 4.7.2 was that optimization of weights was not necessary if the combination of features was optimal. Therefore, we use equal weights for all features. In

some experiments on Resource Management (RM1), we used slightly different combinations of features and warping functions. This is noted where applicable.

5.2.2 Training conditions for the broadcast news (Hub 4) task

The Hub 4 broadcast news task presents a number of conditions that cause state of the art speech recognition systems to deteriorate in performance. Among these conditions, music and competing speech are sometimes present in the background, narrowband channels are sometimes used, and speech is often spontaneous or uttered by non native speakers. Also, there is no control over the amount of speech recorded from each speaker. As a matter of fact, in the test set there is no knowledge of where speech from one speaker ends and another speaker starts.

In Section 3.3.3 we present the official classification of conditions. Roughly, condition F0 applies to clean planned speech, F1 to clean spontaneous speech, F2 to telephone speech, F3 to speech with music in background, F4 to speech with competing speech, F5 to clean speech from non native speakers, and FX to all other speech.

In an attempt to control the audio quality and speech mode, we limited the training set to a subset containing speech under F0 condition alone. Moreover, we used only 8 hours of the available 21 hours on this condition so as to reduce the training cycle. The performance we achieve with this limitation is obviously worse than if we used all data. However, we are concerned with the difference in performance between performing normalization or not.

5.3. Adaptation and normalization

In this section, we study the benefit of jointly applying speaker adaptation and normalization to the recognition system. At a very superficial level, adaptation attempts to modify HMMs so as to make it a more accurate statistical representation of a given speaker. Normalization, on the other hand, attempts to modify incoming speech so as to reduce differences between a given speaker and a canonical one. Therefore, ideally both adaptation and normalization attempt at approaching a speaker dependent system from a speaker independent one, in one case by modifying the models

and keeping data from speakers untouched, in the other case by modifying data from the speakers so as to work with HMMs statistically representing only one prototypical speaker.

Both technologies tackle the same problem, but with a different approach. An issue that naturally arises in this context is whether the joint use of adaptation and normalization can be beneficial in terms of performance improvement.

In Section 2.6, we referred to MLLR, a popular technique of speaker adaptation. In the experiments reported in this section, we used a version implemented at CMU by Parikh and Raj [55]. Adaptation was performed in unsupervised mode, that is, we have used hypothesis generated by a first pass of recognition in the adaptation.

On RM1, however, we have been using a language weight unusually low. The reason for that, as mentioned in Section 3.3.1, was to disable the language model. With word error rates in the order of 40%, the use of transcriptions produced by the decoder might be misleading. For this reason, the estimation of the parameters for the MLLR adaptation was performed using transcriptions produced by a decoder using a fair language weight of 9.5. A language weight in this range is what a real system would require, therefore this implementation detail is more realistic than using the same language weight we have been using throughout this work.

We have combined MLLR with the normalization technique presented in this Thesis. The warping function used in this section is a linear function fitting the points defined by the minimum of the distribution of the first formant F1 as well as the medians of the first two formants F1 and F2. All points were equally weighed. Results of experiments on Resource Management (RM1) involving combinations of MLLR and normalization are presented on Table 12.

It is easy to understand that normalization by warping function simply expands or compresses the spectrum. The FFT coefficients after normalization are in fact a weighted average of the FFT coefficients before normalization, which can be thought of in matrix terms as the multiplication of the

Adaptation/Normalization	WER(%)
Baseline	38.0
Adaptation only	28.2
Normalization only	29.4
Adaptation and Normalization	23.8

Table 12. Word Error Rate (WER) on Resource Management (RM1) for experiments involving combinations of speaker adaptation and normalization. Adaptation was implemented using MLLR in unsupervised mode whereas normalization was implemented with a linear function fitting points defined by the minimal point of F1 and medians of F1 and F2 with equal weights.

vector containing the FFT coefficients before normalization by a matrix that is zero except at some "diagonal" lines. These "diagonal" lines of course are not necessarily parallel to the main diagonal, but rather will depend on the slope of the warping function being used. The important issue is that normalization can be thought of as the multiplication of a vector, the FFT coefficients before normalization, by a sparse matrix.

MLLR on the other hand makes use of all elements of a matrix. The use of all elements of a matrix allows for more flexibility than the use of only some lines in the matrix in that each element of the adapted vector depends on a larger number of elements of the unadapted one. This greater flexibility explains why MLLR alone outperforms normalization when the latest is applied by itself, as we can observe on Table 12.

When MLLR and normalization are jointly applied, however, we achieve a much greater gain than with any of the methods alone. Even if no normalization is employed in the training set, normalization still provides a gain jointly with MLLR. If normalization is applied to both the training and testing data and MLLR is employed, the word error rate goes down by a sizable amount.

Although the two methods tackle the same problem, the approaches are different. Neither of the methods is perfect. Even though speaker normalization and adaptation are successful in reducing

speaker variability, differences still persist. MLLR attempts to reduce differences by maximization of likelihood. The method presented on this Thesis utilizes a different criterion based on acoustic characteristics of speakers alone. The difference in optimization criteria provides a gain when the methods are jointly applied.

We also employed the combination of speaker adaptation and normalization on WSJ. In this case, the warping function for the normalization was implemented through an affine function fitting the set of 7 points defined by the origin and the minimal and median points of the first three formants. All points were equally weighted. Adaptation was unsupervised, i.e., a first run of recognition was deployed and the hypothesis transcriptions were used in the adaptation step. Results are presented on Table 13.

Adaptation/Normalization	WER(%)
Baseline	11.7
Adaptation only	11.4
Normalization only	10.9
Adaptation and Normalization	10.5

Table 13. Word Error Rate (WER) on Wall Street Journal (WSJ) for experiments involving combinations of speaker adaptation and normalization. Adaptation was implemented using MLLR whereas normalization was implemented with an affine function fitting points defined by the minimal and median points of F1, F2, and F3 with equal weights.

The results presented on Table 13 show that the combination of adaptation and normalization is also successful on WSJ. The overall combination did not result in as much improvement as was the case with RM1.However, most of this difference is due to the fact MLLR did not perform as well. We did not attempt to optimize MLLR since our interest was restricted to finding whether a the combination of techniques would be helpful. In fact, the combination of techniques does bring improvements over applying only either one of them.

5.4. Normalization applied to the training data

In this section we examine whether there is a need to apply normalization to all available speech, including training and test data, or whether normalization to the test data alone is enough to achieve improvements using this technique.

We have run experiments on the three databases RM1, WSJ, and Hub 4 comparing the scenarios of normalization on the test set only and on all data. Results of this comparison are presented on Table 14.

In all experiments in this section, the warping function was an affine function fitting the set of 7 points defined by the minimal points in the distribution of the three first formants, the medians of the three distributions, and the origin of the coordinate system. All points were weighed equally. This combination of features gives slightly better results than a mere combination of the medians of the three first formants as pointed out in Section 4.7.1.

On RM1 and WSJ, normalization was performed on a speaker basis, i.e., all speech available from a speaker was utilized to estimate the distributions of formants of the speaker.

On Hub 4, we have no information about who the speakers are. We have performed normalization on a segment basis. A large file with 30 minutes worth of speech was segmented into small parts, and these segments were used for recognition, as well as for the estimation of distribution of formants. Segments were assumed to be independent, that is, not uttered by the same speaker.

In addition to this, we have created models for Hub 4 not from all available speech, but from a subset of the available speech labeled as clean and planned, for reasons mentioned in Section 5.2.2. On Table 14 we present results for this condition, clean planned speech, which is referred to as condition F0, alone.

On RM1, we observe that the normalization, when performed only on the test set data, is beneficial

Normalization of cepstra		WER (%)		
Train	Test	RM1	WSJ	Hub 4 (F0)
No	No	38.0	11.7	31.4
No	Yes	32.5	11.6	28.6
Yes	Yes	30.4	10.9	28.7

Table 14. Comparison of performance when normalization is applied to speakers in the training set and testing set or in the testing set alone. The warping function is an affine function fitting points defined by the minima and medians of the three first formants plus the origin, with equal weights. On RM1 and WSJ normalization was performed on a speaker basis. On Hub 4, it was on a sentence basis. Moreover, on Hub 4 the results concern the focus condition F0 (clean planned speech) alone.

as compared to no normalization at all. However, normalization of the training set data provides further measurable improvement, thus implying that for this task normalization of all data is beneficial.

Ideally, normalization would eliminate all speaker specific differences. Therefore, if speaker normalization were perfect, we would expect all speakers to be mapped exactly to one standard speaker. Essentially, this would transform a speaker independent system into a speaker dependent one, where the standard speaker is the only speaker for whom the system was trained. Therefore, ideally, the scenarios of normalization on training and testing data as opposed to normalization of the testing data alone can be thought of as a comparison of performance of the standard speaker with a speaker independent and a speaker dependent systems. In this context, it does not surprise that normalization of the training set outperforms the other alternatives.

On WSJ, we observe that if models created from speech with no normalization are used, the performance achieved with normalization of the test data as compared to no normalization is not statistically significant. However, normalization of cepstra used for training allows us to create models which are statistically significantly better if used in conjunction with normalization on the test data. Statistical significance was measured with the standard matched pairs test provided by NIST.

On Hub 4, we note that normalization in fact brings improvement as compared to no normalization. However, normalization of speakers in the training set does not seem to be relevant. If normalization is applied to speakers in the test set, there is no statistically significant difference between the results obtained when normalization is applied or not applied to the speakers in the training set.

These experiments show that normalization is in fact effective in all tasks considered, however, to different degrees. On RM1, normalization of the training data is not essential but it helps. On WSJ, it is essential. On Hub 4, it is irrelevant. This difference in how normalization benefits performance in different tasks invites further investigation that takes into account the differences in design of the databases. RM1 contains only speech recorded under excellent audio conditions, and the number of speakers as well as the vocabulary size are fairly small. WSJ differs from RM1 essentially on the number of speakers and on the dictionary size. Hub 4 differs from the previous tasks on several issues, worth mentioning the amount of speech from each speaker, and the lack of knowledge about speaker identification or even speaker change. In the following sections we explore these issues to reach an understanding of the differences in performance raised in this section.

5.5. Tasks with similar recording conditions

We investigate in this section the possible reasons for the difference in performance of the normalization algorithm when applied to RM1 and WSJ. These differences are the relative improvement brought about by the normalization and the fact that normalization of the training data seems to be essential to WSJ, as opposed to being nothing but helpful for RM1. We attribute these differences mainly to the vocabulary size, the number of speakers, and the amount of data per speaker.

