AUTOMATIC DETECTION OF CORRUPT SPECTROGRAPHIC FEATURES FOR ROBUST SPEECH RECOGNITION

Michael L. Seltzer

Department of Electrical and Computer Engineering

Carnegie Mellon University

Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Electrical and Computer Engineering

Pittsburgh, Pennsylvania

May 2000

Abstract

When speech is corrupted by noise, the performance of automatic speech recognition systems degrades significantly. There have been many algorithms proposed that compensate for the negative effects of noise in speech and greatly improve recognition accuracy. However, these methods assume that the corrupting noise is stationary. If the noise is non-stationary, these methods fail. A promising new group of compensation algorithms have recently emerged which do not have this restriction on the noise characteristics. These methods operate on the notion that noise affects different frequency bands of speech differently depending on the relative energies of the speech and the noise at each time-frequency location. In a spectrographic display of noisy speech, regions of low SNR will be more corrupt than regions of high SNR. Low SNR regions of a spectrogram are considered to be “missing” or “unreliable” and are removed from the spectrogram. Noise compensation is carried out by either estimating the missing regions from the remaining regions in some manner prior to recognition, or by performing recognition directly on incomplete spectrograms. These techniques clearly require a "spectrographic mask" which accurately labels the reliable and unreliable regions of a spectrogram. Currently, there are no good techniques for accurately estimating such a mask. The methods that have been used so far rely on the assumptions about the interfering noise such as global SNR or stationarity, and fail when these assumptions do not hold.

In this thesis we have designed a classifier for spectrographic mask estimation that does not make any assumptions about the characteristics of the interfering noise. The classification is performed using features that exploit the intrinsic characteristics of the speech itself. A separate mask-estimation classifier has been designed for voiced and unvoiced spectrographic regions of speech. In voiced regions, features that exploit the inherent harmonicity and the distinctive spectral contour of voiced speech are utilized in mask estimation. In order to derive the features in the voiced speech regions, a new pitch tracking algorithm is proposed which is more robust to noise than other methods. While there is no harmonicity present in unvoiced speech, it also has a distinct spectral contour which forms the basis of the feature set used for the unvoiced spectrographic regions.

Experiments in mask estimation were performed on speech corrupted by white noise and speech corrupted by music. The masks generated by the classifier are evaluated in two ways. The estimated masks are

compared with oracle masks generated from full *a priori* knowledge of the corrupting noise to determine the classifier accuracy. The masks are then passed to two different missing feature reconstruction methods to determine the improvement in recognition accuracy.

The spectrographic masks generated by the classifier result in recognition accuracy that is comparable to the best previously reported method of mask estimation for speech corrupted by white noise. On speech corrupted by music, which is highly non-stationary, the classifier-based masks produce significant and consistent improvements in recognition accuracy. No other reported mask estimation method to has been able to do so.

Acknowledgements

This thesis would not have been possible without the support and encouragement of my advisor, Rich Stern. After spending two years in the semiconductor test industry, I arrived at CMU with no speech recognition background whatsoever. Rich welcomed me into the group and gave me the opportunity to learn and grow as a researcher. He allowed me the freedom to pursue my own ideas and never let me settle for mediocrity.

I also owe a tremendous debt of gratitude to the senior members of the robust speech group. Bhiksha, Juan, and Sam-Joo never grew tired of answering my endless stream of question and were always available for discussions, suggestions, and advice. Over the last two years, their support and camaraderie have been invaluable to me.

Thanks to Tsuhan Chen for graciously agreeing to be the second reader of this thesis, and providing valuable feedback and suggestions.

Finally, I would like to thank my parents, without whom none of this would be possible. They are an unbelievable inspiration to me. I thank them for instilling the love of learning in me and for teaching me about integrity, dignity, and respect. Their boundless love and support has truly been a rock to me.

This thesis is dedicated to my parents.

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

Automatic speech recognition (ASR) is statistical pattern classification problem. A sequence of feature vectors derived from short windowed segments of speech is input to an ASR system where each feature vector is labeled as belonging to one of a set of many possible sound classes. The decision as to which is the most likely sound class is based on the distributions of the feature vectors belonging to each sound class which are learned from a corpus of training speech. Using these distributions, the sound class that is the most likely to have generated the input feature vector is the class label for that input segment.

When speech is corrupted by noise, speech recognition accuracy degrades [1]. The feature vectors generated from noisy speech are no longer similar to the class distributions learned from the training data. If the noise is stationary, this degradation can be minimized by retraining the system on speech that has been corrupted with the same level of noise as the speech being recognized. However, because of the noise, there is inherently more variability in the training data, and as a result, the variance of the distributions of the sound classes increases [23]. Even with retraining, this “broadening”of the class distributions leads to increased classification errors over the case where both the training and test speech are both clean. When the noise is non-stationary, retraining the system on noise-corrupted speech does not help. The non-stationarity of the noise implies that the noise used to degrade the training speech will not necessarily be representative of the noise that corrupts the test speech. As a result, there can still be a large mismatch between the feature vectors of the speech to be recognized and the distributions of the sound classes learned during training.

Techniques which attempt to reduce the effects of noise on speech recognition performance are called *compensation methods*. Several compensation methods proposed in the literature have been quite successful at improving the performance of speech recognition systems on noise-corrupted speech. Some methods such as Codeword Dependent Cepstral Normalization [1] and Spectral Subtraction [5] attempt to “clean” the incoming noisy data before it reaches the recognizer for classification. Others, such as Parallel Model Combination [18] and Maximum Likelihood Linear Regression [24] modify the class distributions within the recognizer to alleviate the negative effects of the noise.

However, all of these methods assume that the corrupting noise is stationary and that the effect of the noise can be represented by a linear transformation of some kind. Therefore, while the above methods are very successful if these conditions are met, they fail when the corrupting noise is non-stationary.

A promising new group of compensation methods have emerged which do not have this stationarity restriction. They are based on the notion that in noisy conditions, the human auditory system preferentially processes the high energy components of the speech signal while suppressing the weaker ones [31]. These new methods mimic this idea by suppressing the low SNR components of a speech signal in favor of the components of high SNR in some fashion.

Missing feature methods [7][27][43] are one such class of compensation techniques. They attempt to emphasize the high SNR components of speech not just in frequency but in time as well. Speech is transformed to a two-dimensional time-frequency display where each pixel location represents the energy at that time-frequency location. When speech is corrupted by additive noise, different pixels in the spectrogram will be affected differently, depending on the relative energies of the speech and the noise at each particular time-frequency location. Missing feature approaches effectively erase the low SNR pixels from the spectrogram. This is done by creating a binary mask based on local SNR that labels each time-frequency location in the spectrogram as “present” (meaning reliable) or “missing” (meaning corrupt). This spectrographic mask is then applied to the spectrogram and recognition is either performed on the remaining incomplete spectrogram, or the erased pixels are reconstructed in some fashion and then recognition is performed on the reconstructed complete spectrogram.

Missing feature methods have been shown to be very successful at compensating for noise in speech in both stationary and non-stationary noise conditions *when the spectrographic mask labeling every time-frequency location as reliable or unreliable is known a priori* [43][10]. However, when the masks are unknown, these techniques are unusable.

In this thesis, we attempt to design a classifier that automatically determines whether each time-frequency location in a spectrogram is reliable or unreliable and then generates a spectrographic mask accordingly. Because the missing feature methods are able to compensate for both stationary and non-stationary noises, the classifier is designed to exploit the intrinsic features of speech itself while making minimal

assumptions about the corrupting noise. We apply the resulting estimated masks to two missing feature compensation methods to evaluate their effectiveness for use with missing feature approaches to robust speech recognition.

This thesis is organized as follows:

Chapter 2 presents the missing-feature reconstruction paradigm. It describes how the effect of noise on speech can be modeled as missing features of a spectrogram and then explains two missing-feature reconstruction methods that will be used in this thesis. Chapter 3 examines the mask estimation problem in more detail and reports some previous work in this area by other researchers. Chapter 4 describes a novel pitch detection algorithm that is required for certain pitch-dependent classifier features. Chapter 5 discusses the feature extraction procedure. The classification features for voiced and unvoiced speech used to estimate the spectrographic masks are described. In Chapter 6, the classification strategy is presented, and mask estimation accuracy is analyzed. Results on speech recognition accuracy when the estimated masks are used with missing-feature methods for noise compensation are reported. Chapter 7 summarizes our conclusions and suggestions for future work are proposed.

Chapter 2
Spectrograms and Missing Feature Compensation

2.1 Introduction

Missing feature methods compensate for noise in speech by utilizing only reliable, high SNR regions of time-frequency representations of the noisy speech while ignoring the corrupt low SNR regions. In this chapter, we will describe the spectrogram, the most common time-frequency representation of speech. We will discuss the effect of noise on a spectrogram and how regions of low SNR can be modeled as missing features by applying a spectrographic mask that effectively removes them from the picture. We will then describe two missing feature methods that enhance the incomplete spectrograms and dramatically improve speech recognition performance if the spectrographic masks are completely known *a priori*.

2.2 The Mel Spectrogram

A spectrogram is a two-dimensional representation of a speech signal. Time is displayed on the horizontal axis and frequency on the vertical axis. Each time-frequency location in the spectrogram represents the power $P_x(l, \omega)$ in the signal at time l and frequency ω, as given by Equation (2.1), where $X(l, \omega)$ is the Short-Time Fourier Transform (STFT) [42] of length $2L + 1$ of the signal $x[n]$.

$$P_x(l, \omega) = \frac{1}{2L + 1}\left|X(l, \omega)\right|^2 \tag{2.1}$$

While the value of $P_x(l, \omega)$ can be computed at every time step, it is more common to compute the $2L + 1$ point STFT at every L^{th} sample in the sequence. When the sequence of STFT vectors computed every L samples are arranged in consecutive columns, they comprise a two-dimensional picture that describes energy of the signal across time and frequency. The intensity of each point represents the value of $\log(P_x(l, \omega))$, *i.e.* the logarithm of the signal power at time l and frequency ω. Figure 2.1 shows a typical spectrogram.

The missing feature compensation methods used to evaluate the work in this thesis all operate on a vari-

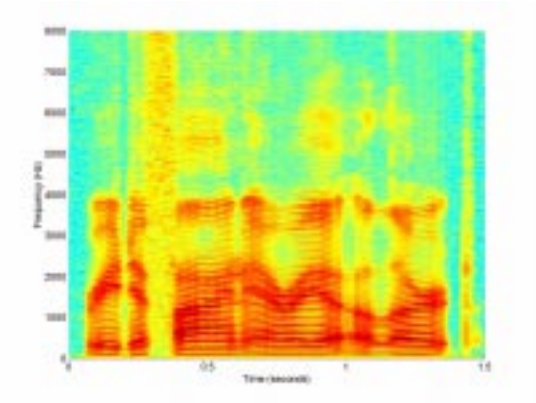


Figure 2.1 This figure shows a spectrogram of the utterance “Redefine Red Alert”. The length of the analysis window was 30ms. Adjacent windows were overlapped by 5ms.

ant of the traditional spectrogram known as the mel spectrogram. The mel spectrum captures the power at the output of a bank of bandpass filters. In practice, the filters are overlapping triangles of unit area with increased bandwidth at higher frequencies applied to the power spectrum of the signal [35]. The composite magnitude response of twenty mel filters is shown in Figure 2.2.

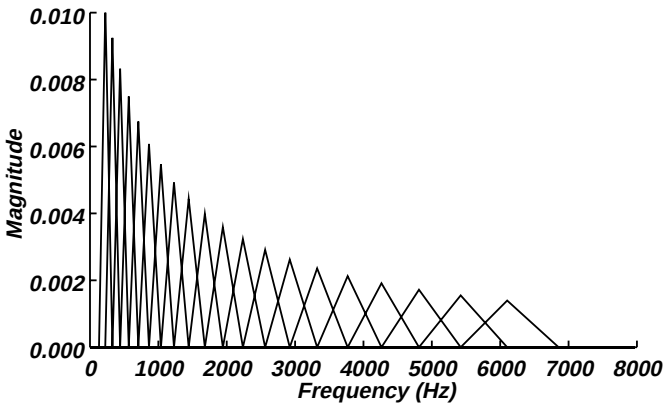


Figure 2.2 The composite frequency response of 20 Mel filters.

In the mel spectrogram, each pixel $P_x(l,k)$ represents the power at time l at the output of the k^{th} mel filter, given by

$$P_x(l,k) = \sum_{j=0}^{2L} m_k(j)\left|X(l,j)\right|^2 \tag{2.2}$$

where $m_k(j)$ are the DFT coefficients of the impulse response of the k^{th} mel filter and $X(l,j)$ is the j^{th}

frequency component of the DFT of the l^{th} analysis window of the speech signal $x[n]$. The value at each time-frequency location in the mel spectrogram is given by $\log(P_x(l,k))$, *i.e.* the logarithm of the power at the output of mel filter *k* in frame *l*. Thus, the mel spectrogram consists of a sequence of log mel-spectral vectors, each of which has K components, where K is the total number of mel filters. The mel spectrogram can be viewed as an spectrally smeared version of the classical spectrogram. Frames are typically 25 ms long, and overlap by 15 ms for the mel-spectral representation. Figure 2.3 shows the mel spectrogram representation of the same utterance used in Figure 2.1.

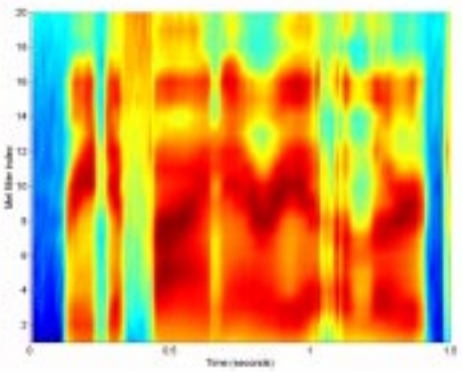


Figure 2.3 The Mel spectrogram of the utterance “Redefine Red Alert”. 20 Mel filters covering the frequency range from 150 Hz to 8 KHz were used. The analysis windows were 25ms long. Adjacent frames were overlapped by 15ms.

2.3 Modeling Noise as Missing Spectrographic Features

Missing feature compensation methods treat spectrographic regions that have been most corrupt by noise as “unreliable” or “missing” and remove them from the spectrogram. This notion therefore requires a measure of “reliability”. A logical choice for such a measure is the local SNR. Regions of high SNR can be considered reliable while regions of low SNR can be considered corrupt. If speech is corrupted with additive noise, we have

$$y[l] = x[l] + n[l] \tag{2.3}$$

where $y[l]$ is the noisy speech signal, $x[l]$ is the clean speech signal, and $n[l]$ is the noise that has been added to the signal. It can be shown [43] that the SNR at every time-frequency location in the mel spectrogram of $y[l]$ can be computed from the mel spectrograms of $x[l]$ and $n[l]$ if $x[l]$ and $n[l]$ are known

a priori. This local SNR is given by Equation (2.4).

$$SNR(l, k) = 10\log 10\left(\frac{P_x(l, \omega_k)}{P_n(l, \omega_k)}\right) \tag{2.4}$$

If the speech and the noise are both completely known *a priori*, we can calculate the local SNR at every (l, k) for the noisy speech. All of the spectrographic elements that have an SNR below a fixed threshold. can then be removed. Figure 2.4 shows a mel spectrogram of an utterance of noisy speech and Figure 2.5 shows the same spectrogram when all time-frequency locations with a local SNR of less than 0dB have been erased.

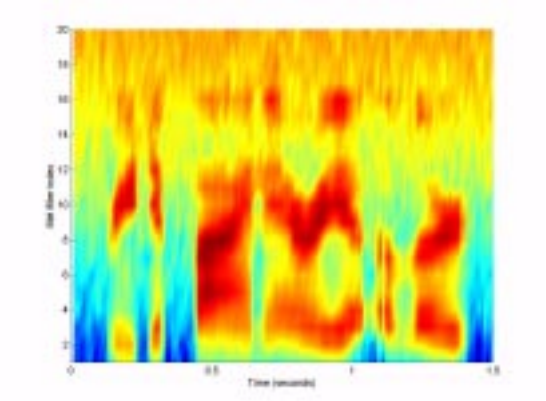


Figure 2.4 Mel spectrogram of the utterance “Redefine Red Alert” when the speech has been corrupted with white noise to 10dB.

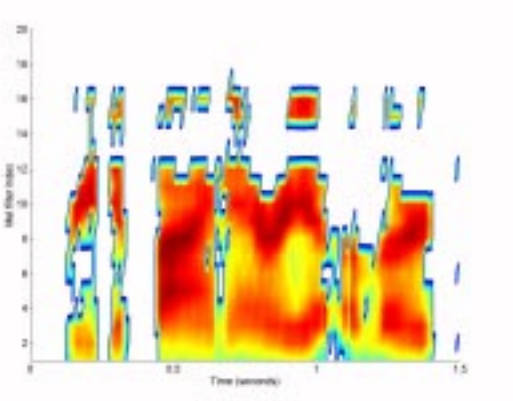


Figure 2.5 The same Mel spectrogram but all regions with a local SNR of less than 0dB have been deleted. The white regions of the figure represent the deleted regions.

Missing feature compensation can now be applied to the *incomplete spectrogram* shown in Figure 2.5. A binary representation of Figure 2.5 represents the spectrographic mask for this utterance. Based on local SNR, it labels which pixels are “reliable” and which are “missing”.

2.4 Reconstruction Missing Spectrographic Features

Once the unreliable pixels in the noisy mel spectrogram have been erased, we can perform missing feature compensation. The missing feature methods that will be used in this thesis are called Cluster-Based Reconstruction and Correlation Based Reconstruction. These methods, developed by Raj [43], both

attempt to reconstruct the missing features by making explicit used of the information contained in the remaining reliable spectrographic elements. Additionally, both utilize *a priori* information gained from a training corpus of clean speech to estimate the missing components of the corrupted spectrographic display. However, the techniques differ in the kind of *a priori* information used and how it is applied in the reconstruction process.

After missing feature reconstruction is performed in the log spectral domain, the features are transformed to the cepstral domain for recognition using the standard inverse DCT transform. This is an important distinction between these methods and other missing feature methods that perform recognition directly on the log spectra, because speech recognition accuracy is much better when cepstral features are used rather than log spectral features [13].

2.4.1 Cluster-Based Reconstruction

In the cluster-based reconstruction method, the log-spectral vectors of clean speech training corpus are grouped into a number of clusters using conventional Expectation-Maximization techniques [14]. The distributions of the vectors within each cluster are assumed to be Gaussian, and the mean, covariance, and *a priori* probability of each cluster are estimated from the training data. To compensate for noisy speech, the missing features are estimated by first identifying the cluster each corrupted log-spectral vector belongs to and then using the distributions of these clusters to estimate the noisy missing elements of the vector. Cluster membership is given by the cluster *k* that has the highest likelihood of generating the noisy vector $S(t)$, as given by

$$k_{S(t)} = \text{argmax}_k\{P(S(t)|k)P(k)\} \tag{2.5}$$

However, because $S(t)$ has missing elements, cluster membership cannot be identified in this way The missing elements must first be marginalized out of the cluster distributions so that cluster membership can be estimated only from the components in vector that are present. This marginalization is a crucial step in properly identifying cluster membership. Because the observed value (considered noisy or corrupt) represents the combined energy of the speech *and* the additive noise, we know that this value is the upper bound on the true value of the speech alone. Therefore, we can use the observed noisy value as the upper bound for marginalization. Cluster membership is now given by Equation (2.6) where $S_m(t)$ is a vector of the

missing elements of vector $S(t)$ and $Y_m(t)$ is the vector of their observed values.

$$\hat{k}_{S(t)} = \text{argmax}_k \left\{ P(k) \int_{-\infty}^{Y_m(t)} P(S(t)|k) dS_m(t) \right\} \tag{2.6}$$

Once the cluster membership *k* of a vector has been determined, missing feature reconstruction is performed using bounded MAP estimates based on the Gaussian distribution of the appropriate cluster and the upper bounds given by the observed corrupt values, as shown in Equation (2.7)

$$\hat{S}_m(t) = \text{argmax}_{S_m} \{ P(S_m | S_o, \mu_{\hat{k}_{S(t)}}, \Sigma_{\hat{k}_{S(t)}}, S_m(t) \le Y_m(t)) \} \tag{2.7}$$

2.4.2 Correlation-based Reconstruction

The correlation-based reconstruction method operates on the premise that the log spectral vectors of speech are generated by a stationary Gaussian random process. Because of the stationarity assumption, the means of the elements in the vector are dependent only on their frequency index, not on time. Furthermore, the covariance between any two elements in a spectrogram is dependent only on their indices and the distance (in time) between them. Because the distribution of the vectors is assumed to Gaussian, the joint distribution of all the individual elements in the spectrogram is Gaussian as well. A training corpus of clean, uncorrupted speech representing samples of this random process is used to estimate its parameters, the mean and covariance.

To compensate for missing features using this method, the entire spectrogram is separated into two parts: the missing components and the observed components. Using the parameters derived from the training corpus, the MAP estimate of the missing components conditioned on the observed components is computed. This estimate is given by Equation (2.8), where *m* represents the missing components in the spectrogram and *o* represented the observed reliable elements in the spectrogram.

$$\hat{S}_m = \mu_m + \Sigma_{mo} \Sigma_{oo}^{-1} (S_o - \mu_o) \tag{2.8}$$

However, jointly estimating *all* the missing components based on *all* the observed components in computationally impractical because of the matrix operations required. Therefore, reconstruction is done on a

vector-by-vector basis, jointly estimating the missing components of each vector separately. Furthermore, the included observed components on which the estimate is conditioned is limited to those in the spectrogram that have a relative covariance of 0.5 or greater with at least one missing element in the vector. As was done in the cluster-based reconstruction method, the MAP estimate of the missing elements can be upper bounded by the observed "corrupt" value that represents the combined energy of the speech and the noise.

2.4.3 Performance of Missing Feature Methods

Throughout the discussion of missing feature methods, the idea of removing elements of low SNR and retaining elements of high SNR has been repeatedly discussed. However, there has been no mention of what this SNR threshold used to label the elements as reliable/unreliable is. This is because this threshold is not fixed and is dependent on the missing feature method applied. For the methods of Cooke et. al., the threshold has been shown to be 15 dB [8], while the methods used in this work use an SNR threshold of -5dB. This number was determined experimentally [43]. To evaluate the effectiveness of the methods described above, a series of experiments was performed using clean speech corrupted with known noise sources at various known SNRs. Because the speech signal and the noise signal were both known *a priori*, we could construct "oracle" spectrographic masks based on full knowledge of the local SNR and the -5db threshold. Figure 2.6 shows recognition accuracy as a function of SNR when speech is corrupted with white noise. Figure 2.7 shows the same plot for speech corrupted with music (which is highly non-stationary). For comparison, the recognition accuracy when spectral subtraction is used for noise compensation

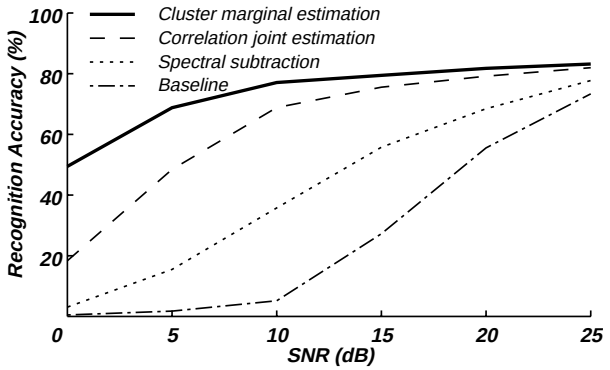


Figure 2.6 Recognition accuracy obtained missing feature reconstruction methods are used with oracle masks on speech corrupted with white noise.

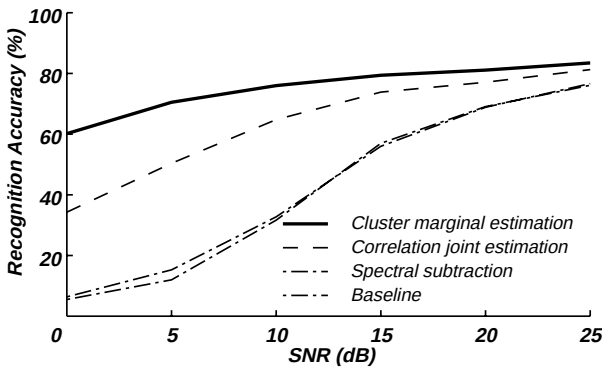


Figure 2.7 Recognition accuracy obtained missing feature reconstruction methods are used with oracle masks on speech corrupted with music.

isi also shown.

Clearly, these missing feature methods are very powerful *if the spectrographic masks are perfectly identified*. There is improvement in the white noise case at every SNR, and more importantly, significant improvement in the case where speech has been corrupted with music.

2.5 Summary

In this chapter, we have described the effects of additive noise on a spectrographic representation of speech. We have shown how regions of a spectrogram with low SNR can be considered “corrupt” and can be removed from the spectrogram to create an incomplete spectrogram. Two missing feature compensation methods have been presented that attempt to reconstruct the missing elements of an incomplete spectrogram based on the remaining elements and *a priori* information collected from training data. Finally, we have demonstrated that these techniques are tremendously effective at reducing the degradation in speech recognition performance of noisy speech if full “oracle” knowledge of the spectrographic masks is available.

Of course, in a real situation, these “oracle” masks are not available. In the next chapter, and the remainder of this thesis, we will describe a classification system to estimate these spectrographic masks.

Chapter 3
Classifier-Based Mask Estimation

3.1 Introduction

As shown in Chapter 2, missing feature methods are very powerful noise compensation techniques if the spectrographic masks identifying all of the corrupt, and therefore "missing" features are known. In a real situation, we do not have these masks, nor do we have access to the speech and the noise signals individually. Therefore, it is necessary to develop a method to estimate these masks using only the available noisy speech signal.

In this chapter, we describe the mask estimation problem in further detail and propose a solution using a Bayesian classification strategy. We also examine previous related work by other researchers.

3.2 The Mask Estimation Problem

In Chapter 2, we described how local SNR can be used as a measure of the "reliability" of the content of a spectrographic element. However, without full access to the clean speech and the corrupting noise signal, reliably estimating the local SNR, or even the global SNR of an utterance, is very difficult in some situation, especially when the corrupting noise is non-stationary.

However, we do not actually need to know the local SNR. It is important to recognize that we are not trying to estimate the local SNR at every pixel location, but rather we are simply trying to make a binary decision about every pixel's reliability: either is it usable or it is not. While local SNR is a convenient measure of "reliability", it is not the only relevant piece of information. There are perhaps other features that are easier or more reliable to compute that can be used to distinguish usable and corrupt spectrographic pixels. Ideally, we could take all the pieces of information that help decide a pixel's reliability and combine them to make a single decision. This is a two-class classification problem, where the possible outcomes are (1) that the pixel is reliable or (0) that the pixel is corrupt. With this strategy, we can combine any and all useful information into the decision process.

However, the situation is a bit more complicated than it seems at first. Because missing feature methods do not make any assumptions about the nature of the corrupting noise, we would like our mask estimation

procedure to be free of assumptions about the noise as well. If we cannot make any assumptions about the noise signal, we are only left with what we know about speech itself. Therefore, the features used by the classifier should be based on the intrinsic characteristics of the speech and make few or no assumptions about the noise.

3.3 Previous Work

Missing feature methods for robust speech recognition are relatively new, first appearing in 1994 [7]. As a result, the work in mask estimation for missing feature compensation is limited. However, there are some related fields that have work that is useful to us.

3.3.1 Mask Estimation for Missing Feature Compensation Methods

Other researchers working with missing feature methods have all attempted to estimate the spectrographic masks using a running estimate of the noise spectrum [8]. One such technique for obtaining a running noise estimate is spectral subtraction [5]. A similar method presented in [43] uses the vector Taylor series algorithm (VTS) to obtain the running noise estimate. Another method of identifying spectrographic masks is based on the hypothesis that the energy of highly noisy elements of spectral vectors is significantly different from those with low noise[21]. The histogram of spectral elements in any frequency band over a given time window would therefore exhibit two peaks, one representing the noisy elements and the other representing the clean elements. Spectrographic masks are derived based on estimates of the noise spectra obtained as the difference in the positions of the two peaks[10]. No other method has been employed to identify masks to the best of our knowledge. All of these mask estimation methods perform well when the corrupting noise is stationary. However, when the noise is non-stationary, they fail and the mask estimation is very poor. This is illustrated in Figures 3.1 and 3.2. In Figure 3.1, the recognition accuracy is plotted versus SNR for speech that has been corrupted by white noise. Two methods of obtaining a running noise estimate, spectral subtraction and VTS, were used to estimate the spectrographic masks that were used for missing-feature reconstruction. The masks are quite effective and significant improvements can be seen over baseline performance. Figure 3.2 shows the recognition accuracy when the same two mask estimation methods are used for speech that has been corrupted with music. Because the music is non-stationary, these mask estimations techniques fail and there is no improvement over baseline recogni-

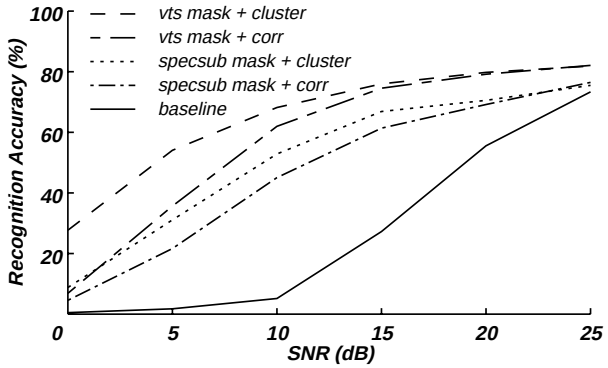


Figure 3.1 Recognition accuracy vs.SNR on speech corrupted by white noise when spectral subtracted-based or VTS-based mask estimation is used with missing-feature reconstruction.

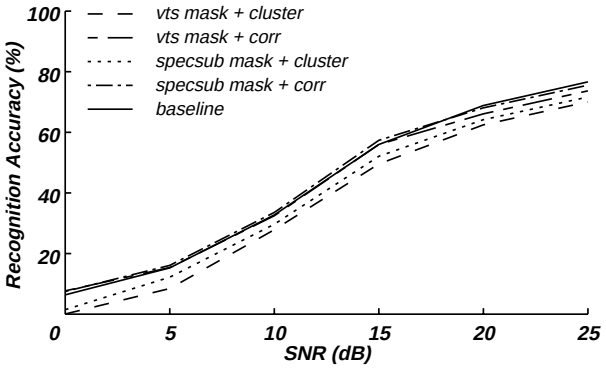


Figure 3.2 Recognition accuracy vs.SNR on speech corrupted by music when spectral subtracted-based or VTS-based mask estimation is used with missing-feature reconstruction.

tion accuracy. In fact, the when these masks are used with the cluster-based missing feature reconstruction technique, the performance is actually worse than baseline.

We must recall that the benefit of the missing feature paradigm is that it requires no assumption about the stationarity characteristics of the noise. It does, however, require an accurate spectrographic mask identifying reliable and corrupt regions. Therefore, the ability to obtain such a mask when the speech is corrupted with non-stationary noise is critical. To this point, no solution to this problem has been proposed.

3.3.2 Computational Auditory Scene Analysis

Humans are amazingly capable at distinguishing competing audio streams from each other and isolating one or more of these streams from the rest. This ability, dubbed the “cocktail party problem” has motivated extensive research into the perceptual segregation of sound. This research has resulted in much theoretical and experimental work in so-called *auditory scene analysis* by Bregman [6] and others. This led to the development of early computational models of the auditory system, such as [3] and [50] that attempt to emulate our ability to separate concurrent streams of sound from on another. The more recent work in this field has been done by Brown and Cooke [4] and Ellis [16]. Brown and Cooke utilize various features derived from grouping and transition cues to separate and organize the individual elements of an auditory map, a representation of the higher auditory pathways not unlike a spectrogram display of speech. The elements are tagged as belonging to a particular audio stream through masks which assign a binary label to each element in the auditory map. However, the primary goal of this work is to model the segregation of sound in the auditory system as accurately as possible, not to create a spectrographic masks of reliable and

corrupt elements. As a result, there is a tremendous amount of complexity in these models that are not needed for our purposes. Furthermore, the auditory maps that are produced identify all the elements that belong to a particular sound stream. The notion of “reliable” or “corrupt” is not considered in these models.

3.3.3 Co-Channel Speech Separation

One particularly difficult incarnation of the cocktail party effect is the isolation of a single speaker from a mixture of one or more other speakers. When there are two speakers on a single channel, this is known as co-channel speech separation. Extensive research has been done in this field, but without a large amount of success. Recently, the work of Morgan et al. [33] and Quatieri [39] has had some promising results. These algorithms, however, are primarily focused on the perceptual separation of the two speakers. That is, the end goal is a signal that sounds more separated to a human listener. Because humans and speech recognizers process information in different ways, a perceptual improvement does not necessarily translate into an improvement in recognition accuracy.