As we can see from Table 14, the improvement obtained on RM1 is much larger than that on WSJ or Hub 4. A direct comparison between RM1 and Hub 4 would not shed much light onto the reasons for this discrepancy, since the tasks differ so much. However, we can try to understand some of these reasons by comparing RM1 and WSJ.

RM1 and WSJ have been recorded under very similar conditions, with speakers reading from a prepared text, and speech is recorded using close talking microphones in a quiet office environment. Speakers were neither professionals nor naive speakers

The main differences between RM1 and WSJ regard the number of speakers, the amount of data per speaker, and the vocabulary size. Of course there are other differences, like the complexity of the texts, and therefore of language models, which contribute to differences in the baseline performance. However, we intend to investigate reasons why the relative improvement resulting from the speaker normalization differs, thus our choice to restrict the issues investigated to speaker and vocabulary issues.

RM1 has 120 speakers in the training set, whereas WSJ has 284. The RM1 system uses a dictionary with around 1000 words, whereas the WSJ system employs a dictionary with more than 20000 words. The effects of these differences are investigated next.

5.5.1 Influence of the number of speakers

In this section, we study how the number of speakers and the amount of data per speaker can affect the need for normalization in the training set. In order to study it, we have created two sets of HMMs. One of these sets was created from a subset of WSJ containing only 124 speakers, and the other set with all 284 speakers. We limited the amount of data for these 284 speakers so that the total amount of speech for both sets was roughly the same. We have also employed another set of models created from all available data from all 284 speakers. Results are presented on Table 15.

Although differences are not large, this experiment suggests that normalization provides more improvement for a system trained with a set of speakers with less speakers. The relative improvement is around 8% in the system trained with a set containing 124 speakers and around 5% in the system trained with a set of 284 speakers using half the available data. When the amount of data is the same, normalization seems to be more effective. Both differences are statistically significant,

Normalization of cepstra		Number of speakers in the training set / (Amount of data per speaker - total available = 1)		
Training	Test	124 (1)	284 (1/2)	284 (1)
No	No	12.5	12.1	11.7
No	Yes	12.2	11.6	11.6
Yes	Yes	11.5	11.5	10.9

Table 15. Word error rate for Wall Street Journal with different number of speakers in the training set. The data used for training was constrained so that the sets of speakers in the first two columns had roughly the same amount of speech, roughly half as much as in the last column. The overall relative improvement between no normalization and normalization of both training and testing data is 8% for models created with 124 speakers, 5% for models created with 284 speakers using half the data, and 7% with 284 speakers using all the data. Difference is significant in all cases. Normalization of the training data becomes irrelevant when less data is available per speaker.

although with different levels of confidence, namely less than 1% for the former and 2.6% for the latter.

A possible reason for this difference in performance for systems trained with different number of speakers is that the output probability density functions (pdfs) of the HMMs are going to be more spread out if the models are created from a larger number of speakers. This smearing occurs as a consequence of the spreading of acoustic characteristics we encounter in a larger set of speakers. A new speaker presented to the system will cause the system to have a worse performance if this new speaker has characteristics very dissimilar to the average speaker in the training set, since there is a tendency for speech from this speaker to be located in the tails of the output pdfs. This degradation will be worse if the output pdfs are sharper, and this will be the case if the models were trained from fewer speakers.

On the other hand, we can see that if we have a large amount of data for each speaker, normalization of the training data is essential. Normalization seems to work well if the HMMs are trained with a smaller amount of data per speaker, regardless of normalization of the training data. A larger number of speakers results in HMMs less well tuned to any particular speaker. The system is

farther from a speaker dependent one in that the performance will not be the best possible for any particular speaker. However, a new speaker presented to the system will achieve a reasonable performance. On the other hand, more data per speaker will cause the opposite effect, resulting in HMMs more well tuned to the speakers in the training set.

Normalization reduces variability between speakers. If the models are not well tuned to a set of speakers, normalization will have little effect in performance. If the models are somehow well tuned, normalization will help, in that normalization of the training data will approximate speakers on the training set an average speaker, the same to which normalization will approximate a new speaker.

Speaker normalization is an attempt to reduce variability between speakers. If this attempt is successful, its effect will be felt less dramatically on a system trained from more speakers than on a system trained from fewer speakers because the latter has a tendency of performing worse for a generic new speaker. This effect, although less dramatic, will only be achieved when normalization is also performed to speakers in the training set if the amount of data per speaker is very large. If the amount of speech per speaker is small, normalization is beneficial even when applied to the test set alone.

5.5.2 Effects of vocabulary size

In this section we study the influence of dictionary size on the performance of a system. Instead of using a typical dictionary for this task, containing around 20000 words, we have created a dictionary containing only words present in the test set. Moreover, this dictionary contained all words from the test set, that is, the system had no out of vocabulary (OOV) words. In general, the presence of OOVs in the test set causes degradation in performance because the system is restricted to finding a match for the acoustic signal from words in the dictionary and, since the correct word is not there, any outcome will be a failure. Experiments using this dictionary will obviously measure the influence of the vocabulary size as well as of OOV. However, we reason that the effect on the search of a dictionary with no OOVs is the same as of a smaller number of words, since we just

reduce the search space. A well designed dictionary will have a small percentage of OOVs in any case. Therefore, our conclusions will hold true in any case.

Additionally, the language model weight plays a role in the word error rate achieved with a system. To assess it, we have compared two different scenarios. In one of them, we have set the language model weights so that the word error rate with no normalization would be the same for all tasks and all dictionary sizes. Since the baseline numbers are the same, a comparison between improvements is more easily accomplished. In the second scenario, the language model weights are set to realistic numbers, i.e., language weights that would actually be used in a real system, Results for the first scenario, with the same baseline word error rates, are presented on Table 16. Results for the second scenario, realistic settings, are presented on Table 17.

Normalization of cepstra	WER(%)		
	WSJ - 2k dict	WSJ - 20k dict	RM1
No	11.7	11.7	11.7
Yes	9.5	10.9	9.6
Improvement (%)	18.8	6.8	17.9

Table 16. Word error rate comparing systems with different vocabulary sizes. Wall Street Journal was decoded with a small dictionary of 2000 words and a large vocabulary of 20000 words. The small dictionary was derived from the test set, thus the system has no out of vocabulary (OOV) words. The language model weight was set so the baseline WER would be the same for all tasks. The language weight is 0.31 for WSJ with a 2k dictionary, 9.5 for WSJ with a 20k dictionary, and 4.85 for RM1. Normalization results in statistically significantly better performance in all cases. Normalization was performed with an affine function fitting the minimal points and the medians of the three first formants.

In both scenarios we can clearly see that normalization performs comparably well on RM1 and WSJ 2k dict,, which are both small vocabulary systems, and in both cases normalization performs better than on WSJ 20k dict, which is a larg vocabulary system. The smaller dictionary affects the performance because it decreases the search space. An HMM based system concatenates models

Normalization of cepstra	WER(%)		
	WSJ - 2k dict	WSJ - 20k dict	RM1
No	2.9	11.7	6.6
Yes	2.5	10.9	5.5
Improvement (%)	13.8	6.8	16.7

Table 17. Word error rate comparing realistic systems with different vocabulary sizes. Wall Street Journal was decoded with a small dictionary of 2000 words and a large vocabulary of 20000 words. The small dictionary was derived from the test set, thus the system has no out of vocabulary (OOV) words. The language model weight was set to 9,5 for al tasks. Normalization was performed with an affine function fitting the minimal points and the medians of the three first formants.

of phonetic or subphonetic units so as to assemble the words contained in the dictionary. If the dictionary is larger, the system will have to look for a larger number of alternatives. In other words, with a larger dictionary the decoder will have to follow a larger number of paths defined by the concatenation of a larger variety of models, and this superset of paths might include paths containing wrong words with higher likelihood than the path with the correct sequence of words. The recognition engine will commit an error if it assigns a higher probability to a sequence of words containing an incorrect word. We might happen to eliminate this incorrect word when we decrease the dictionary size, thus rendering this wrong sequence of words inexistent.

Moreover, the acoustic observations of the normalized and unnormalized speech will be slightly different, so the probabilities assigned to the same sequence of words will be slightly different. Better acoustics will cause the correct sequence to have higher probability. When we reduce the dictionary, we throw away incorrect words, but the observation with worse acoustics might still lead to another sequence of words containing incorrect words, whereas the observation with better acoustics will cause the correct sequence of words to be assigned highest probability. The results reported above suggest that normalization indeed results in better acoustical models.

5.6. Issues related to availability of speaker specific data

In this section, we examine how the improvement brought about by normalization varies with the

amount of data available for each speaker. Although this issue is not relevant in the context of RM1 or WSJ tasks, the fact that we have enough data for each speaker in these databases allows us to control the amount of speech we utilize in the normalization process.

The normalization process involves estimation of formants, of histograms representing the distribution of these estimates, and of statistics from these histograms. As in any statistical approach, having too little data might lead to skewed estimates of the medians, means, minimal points, etc.

Moreover, we do not attempt to restrict the collection of formant estimates to any particular phoneme. Doing so would require a run of decoder so as to segment the incoming data, i.e., so as to locate the different phonemes in the incoming utterances. We thus rely on a distribution which will contain data from as large a variety of phonemes as possible. Shorter utterances will have fewer phonemes. The distribution of formants taking into account only the phonemes present in short utterances are therefore likely to be biased.

We investigate how limiting the amount of data available affects normalization. To approach this issue, we perform normalization on a speaker-by-speaker basis using increasing amounts of data for each speaker. Statistics were computed on increasingly larger subsets of the total amount of speech available per speaker, and at each step all speech from each speaker was normalized based on statistics computed from these subsets subsequently. Therefore, we assess the minimum amount of data needed for normalization to be effective.

We have performed these experiments on RM1, where we have complete knowledge about who the speakers are and control over how much speech we employ from each one. We compare these results with those obtained with Hub 4, where we have no knowledge of speaker identification, and are thus restricted to performing normalization with however little speech we are given. The comparison is accomplished by measuring the duration of segments and computing the average word error rate for subsets of segments with similar duration.

In an attempt to cope with segments of short duration on Hub 4, we studied the possible benefits of clustering segments together. Ideally, clustering would group different segments uttered by the same speaker in the same cluster. However, clustering is automatically performed. Segments from different speakers can be grouped in the same cluster. Moreover, it is not guaranteed to produce cluster with enough data per cluster. We examine the results obtained in our experiments.