A pilot experiment was performed to examine the methods in [33] for recognition. The utterances in the TIMIT speech corpus [24] were resynthesized at a known constant pitch. The recognition accuracy on the resynthesized clean speech was the same as the original clean speech. A second utterance was added to the resynthesized speech at 10 dB to create a corpus of co-channel speech where the pitch of the primary speaker was known and constant. The co-channel separation algorithm described in [33] was applied to this corpus utilizing the *a priori* knowledge of the pitch of the resynthesized speaker. Informal listening tests conducted on a small audience showed a perceptual improvement in the speech after the separation algorithm was applied. However, the improvement in speech recognition accuracy was small. The word error rate (WER) of the clean resynthesized speech was 9.5%. When the speech was corrupted with another utterance to 10 dB, the error rate increased to 44.6%. After the separation algorithm, error rate was reduced to 40.1%. This reduction in error rate is quite small, especially considering the constrained conditions and amount of *a priori* knowledge available because the speech was resynthesized and monotone.

In a real co-channel situation, one would be faced with the need to obtain a pitch estimate of one or both of the speakers as a starting point for separating the two speech signals. Reliably estimating pitch of one speaker in the presence of another is a very difficult task. Nonetheless, the idea of using pitch information is a good one and will be examined in greater detail in this thesis.

3.4 Summary

In this chapter, we have described the challenge of estimating the spectrographic masks required by missing feature compensation methods. We have proposed a classifier based mask estimation solution and established the guidelines for its design. Some of the previous work of other researchers in this and other similar fields has been reviewed. Previous mask estimation methods have been successful if the corrupting noise is stationary, but unsuccessfully when the noise is not. This is a very important limitation to these methods. The key benefit of missing feature methods for noise compensation is they do not require any assumption about the noise and therefore can compensate for any type of additive noise. However, if masks cannot be estimated in non-stationary noise conditions, this benefit will be lost, and missing feature methods will become yet another compensation method that requires the assumption of noise stationarity. In the related fields of auditory scene analysis and co-channel speech separation, we have discussed some ideas that might be relevant to the mask estimation problem, but also highlighted important differences in the problems they are trying to solve and the problem of spectrographic mask estimation.

In this thesis, we set out to design a classifier that can generate spectrographic masks for speech corrupted by noise. We will not place any constrains on the noise, nor make any assumptions about its characteristics. Rather, our classifier will rely on the intrinsic characteristics of the speech signal itself. In the next chapter, we will examine pitch, one of the key intrinsic characteristics of speech, and present a new robust pitch detection algorithm. Reliable estimates of the pitch will play a significant part in the classification system for mask estimation.

Chapter 4
Histogram-Based Pitch Detection Algorithm

4.1 Introduction

Human speech can be categorized in many ways. One popular way is to distinguish voiced speech from unvoiced speech [41]. The main difference in voiced and unvoiced speech is the periodicity present in voiced speech. This results in local peaks in the spectrum of voiced speech at the fundamental frequency, F_0, and its harmonics. This harmonic structure is also a key difference between voiced speech and most interfering noise signals. Because of this periodicity, most of the signal energy of voiced speech is contained within its harmonics [33]. Interfering noise, however, follows no such pattern. This energy pattern in voiced speech can be useful for estimating the noise level present at spectrographic locations. However, we can only take advantage of this knowledge if we know the fundamental frequency, or pitch, of the speech signal, which of course, we do not. Furthermore, if we wish to build separate classifiers for voiced speech and unvoiced speech, we also need to accurately label all frames of an utterance as voiced or unvoiced.

There has been extensive work done in pitch detection [20][40]. These pitch detection algorithms (PDA) all work quite well when the speech signal is relatively noise free, but the accuracies of their pitch estimation and voiced/unvoiced labeling decrease significantly as noise is added to the signal. We have attempted to address this problem by developing a new pitch detection algorithm that is more robust to additive noise than previous methods. In this chapter we will describe the details our pitch detection algorithm and compare its performance to a well-known and widely used pitch algorithm, RAPT (Robust Algorithm for Pitch Tracking) [48], in both the clean and noise-corrupted speech conditions.

4.2 The Algorithm

The human auditory system is remarkably robust to environmental noise. We are easily able to isolate a single auditory stream from a several interfering streams. For instance, we can focus on a single speaker in a crowded room (the well-known “cocktail party” effect) or a single instrument or group of instruments playing in a symphony. An extensive amount of work, such as [30][4], has been done to model the signal processing of the peripheral auditory system. The peripheral auditory system can be represented as a bank

of band pass filters whose filter spacing is approximately linear at lower frequencies and logarithmic at the higher frequencies, with bandwidth that increases with increasing center frequency. We have attempted to use this knowledge as a basis for our pitch detection algorithm. This multi-band approach is similar to that used in the model of the auditory periphery by Meddis and Hewitt [30] and more recently in the computational auditory scene analysis work of Cooke[4] and Ellis [16]. However, in these methods, information in the each of the subbands is pooled together into a "summary" function to make a single pitch estimate. In the algorithm presented here, a pitch estimate is computed in every subband and the multiple pitch estimates are used to determine the final pitch estimate. The pitch estimation algorithm has four main steps: band-pass filtering, autocorrelation, creating the pitch period histogram, and smoothing.

4.2.1 Band-Pass Filtering

The Seneff filterbank [45] is one model of the filtering done by the peripheral auditory system. It is comprised of 40 filters, implemented with a cascade/parallel network as shown in Figure 4.1 (from [46]).

FILTER A is an FIR filter with eight zeros

Each of the 40 ***FILTER B*** filters is FIR with 2 zeros

Each of the 40 ***FILTER C*** filters is IIR with 4 poles and 4 zeros.

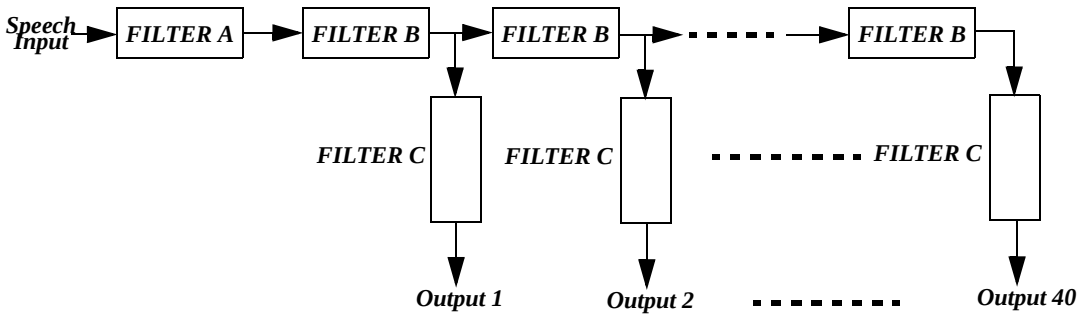


Figure 4.1 Cascade/parallel implementation of the Seneff filterbank.

The composite frequency response of the filterbank is shown in Figure 4.2.

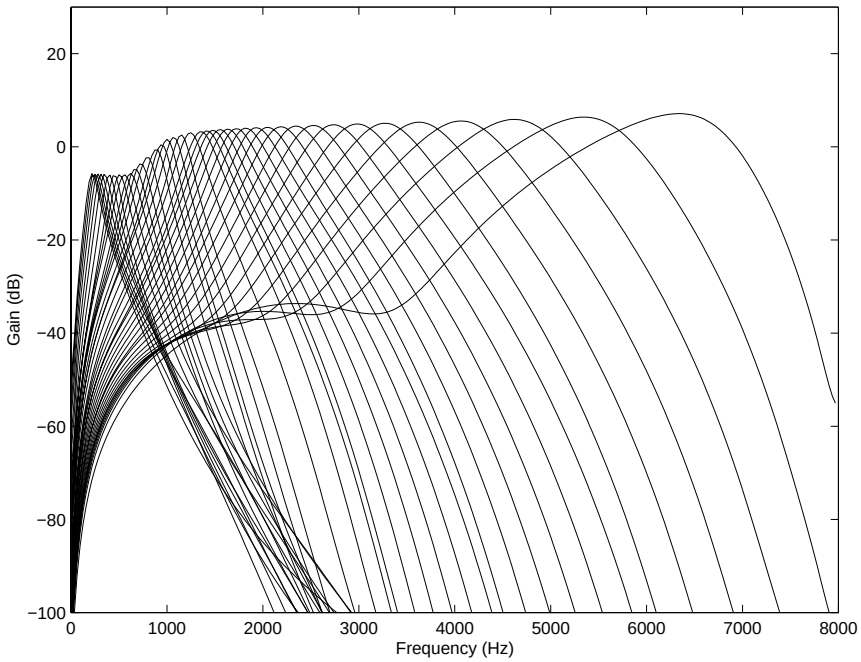


Figure 4.2 Composite Frequency Response of the Seneff filterbank.

4.2.2 Autocorrelation

After band-pass filtering, we have forty subband representations of the speech signal. Each of the forty subbands of speech contains the harmonics of F_0 that reside within the bandwidth of that subband. The fundamental period of the speech in each subband is determined by autocorrelation, smoothing and peak detection. To eliminate spurious peaks from the signal and detect peaks more reliably, the autocorrelation outputs are passed through a low-pass filter and an envelope detector. The distance between the two largest peaks represents the fundamental period of the signal. At the end of the autocorrelation step of the algorithm, we have forty estimates of the pitch period for a frame of speech, one for each subband in the filterbank.

4.2.3 Creating the Pitch Period Histogram

The forty pitch period estimates for the frame of speech are pooled into a pitch period histogram. For voiced speech, a single period estimate dominates the distribution of the forty candidate values. This is illustrated in Figure 4.3. However, for unvoiced speech, this distribution is roughly uniform, as shown in Figure 4.4. The initial pitch period estimate for the frame is determined by majority rule. If a single period

dominates 25% or more of the frequency bands, the frame is labelled as voiced, with a fundamental frequency determined by the winning pitch period and the sampling rate. Otherwise, the frame is initially labelled as unvoiced, and the majority pitch candidate is stored for later processing.

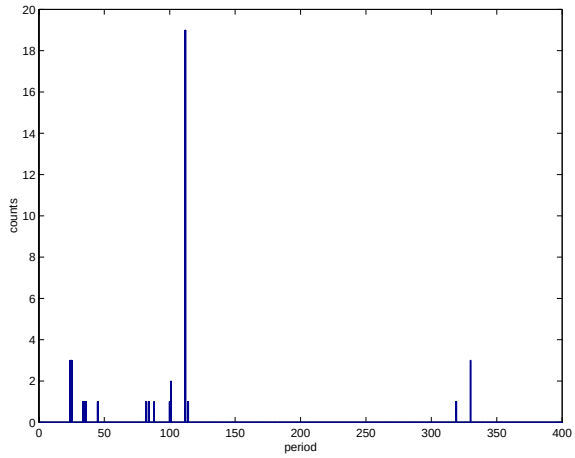


Figure 4.3 Pitch period histograms for voiced speech.

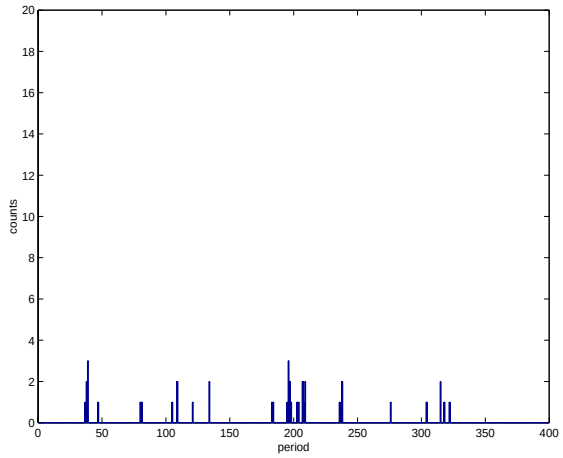


Figure 4.4 Pitch period histograms for unvoiced speech.

4.2.4 Smoothing

After band-pass filtering, autocorrelation and histogramming are performed on all frames of an utterance, we have an initial labeling of each frame as either voiced or unvoiced, as well as a pitch estimate for the voiced frames. Additionally, we have retained a potential pitch estimate for each unvoiced frame. This information is processed by a rule-based smoothing algorithm. The following rules are used to smooth the estimated pitch contour:

- a voiced segment of speech must consist of at least 3 consecutive frames. Any voiced segments less than 3 frames in length are relabeled as unvoiced
- an unvoiced segment must also last at least 3 frames. Any unvoiced segment that is less than 3 frames is considered incorrectly labeled and changed to voiced speech. The pitch estimates for these frames are determined by linearly interpolating between the pitch estimates of the bounding adjacent voiced frames.
- At voiced/unvoiced or unvoiced/voiced boundaries, the candidate pitch estimates of the unvoiced regions are re-evaluated. If the unvoiced frames at the boundary have a pitch estimate that is within a fixed threshold of the neighboring voiced frames, those frames are relabeled as voiced. This threshold was empirically set at 8Hz.

4.3 Performance Evaluation

The pitch detection algorithm presented here was evaluated using a corpus of speech waveforms with simultaneously recorded laryngograph data [12]. The laryngograph data was then processed using methods reported in [2] to develop an oracle pitch contour for each waveform. Our pitch detection algorithm was compared to another widely used pitch detection algorithm, RAPT. Four criteria were used to evaluate the pitch detection algorithm. Frames of voiced speech erroneously labeled as unvoiced speech and frames of unvoiced speech erroneously labeled as voiced speech were accumulated over the entire test set. The root mean squared error of the pitch estimates was computed over all regions where the oracle pitch and the pitch estimate were both voiced. Additionally, the number of frames where the pitch estimate was more than 20% away from the reference value was also computed. These were considered gross errors. Similar pitch detection algorithm performance metrics were reported in [2]. The test set consisted of the first 100 utterances from the oracle corpus. The pitch detection algorithms were run on both clean speech and speech corrupted with white noise to various SNRs. The results are shown in Table 4.1. The histogram-based pitch detection algorithm is generally more accurate than RAPT. However, the RAPT pitch tracking algorithm is significantly faster than the histogram-based pitch algorithm. An utterance that was 3.52 seconds long was processed by both pitch detection algorithms. Each algorithm generated pitch estimates for the utterance fifty times, and the average computation time was computed. The RAPT algorithm generated the pitch estimates in 3.5 seconds on average, while the histogram-based method took an average of 58.8 seconds to generate the pitch estimates for the same utterance.

	clean speech		20 dB AGWN		10 dB AGWN		0 dB AGWN	
	RAPT	HB	RAPT	HB	RAPT	HB	RAPT	HB
% error voiced	1.4	3.7	5.8	6.4	16.7	11.6	55.0	30.1
% error unvoiced	15.5	5.5	9.1	4.6	5.2	3.5	1.5	1.8
% gross error	0.3	0.05	0.2	0.03	0.08	0.01	0.0	0.1
RMS voiced error	3.6	1.8	3.2	1.8	2.3	1.8	1.6	1.8

Table 4.1. A comparison of two pitch detection algorithms, RAPT, and the Histogram-Based Pitch Detection Algorithm (HB), for clean speech and speech corrupted with white noise.

4.4 Summary

In this chapter, we presented a new pitch detection algorithm that is more accurate and more robust to additive noise than other methods. However, the algorithn is also slower than other methods. We now have a reliable method of labeling frames of speech as voiced or unvoiced, and estimating the pitch in voiced frames. This enables us to create a separate classifier for voiced and unvoiced speech. Furthermore, we will see in Chapter 5 how this pitch information can be used to develop pitch-related classification features to analyze noise-corrupted speech.

Chapter 5
Feature Extraction

5.1 Introduction

Perhaps the most important step in developing a classification scheme is the feature extraction. Regardless of the sophistication of the classifier, poor features will result in a poor classifier. Ideal features are those that maximize the discriminability between the classes. Features that result in good separation of the classes make the job of the classifier much easier.

Because voiced speech and unvoiced speech are generated by different production mechanisms, they have very different characteristics. As a result, we want to make a distinction between the features used to classify the reliability of spectrographic locations in voiced speech and those used to classify the reliability of the spectrographic locations in unvoiced speech. We will see in Chapter 6 that the use of different features for voiced and unvoiced speech necessitates the use of a separate classifier for each type of speech. Additionally, we want to create features that make minimal assumptions about the corrupting noise signal by relying on the inherent characteristics of the speech signal itself.

In this chapter we discuss the characteristics of voiced speech and then develop a series of features that exploit these characteristics to capture the influence of noise on each spectrographic location. We then focus on the characteristics of unvoiced speech and describe the features that are used to identify noisy elements in unvoiced spectrographic regions.

5.2 Features for Voiced Speech

We would like to develop features for our classifier that exploit the intrinsic structure of speech itself, rather than depending on assumptions about the noise characteristics. The two key characteristics of voiced speech that we utilize in the design of our classification features are the presence of a strong fundamental frequency (pitch), and all of its harmonics, and the distinctive spectral contour of voiced speech across frequency.

5.2.1 Comb Ratio

Because of the harmonic nature of voiced speech, the majority of the energy of a clean voiced speech signal resides in its harmonics [33]. Additive noise does not typically have this characteristic. Therefore when additive noise is mixed with voiced speech, the overall signal energy will increase both at the harmonics of the pitch and at the frequencies in between. Therefore, a measure that compares the energy in the harmonics of voiced speech to the energy outside the harmonics would be a good indicator of noise present in the signal.

The amount of energy present in the harmonics can be captured using a comb filter. A comb filter has peaks at the harmonics of the fundamental frequency ($n*F_0$) and nulls in between ($n*F_0 + F_0/2$). Traditional comb filters are simply delay-and-add filters. The signal is shifted by a single period, added to the original signal, and normalized. The frequency response is very soft and only has significant attenuation exactly halfway between harmonics. For a sharper response, an IIR implementation can be used [21]. In this implementation, given by the transfer function in Equation (5.1), p is the pitch period and g is a tunable parameter which sets the distance of the poles from the unit circle, and hence, the sharpness of the teeth of the comb.

$$H_{comb}(z) = \frac{z^{-p}}{1 - gz^{-p}} \tag{5.1}$$

It was determined empirically that setting $g = 0.7$ captures most of the harmonic information of voiced speech. The frequency response of this comb filter for a pitch of 160 Hz ($p = 100$ for a sampling rate of 16 KHz) is shown in Figure 5.1.

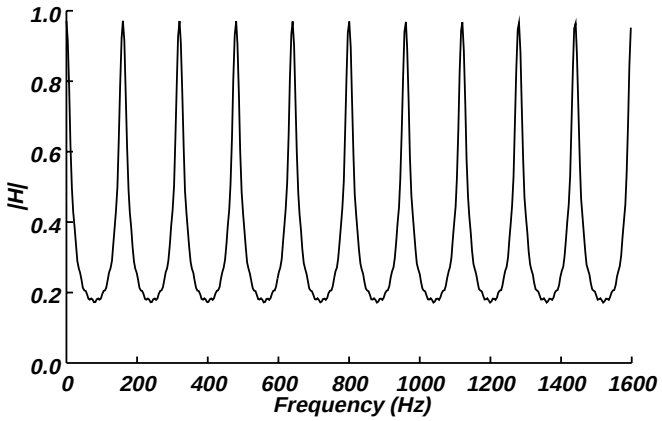


Figure 5.1 The magnitude frequency response of the IIR comb filter when $g = 0.7$ and p=100. The peaks are at the harmonics of 160 Hz.

To capture the energy of the components of the signal that fall in between the harmonics, the comb filter is simply shifted by $F_0/2$. The transfer function for this shifted comb filter is given by Equation (5.2).

$$H_{combshift}(z) = \frac{-z^{-p}}{1 + gz^{-p}} \tag{5.2}$$

If we assume that the voiced speech resides at the harmonics of the fundamental frequency while noise may reside in all frequency bands, the comb filter is a measure of speech and noise energy while the shifted comb filter is a measure of noise energy only. Thus, the log ratio of the energies of the speech signal passed through the comb and shifted comb filters is a measure of speech plus noise to noise. The cleaner the speech signal is, the larger this ratio will be. We call this metric the *comb ratio*. The comb ratio, $CR(n, \omega_i)$, is given by Equation (5.3), where y_{comb} and $y_{combshift}$ are the outputs when the speech signal in frame n and subband ω_i have been passed through the comb and shifted comb filters, respectively,

$$CR(n, \omega_i) = 10\log_{10}\left(\frac{\sum_k y_{comb}[k, \omega_i]^2}{\sum_k y_{combshift}[k, \omega_i]^2}\right) \tag{5.3}$$

The comb ratio can be used as a measure of SNR. Figure 5.2 shows a plot of the average comb ratio over all voiced frames and all sub-bands vs. global SNR for an utterance corrupted with white noise and with music. In both cases, the comb ratio tracks with SNR and the two lines are very similar even though the corrupting signals had very different characteristics.

5.2.2 Autocorrelation Peak Ratio

The pitch algorithm described in Chapter 4 used autocorrelation to determine the fundamental period of the speech in each subband by finding the lag between the two largest peaks. If a signal is purely periodic, then the peaks of the autocorrelation function will all be of uniform height, determined by the energy of the signal. If a signal is quasi-periodic, such as voiced speech, the secondary peaks in the autocorrelation function will be less than or equal to the height of the main peak. The less periodic the signal is, the smaller the

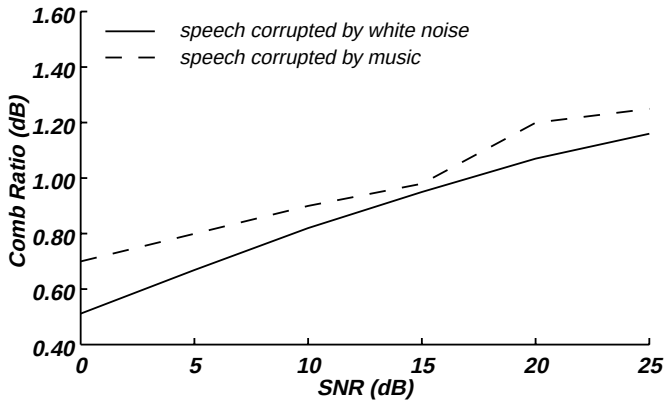


Figure 5.2 Average Comb Ratio vs. SNR for all voiced bands in an utterance for speech corrupted with noise and speech corrupted with music.

secondary peaks will be. Adding uncorrelated noise to a signal effectively reduces its periodicity, increasing the difference in the heights of the main peak and the secondary peaks. Therefore, we can use the ratio of the height of largest secondary peak to the height of the main peak as a measure of periodicity. Because voiced speech is quasi-periodic, the autocorrelation peak ratio will be close to one for clean speech and decrease as the signal is increasingly corrupted by noise.

5.2.3 Subband Energy to Fullband Energy Ratio

In addition to its characteristic harmonicity, voiced speech has a distinct spectral shape. The energy of voiced frames is concentrated at the lower frequencies and tails off at higher frequencies. In Figure 5.3, the solid line shows the smoothed log spectrum of the vowel “EH” derived from its LPC coefficients. The 3 local peaks represent the first 3 formants, or resonant frequencies of the vocal tract. As noise is added to the speech, its spectral shape will change as a function of the spectral characteristics of the noise. The dashed line in Figure 5.3 shows the same vowel spectrum when white noise has been added to the vowel. It is evident that additive noise has a more significant impact on the valleys of the spectrum than the peaks. We can use the log ratio of the energy in a subband to the overall frame energy as a measure of effect of additive noise on a particular subband and on the overall contour.

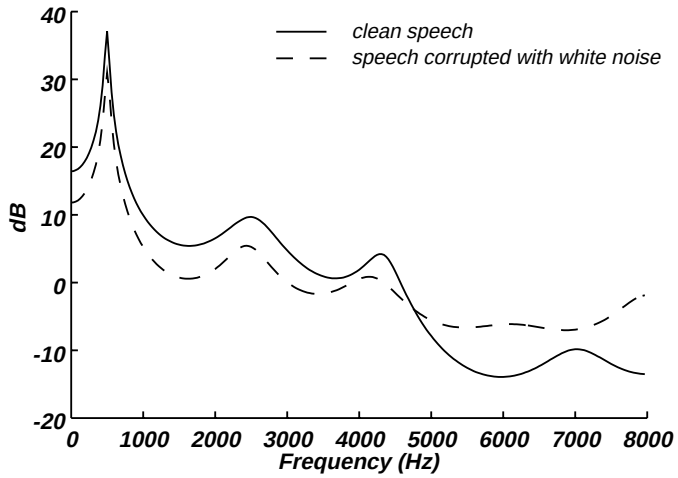


Figure 5.3 The smoothed spectrum of a vowel “EH” derived from its LPC coefficients. The solid line shows the spectrum for clean speech and the dashed line show the spectrum when the speech has been corrupted with white noise to 10db.

5.2.4 Subband Energy to Fullband Noise Floor Ratio

Having knowledge of the noise floor of a noise-corrupted speech signal is obviously very useful for estimating the SNR. However, an accurate measure of the noise floor is difficult to obtain. If we assume that the corrupting noise is stationary, we can coarsely estimate the level of the noise floor by looking at the distribution of the energy of the frames in an utterance. These distributions typically have two modes, one at a low energy value representing the silence and low energy speech regions and one at a higher energy representing high energy speech regions. The idea of statistically modeling the energy distributions of speech has been used for speech endpoint detection using HMMs. We have used a much simpler technique to get a rough estimate of the noisefloor. The energies of all frames of an utterance are put into a histogram and the lower energy peak is found. The energy bin in the histogram corresponding to this peak value is considered the noise floor of the noisy speech signal. We can compare the energies of a subband of a frame of speech and the overall noise floor of an utterance to help determine the likelihood that a specific spectrographic location has been corrupted by noise. It is important to point out that using the energy of the silence frames to estimate the noise floor of the entire utterance implies stationarity of the interfering noise. If the noise is highly non-stationary, the noise floor estimate will not necessarily be accurate.

5.2.5 Subband Energy to Subband Noise floor Ratio

While the subband energy to global noise floor ratio is a useful feature, we can gain more local noise information by repeating the noise floor estimation technique in each subband. In this case, the energy of a subband of a frame of speech is compared to the estimate of the noise floor of the utterance in that particular subband.

5.2.6 Flatness

As was noted earlier, voiced speech exhibits a very definitive trajectory across frequency, and when noise is added to the speech utterance, this spectral shape will change. As shown above in Figure 5.3, the valleys in the spectrum tend to flatten as noise is added to a speech signal. This "flatness" can be characterized by the variance $\sigma_{flat}^{\ 2}$ of the subband energy in a neighborhood of spectrographic locations around a given pixel. For an 3x3 neighborhood of pixels, the flatness is given by Equation (5.4), where $s(n, \omega_i)$ represents the subband energy of frame n and subband ω_i, and $\mu_s(n, \omega_i)$ is the mean of the subband energy values in a 3x3 neighborhood around frame n and subband ω_i.

$$\sigma_{flat}^{\ 2}(n, \omega_i) = \frac{1}{9} \sum_{k=i-1}^{i+1} \sum_{j=n-1}^{n+1} (s(j, \omega_k) - \mu_s(n, \omega_i))^2 \tag{5.4}$$

For a given subband, a signal corrupted with noise tends to have shallower, flatter valleys than its uncorrupted counterpart. Therefore, we expect noise-corrupted spectrographic locations will have a lower variance than cleaner ones.

5.3 Features for Unvoiced Speech

Unvoiced speech is much more difficult to characterize than voiced speech. Because it is generated by air passing over relaxed vocal chords, the excitation signal is essentially random. There is no harmonicity or other regularity as in voiced speech. As a result, the pitch-related features developed for voiced speech will be ineffective for unvoiced speech. Unvoiced speech also has lower energy than voiced speech and is therefore more affected by noise than voiced frames. However, it does have a general spectral shape that is

unlike voiced speech and most naturally occurring noises. Unvoiced speech energy is concentrated at the higher frequencies and tails off at lower frequencies. Figure 5.4 shows the log spectrum of the “SH” sound. derived from its LPC coefficients. The voiced speech features that do not rely on pitch characterize a frame of speech in terms of the relative energy levels in each of the subbands, and the overall and local spectral shape. They are useful features because we know that adding noise to a speech signal alters both the relative subband energy levels and the spectral shape. This is true for both voiced and unvoiced speech. While the energy distribution of unvoiced speech across frequency is very different from that of voiced speech, it too will be altered by additive noise. As a result, we can use the remaining four non-pitch dependent features to characterize unvoiced speech.

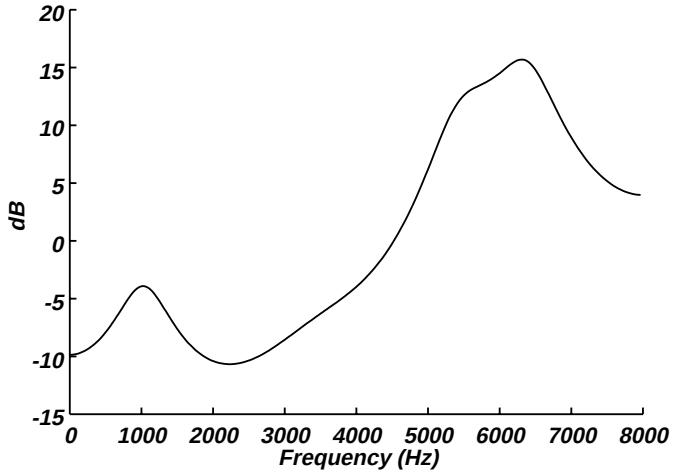


Figure 5.4 The smoothed spectrum of the unvoiced phoneme “SH” derived from its LPC coefficients.

5.4 Summary

In this chapter, we have described the features we will use in our classification scheme to detect corrupt spectrographic locations. The pitch algorithm described in Chapter 4 enables us to develop two pitch-related features: the comb ratio and the autocorrelation peak ratio. These features exploit the inherent periodicity and harmonicity of voiced speech to estimate the noise levels in each subband. We have also developed four additional features to characterize the influence of noise in a particular subband of speech. These features exploit the characteristic energy distributions of voiced and unvoiced speech across frequency to capture the level of noise corruption in each spectrographic location.

In the next chapter, we will describe the classification scheme that uses these features to decide if a spectrographic element is reliable or corrupt.

Chapter 6
Classification Strategy

6.1 Introduction

To estimate a spectrographic mask, we need to decide whether each pixel in the spectrogram is reliable or corrupt. In the previous chapter, we generated a feature set to characterize the distinctions between good pixels and bad pixels. This feature set can now be used to make a decision about the reliability of the information in the spectrogram at every time-frequency location. In this chapter, we present the decision strategy we will use, Bayesian classification. We construct a two-class classifier using the features described in Chapter 5. Through a series of experiments, we describe the performance of the classifier with regard to mask estimation accuracy and speech recognition accuracy achieved when the estimated masks are partnered with missing feature compensation methods.

6.2 Bayesian Classification

For every spectrographic location, we would like to decide, based on a set of features, if the location is more likely to be reliable or corrupt. That is, given an input feature vector, the classifier should estimate which class, reliable or corrupt, is more likely to have produced these features. Mathematically, the classifier should assign the feature vector $\boldsymbol{x}$ to the class ω_j that has the highest *a posteriori* probability $P(\omega_j|\boldsymbol{x})$ over both classes j={0,1}. This is expressed in Equation (6.1), where $\overset{\omega_1}{\underset{\omega_0}{\gtrless}}$ indicates that we choose class 1 if $P(\omega_1|\boldsymbol{x}) > P(\omega_0|\boldsymbol{x})$ and class 0 if $P(\omega_1|\boldsymbol{x}) < P(\omega_0|\boldsymbol{x})$.

$$P(\omega_1|\boldsymbol{x}) \overset{\omega_1}{\underset{\omega_0}{\gtrless}} P(\omega_0|\boldsymbol{x}) \tag{6.1}$$

Because the *a posteriori* probabilities are usually not directly known, we use Bayes Rule, shown in Equation (6.2), to express this *a posteriori* probability as a function of the class conditional probability $P(\boldsymbol{x}|\omega_j)$, the class prior probability $P(\omega_j)$, and the prior probability of the feature vector $P(\boldsymbol{x})$.

$$P(\omega_j|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\omega_j)P(\omega_j)}{P(\boldsymbol{x})} \tag{6.2}$$

The class conditional probability and the class prior probability for each class can be determined from the distributions of the classes learned from training data. Using the *a posteriori* probabilities of each class and Bayes rule, we can derive a likelihood ratio, shown in Equation (6.3), that describes the decision boundary between the two classes.

$$\frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_0)} \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_0)}{P(\omega_1)} \tag{6.3}$$

Note that the prior probability of the feature vector $P(\boldsymbol{x})$ has been removed because it is constant over both classes. In our classifier, we assume that the distributions of the features are Gaussian. The likelihood ratio now describes the decision boundary between two multivariate Gaussian distributions. After taking the log of both sides and some manipulation, the log likelihood ratio is given by

$$0.5\log\left(\frac{|\Sigma_0|}{|\Sigma_1|}\right) - 0.5(\boldsymbol{x}-\mu_1)^T\Sigma_1^{-1}(\boldsymbol{x}-\mu_1) + 0.5(\boldsymbol{x}-\mu_0)^T\Sigma_0^{-1}(\boldsymbol{x}-\mu_0) - \log\left(\frac{P(\omega_0)}{P(\omega_1)}\right) \underset{\omega_0}{\overset{\omega_1}{\gtrless}} 0 \tag{6.4}$$

where μ_j and Σ_j are the mean and covariance matrix, respectively, of the feature vectors in class *j*. These parameters are estimated from the training data.