A comparison between RM1 and Hub 4 only makes sense if we restrict it to the focus conditions on Hub 4 which have similar acoustic characteristics to RM1. Therefore, our comparisons are between RM1 and the focus conditions F0 and F1 on Hub 4.

5.6.1 Amount of data needed for normalization

In this section, we investigate the issue of how little data would be enough on average for the normalization to be effective. So far, we have been using as much data from each speaker as we can. We do not have a clear sense, however, of whether the amount of data we have been using is more than necessary or whether we could still have further improvements if we had more data. To study this issue, we computed features from distributions of formants extracted from increasingly larger amounts of data. With these features, we performed normalization and analyzed the results.

Figure 16 displays the results of these experiments. Normalization was implemented with a linear warping function fitting points defined by the medians of the distributions of the first three formants F1, F2, and F3. Weights were chosen with MDLR. We used an increasing number of sentences per speaker. Features, that is, the medians of the distributions of formants, were computed from these subsets of sentences subsequently. The average total duration of speech used in each case, i.e., the average over speakers of the amount of speech employed for each speaker, is displayed on the panel, together with the resulting word error rate.

These results show that improvement in performance achieved with speaker normalization reaches a stable level when the average amount of data from each speaker reaches around 12 seconds of speech. This observation suggests that this amount of speech on average would be enough to reach

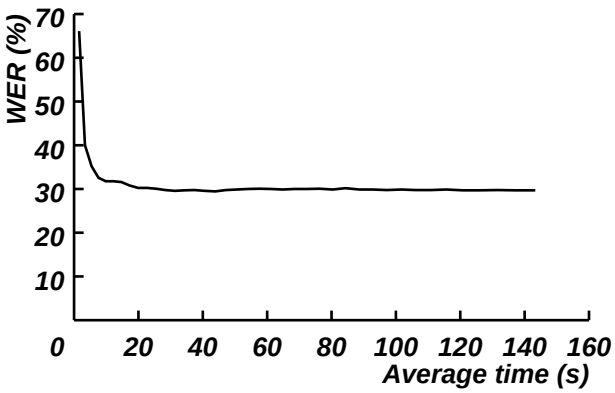

Figure 16. Word error rate on Resource Management (RM1) when normalization is performed based on features computed from varying amounts of speech. The features are the medians of the three first formants with weights set using MDLR. Normalization was implemented with a linear warping function.

saturation in the benefit achieved with normalization. The best case, that is, the speaker for whom the least speech was necessary to achieve saturation in performance improvement, required 1.2 seconds of speech, whereas the worst case required 114 seconds. Saturation in performance in this context means that the word error rate reaches a range within 10% of the word error rate obtained when all speech is used.

5.6.2 Normalization on short duration sentences

In this section, we present results concerning relation between sentence or segment duration and relative improvement in performance. Section 5.6.1 suggests that speaker normalization will perform poorly with short segments. For most segments extracted from Hub 4, we cannot select increasingly larger amounts of speech to perform a similar experiment since most segments are already very short. However, we can measure durations of segments and the word error rate with and without normalization. For comparison purpose, we performed a similar measurement on RM1.

Figure 17 shows the relative improvement in performance for segments of different duration. The improvement considered on this panel refers to relative variation of the average word error rate (WER) of sentences whose duration lies within a certain interval.

We have performed similar experiments on Hub 4. We have limited the study to focus conditions

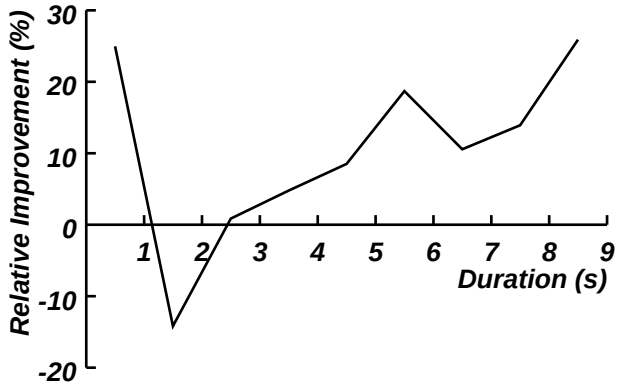

Figure 17. Relative improvement in WER on RM1 sentences. WER was computed as the average WER for sentences whose duration lies within a certain interval.

containing only clean speech. Figure 18 presents results for focus condition F0, clean planned

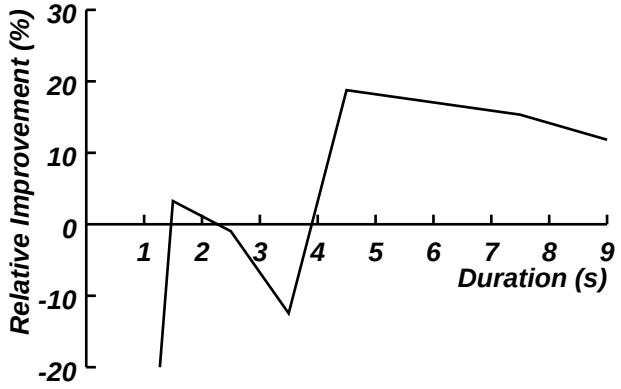

Figure 18. Relative improvement in WER on Hub 4 - F0 sentences. WER was computed as the average WER for sentences whose duration lies within a certain interval.

speech, whereas Figure 19 presents results for F1, clean spontaneous speech.

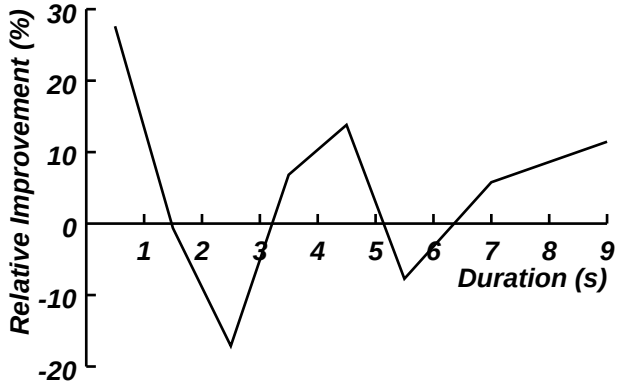

Figure 19. Relative improvement in WER on Hub 4 - F1 sentences. WER was computed as the average WER for sentences whose duration lies within a certain interval.

Sentences of less than one second of duration are usually one-word sentences, for which the WER

does not bear a meaningful value. The initial point in each of these panels is not very meaningful for this reason. However, we can see the tendency in all three tasks for the normalization to not work well for short utterances which last for less than five seconds. On the other hand, the algorithm works on longer utterances, providing improvements on average.

We hypothesize that the reason why the normalization fails for very short segments is related to the contents of the utterance rather than to acoustic characteristics. The best warping function for a given speaker would arguably be the same no matter how short the utterance is. If we can not achieve the same gain in performance when less data are available, it is because the features which define the warping function have been affected. The features defining the warping function are based on statistics computed from distributions of formant estimates. Very short utterances do not have a large variety of different phonemes. Consequently, the distribution produced by estimating formants from these utterances will be biased towards some frequency bands.

The standard speaker has been defined as an average over a large number of speakers, each providing a fairly large amount of data. The average distribution will therefore statistically represent a large variety of phonemes. The normalization attempts to map distributions of formants from a new speaker to the standard speaker by mapping statistics of these distributions. If a small amount of speech is available, these utterances will contain only a small subset of the possible phonemes. The distribution of formants extracted from these data will be biased around the formants of that particular set of phonemes. When we compute statistics from these biased distributions, we produce features which do not exactly correspond to the features of the standard speaker. As we increase the amount of data, we increase the variety of phonemes in the utterance and therefore increase the match between the set of phonemes uttered by the new speaker and the standard speaker. We therefore expect the relative improvement in word error rate to increase as we increase the variety of phonemes in the utterance.

Figure 20 illustrates this point. We estimated the relative improvement in word error rate, as compared to the word error rate achieved with no normalization, for a collection of utterances. Normalization was performed on a segment-by-segment basis. For each segment, we counted the number of different vowels. Th ensemble of utterances from many speakers will contain a large variety of

phonemes. Small amounts of speech will contain a smaller collection of phonemes. Therefore, their distribution will tend to be slightly different from the average distribution. By counting the number of different vowels, we measure this difference in the variety of phonemes. We consider vowels only because voiced regions of speech, as well as regions where formants are more stable, are more likely to be vowels. We can indeed observe an improvement in word error rate as the number of phonemes increases.

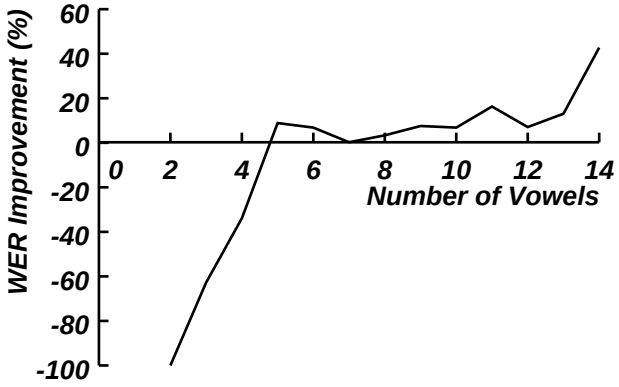

Figure 20. Relative improvement in word error rate for Resource Management (RM1) utterances with different number of vowels. Normalization performed on sentence-by-sentence basis using an affine warping function fitting the minimal points and the medians of F1, F2, and F3 with equal weights.

Figures 21 and 22 show similar plots for the focus conditions F0 and F1 of Hub 4. We can see the same tendency as observed on RM1. Interestingly, neither RM1 nor Hub 4-F0 have segments with less than 4 unique vowels. Hub 4-F1, however, presents a number of utterances with less than 4 vowels, some of them with only one vowel in fact. As mentioned before, there is not much meaning in word error rate in these situations. Discarding the mavericks, we can see that the word error rates also increase with the number of phonemes, thus confirming our initial hypothesis that the contents of the utterances affect the performance of the normalization algorithm.

5.6.3 Normalization on a cluster basis

A possible alternative to the normalization on a sentence basis in the context of Hub 4 is the automatic clustering of segments. Ideally, clusters would group together different segments containing speech from a single speaker. In this section, we examine the effectiveness of such a technique.