We will use this multivariate Gaussian classification strategy with full covariance matrices to generate the spectrographic masks. Because the feature vectors are different for voiced and unvoiced speech, we will use a different classifier for each type of speech. In addition, the values of the features themselves may vary significantly from subband to subband *within* each class. For example, Figure 6.1 shows the mean value of the comb ratio feature across all twenty subbands for each class. The mean value for the “reliable” class, indicated by the solid line in the figure, varies from 2.2 dB in the first mel filter to 0.12 dB in the twentieth mel filter. If we pool the data from all the subbands into two large classes, the class distributions will be broader and discrimination between the classes will be more difficult. Therefore, we will also build a separate classifier for each subband of the spectrogram.

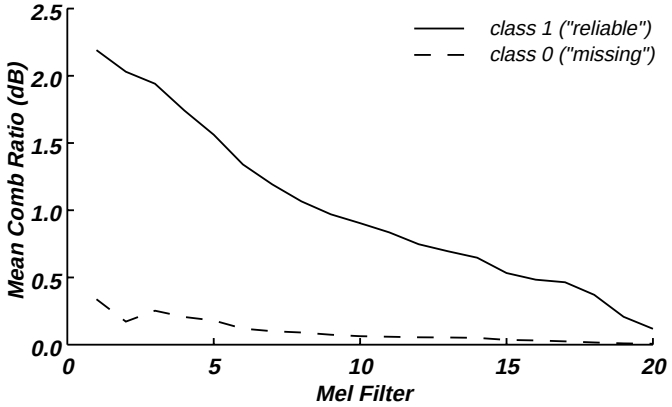


Figure 6.1 The mean value of the comb ratio for each class as a function of the mel spectrum subband. The large disparity in the values from filter to filter suggests that using a separate classifier for each subband would be appropriate.

6.3 Mask Evaluation Criteria

In order to determine the effectiveness and the accuracy of the masks estimated by the classifier, certain performance criteria must be established by which the masks can be judged. The performance of the classifier will be evaluated in two ways. First, the classification accuracy of the estimated masks will be considered. Secondly, we will determine the improvement in recognition accuracy achieved when the classifier-generated masks are used in conjunction with missing feature compensation techniques.

It has been shown that the missing feature reconstruction methods described in Chapter 2 perform optimally when the SNR threshold that determines a spectrographic element's reliability is -5 dB [43]. That is, elements with a local SNR greater than -5 dB are considered reliable and those with a local SNR less than -5 dB are considered corrupt. If perfect knowledge of the noise is available, the local SNR can be computed for every spectrographic location and the optimal mask can be constructed. We refer to this mask as the *oracle* mask.

We measure the accuracy of the classifier by comparing the masks estimated by the classifier to the oracle masks. In a two-class problem such as this one, there are two types of errors the classifier can make: misses and false alarms. We define a *miss* as the incorrect labeling of a corrupt spectrographic element as reliable and a *false alarm* as the incorrect labeling of a reliable element as corrupt. Similarly, there are two types of correct identifications the classifier can make: hits and correct rejections. We define a *hit* as the correct labeling of a corrupt spectrographic element and a *correct rejection* as the correct labeling of a reli-

able spectrographic element. The counts of all four possible classifier outcomes can be accumulated and used to estimate the probabilities of a hit, miss, false alarm and correct rejection for each classifier. Clearly, a classifier which maximizes the hit probability and minimizes the false alarm probability is optimal and will result in the most accurate spectrographic masks.

Performance of the classifier was also measured in terms of the speech recognition accuracy achieved when the estimated masks are combined with missing feature compensation methods. The upper bound for recognition performance is considered to be the accuracy attained when missing feature reconstruction is performed using the oracle spectrographic masks. This accuracy, shown in Figure 6.2, is considered the best possible performance. In addition, the relative importance of mask estimation accuracy in the voiced regions and the unvoiced regions was also determined. Figure 6.2 also shows the recognition accuracy when the oracle mask is only applied to the voiced regions and the unvoiced regions are left untouched, and the reverse case, when compensation is applied only to the unvoiced regions. The recognition accuracy is considerably higher when compensation is applied only to the unvoiced regions as compared to only the voiced regions. Based on these plots, we can infer that the accuracy of the mask estimation for the unvoiced speech regions is more important for recognition than that of the voiced speech regions. This is logical when we consider that unvoiced speech is typically of lower energy than voiced speech, so that for a given global SNR, the unvoiced regions of the spectrogram will be more corrupt that the voiced regions.

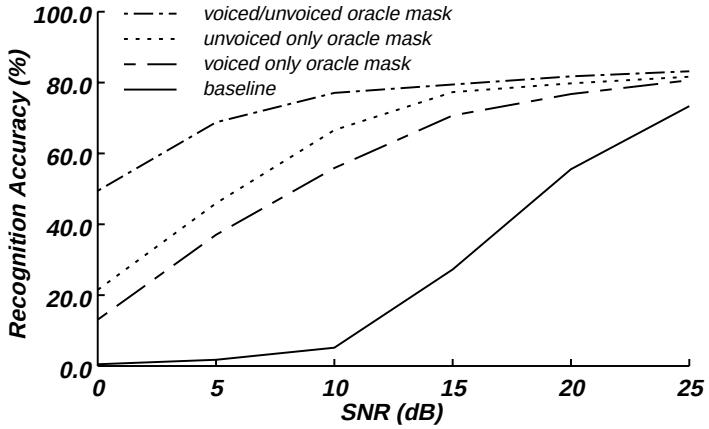


Figure 6.2 Recognition accuracy as a function of SNR on speech that has been corrupted by white noise. Missing feature compensation was applied using oracle spectrographic masks applied both the voiced and unvoiced regions, only the unvoiced regions and only the voiced regions.

6.4 Effects of Mask Estimation Error on Missing Feature Methods

Errors in the mask estimation will certainly affect the performance of the missing feature reconstruction algorithms. Because missing feature methods reconstruct the missing elements from the remaining reliable ones, the two kinds of errors, misses and false alarms, will not have the same effect. This becomes apparent if we consider the limiting case for both errors. If every pixel is declared reliable, there will be no false alarms. No features will be removed from the spectrogram, and no reconstruction of missing features is required. Speech recognition performance will be the same as that of noisy speech with no compensation. On the other hand, every pixel is labeled as unreliable, there will be no misses. All the elements in the spectrogram will be considered missing and erased. With no features remaining in the spectrogram, both missing feature reconstruction and recognition are impossible. The effect of false alarms and misses on the performance of the missing feature methods is illustrated in Figures 6.3 and 6.4 (from [43]). In Figure 6.3, random false alarms were introduced into the spectrographic masks of speech corrupted to 15 dB by white noise. There were no misses introduced. The figure shows how recognition accuracy is affected by an increasing percentage of false alarms. Similarly, Figure 6.4 shows the recognition accuracy as a function of the percentage of random misses introduced into the spectrographic masks. Clearly, the performance of the

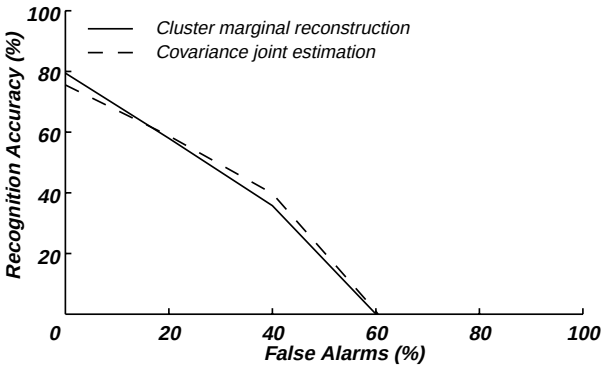


Figure 6.3 Recognition accuracy derived from reconstructed spectrograms, as a function of the fraction of reliable elements in the spectrogram that were erroneously tagged as being unreliable

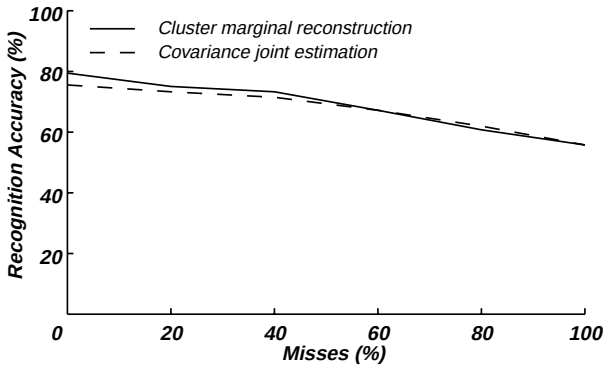


Figure 6.4 Recognition accuracy derived from reconstructed spectrograms, as a function of the fraction of unreliable elements in the spectrogram that were erroneously tagged as being reliable

missing feature reconstruction methods, and hence, recognition accuracy, degrades much more quickly with increasing numbers of false alarms than misses. The classifier therefore should favor misses over false alarms.

We can bias the classifier to favor misses over false alarms by incorporating a cost factor λ_{ij} into the classifier. λ_{ij} represents the cost of choosing class *i* when the test vector really belongs to class *j*. Using this notation, we can assign λ_{01} as the cost of a false alarm, λ_{10} as the cost of a miss and λ_{11} and λ_{00} the costs of a correct assignment. It can be shown [15] that this cost factor changes the likelihood ratio in Equation (6.3) to:

$$(\lambda_{01} - \lambda_{11})P(\omega_1|\boldsymbol{x}) \underset{\omega_0}{\overset{\omega_1}{\gtrless}} (\lambda_{10} - \lambda_{00})P(\omega_0|\boldsymbol{x}) \tag{6.5}$$

$$= \frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_0)} \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \frac{(\lambda_{10} - \lambda_{00})P(\omega_0)}{(\lambda_{01} - \lambda_{11})P(\omega_1)} \tag{6.6}$$

$$= \frac{P(\boldsymbol{x}|\omega_1)}{P(\boldsymbol{x}|\omega_0)} \underset{\omega_0}{\overset{\omega_1}{\gtrless}} \frac{P_\lambda(\omega_0)}{P_\lambda(\omega_1)} \tag{6.7}$$

As shown in Equations (6.6) and (6.7), by incorporating decision costs into the classifier, we effectively alter the prior probabilities of the two classes to favor one type of error over the other.

6.5 Experimental Results

To determine the effectiveness and accuracy of the classifier-based mask estimation, experiments were performed on speech corrupted by noise. In the first series of experiments, we corrupt speech with Gaussian white noise. To measure the performance of the classifier when the corrupting noise is non-stationary, we perform another series of experiments on speech that has been corrupted with music.

6.5.1 Mask Estimation on Speech Corrupted by White Noise

To estimate spectrographic masks for speech corrupted with white noise, the classifier was trained on 2880 utterances from the DARPA Resource Management speech corpus [38]. The speech was corrupted with white noise to various SNRs between 0 dB and 25 dB. The pitch estimates, necessary for the pitch-dependent features of the voiced regions, were estimated from the noise-corrupted speech using the histogram-based pitch detection algorithm described in Chapter 4. Because we had access to both the clean speech signal and the noise signal, the training vectors could be correctly labeled based on the local SNR.

In all subbands, training vectors with a local SNR of less than -5 dB were assigned to class 0 (missing) and training vectors with a local SNR greater than -5 dB were assigned to class 1 (reliable). Using this labeled training data, the mean vector and covariance matrix of each class were estimated for each subband and type of speech.

We expect the prior probabilities of corrupt and reliable elements to vary with the global SNR. That is, at higher noise levels, more spectrographic elements will be corrupt than at lower noise levels. However, because we do not know the global SNR, we need to choose constant prior probabilities that yield the best recognition accuracy over all SNRs. The appropriate values for the prior probabilities $P_\lambda(\omega_0)$ and $P_\lambda(\omega_1)$ were determined through a series of experiments conducted on a cross-validation set. The cross validation set consisted of 200 utterances from the Resource Management corpus. There is no overlap between the cross validation set and the training or test sets. The cross validation set was corrupted by white noise to a range of SNRs and mask estimation was performed using various values of $P_\lambda(\omega_0)$ and $P_\lambda(\omega_1)$. Figures 6.5 and 6.6 show Receiver Operating Curves (ROC) for various prior probabilities for the voiced and unvoiced classifiers. The number labeling each data point in the figures is the prior probability of a corrupt element, $P_\lambda(\omega_0)$ used in each trial. The prior probability of a reliable element $P_\lambda(\omega_1)$ is therefore $1-P_\lambda(\omega_0)$.

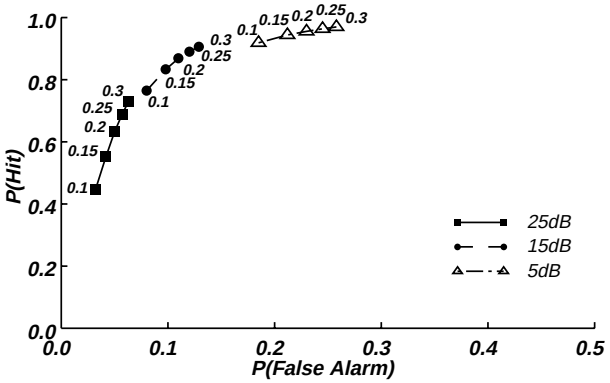


Figure 6.5 ROC of the mask estimation for the voiced regions of speech corrupted with white noise. The value next to each data point is the prior probability of a corrupt spectrographic element.

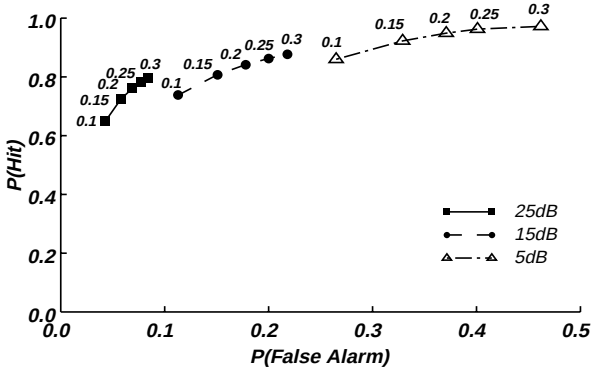


Figure 6.6 ROC of the mask estimation for the unvoiced regions of speech corrupted with white noise. The value next to each data point is the prior probability of a corrupt spectrographic element.

The estimated masks and the log mel-spectral vectors were passed to the cluster-based missing feature

algorithm described in Chapter 2 and the log mel-spectral elements declared missing by the masks were reconstructed. The twenty-dimensional reconstructed log mel-spectral vectors were transformed to thirteen-dimensional cepstral vectors for recognition. Speech recognition was performed using Sphinx III, an Hidden Markov Model (HMM) based large vocabulary speech recognition system. Continuous context-dependent HMMs with single Gaussian state distributions were trained on clean speech using the 2880 utterances from the Resource Management training set. Figure 6.7 shows the recognition accuracy obtained using the reconstructed spectrograms. Based on the results from the cross validation data, prior probabilities of $P_\lambda(\omega_0) = 0.2$ and $P_\lambda(\omega_1) = 0.8$ provided the best overall recognition accuracy. It is interesting to note that although a best prior probability value can be found from the cross validation data, the recognition accuraracies are remarkably close over the range of prior probabilities shown.

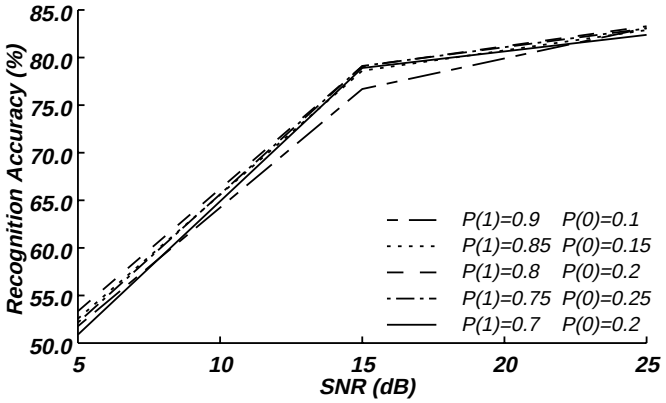


Figure 6.7 Recognition accuracy of the cross validation set vs. SNR using the classifier to estimate the spectrographic masks and then applying cluster-based missing feature compensation.

Using these values for the *a priori* probabilities, a series of experiments was performed on the test data. The test set consisted of 1600 utterances from the Resource Management corpus. None of the speakers in the test set appeared in the training or cross validation sets. In each experiment, the test set was corrupted with white noise at a fixed SNR.

The spectrographic masks for all utterances in the test set were estimated by the classifier. Figure 6.8 shows a typical estimated spectrographic mask generated by the classifier. The oracle mask for the same utterance is shown in Figure 6.9. The estimated mask is quite successful at capturing the trends of the oracle mask. The estimated mask does not fully capture the block nature of the corrupt regions of the spectrogram but most corrupt regions have been accurately identified. The estimated masks were compared to the oracle masks to measure the performance of the classifier. The classifier accuracy is shown in Figure 6.10

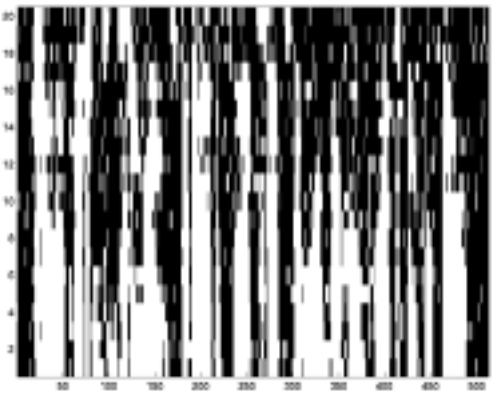


Figure 6.8 Estimated mask for an utterance corrupted with white noise to 10dB. The horizontal axis is frame number and the vertical axis is Mel filter number. The black pixels indicate corrupt or "missing" features. The white pixels indicate reliable features.

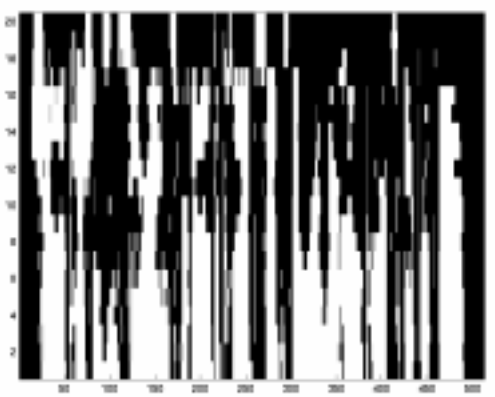


Figure 6.9 Oracle mask for an utterance corrupted with white noise to 10dB. The horizontal axis is frame number and the vertical axis is Mel filter number. The black pixels indicate corrupt or "missing" features. The white pixels indicate reliable features.

in terms of hit probability and false alarm probability of the mask estimation for the voiced regions over a range of SNRs. The classifier performance in the unvoiced regions is illustrated in Figure 6.11.

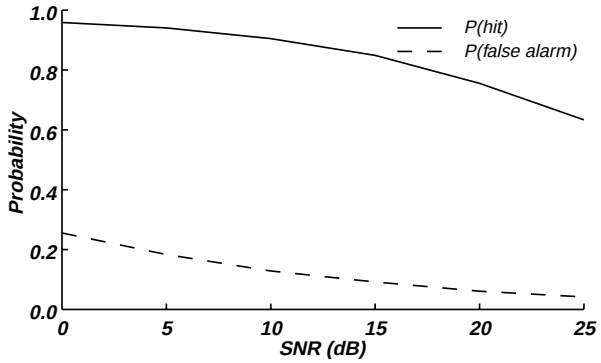


Figure 6.10 Mask accuracy for the voiced speech regions as a function of SNR for speech corrupted with white noise. The prior probabilities were constant over all SNRs and subbands. Corrupts elements labeled correctly are "hits". Reliable elements labeled as corrupt are "false alarms".

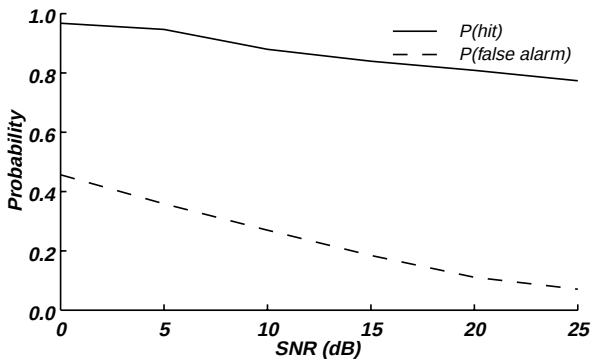


Figure 6.11 Mask accuracy for the unvoiced speech regions as a function of SNR for speech corrupted with white noise. The prior probabilities were constant over all SNRs and subbands. Corrupts elements labeled correctly are "hits". Reliable elements labeled as corrupt are "false alarms".

As before, missing feature reconstruction was performed on the elements declared corrupt by the estimated masks and the reconstructed vectors were converted to cepstra for recognition. Figure 6.12 shows the recognition accuracy as a function of global SNR for speech corrupted with white noise when the estimated masks are used with the cluster-based missing feature reconstruction method. For comparison, the figure also shows the recognition accuracy obtained when spectral subtraction is used for noise compensation and when cluster-based reconstruction is performed using VTS-based masks. Figure 6.13 shows the

same plots when the correlation-based method is used for missing feature reconstruction. All of the mask/ missing feature compensation methods clearly outperform spectral subtraction for noise compensation. The performance of the classifier-based masks is almost identical to, but not better than VTS-based mask estimation, the best mask estimation technique reported in the literature [43]. Both mask estimation techniques are close to the oracle mask performance, especially at high SNRs. These results are very encouraging, as classifier-based mask estimation has achieved comparable performance to the best reported mask estimation method, yet requires none of the stationarity assumptions about the noise used by other methods.

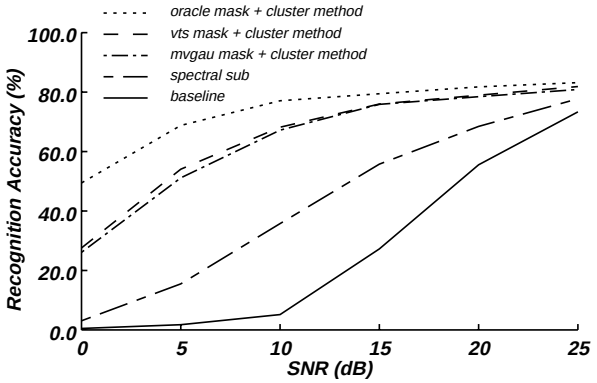


Figure 6.12 Recognition accuracy using cluster-based reconstruction vs. SNR for speech corrupted by white noise.

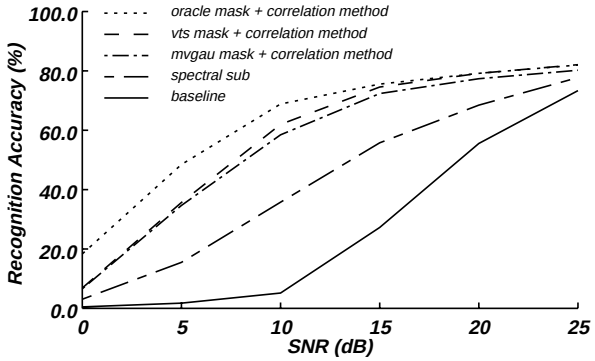


Figure 6.13 Recognition accuracy using correlation-based reconstruction vs. SNR for speech corrupted by white noise.

6.5.2 Mask Estimation on Speech Corrupted by Music

Similar experiments were performed on speech corrupted with music. The training and test set utterances from Resource Management were corrupted with music from the “Marketplace” radio show. This music is highly non-stationary. Because music also has a strong harmonic structure, the performance of the pitch detection algorithm was poor on speech corrupted with music. Therefore, for the purposes of these experiments, pitch estimates extracted from clean speech were used to derive the pitch-dependent features to train and test the classifier. Again, a cross validation set also corrupted with music was used to determine the optimal prior probabilities of the two classes. As in the white noise case, mask estimation was performed using several different prior probability values, followed by cluster-based missing feature reconstruction and recognition. Figure 6.14 shows the recognition accuracy over three different SNRs for various prior probabilities. In this case, the prior probabilities that gave the best recognition performance

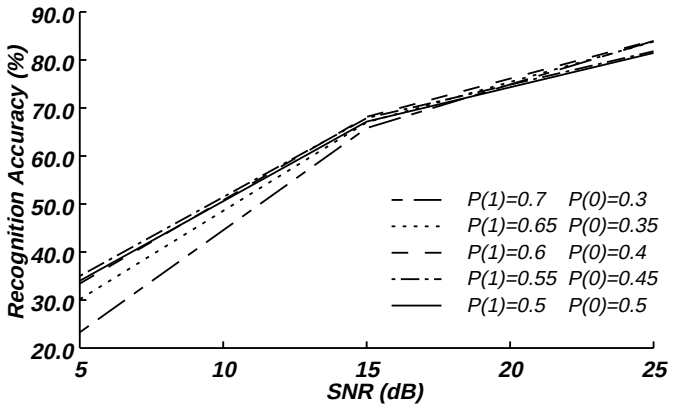


Figure 6.14 Recognition accuracy of the cross validation set vs. SNR for speech corrupted with music, using variou prior probabilities in the classifier to estimate the spectrographic masks and then applying cluster-based missing fea ture compensation.

were $P_\lambda(\omega_0) = 0.4$ for corrupt elements and $P_\lambda(\omega_1) = 0.6$ for reliable elements.

Using these values for the prior probabilities, mask estimation was performed on the test set of music-corrupted speech. The accuracy of the mask estimation for the voiced and unvoiced speech regions is shown in Figure 6.15 and Figure 6.16, respectively. Masks generated by the classifier were passed to the

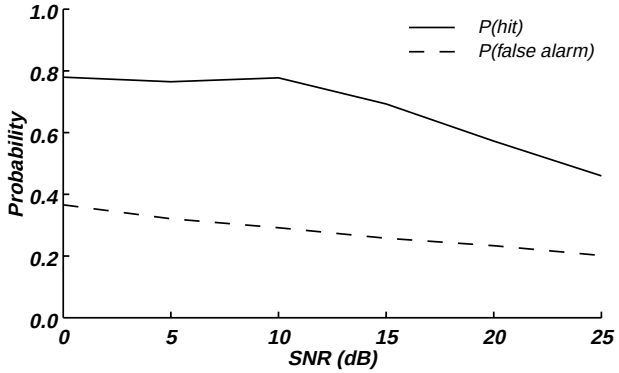


Figure 6.15 Mask accuracy for the voiced speech regions as a function of SNR for speech corrupted with music. The prior probabilities were constant over all SNRs and subbands. Corrupts elements labeled correctly are “hits”. Reliable elements labeled as corrupt are “false alarms”.

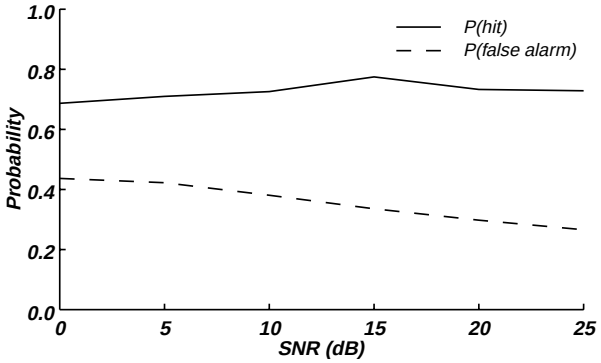


Figure 6.16 Mask accuracy for the unvoiced speech regions as a function of SNR for speech corrupted with music. The prior probabilities were constant over all SNRs and subbands. Corrupts elements labeled correctly are “hits”. Reliable elements labeled as corrupt are “false alarms”.

missing feature algorithms and the spectrograms of the music-corrupted speech were reconstructed. The reconstructed log spectra were converted to cepstra and recognition was performed. Figure 6.17 shows the recognition accuracy as a function of SNR for speech corrupted with music when the cluster-based reconstruction method is used. Again, the recognition results obtained using spectral subtraction for noise compensation and using VTS-based masks with cluster based reconstruction are shown for comparison. Figure

6.18 shows the same results when correlation-based methods are used for reconstruction. In these experiments, spectral subtraction compensation fails because the corrupting noise (music in this case) is highly non-stationary. For the same reason, VTS-based mask estimation also fails. In fact, the VTS-based masks results in recognition accuracy that is worse than the baseline performance with no compensation. However, the classifier-based masks give a consistent improvement over all SNRs with both missing feature reconstruction techniques. This is a very significant result, as we have now shown that a single classifica-

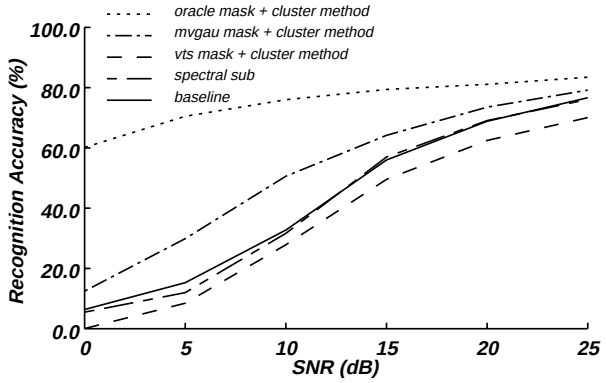


Figure 6.17 Recognition accuracy using cluster-based reconstruction vs. SNR for speech corrupted by music.

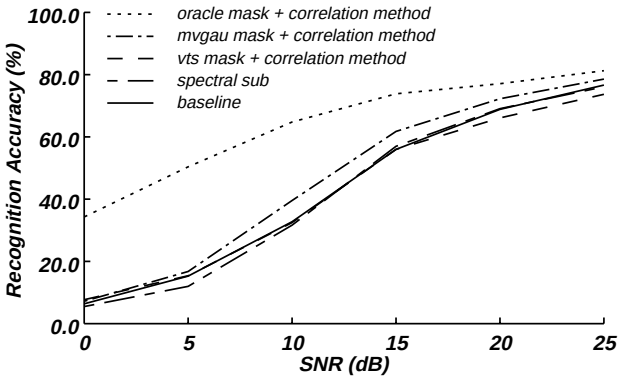


Figure 6.18 Recognition accuracy using correlation-based reconstruction vs. SNR for speech corrupted by music.

tion-based mask estimation method can successfully generate spectrographic masks for speech that is corrupted with both stationary and non-stationary noise. No other previously reported method has been able to do so.

6.5.3 Extensions to the Classification Strategy

These experiments show that classifier-based mask estimation can successfully generate spectrographic masks for speech corrupted by both stationary and non-stationary noises. This is a significant improvement over previous methods which fail in the presence of non-stationary noise. However, the recognition accuracy is still below the upper bound attainable with oracle masks. In this section, we will explore extensions to the multivariate Gaussian Bayesian classification scheme to improve the performance of the classifier and ultimately the recognition accuracy. In any pattern recognition task, changes can be made to the classification strategy in one of three areas: the feature set, the classifier output, and/or the models of the class distributions. Experiments were performed using the original classifier modified in each of these three areas.

6.5.3.1 Incorporating Neighboring Features

In the basic multivariate Gaussian classification used, each spectrographic element was treated as an independent entity. Clearly though, there is some correlation between the reliability of a given pixel and that of its neighbors. This accounts for the “block” nature of the oracle spectrographic masks. We can incorporate the information contained in the neighborhood around a spectrographic element in several ways. One method is to extend the feature vector of each element to include the feature of the surrounding pixels. We can concatenate the features for any or all of the nine surrounding pixels to the feature vector of the target location. There is a computational cost associated with elongating the feature vector as an N-dimensional multivariate Gaussian classifier becomes an $(L+1)*N$-dimensional classifier, where L is the total number of neighboring pixels.

To strike a balance between additional neighboring information and computational cost, we chose to add the features of the neighboring left, right, top, and bottom neighbors to the feature vector of every location, making each feature vector $5*N$ elements long.