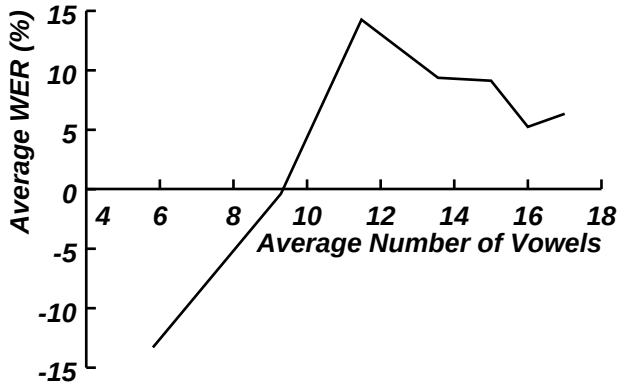

Figure 21. Relative improvement in word error rate for segments from Broadcast News Hub 4, focus condition F0, with different number of vowels. Normalization performed using an affine warping function fitting the minimal points and the medians of F1, F2, and F3 with equal weights.

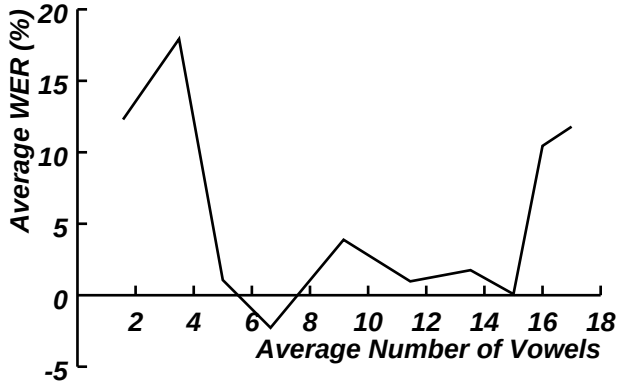

Figure 22. Relative improvement in word error rate for segments from Broadcast News Hub 4, focus condition F1, with different number of vowels. Normalization performed using an affine warping function fitting the minimal points and the medians of F1, F2, and F3 with equal weights.

With RM1, we have perfect knowledge of utterance boundaries and speaker identification, making it a good choice for a controlled experiment. We can employ each utterance separately or group them together so that all speech from one single speaker is employed.

Of course, there are differences between the two tasks. With RM1, we know that each sentence contains utterance from one single speaker, and we know which sentences can be pooled together when we compute distributions of formants on a speaker-by-speaker basis. With Hub 4, segments are automatically selected. This process is not perfect, and it might render segments which contain speech from two different speakers.

Moreover, we can easily group sentences from RM1 on a speaker basis. With Hub 4, we can cluster segments together, but we are restricted to automatic clustering. This automatic clustering might join together segments containing speech from two or more different speakers, thus causing the distribution of formant estimates extracted from these clusters to be inaccurate statistical representations of any of those speakers.

We compare the results obtained with RM1 for recognition employing normalization on a sentence basis to normalization on a speaker basis. These results are compared to those obtained on Hub 4 with normalization on a segment basis and on a cluster basis, although keeping in mind the differences between the perfect knowledge we have on RM1 and the lack of knowledge we have on Hub 4.

With RM1, we have run experiments comparing normalization performed with features extracted on a sentence-per-sentence basis and normalization performed with features extracted on a speaker basis, utilizing all speech available from each speaker. The warping function is a linear function fitting points defined by the medians of the three first formants with weights optimized by MDLR. Results are reported on Table 18.

Normalization of cepstra	WER(%)
No	38.0
Sentence basis	36.4
Speaker basis	29.7

Table 18. Word error rate (WER) on Resource Management (RM1) for normalization on a speaker basis and on a sentence basis. Warping function is the linear function fitting points defined by medians of the first three formants. Weights are optimized by MDLR.

A comparison between normalization on a speaker basis and on a sentence basis shows that we achieve a significant gain by using a larger amount of speech from each speaker. If we employ only one utterance to estimate formants and compute statistics, we achieve a small benefit compared to

no normalization at all, but not as much as if we employ all available speech from each speaker. This result suggests that if we do not have control over how much speech we can use from each speaker and we are forced to perform normalization on a sentence-per-sentence basis the improvement in performance achieved by employing normalization will be limited.

A possible reason for this difference in performance is that the warping function we find with limited data might be far from optimal. The best warping function for each speaker is arguably the same no matter what the utterance might be. However, if there are not enough data in the utterance to provide reliable estimates of formants or estimates far from the speaker's average, the warping function for that particular utterance will be far from the ideal, thus causing the normalization not to perform equally well on average.

We have run experiments with Hub 4 comparing the results with those obtained with RM1, keeping in mind the differences in design between the two tasks. We have examined the benefits of performing normalization with statistics collected from each segment individually or from automatically generated clusters of segments as well as from clusters manually created based on speaker identification. The segments were automatically extracted from the test data, and might be too short to provide reliable statistics. Automatic clustering attempts to cluster segments based on acoustic similarity. Clusters are created without using information about speaker identification. However, automatic clustering might join together segments with speech uttered by different speakers. Normalization per speaker eliminates this problem. However, information about speaker identifications is not available in normal conditions in the Broadcast News evaluation. In any case, we have no control over how much data are available from each speaker. Normalization on speaker basis can only help is there is a large enough amount of speech from each speaker, as explained in Sections 5.6.1 and 5.6.2. The conditions are not exactly the same as the ones for RM1, but we expected to come to an understanding of the issues involved by the comparison.

Normalization for the experiments with Hub 4 was performed with an affine function fitting the points on a plane defined by the medians and minimal points of the first three formants F1, F2, and

F3, plus the origin used as a data point to a total of 7 data points. HMMs were trained using only a subset of the speech on the F0 focus condition for reasons explained in Section 5.2.2. Results are presented on Table 19. Statistical significance was examined within each condition. Results surrounded by thick lines have no statitsically significant difference.

Normalization of cepstra	Condition							
	Overall	F0	F1	F2	F3	F4	F5	FX
No	47.3	31.4	39.7	73.6	54.0	42.9	62.5	67.3
Per sentence	45.5	28.7	39.4	77.1	50.3	39.3	54.0	64.0
Per cluster	45.8	28.8	38.6	76.8	52.2	39.5	55.5	66.5

Table 19. Word error rate on Hub 4 with models created from speech under F0 condition alone. Thick borders identify results with no statistically significant difference over each condition. Normalization per cluster employed clustering performed automatically.

Regarding the results of normalization on a sentence basis and on a cluster basis on Table 19, we see that all sets of results are very similar, contrasting greatly with the comparison between normalization on a sentence basis and on a speaker basis on RM1. This difference suggests that the automatic clustering is ineffective for normalization purposes. Segments are not guaranteed to contain speech from one single speaker. Automatic clustering joins together these less than ideal segments into groups that might contain speech from more than one speaker, not necessarily with similar acoustic characteristics. Automatic clustering might as well result in clusters containing too little speech to allow for a reliable estimation of formant distribution. These experiments indicate the problems we point out render automatic clustering useless in this context.

On the other hand, normalization per speaker does help normalization. Table 20 presents a comparison between the best result under each condition extracted from Table 19 and speaker normalization performed with speaker specific features. Segments of speech were hand selected to assure proper clustering. The result with speaker normalization is systematically better than the best rsult from the previous table for each condition. The overall result, not surprisingly, reflects this ten-

Speaker normalization	Condition							
	Overall	F0	F1	F2	F3	F4	F5	FX
Best from Table 19	45.5	28.7	38.6	73.6	50.3	39.3	54.0	64.0
Hand clustered per speaker	44.4	27.7	38.3	77.4	45.7	37.9	53.7	63.5

Table 20. Word error rate on Hub 4 with models created from speech under F0 condition alone. First line identifies the best result extracted from Table 19. Second line represents normalization performed with speaker specific speech. Identification of speakers and selection of speech segments was manually accomplished.

dency, showing that normalization on a speaker basis is better than normalization on a segment basis. Improvement in performance therefore can be gained by better methods of segment clustering.

Particularly, it is interesting to notice that manual clustering, or normalization on a speaker basis, did help one of the conditions much more than the others. This focus condition is defined as speech with music in background and referred to as F3. We attempted to assess the reasons why clustering particularly benefited this condition. Table 21 presents the number of segments on each condition for each speaker for whom there is some speech under the F3 focus condition. We observe that for most speakers, there is a larger number of segments under other cleaner conditions. Some speakers have speech only under F3 condition, but these speakers have very little speech overall anyway.

We hypothesize that normalization on the focus condition F3 is affected mostly by the systematic introduction of harmonics of music into the collection of formant estimates for the speaker. The availability of formants estimated on cleaner conditions makes the influence of these erroneous formant estimates less relevant. We ellaborate on this issue in Section 5.6.5.

As a comparison, we found the number of segments for all speakers for whom there is speech under the F2 focus condition. These numbers are presented on Table 21. In this case, we can observe that most speakers have speech only on condition F2. Contrary to the previous case, the collection of formant estimates cannot be improved since there are no estimates of formants

Speaker	F0	F1	F2	F3	F4	F5	FX
ABC_PRT_ANNOUNCER	0	0	0	2	0	0	0
BERNARD_SHAW	2	1	0	1	0	0	0
BYRON_MIRANDA	3	0	0	2	0	0	0
DAVID_BRANCACCIO	3	7	0	5	3	0	1
DIANE_SAWYER	0	0	0	2	0	0	0
DONNA_KELLY	7	0	0	1	6	0	1
FILE4_JANEDOE001	0	0	0	2	0	0	0
I960711P_ANNOUNCER1	0	0	0	2	0	0	0
JUDY_WOODRUFF	2	0	0	1	0	0	0
LISA_MULLINS	3	6	0	1	0	0	0
MARY_AMBROSE	6	2	0	3	0	0	4
N960715P_J_DOE001	0	0	0	1	0	0	0
NPR_MKP_ANNOUNCER	0	0	0	1	0	0	0
O960710P_JANEDOE001	1	0	0	2	0	0	2
O960710P_JANEDOE002	0	2	0	2	0	0	0

Table 21. Number of segments by each speaker under different conditions. The speakers on this table are those for whom there was any amount of speech under focus condition F3, defined as speech with music in background. Most of the speakers have a large amount of speech, measured as number of segments, under clean speech conditions in addition to speech under F3 condition.

extracted from clean speech. Systematic errors in the estimation of formants persist when we cluster segments, since these errors are present in the formant estimates of all segments.