The classifier was retrained using these extended feature vectors. A full covariance matrix was maintained. Everything else was identical to the original classification setup. The recognition results for speech corrupted with white noise are shown in Figures 6.19a and 6.19b, for the cluster-based and correlation-based reconstruction methods, respectively. Compared to the original classifier, there was no improvement using the cluster-based method and a slight improvement in the correlation based method. The results for speech corrupted with music are shown in Figures 6.19c and 6.19d. Again, there is no improvement in recognition accuracy with the cluster-based method, but a small improvement when the correlation-based method is used.

It is somewhat surprising that using these extended feature vectors did not improve the recognition accuracy at all for the cluster-based method and only marginally with the correlation-based method. However, it is possible that the classifier using these elongated feature vectors has a different bias than the original classifier. If this is the case, the prior probabilities established by the cross-validation data may not be optimal for the extended feature vector case, and that further improvements may be seen if cross validation is repeated for this classifier.

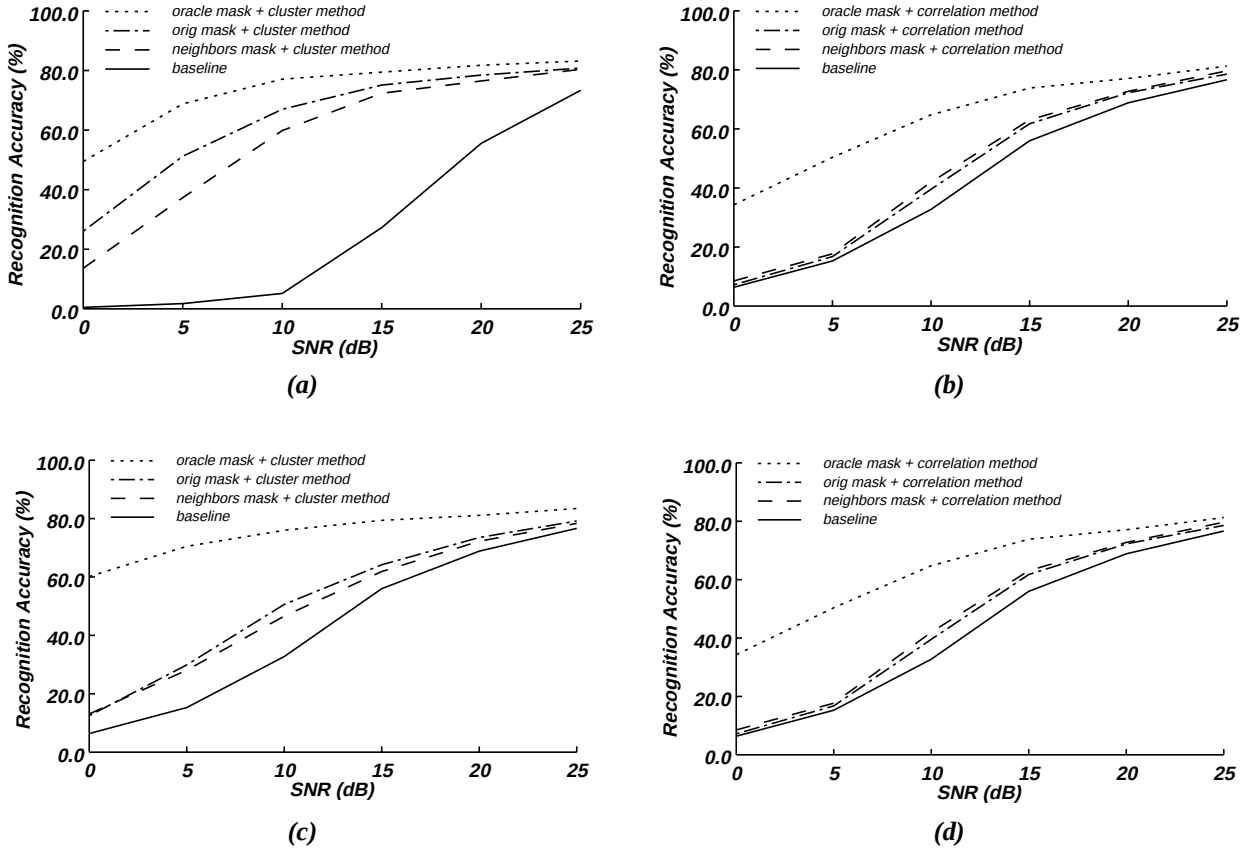
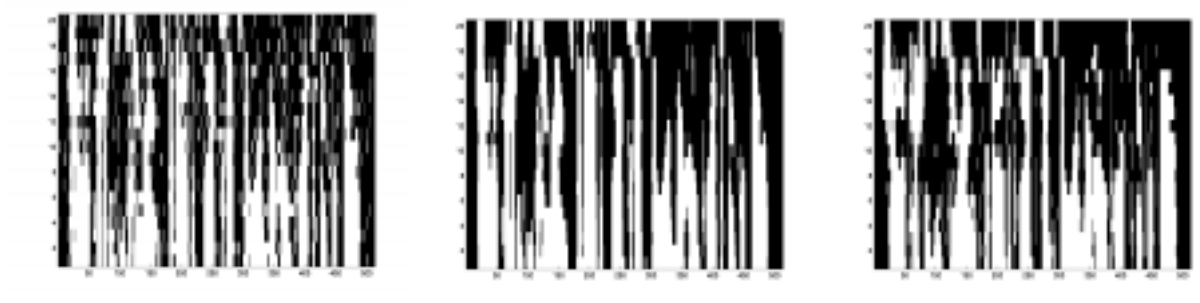


Figure 6.19 Recognition accuracy vs. SNR when features of neighboring pixels are included in the feature vector. *(a)* speech corrupted by white noise, cluster-based reconstruction, *(b)* speech corrupted by white noise, correlation-based reconstruction, *(c)* speech corrupted by music, cluster-based reconstruction, *(d)* speech corrupted by music, correlation-based reconstruction.

6.5.3.2 Median Filtering the Mask

Another simple way to capture the “block” nature of the spectrographic masks is through post processing. Two-dimensional median filtering is used in image processing to reduce impulsive noise and salt-and-pepper noise [26]. It also preserves edges in an image while reducing random noise. In median filtering, a window is moved around the image and the pixel in the center of the window is replaced the median value of the pixels contained within the window. The estimated spectrographic masks generated by the original classifier were median-filtered using a 3x3 pixel window. Figure 6.20 shows the oracle mask, the original estimated mask, and the mask after median smoothing. The filter is very effective at “solidifying” the estimated unreliable regions.

Reconstruction using the missing feature methods was performed using the masks that were generated



(a) *(b)* *(c)*

Figure 6.20 A comparison of the estimated mask to the oracle mask: *(a)* the original estimated mask *(b)* the estimated mask after median smoothing *(c)* the oracle mask

by the original classifier and then median filtered.The recognition results for speech corrupted with white noise are shown in Figures 6.21a and 6.21b, for the cluster-based and correlation-based reconstruction methods, respectively. There was no improvement for the cluster-based method, but median filtering the masks did improve the recognition accuracy slightly when correlation-based reconstruction methods were applied. Improvements were also only seen using the correlation-based methods when the speech was corrupted with music, as shown in Figures 6.21c and 6.21d. No improvement was achieved with the cluster-based method.

Here too, the results are somewhat disappointing. Median-filtering smooths the masks, effectively reducing transients and random “noise”. If the masks are accurate to begin with and classifier is relatively free of bias, median filtering will properly smooth the mask as expected. However, both high estimation error and classifier bias can cause the median filter to “incorrectly” smooth the mask. For example, in a region of the spectrogram that is corrupt, median filtering a correctly labeled pixel that is surrounded by pixels that have been incorrectly labeled as reliable will result in the properly identified corrupt pixel getting relabeled as reliable. Furthermore, in situations where the noise is very transient, the smoothing effect of median filtering may actaully reduce the mask accuracy.

6.5.3.3 Classifier Adaptation

The goal of classifier adaptation is to use the test data itself to improve the models of the class distributions that were derived from training data. We implemented a very basic adaptation scheme in our classifi-

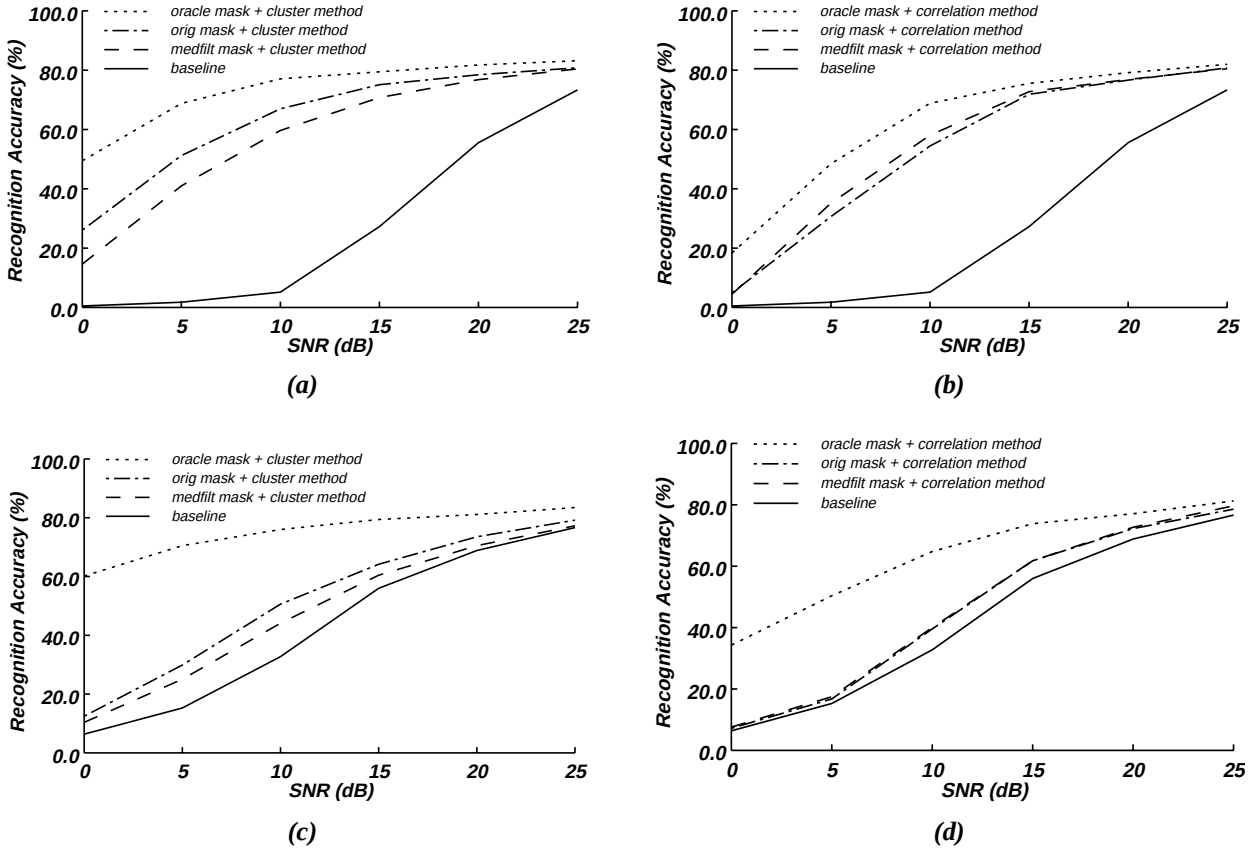


Figure 6.21 Recognition accuracy vs. SNR when the original masks are median filteried. *(a)* speech corrupted by white noise, cluster-based reconstruction, *(b)* speech corrupted by white noise, correlation-based reconstruction, *(c)* speech corrupted by music, cluster-based reconstruction, *(d)* speech corrupted by music, correlation-based reconstruction.

cation system. Classification was performed as described in Chapter 6. Now each spectrographic element was initially labelled as corrupt or reliable. Using this initial class assignment, the means of the distributions of each feature in each class were re-estimated. If any feature in any class in any subband had no samples in the utterance, the original mean of that distribution was retained. Otherwise, the new means replaced those derived from the training data. Then, a second pass through the utterance was performed, re-estimating the spectrographic mask using the newly adapted means. This was done separately for each utterance. That is, for a given utterance, the first pass of mask estimation was performed using the class distributions of the features derived from the training data. The means were adapted based on the initial mask estimate. A second pass of mask estimation was then performed using the adapted means. This was repeated for every utterance. Because the accuracy of the initial mask was not verified, corrected or adjusted in any way, this is an *unsupervised* adaptation scheme. The recognition results for speech cor-

rupted with white noise are shown in Figures 6.22a and 6.22b, and in Figures 6.22c and 6.22d for speech corrupted with music. No improvement was seen with this adaptation scheme.

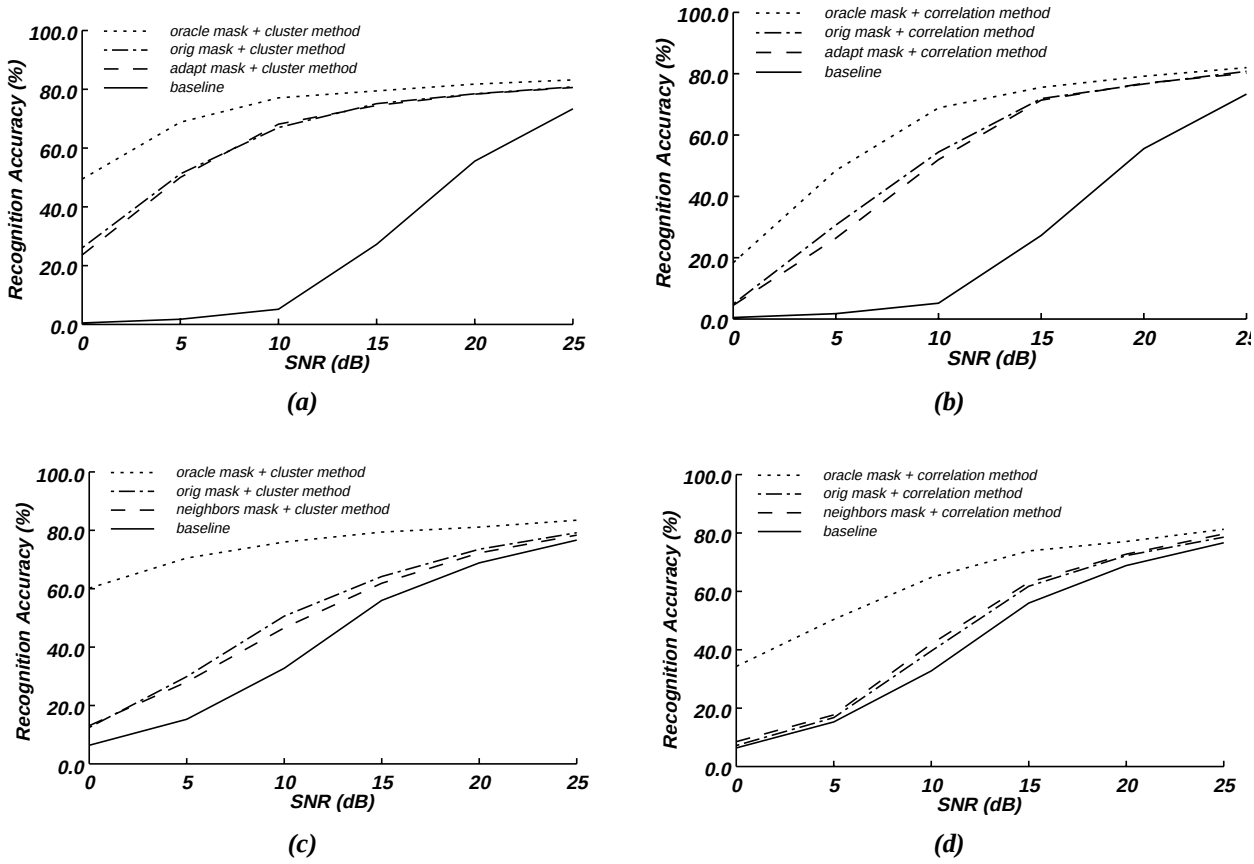


Figure 6.22 Recognition accuracy vs. SNR when unsupervised adaptation is performed. *(a)* speech corrupted by white noise, cluster-based reconstruction, *(b)* speech corrupted by white noise, correlation-based reconstruction, *(c)* speech corrupted by music, cluster-based reconstruction, *(d)* speech corrupted by music, correlation-based reconstruction.