5.6.4 Normalization on different Hub 4 focus conditions

Table 19 in Section 5.6.3 presents results regarding different focus conditions that grant further comments. We observe the algorithm brings about an improvement comparable to WSJ on clean planned speech, namely focus conditions F0 and F5. The normalization fails where we would expect it to fail, namely conditions F2 and F3, which are respectively bandlimited speech and

Speaker	F0	F1	F2	F3	F4	F5	FX
BETSY_KEEFER	0	0	1	0	0	0	0
DAVID_SMITH	0	0	1	0	0	0	0
FILE2_JOHNDOE002	0	2	4	0	0	0	0
FILE2_JOHNDOE003	0	0	5	0	0	0	0
I960711P_JOHNDOE002	0	0	4	0	0	0	0
I960711P_JOHNDOE005	0	0	7	0	0	0	0
JOANNE_MILES	0	0	2	0	0	0	0
O960710P_JOHNDOE002	0	0	1	0	0	0	0
P960712_JANEDOE001	0	0	2	0	0	0	0
SUSAN_SWAIN	0	3	1	0	0	0	0
THOMAS_BUCKLY	0	0	1	0	0	0	0

Table 22. Number of segments by each speaker under different conditions. The speakers on this table are those for whom there was any amount of speech under focus condition F2, defined as narrowband speech. A few speakers have any speech at all under clean speech conditions in addition to speech under F2 condition.

speech with music in the background. The other conditions are less affected by the normalization, i.e., the normalization neither brings about a measurable benefit nor fails.

Under conditions F2 and F3, the formant tracker might easily be led into false estimates. Music in background will introduce harmonics to the signal, and the formant tracker will easily mistake these harmonics for formants. The formant estimates in this case remain fairly constant for several frames of speech, thus introducing a large amount of points that do not belong in the distribution. Bandlimited speech will also cause the formant tracker to fail. Specially for female speakers, who tend to have higher formants, it is possible that the third formant will be located in a region that is attenuated by the channel. In this case, a systematic failure of the formant tracker will lead to poor estimates of distributions, which will in turn result in meaningless features for normalization.

Focus condition F4 represents competing speech. Speech under F4 obviously represents a chal-

lenge to the formant tracker. However, differently than speech under F3, the outliers introduced by competing speech are not systematically around the same value. The distribution of formants will be affected, and the fact that normalization fails to perform well confirms it, however, the influence of outliers is not as dramatic as under F3.

Focus condition FX represents speech that cannot be classified into any of the other conditions. It includes narrowband speech, noise in background, planned or spontaneous speech. The fact that normalization does not work is as much within explanation as if it had worked.

We note that normalization does not seem to provide the same improvement on the focus condition F1, spontaneous speech, as it does on F0, planned speech, although both focus conditions are considered clean speech.

As pointed out in Section 5.6.2, normalization will tend to be less effective on shorter sentences. Figure 23 displays the different distributions of segment duration for both focus conditions. The panel includes only segments shorter than 20 seconds.

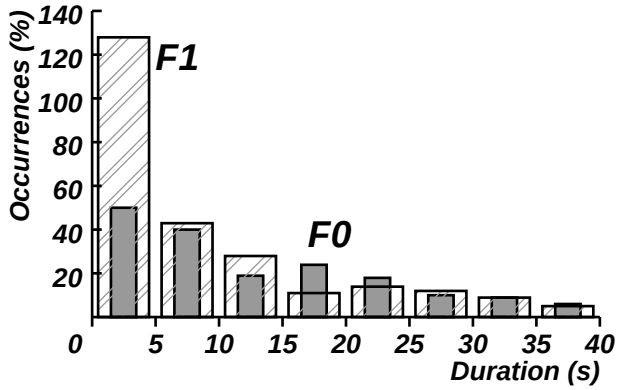

Figure 23. Distribution of segment duration for focus conditions F0 and F1 on Hub 4. Only segments shorter than 20 seconds are included.

We can clearly see that F1 contain a much larger number of sentences with less than 5 seconds of duration. Indeed, the distribution of segment duration seems completely different for both focus conditions. F0 contains a larger percentage of longer segments than F1. Normalization tends to be more effective on longer segments. Moreover, longer segments obviously contain more words,

thus being more relevant for the overall word error rate.

5.6.5 Reliability of formant tracker

The normalization algorithm presented in this thesis depends inherently on the reliability of the formant tracker. In this section, we study the influence of some of the possible kinds of mistakes committed by the formant tracker. We claim that systematic, repetitive errors cause degradation to the system more than sporadic errors. Moreover, we claim that the presence of errors in a collection of formant estimates containing more information about the signal is preferable to an error-free collection containing little information about the signal.

The formant tracker can introduce outliers spread all over the possible range of frequency values, or it can introduce a repetition of a given estimated value, or it can discard correct formant estimates or a frequency range altogether. These errors have different causes and can appear together.

Formant tracking is an errorful process. Automatic estimation of formants produces outliers unless very sophisticated methods are employed. The presence of outliers can be counteracted by carefully hand-selecting frames for which the formant estimates are definitely correct. Hand-selection ensures that no outliers are present in the ensemble of collected data, although it can also cause some valid points to be discarded.

The presence of strong non-varying tones (which can occur, for example, with music in the background) can mislead the formant tracker into selecting the tone as a formant. Since the tone lasts for a long duration compared to the duration of phonemes, this almost-constant value appears in the ensemble of collected data many times, thus skewing the distribution. Moreover, as the value produced by the formant tracker is almost constant, the variance of the data decreases.

The third kind of mistake, discarding formant estimates, appears in two different contexts. It can be a consequence of a very strict outlier detection method, in which case correct estimates are also discarded with most of the outliers. It can also be a consequence of signal characteristics more

than a defect of the formant tracker. In narrowband of telephone speech, for example, some of the "actual" formant tracks go beyond the signal frequency range. Under these circumstances, the formant tracker can label, for example, the second frequency peak (which should be labelled as the second formant) as "third formant", and so on. Even though only one of the formants goes beyond the signal bandwidth, all formant assignments can be wrong.

We ran experiments that tried to assess the influence in recognition performance of each of these types of errors. The first experiment examined the consequence of very strict criteria for detection of outliers. Decisions of whether a given formant estimate was correct or not were based on visual inspection. An automatically-generated estimate was accepted only if it perfectly fit a formant in the spectrogram. We are confident that the final collection of formant estimates contains no outliers, even though many arguably-correct estimates have been discarded. The experiment was performed on the non-English portion of the Broadcast News task. The non-English part is simpler than the English one in that speech is not classified into focus conditions. The results are presented on Table 23.

Normalization	Formant estimate selection	WER(%)
None	N/A	23.2
Per sentence	Automatic	22.9
Per cluster	Automatic	23.0
Per speaker	Manual	28.5

Table 23. Word error rate (WER) on Broadcast News (Hub 4) non-English using automatic or manual selection of frames where formant estimates were considered correct. Automatic selection considers formant estimate to be correct if the probability of the frame being voiced is greater than a threshold. Manual selection is based on visual inspection.

We believe that manual selection of formant estimates performs much worse than automatic classification for reasons similar to those presented in Section 5.6.2. There, we showed that the performance of the normalization algorithm is related to the contents of the utterance. Mismatches between the variety of phonemes uttered by a new speaker and by the ensemble of speakers will

result in distributions of formant estimates that cannot be mapped.

A human selecting “correct” formant estimates based on visual inspection of formant tracks over spectrograms tends to select frames corresponding to some phonemes much more than others, since it is clearer for some phonemes that formant tracks are correct in that region. As a consequence, the collection of phonemes from which formant estimates have been selected is small, resulting in skewed distributions.

Ideally, we should be able to obtain formant estimates that represent all phonemes uniformly, with no outliers whatsoever. However, the interpretation of what an outlier is affects the rigorousness with which we discard estimates. If we have a very strict definition of outliers, the implementation of criteria for detection of outliers will also discard some correct estimates. If we have a loose definition of outliers, we are less likely to discard correct estimates, but we have to cope with the fact that some of the points are wrong estimates. The experiments on Table 23 show that laxer standards work better than strict ones. We restrict these conclusions to outliers that appear sporadically over the whole frequency range.

We hypothesize that occasional outliers do not affect the features, that is, the statistics computed from distributions of formants. Systematic errors, however, cause the distributions to represent something other than the real contents of the utterance. Among examples of situations that can cause systematic errors, we point out narrowband speech and speech over music. As pointed out before, narrowband speech can cause the formant tracker to make completely mistaken assignments of labels, such as assigning the label “second formant” to the “actual” first formant, etc, if one of the “actual” formant goes beyond the frequency range of the signal. Speech over music, on the other hand, can confound the formant tracker. It captures tones, thus altering the distribution of formant estimates. In Section 5.6.3 we pointed out how clustering speech with music in the background with clean speech ameliorates this problem.

In both cases, the distribution of formant estimates is skewed. Table 24 presents the average vari-

ance of the formant estimates of the first three formants for each focus condition on Hub 4 as well as the relative improvement achieved on each condition using the whole development test data provided for the 1997 evaluation.

We observe some degree of dependency between the improvement achieved with normalization and the average variances of formants. A larger variance suggests a larger variety of phonemes in the utterance or ensemble of utterances, since it implies that formant estimates have been produced at a larger frequency range. As pointed out before, since we do not restrict the training set to any small subset of phonemes, a larger number of phonemes in the test set implies a better match between the varieties of phonemes.

Smaller variance on the other hand suggests a smaller variety of phonemes. Additionally, it might be a consequence of systematic errors in the formant estimation. Note that narrowband speech, usually referred to as condition F2, and speech over music, referred to as condition F3, are the two focus conditions with smallest average variance of the third formant and among the smallest of the second formant. Narrowband speech has a cap on how far the third formant can go, and it is not surprising that the variance of the third formant is the smallest for F2. On the other hand, systematic insertion of an almost constant value for a relatively large period of time, as is often the case with music in the background, also causes the variances to decrease. These results confirm the fact that systematic insertion of errors can cause degradation as compared to the presence of sporadic outliers.

We have to deal with the limitations of formant trackers. An extremely strict definition of “outlier” causes important information from the signal to be lost. Therefore, presence of outliers is preferable to loss of correct formant estimates. Systematic errors, however, cause larger degradation. Systematic errors are most often consequence of the signal characteristics rather than a failure of the formant tracker.

5.7. Systems with realistic settings

Condition	Segments	F1	F2	F3	Relative improvement(%)
F0	230	57,862.57	290,601.84	268,164.98	5.6604
F1	346	46,835.86	263,590.26	270,569.70	4.0678
F2	108	45,179.87	246,148.22	179,956.69	2.1912
F3	103	64,129.58	238,779.37	195,188.29	4.3353
F4	133	49,098.91	247,960.03	211,292.09	4.6729
F5	17	59,912.05	263,200.51	380,972.82	8.7533
FX	113	57,700.28	208,731.82	232,016.71	2.6101
Overall	1050	52,444.92	257,389.26	243,458.14	3.8576

Table 24. Mean of variances of formants - Hub 4-97 PE

So far, we have artificially bounded performance of the system so as to study particular issues. We have been limiting the language weight on experiments on Resource Management (RM1) so as to concentrate on the influence of the speaker normalization on the acoustic model. We have limited the training set of Hub 4 so as to take into account speech with high audio quality alone. These artificial settings might create conditions where the normalization might work whereas it might be ineffective in a realistic setting. In this section, we present comparative results between normalization and baseline systems for a realistic system setting.

On RM1, the limiting setting was the language weight set to a very low number so as to disable the language model. In Chapter 4, we learned that we can achieve the best performance with a linear function fitting a set of points with appropriate weights. The set of points or the set of weights can be chosen carefully so as to achieve the best performance. The best set of weights can be selected with MDLR. Alternatively, the best combination of features can be chosen and equally weighed. Normalization was performed on a speaker basis in all speakers for the training and testing data sets. Table 25 presents results with a realistic language weight of 9.5. From the table, we indeed observe that speaker normalization provides a measurable improvement in word error rate for a realistic system.

Points	Weights	WER(%)
None	None	6.6
Min F1, Med F1, F2	Equal	5.7
Med F1, F2, F3	MDLR	5.6

Table 25. Word error rate (WER) on Resource Management (RM1) for normalization using a realistic language weight. Warping function is a linear function. Points and weights are indicated.

With Hub 4, the limiting parameters were the amount and quality of the data used for training, which we restricted to the focus condition F0, clean planned speech. The main findings were that normalization of the training data does not seem to be relevant, and normalization on a segment basis or on a cluster basis yield similar performances. Table 26 presents results of an experiment where we utilized HMMs created from all available speech on the Hub 4 training set. Speech for training did not undergo speaker normalization. Normalization was applied to segments on the test set on a segment by segment basis.

Normalization	Condition							
	Overall	F0	F1	F2	F3	F4	F5	FX
No	29.2	18.4	25.8	43.8	28.1	25.5	36.3	47.6
Yes	28.7	17.0	25.1	45.3	28.1	26.7	28.2	48.1

Table 26. WER on H4 trained on all data

The results for focus conditions F0 and F5 are statistically significantly better with normalization than without it. Results for the other conditions do not present a large difference. We can observe though that the observations set forth in Section 5.6.4 hold true.

5.8. Conclusions

In this chapter we have investigated issues related to the performance of the speaker normalization

algorithm across different tasks. We examined the possible causes of differences in performance. Most experiments on this Thesis were performed using Resource Management (RM1). In this chapter, we also employed Wall Street Journal (WSJ) and Broadcast News (Hub 4).

Initially, we examined speaker adaptation as an alternative approach to reducing speaker variability. Both speaker adaptation and normalization are effective. We focused our attention in finding how well the two approaches would work together. In fact, we found that the combination of speaker normalization and adaptation is beneficial.

WSJ is similar to RM1 in audio quality but rather different in complexity of language and amount of data. We have examined how these differences affect normalization. By limiting the amount of data from which we created HMMs, we have found that with the same amount of speech, normalization tends to be more effective with a smaller number of speakers. A larger number of speakers will produce HMMs with broader output pdfs. New speakers presented to the system will perform better than if the output pdfs were narrower. Speaker normalization has a more effective contribution to performance since it reduces variability between speakers.

On the other hand, we found that with the same number of speakers, normalization of the training data as well as the test data becomes essential if the amount of data per speaker is larger. In this case, reducing variability of speakers in the training set has the advantage of producing models which are not so well tuned to the particular set of speakers in the training set.

Our experiments also show that vocabulary size contributes to the difference in performance between WSJ and RM1. A smaller dictionary reduces the search space by throwing away several paths leading to wrong transcriptions. Speaker normalization successfully transforms the original data so that most of these wrong paths are replaced by correct ones as the top likelihood in the reduced search space.

On Hub 4, however, we need to use a dictionary which is as large as possible, as well as as much

data as available from as many speakers as feasible. Nonetheless, we found similar tendencies between RM1 and Hub 4 regarding a better performance for longer segments of speech. We attempted to cope with the presence of short duration segments by clustering them. This approach proved to be fruitless, since some of the clusters still contained very little data and some contained data from more than one speaker.

Our experiments have shown that speaker normalization brings about some improvement on the Hub 4 focus conditions containing clean planned speech, like F0 and F5. Normalization was ineffective or in fact brought degradation in some conditions involving systematic disturbances to the formant tracking process, like narrowband speech and speech with music in the background. Under condition F1, which is also clean speech, normalization seems to be ineffective on the subset of the data that we employed as the test set. F1 contains a very large number of very short segments. Normalization tends not to work on short segments of data, thus explaining the difference in performance of the normalization algorithm between F0 and F1.

Chapter 6
Summary and Conclusions

6.1. Summary

This Thesis presents a study on speaker normalization using warping functions. We present a method for speaker normalization and examine the effectiveness of this method under several conditions using different databases. Although the experiments on different tasks were performed with our method, we believe the conclusions can be extended to other speaker normalization techniques.

Our method makes use of acoustic features, specifically formants, to define the warping function. We initially compute statistics from distributions of formants. These statistics are the coordinates of points which define the warping function. The warping function thus defined maps the frequency axis of a standard speaker to the frequency axis of a new speaker. No constraint is imposed on the shape of the warping function or on the statistical measurements we use, thus making it a flexible framework to benchmark different choices of shapes and features.

Regarding shapes, our experiments indicate that the specifics of the shape of warping function are largely irrelevant. We found no statistically significant differences when using a linear function or other different shapes. This is not surprising, since other authors have presented speaker normalization techniques using both linear [14][73] and nonlinear [17] warping functions without either function revealing any clear advantage of one over the other.

For choosing features and which statistical measurements to use, we present a method based on Multi Dimensional Linear Regression (MDLR) which allows us to optimize the weights of each feature. As a special case, this method allows us to select a set of optimal features. With this method, we have observed the somewhat curious result that the best features taken in isolation do not necessarily yield the best combinationof features.

We have performed experiments using this technique on the ARPA Resource Management database (RM1) with good results. We observed relative improvements of up to 22% in word error rate (WER), while relative improvements obtained by our implementation of techniques proposed by other authors were between 9% and 17%.

For the Wall Street Journal task (WSJ) and Broadcast News task (Hub 4) we achieved more modest improvements. The investigation of possible reasons for this diminution in improvement in performance for WSJ and Hub 4 brought about some interesting conclusions.

The most obvious difference between RM1 and WSJ is the amount of data, both in terms of number of speakers and amount of data per speaker. We have found that speaker normalization is less effective with a larger number of speakers. A given system will perform better with a new speaker if the output probabilities of Hidden Markov Models (HMMs) are less well tuned to the speakers in the training set. A speaker independent system, for example, will have a better performance for a completely new speaker than a speaker dependent system not trained for that particular speaker. The HMMs are less well tuned to any given speaker if the number of speakers is larger. The system will perform better with a new speaker with these less well tuned models, thus causing the speaker normalization to be less helpful.

We also found that normalization of the training data becomes more important if the amount of data per speaker is large. Speaker normalization attempts to map speakers to a standard speaker. Ideally, if this attempt is fully achieved, a speaker-independent system would in fact be a speaker dependent system exclusively trained for the standard speaker. Therefore, it is not surprising that normalization of the training data makes speaker normalization more effective. It is counterintuitive, however, that normalization of the training data is not helpful if the amount of data per speaker is smaller, as we found in our experiments with WSJ. More data per speaker causes the HMMs to become better tuned to the speakers in the training set. Speaker normalization reduces variability between speakers, reducing the tuning of the HMMs to that set of speakers, and improving performance for new speakers.

Somewhat surprisingly, we found that speaker normalization is more effective with smaller vocabularies. Speaker normalization deals with acoustic characteristics of speakers. The contents of an utterance should be irrelevant. However, the performance of the recognition system improves in the reduced search space obtained with the smaller dictionary. An effective transformation of the data will replace the wrong path from the original search space by a correct path more often than by another wrong path in the reduced search space. A comparison of the relative improvement in performance with speaker normalization between a large and a small vocabulary searches shows that normalization is in fact an effective transformation of the original data.

Some issues, *e.g.* the availability of data from each speaker, are specific to Hub 4. We have made comparative studies where we limit the amount of data used for normalization with RM1 so as to have a sense of how much data we needed for an effective normalization. Not surprisingly, performance increased with the amount of data. In fact, performance achieved with speaker normalization based on very little data tends to be much worse than performance without normalization.

We found a strong correlation between the improvement in performance and the number of unique vowels in the utterance. The longer the utterance, the larger the set of vowels present in the utterance, which is consistent with the improvement being larger when we have more data. A possible reason behind this fact is that the match between the set of vowels in the utterance and the set of vowels “uttered” by the standard speaker increases. The standard speaker represents the average of all speakers collected with a large amount of data. Therefore, we can interpret this averaging as if the standard speaker had “uttered” all possible vowels in the sense that all vowels are statistically represented in the standard speaker.

Better matching of the set of vowels uttered by a new speaker and the set uttered by the standard speaker can be obtained by requiring all speakers to say a particular standard utterance. In this case, the mapping between standard and new speaker will be based on statistics of histograms of formants representing the same sets of phonemes, potentially allowing for shorter utterances.

Clustering of utterances can improve speaker normalization performance in the Hub 4, increasing the amount of usable data from each speaker. While automatic clustering was not effective for us, hand labelled that clustering using speaker identification was helpful. The increase of the amount of data helps speaker normalization, as should be no surprise.

Speaker adaptation is an alternative approach to solving the same problem of speaker variability. We have examined whether the joint employment of speaker adaptation and normalization would be beneficial. Speaker normalization attempts to bring all speakers close to the same standard speaker. Speaker adaptation, on the other hand, attempts to bring the HMMs farther from the average and closer to each particular speaker. The combination of both approaches brings a sizable improvement compared to either approach applied separately. Obviously, neither of the techniques achieves perfect results. Normalization reduces variability between speakers but this variability still exists, thus leaving scope for the adaptation to work. Similarly, adaptation does a good but not perfect job at transforming speaker-independent HMMs into speaker-dependent ones.