Adaptation requires that the initial estimate used to adapt the model parameters be accurate. If the first pass of mask estimation is poor and there is non-negligible estimation error, adaptation will actually move the means of the two classes toward each other resulting in less separable distributions and poor classification. Because the mask estimates in the original classifier contained significant estimation error, as seen in Figures 6.10, 6.11, 6.15, and 6.16, adaptation did not improve mask estimation performance or recognition accuracy.

6.6 Summary and Conclusions

In this chapter we have described the classification scheme that uses the features extracted from the noisy speech and the distributions of these features estimated from the training data to determine whether a spectrographic element is more likely to be reliable or corrupt. We have demonstrated how we can choose the class prior probabilities to bias the classifier to favor misses over false alarms. Finally, we have shown how these estimated masks, in conjunction with the missing feature methods, can increase the speech recognition accuracy on noise-corrupted speech. The recognition performance obtained using these masks is comparable that obtained using the best mask estimation technique reported in the literature for speech corrupted by white noise. However, the recognition performance for speech corrupted by music using these classifier based-spectrographic masks is better than that obtained using other reported methods. In fact, other reported methods produced no improvement in recognition accuracy for speech corrupted by music. Because no assumptions about the noise itself were made in the design of the classifier and the feature set, the classifier can estimate spectrographic masks when speech has been corrupted by both stationary and non-stationary noises.

Extensions to the original classification scheme were also implemented. Experiments were performed in which the feature vector was elongated by incorporating the features of neighboring spectrographic locations, the masks were post processed using a median filter, and the class distributions were adapted in an unsupervised manner. Using the features of the neighboring pixels and median filtering resulted in small improvements in recognition accuracy when the correlation-based missing feature method was used to reconstruct the spectrograms. Adaptation did not result in any improvements in the recognition accuracy with either missing feature method.

In the next chapter, the major findings of this thesis are summarized and discussed.

Chapter 7
Summary and Conclusion

7.1 Summary and Conclusions

The missing feature noise compensation paradigm operates on the principle that noise affects different regions of a spectrographic display of speech differently depending on the relative energies of the speech and the noise at each time-frequency location. The regions of low SNR will be more corrupt that those of high SNR. Noise-corrupted regions of a spectrogram are deleted to minimize the effect of the noise on the speech resulting in incomplete spectrograms. Recognition is then either performed directly on the incomplete spectrograms or the missing regions are reconstructed prior to recognition. Missing feature methods make no assumptions about the stationarity of the corrupting noise. This is a significant advantage over previous noise compensation methods that require that the noise be stationary. However, missing feature methods require a mask that labels every spectrographic element as either corrupt or reliable.

Previous methods of mask estimation rely on the same stationarity assumption that plague earlier noise compensation methods. As a result, they can successfully estimate spectrographic masks when the speech is corrupted by stationary noise, but fail completely when the noise is non-stationary. This is a sizable limitation, as the key benefit of missing feature methods is their successful compensation of speech corrupted stationary *or* non-stationary noise if the spectrographic masks are known.

In this thesis we have designed a classifier to automatically generate spectrographic masks for noise corrupted speech. We have shown that classifier-based mask estimation is a consistent and reliable method of estimating spectrographic masks for noisy speech, and because no assumptions about the noise itself were made in the design of the classifier and the feature set, it performs well on both stationary and non-stationary noises.

The experiments performed with speech corrupted by white noise showed that the use of masks estimated by the classifier results in recognition accuracy that is close to the accuracy possible when oracle masks that assume perfect knowledge of the noise are used. The recognition accuracy achieved using VTS-based mask estimation techniques described in [43] is marginally better than when the classifier-based

mask are used. While experiments with stationary noise were only conducted with white noise, similar performance can be expected on other stationary or quasi-stationary noise signals, such as automobile or cockpit noise.

The classifier-based mask estimation method also resulted in a consistent improvement in recognition accuracy when the speech was corrupted by music. While the recognition accuracy was not as close to the oracle mask recognition accuracy as in the white noise case, the improvement was consistent across all SNRs and both missing-feature reconstruction methods. In addition, it was the only mask estimation technique that yielded any improvement in recognition accuracy. Using VTS-based masks actually resulted in accuracies that were worse for background music than baseline performance when no compensation is applied.

Three extensions to the basic classification scheme were tried in an effort to further improve the mask estimation performance. Elongating the feature vector by incorporating the features of neighboring pixels, and post-processing the masks with a median filter both resulted in small improvements in recognition accuracy when the correlation method of missing feature compensation was applied. It is interesting to note that none of these extensions to the classification strategy improved the recognition accuracy when the cluster-based method was used for spectrogram reconstruction. This is perhaps because both median filtering and elongating the feature vector attempt to capture information about the pixels surrounding a given pixel across both the time and frequency dimensions, much in the same way that the correlation-based reconstruction method uses the neighboring pixels across time and frequency to estimate the value of a missing pixel. The cluster-based reconstruction method, on the other hand, uses only information along the frequency axis.

The features used by the classifier for voiced regions of speech depend on an accurate pitch estimate to assess the reliability of each spectrographic location. The classification scheme also required an accurate labeling of voiced and unvoiced frames as distinct classifiers were used for each. These requirements were addressed by the histogram-based pitch detection algorithm presented in this work. A new multi-band approach to pitch detection was presented. This method was shown to be generate more reliable pitch estimates that the widely used RAPT pitch detection algorithm for both clean and noise-corrupted speech.

It should be noted that the pitch estimates used in these experiments when speech was corrupted with

music were extracted from clean speech, rather than from the music-corrupted speech. Of course, this would not be possible in a real situation. However, the pitch extraction failed not because of any stationarity (or lack thereof) assumption about the noise, but rather because the corrupting signal (music) had a harmonic structure similar to speech which causes false pitch estimates. However, there are few “real world” situations, such as television and radio news broadcasts, where speech is corrupted by music. It is more common for the noise to be both non-stationary and non-musical, such as transients, street noise, and factory noise. As long as the noise is non-musical (or non-harmonic), we expect that the pitch detection algorithm will provide accurate pitch estimates in the presence of non-stationary noise and the classifier-based masks will produce improvements in recognition that are comparable or better to those seen in this thesis when speech was corrupted by music.

Because a classifier is used to estimate the spectrographic masks, we need to have enough training examples of the noise to adequately determine the distributions of the features for each class. This may not be readily available. However, the VTS-based mask estimation method requires training examples as well. For stationary noises, there are other mask estimation methods that do not require any training data such as those described in [8] which use spectral subtraction to obtain a running estimate of the noise. However, the recognition accuracy obtained with these methods is significantly worse than that achieved by either classifier-based masks or VTS-based masks. And again, spectral subtraction mask estimation completely fails for non-stationary corrupting noises.

Missing-feature methods for noise compensation in speech recognition are gaining popularity both because of the significant improvements in recognition accuracy they are capable of and because the concept makes logical sense based on our knowledge of the human auditory system. Similarly, building a classifier that uses features that are based on the intrinsic characteristics of the speech signal itself is also intuitively satisfying. Because no assumptions about the noise are made, it is logical that the classifier will be able to estimate spectrographic masks for many, if not all, noise types. This is a significant improvement over previous mask estimation methods. While they performed will in some situations, they performed very poorly in others. The classifier-based mask estimation method presented here is a much more general purpose solution.

7.2 Suggestions for Future Work

While the methods presented in this thesis produce effective spectrographic masks that significantly improve recognition accuracy when speech has been corrupted by noise, there is much additional work that can be done to improve the mask estimation.

The pitch estimation algorithm is very accurate at estimating the pitch contour of a utterance in both clean and noisy conditions. However, it is quite slow. There is room for efficiency and optimization in both the algorithm and the implementation. For example, an interesting question is whether the Seneff filter-bank actually play a part in the success of the algorithm, or if any other, perhaps simpler bank of band-pass filters would be as effective and computationally simpler. Also, pitch estimation in the presence of musical or otherwise harmonic noise is a large on-going research problem. The histogram-based pitch detection algorithm was not able to reliably estimate the pitch when speech was corrupted by such noise. The pitch detection algorithm and the classification scheme share some operations. However, currently they are two independent entities. Efficient integration of the pitch detection algorithm into the feature extraction would also dramatically reduce the computational load. For example, the autocorrelations that are used to estimate the pitch could also be used to extract the autocorrelation peak ratio feature for the classifier. However, currently there is a disparity in the band-pass filtering. The pitch detection algorithm uses forty Seneff filters and the classifier estimates masks using twenty Mel filters. These inconsistencies would need to be resolved to smoothly integrate the pitch detection and the feature extraction.

In this thesis, we have presented experiment using speech corrupted by white noise and by music. However, these are probably the two extreme cases of stationarity and non-stationarity. The performance of the classifier when the speech is corrupted by other more realistic noises, such as automobile or cockpit noise, or by transients, such as door slams or factory noise, needs to be determined. Additionally, in these other noise cases, other features might be useful. For example, derivative based features would be helpful at estimating sudden changes in energy or noise level across time. They could detect the onset or offset of transient noise.

Finally, the classification strategy in this work centered around a multivariate Gaussian classifier. However, a more complex classifier based on a higher order Hidden Markov Model or two-dimensional Markov field would enable us to model both the feature distributions of the two classes and the relationships between neighboring spectrographic elements.

References

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

[2] P. Bagshaw, S. M. Hiller, and M.A. Jack, "Enhanced Pitch Tracking and the Processing of F0 Contours for Computer Aided Intonation Teaching," *Proc. Eurospeech '93*, p.1003-1006.

[3] Beauvois, M. W., Meddis R., "A computer model of auditory stream segregation," *Quarterly Journal of Experimental Psychology*, vol. 43A no. 3, Aug. 1991, p. 517-541.

[4] Brown, G.J., Cooke, M., "Computational Auditory Scene Analysis," *Computer Speech and Language*, vol. 8 no. 4, Oct. 1994. p. 297-336.

[5] Boll, S.F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 27 no. 2, April 1979, p. 113-120.

[6] Bregman, A., *Auditory Scene Analysis*, London, England: MIT Press, 1990.

[7] Cooke, M.P., Green, P.G., Crawford, M.D., "Handling missing data in speech recognition," *Proc. ICSLP '94*, p. 1555-1558.

[8] Cooke, M.P., Morris, A., Green, P.D., "Missing data techniques for robust speech recognition," *Proc. ICASSP '97*, p. 863-866.

[9] Cooke, M., Green, P., Josifovski, L., Vizinho, A., "Robust ASR with Unreliable Data and Minimal Assumptions," *Proc. Robust '99.*

[10] Cooke, M., Green, P., Josifovski, L., Vizinho, A., "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data," to be published in *Speech Communication.*

[11] Cooke, M., Green, P., "Auditory Organization and Speech Perception: Pointers for Robust ASR," to appear in *Listening to Speech*, editors Greenberg and Ainsworth, Oxford University Press.

[12] Centre for Speech Technology Research, University of Edinburgh, http://www.festvox.org/dbs/dbs_kdt.html

[13] Davis, S., Mermelstein, P., "Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 28 no. 4, August 1980, p. 357-366.

[14] Dempster, A.P, Laird, N.M, Rubin, D.B., "Maximum likelihood from incomplete data via the EM algorithm (with discussion)," *Journal of the Royal Statistical Society*, Series B, Vol. 39, 1977, p. 1-38.

[15] Duda, R.O., Hart, P.E., *Pattern Classification and Scene Analysis*, New York, NY: John Wiley and Sons, 1973.

[16] Ellis, D.P.W., *Prediction-driven Computational Auditory Scene Analysis*, Ph.D. disseration, MIT, June, 1996.

[17] Fletcher, H., *Speech and Hearing in Communication*, Van Nostrand: New York, NY, 1953.

[18] Gales, M.J.F., Young, S.J., "HMM recognition in noise using parallel model combination," *Proc. Eurospeech '93*, p 837-840.

[19] 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*, vol. 2, no. 2, April, 1994. p. 291-298.

[20] Hess, W.H., *Pitch Determination of Speech Signal: Algorithms and Devices*, Heidelberg, Germany: Springer-Verlag, 1983.

[21] Higgins, R.J., *Digital Signal Processing in VLSI*, Englewood Cliffs, NJ: Prentice Hall, 1990.

[22] Hirsch, H.G., Ehrlicher, C., "Noise estimation techniques for Robust Speech Recognition," *Proc. ICASSP '95*, p. 153-156.

[23] Kay, S.M., *Modern Spectral Estimation: Theory And Application*, Englewood Cliffs, NJ: Prentice Hall, 1988.

[24] Lamel, L., Kassel, R., Seneff, S., "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, 1986, p. 100-109.

[25] Leggetter, C. J., Woodland, P. C. (1994), "Speaker Adaptation Of HMMs Using Linear Regression," *Technical Report CUED/F-INFENG/ TR. 181*, Cambridge University Engineering Department, Cambridge, June 1994.

[26] Lim, J., *Two-Dimensional Signal and Image Processing*, Englewood Cliffs, NJ: Prentice Hall, 1990.

[27] Lippmann, R.P., "Speech recognition by machines and humans," *Speech Communication*, vol. 22 no. 1, July 1997, p. 1-16.

[28] Lippmann, R.P., Carlson, B.A., "Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise," *Proc. Eurospeech '97*, p. KN37-40.

[29] Little, R.J.A., Schluchter, M.D, "Maximum likelihood estimation for mixed continuous and categorical data with missing values," *Biometrika*, vol. 72, 1985, p. 497-512.

[30] Meddis, R., Hewitt, M.J.,"Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. Pitch identfication," *Journal of the Acoustical Society of America*, vol. 89 no. 6, June, 1991, p. 2866-2882.

[31] Moore, B.C.J., *An introduction to the Psychology of Hearing*, San Diego, CA: Academic Press, 1997.

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

[33] Morgan, D.P., George, E.B, Lee, L.T., Kay, S.M., “Cochannel speaker separation by harmonic enhancement and suppression,” *IEEE Trans. on Speech and Audio Processing*, vol. 5 no. 5, Sept. 1997, p. 407-424.

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

[35] O’Shaughnessy, D., *Speech Communication - Human and Machine*, Reading, MA: Addison-Wesley, 1987.

[36] Papoulis, A., *Probability, Random Variables, and Stochastic Processes*, New York, NY: McGraw Hill, 1991.

[37] Porter, J.E., Boll, S.F., “Optimal estimators for spectral estimators of noisy speech,” *Proc. ICASSP ’84*, p.18A.2.1-4.

[38] Price, P., Fisher, W.M., Bernstein, J., Pallet, D.S., “The DARPA 1000 word Resource Management database for continuous speech recognition,” *Proc. ICASSP ‘88*, p. 651-654.

[39] Quatieri, T.F., “An Approach to Co-Channel Talker Interference Suppression Using a Sinusoidal Model for Speech,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, Jan. 1990, p. 56-69.

[40] Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., McGonegal, C.A., “A Comparitive Study of Several Pitch Detection Algorithms,” *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-24, p. 399-417, Oct. 1976.

[41] Rabiner, L.R., Juang, B.-H., *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ: Prentice-Hall, 1993.

[42] Rabiner, L.R., Schafer, R.W., *Digital Processing of Speech Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1978.

[43] Raj, B., *Reconstruction of Incomplete Spectrograms for Robust Speech Recognition*, Ph.D. Dissertation, Carnegie Mellon University, May 2000.

[44] Raj, B., Parikh, V., Stern, R.M., “The Effects of Background Music on Speech Recognition Accuracy,” *Proc. ICASSP ’97*, p 851-854.

[45] Seneff, S., “A Joint Synchrony/Mean-Rate Model of Auditory Speech Processing,” *Journal of Phonetics*, vol. 16 no. 1, January, 1998.

[46] Sullivan, T., *Multi-Microphone Correlation-Based Processing for Robust Automatic Speech Recongition*, Ph.D. Dissertation, Carnegie Mellon University, August, 1996.

[47] Stark, H., Woods, J.W., *Probability Theory, Random Processes, and Estimation Theory for Engineers*, Englewood Cliffs, NJ: Prentice Hall, 1994.

[48] Talkin, D., “A Robust Algorithm for Pitch Tracking (RAPT)”, in *Speech Coding and Synthesis*, Amsterdam, NL: Elsevier Science, 1995, p 495-518.

[49] Therrien, C.W., *Discrete Random Signals and Statistical Signal Processing*, Englewood Cliffs, NJ: Prentice Hall, 1992.

[50] Weintraub M., *A Theory and computational model of monaural auditory sounds separation*, Ph.D. Dissertation, Stanford University, 1985.