Gender normalization, which is nothing but a simplified version of the speaker normalization, has also been studied and shown to be beneficial. We found gender normalization more effective than gendr dependent models in fact. Gender dependent models are restricted to using only part of the training data. Gender normalization reduces variability between speakers and still allows the use of all data to create a single set of models.

We have also investigated the performance of the normalization based on acoustic features applied to narrowband speech. While we were concerned that the frequency range of the narrowband data might be smaller than the range of variation of the formants, thereby attenuating formants in some frames, we have found that normalization does perform well even in these cases.

6.2. Future directions

Every piece of work has to deal with a trade off between cost and performance. The best perfor-

mance we can achieve usually also implies a higher cost in terms of time or computational cost. A sub-optimal solution with respect to performance might be the best solution in some contexts.

In this work, we have attempted to limit the computational cost. One important constraint we imposed is that normalization in this context should be attempted without requiring any run of recognition. However, we believe some improvement can be achieved if we can use the information of location of phonemes or at least contents of an utterance.

We pointed out in Chapter 5 that there is a relation between improvement in performance and duration as well as number of unique vowels in the utterance. It is at least intuitive that the contents of the utterance in terms of variability of phonemes is more relevant than duration. This suggests that with a better match of the set of phonemes uttered by a new speaker and by the standard speaker we can have better choices of warping function.

The warping function is selected based on acoustic features, which are statistical measurements of speaker specific formant distributions. A set of sentences uttered by a speaker might contain only a small number of different phonemes. In this case, the distribution of formants will be skewed. However, if all the speaker utter the same utterance, or if we take into account only one specific set of phonemes, even if the distributions are distorted, they are distorted due to the same influence. Statistics computed from these distorted distributions map to each other better than if we do not consider these effects.

The standard speaker is an average of speaker. Therefore, the phonemes “uttered” by the standard speaker are of course an ensemble of the phonemes uttered by all speakers in this average. A better match between the set of phonemes uttered by a new speaker and the set from the standard speaker can be achieved by requesting all speakers to say the same utterance, for example, an enrollment utterance. In contexts where this is not possible, we can make use of a first initial pass of recognition where we locate the phonemes in an utterance and select a subset of the phonemes present in the utterance.

Potentially, this approach can benefit the normalization tchnique as it is now. Moreover, since we will be attempting to collect only a specific set of phonemes, we might be able to achieve as large an improvement with very short sentences, provided they contain the required set of phonemes.

Another development which makes use of an initial pass of recognition regards normalization diversely applied to different regions of speech. Since normalization attempts to model differences due to different vocal tract lengths, it intuitively does not make much sense to apply it to phonemes produced in obstructions of the vocal tract far from the vocal chords, such as fricatives. An initial phoneme-labelling of the utterance could flag regions where normalization should not be applied.

An additional extension of this work makes use of a Bayesian approach to the linear regression estimates. We found in Chapter 4 that gender normalization provides substantial improvement as compared to no normalization. A Bayesian approach would take advantage of these gender specific warping functions. If no further speaker specific information is available, the linear regression estimates of the slope of the linear warping function would be simply the gender specific slope. As more data become available, the linear regression estimates gets modified so as to make use of this new piece of information. The Bayesian approach has importance under a practical point of view. Since gender normalization gives improvement at almost no cost, it can be used if no other source of information is available.

Appendix A
Derivation of Linear Regression Equations

A.1. Least-squares solution to warping function

In this section, we present a brief description of the main mathematical tool employed in the previous chapters, linear regression. Linear regression gives an estimate of the coefficients of a straight line that best fits a collection of datapoints. Although this is a well known technique, we present the derivation in this appendix since we use one of the results often.

The warping function is a curve that will fit points corresponding to features on a plane. Let x_k be the k-th feature of a standard speaker, and let y_k be the k-th feature of a speaker we intend to normalize to the standard speaker. Assuming that each feature is independent of the others, and that the distribution of measurements of a feature is Gaussian, we can compute the probability of observing the feature y_k as

$$p(y_k) = N(ax_k + b, \sigma_k^2) \tag{29}$$

where $ax_k + b$ is the expected value of y_k, that is, the estimated value of y_k according to the linear model we intend to fit. Because the features are independent, the joint probability is the product of the corresponding probabilities. This product can be written as the product of two factors, with one of them independent of the parameters a or b:

$$p(y_1, y_2, y_3, \ldots) = C\prod_k \exp\left(\frac{-[y_k - (ax_k + b)]^2}{2\sigma_k^2}\right) \tag{30}$$

To maximize the previous expression, it is more convenient if we take the logarithm on both sides, thus converting the product of terms into a summation of terms, and getting rid of the exponentiation. Considering that we intend to maximize the expression with respect to a and b, we can also eliminate constants and variables that do not depend on these variables. Maximization of the previous expression therefore corresponds to the minimization of:

$$\sum_k \frac{[y_k - (ax_k + b)]^2}{\sigma_k^2} \tag{31}$$

Alternatively, we could have chosen to minimize the distance between the estimate $\hat{y}_k$ according to the model $y = ax + b$ and the measure y_k, weighing the distance between the estimate $\hat{y}_k$ and the point (x_k, y_k) with a weight W_k. In this case, we could find the parameters a and b by minimizing:

$$\sum_k (y_k - \hat{y}_k)^2 W_k \tag{32}$$

or explicitly in terms of the variables and data points:

$$\sum_k [y_k - (ax_k + b)]^2 W_k \tag{33}$$

The similarity between Equations (31) and (33) leads us to interpret the former as a special case of the latter when we set the weight W_k to the inverse of the variance $1/\sigma_k^2$.

We continue the minimization of Expression (33). Taking the derivatives with respect to a and b and equating the derivatives to zero, we obtain:

$$a\sum_k x_k^2 W_k + b\sum_k x_k W_k = \sum_k x_k y_k W_k \tag{34}$$

and

$$a\sum_k x_k W_k + b\sum_k W_k = \sum_k y_k W_k \tag{35}$$

Notice that the weights W_k are not necessarily derived from statistical treatment of the data, but may be obtained by other criteria. These criteria depend entirely on the application.

The solutions to the previous set of equations can be easily shown to be:

$$a = \frac{\sum_i x_i \sum_j x_j y_j W_j - \sum_i x_i W_i \sum_j y_j W_j}{\sum_i W_i \sum_j x_j^2 W_j - \left(\sum_i x_i W_i\right)^2} \tag{36}$$

$$b = \frac{\sum_i x_i^2 W_i \sum_j y_j W_j - \sum_i x_i W_i \sum_j x_j y_j W_j}{\sum_i W_i \sum_j x_j^2 W_j - \left(\sum_i x_i W_i\right)^2} \tag{37}$$

An important special case is the model $y = ax$, obtained by setting b equal to zero in Equations (29) and (30). The solution for a turns out to be:

$$a = \frac{\sum_k x_k y_k W_k}{\sum_k x_k^2 W_k} \tag{38}$$

This special case is used throughout this work. Based on Equation, we derive weights and optimal combinations of features as described in Chapter 4.

References

[1] Acero, A., *Acoustic and Environmental Robustness in Automatic Speech Recognition*, Kluwer Academic Publishers, Boston, 1993.

[2] Alleva, F., Huang, X., Hwang, M., "An Improved Search Algorithm for Continuous Speech Recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 2, p. II 307-310, May 1993.

[3] Andrews, D.R., Atrubin, A.J., Hu, K.C., "The IBM 1975 optical page reader Part III: Recognition logic development", *IBM Journal of Research and Development*, v. 12, n. 5, p. 364-371, Sept. 1968.

[4] Atal, B.S., "Automatic Recognition os Speakers from Their Voices", *Proceedings of the IEEE*, v. 64, n. 4, p. 460-475, April 1976.

[5] Bahl, L., Jelinek, F., Mercer, R., "A Maximum Likelihood Approach to Continuous Speech Recognition" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. PAMI-5, n. 2, p. 179-190, March 1983.

[6] Baird, J. C., Noma, E., *Fundamentals of Scaling and Psychophysics*, Wiley Series in Behavior, John Wiley and Sons, Inc., New York, 1978.

[7] Baker, J., *Stochastic Modeling as a Means of Automatic Speech Recognition*, Ph.D. Dissertation, Carnegie Mellon University, April 1975.

[8] Bakis, R., "Continuous Speech Recognition via Centisecond Acoustic States", *91st Meeting of the Acoustical Society of America*, April, 1976.

[9] Baum, L., "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes", *Inequalities*, v. 3,p. 1-8, 1972.

[10] Bevington, P., *Data Reduction and Error Analysis for the Physical Sciences*, McGraw Hill, New York, 1992.

[11] Broad, D. J., Clermont, F., "Formant Estimation By Linear Transformation Of The LPC Cepstrum", *J. Acoust. Soc. Am.*, v. 86, n. 5, p. 2013-2017, 1989.

[12] Chow, Y., Dunham, M., Kimball, O., Krasner, M., Kubala, F., Makhoul, J., Price, P., Roucos, S., Schwartz, R., "BYBLOS: The BBN Continuous Speech Recognition System", *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p.89-92, April 1987.

[13] Claes, T., Dologlou, I, ten Bosch, L., Van Compernolle, D., "New Transformations of Cepstral Parameters for Automatic Vocal Tract Length Normalization in Speech Recognition", *5th European Conference on Speech Communication and Technology*, v. 3, p. 1363-1366, Sept. 1997.

[14] Cohen, J., Kamm, T., Andreou, A.,"An Experiment In Systematic Speaker Variability", *Final Day Rev., DoD Speech Workshop on Robust Speech Recognition*, 1994.

[15] Davis, S., Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP-28, n. 4, p. 357-366, Aug. 1980.

[16] Duda, R. O., Hart, p. E., *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 1973.

[17] Eide, E., Gish, H., "A Parametric Approach To Vocal Tract Length Normalization", *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p. 346-348, May 1996.

[18] Entropic Research Laboratory, Inc., *Waves+*, 1997.

[19] Fant, G., *Speech Sounds and Features*, MIT Press, Cambridge, 1973.

[20] Fukada, T., Sagisaka, Y., "Speaker Normalized Acoustic Modeling Based on 3-D Viterbi Decoding", *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p. 437-440, May 1998.

[21] Garofolo, J., Fiscus, J., Fisher, W., "Design And Preparation Of The 1996 Hub-4 Broadcast News Benchmark Test Corpora", *Proc. DARPA Speech Recognition Workshop*, p. 15-21, Feb. 1997.

[22] Gauvain, J.-L., Lee, C.-H., "Maximum A Posteriori Estimation For Multivariate Gaussian Mixture Observations Of Markov Chains", *IEEE Trans. on Speech and Audio Processing*, v. 2, n. 2, p. 291-298, April 1994.

[23] Gillick, L., Cox, S., "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p. 532-535, May 1989.

[24] Goldstein, U., "Speaker-identifying Features Based on Formant Tracks", *Journal of the Acoustical Society of America*, v. 59, n. 1, p. 176-182, Jan. 1976.

[25] Gouvêa, E. B., Stern, R. M., "Speaker Normalization through Formant-Based Warping of the Frequency Scale", *5th European Conference on Speech Communication and Technology*, v. 3, p. 1139-1142, Sept. 1997.

[26] Gouvêa, E. B., Moreno, p. J., Raj, B., Stern, R. M., "Adaptation and Compensation: Approaches to Microphone and Speaker Independence in Automatic Speech Recognition", *Proc. DARPA Speech Recognition Workshop*, p. 87-92, Feb. 1996.

[27] Graff, D., "The 1996 Broadcast News Speech And Language-model Corpus", *Proc. DARPA Speech Recognition Workshop*, p. 11-14, Feb. 1997.

[28] Huang, X., Alleva, F., Hon, H., Hwang, M., Lee, K., Rosenfeld, R., "The SPHINX-II Speech Recognition System: An Overview", *Computer Speech and Language*, v. 2, p. 137-148, 1993.

[29] Hwang, M., Huang, X., "Shared-Distribution Hidden Markov Models for Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, v. 1, n. 4, p. 414-420, 1993.

[30] Jelinek, F., *Statistical Methods For Speech Recognition*, MIT Press, Cambridge, 1998.

[31] Juang, B., Rabiner, L., "Mixture Autoregressive Hidden Markov Models for Speech Signals", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP-33, p. 1404-1413, 1985.

[32] Kang, G. S., Coulter, D. C., "600 Bits/Second Voice Digitizer (Linear Predictive Formant Vocoder)", Naval Res. Lab., 1976.

[33] Kubala, F., “Design of the 1994 CSR Benchmark Tests”, *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, p. 41-46, Jan. 1995.

[34] Lasry, M. J., Stern, R. M., “A Posteriori Estimation Of Correlated Jointly Gaussian Mean Vectors”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, v. 6, n. 4, p. 530-535, July 1984.

[35] Lee, C.-H., Lin, C.-H., Juang, B.-I., “A Study On Speaker Adaptation Of The Parameters Of Continuous Density Hidden Markov Models”, *IEEE Trans. on Signal Processing*, v. 39, n. 4, p. 806-814, April 1991.

[36] Lee, C., Rabiner, L., Pieraccini, R., Wilpon, J., “Acoustic Modeling for Large Vocabulary Speech Recognition” *Computer Speech and Language*, v. 4, n. 2, p. 127-165, April 1990.

[37] Lee, K., *Automatic Speech Recognition : The Development Of The Sphinx System*, Kluwer Academic Publishers, Boston, 1989.

[38] Lee, K. Hon, H., “Large-Vocabulary Speaker-Independent Continuous Speech Recognition”, *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p. 123-126, April 1988.

[39] Lee, K., Hon, H., Reddy, R., “An Overview Of The SPHINX Speech Recognition”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP-28, n. 1, p. 35-45, Jan. 1990.

[40] Lee, L., Rose, R. C., “Speaker Normalization Using Efficient Frequency Warping Procedures”, *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p. 353-356, May 1996.

[41] Leggetter, C. J., Woodland, p. C., “Speaker Adaptation Of HMMs Using Linear Regression”, *Technical Report CUED/F-INFENG/ TR. 181*, Cambridge University Engineering Department, Cambridge, June 1994.

[42] Levinson, S., Rabiner, L., Sondhi, M., “An Introduction to the Application of the Theory of Probabilistic Function on a Markov Process to Automatic Speech Recognition”, *Bell System Technical Journal*, v. 62, n. 4, April 1983.

[43] Lin, Q., Che, C., “Normalizing The Vocal Tract Length For Speaker Independent Speech Recognition”, *IEEE Signal Processing Letters*, v. 2, n. 11, p. 201-203, Nov. 1995.

[44] Lincoln, M., Cox, S., Ringland, S., “A Fast Method of Speaker Normalisation using Formant Estimation”, *5th European Conference on Speech Communication and Technology*, v. 4, p. 2095-2098, Sept. 1997.

[45] Macdonald, J. R., Thompson, W. J., “Least-Squares Fitting when Both Variables Contain Errors: Pitfalls and Possibilities”, *Am. J. Phys.*, v. 60, n. 1, p. 66-73, 1992.

[46] Makhoul, J., “Linear Prediction: a Tutorial Review”, *Proceedings of the IEEE*, v. 63, n. 4, p. 561-580, April 1975.

[47] Markel, J., Gray, A., *Linear Prediction of Speech*, Springer-Verlag, New York, 1976.

[48] McCandless, S. S., “An Algorithm for Automatic Formant Extraction Using Linear Prediction Spectra”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP-22, n. 2, p. 135-141. April 1974.

[49] Moreno, P. J., *Speech Recognition in Noisy Environments*, Ph.D. Dissertation, Carnegie Mellon University, May 1996.

[50] Nilsson, N., *Principles of Artificial Intelligence*, Tioga Publishing Co., Palo Alto, 1980.

[51] Oppenheim, A. V., Schafer, R. W., *Discrete-Time Signal Processing*, Prentice Hall Signal Processing Series, Englewood Cliffs, 1989.

[52] Oppenheim, A. V., Johnson, D. H., "Discrete Representation of Signals", *Proceedings of the IEEE*, v. 33, p. 681-691, 1972.

[53] Pallet, D.S., Fisher, W.M., Fiscus, J.G., "Tools for the Analysis of Benchmark Speech Recognition Tests", *1990 IEEE International Conference on Acoustics, Speech and Signal Processing Conference Proceedings*, v. 1, p. 97-100, April 1990.

[54] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, Third Edition, McGraw Hill, Inc., New York, 1991.

[55] Parikh, V., Raj, B., Stern, R., "Speaker Adaptation and Environmental Compensation for the 1996 Broadcast News Task", *Proc. DARPA Speech Recognition Workshop*, p. 119-122, Feb. 1997.

[56] Parsons, T., *Voice and Speech Processing*, McGraw-Hill Book Company, New York, 1986.

[57] Paul, D., Baker, J., "The Design of the Wall Street Journal-based CSR Corpus", *Proceedings of ARPA Speech and Natural Language Workshop*, p. 357-362, Feb. 1992.

[58] Peterson, G. E., Barney, H. L., "Control Methods Used In A Study Of Vowels", *J. Acoust. Soc. Am.*, v. 24, p. 175-184, 1952.

[59] Picone, J., Doddington, G., Pallett, D., "Phone-mediated Word Alignment for Speech Recognition Evaluation", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP-38, n. 3, p. 559-562, March 1990.

[60] Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., Stern, R., Thayer, E., "The 1996 Hub-4 Sphinx-3 System", *Proc. DARPA Speech Recognition Workshop*, p. 85-89, Feb. 1997.

[61] Rabiner, L., Juang, B., *Fundamentals of Speech Recognition*, Prentice Hall Signal Processing Series, Englewood Cliffs, 1993.

[62] Ravishankar, M., "Some Results on Search Complexity vs. Accuracy", *Proc. DARPA Speech Recognition Workshop*, p. 89-92, Feb. 1997.

[63] Reed, B. C., "Linear Least-Squares Fits with Errors in Both Coordinates. II: Comments on Parameter Variances", *Am. J. Phys.*, v. 60, n. 1, p. 59-62, 1992.

[64] Sambur, M., "Selection of Acoustic Features for Speaker Identification", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, v. ASSP-23, n. 2, p. 176-182, April 1975.

[65] Schiffman, S. S., Reynolds, M. L., Young, F. W., *Introduction to Multidimensional Scaling*, Academic Press, New York, 1981.

[66] Schwartz, R., Chow, Y., "The Optimal N-Best Algorithm: An Efficient Procedure for Finding Multiple Sentence Hypotheses", *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p. 81-84, April 1990.

[67] Schwartz, R., Chow, Y., Roucos, S., Krasner, M., Makhoul, J., "Improved Hidden Markov

Modeling of Phonemes for Continuous Speech Recognition", *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 3, p. 35.6.1-35.6.4, March 1984.

[68] Schwartz, R., Jin, H., Kubala, F., Matsoukas, S., "Modeling Those F-Conditions – Or Not", *Proc. DARPA Speech Recognition Workshop*, p. 115-118, Feb. 1997.

[69] Shahshahani, B. M., "A Markov Random Field Approach to Bayesian Speaker Adaptation", *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 2, p. 697-700, May 1996.

[70] Stevens, S. S., Volkmann, J., "The relation of Pitch of Frequency: A Revised Scale", *Am. J. Psychol*., v. 53, p. 329-353, 1940.

[71] Takahashi, J.-I., Sagayama, S., "Vector-Field-Smoothed Bayesian Learning for Incremental Speaker Adaptation", *1995 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p. 696-699, May 1995.

[72] Viterbi, A., "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm", *IEEE Transactions on Information Theory*, v. IT-13, p. 260-269, 1967.

[73] Wegmann, S., McAllaster, D., Orloff, J., Peskins, B., "Speaker Normalization On Conversational Telephone Speech", *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p. 339-341, May 1996.

[74] Young, S. J., Woodland,p. C., *HTK Version 1.5: User, Reference and Programmer Manual*, Cambridge University Engineering Dept., Speech Group, 1993.

[75] Zahorian, S. A., Jagharghi, A. J., "Speaker Normalization Of Static And Dynamic Vowel Spectral Features", *J. Acoust. Soc. Am.*, v. 90, n. 1, p. 67-75, 1991.

[76] Zavaliagkos, G., *Maximum A Posteriori Adaptation Techniques For Speech Recognition*, Ph.D. Thesis, Northeastern University, May 1995.

[77] Zhan, P., Westphal, M., "Speaker Normalization Based On Frequency Warping", *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, v. 1, p. 1039-1042, May 1997.