RECONSTRUCTION OF INCOMPLETE SPECTROGRAMS FOR ROBUST SPEECH RECOGNITION

Bhiksha Raj Ramakrishnan

Department of Electrical and Computer Engineering

Carnegie Mellon University

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in Electrical and Computer Engineering

Pittsburgh, Pennsylvania

April 2000

Dedicated to my parents:

Smt. Lalitha Ramakrishnan

and

Shri. K.S. Ramakrishnan

ABSTRACT

The performance of automatic speech recognition (ASR) systems degrades greatly when speech is corrupted by noise. Missing feature methods attempt to reduce this degradation by deleting components of a time-frequency representation of speech (such as a spectrogram) that exhibit low signal-to-noise ratio (SNR). Recognition is then performed using only the remaining components of the incomplete spectrogram. These methods have been shown to result in recognition accuracies that are very robust to the effects of additive noise. However, conventional missing feature methods, which modify the classifier used to perform the recognition, suffer from the drawback that they are constrained to use the log-spectral vectors of the spectrogram as features for recognition. It is well known recognition systems that use log-spectral features perform poorly compared to systems that use cepstral features.

In this thesis we propose two new approaches that recast the missing feature paradigm as a data compensation problem, by reconstructing missing elements to obtain complete spectrograms. In the first approach, referred to as cluster-based reconstruction, incoming log-spectral vectors from clean speech are clustered. Missing spectrographic features from noisy data are recovered by first identifying the closest cluster based on the values of the features that are present, and then estimating the missing values using MAP procedures. The second approach, referred to as covariance-based reconstruction, uses MAP procedures to estimate the value of the missing components of the spectrogram based on their correlations with the elements that are present. Both methods take into account the bounds on the clean spectrogram imposed by additive noise. In either case, cepstral features are computed from the reconstructed spectrograms and used for recognition without any modification of the speech recognition system.

When corrupt regions of the spectrogram are known *a priori*, recognition accuracies resulting from reconstruction methods are seen to be much higher than those obtained with the best current missing feature methods based on modification of the recognition system. The proposed spectrogram reconstruction methods are also computationally less expensive than the best conventional missing feature methods.

We also propose two methods that attempt to identify corrupt regions of the spectrographic representations of incoming speech. The first method utilizes noise spectrum estimates of vector Taylor series (VTS) compensation for noise-corrupted speech, while the second method treats the identification task as a classic Bayesian classification problem. Combination of the best method to identify corrupt regions with the

best method to reconstruct them produces recognition accuracies better than any other known algorithm for speech in additive white noise. We also observe significant improvement in recognition accuracy for speech in the presence of background music if the locations of corrupted spectrographic regions are known *a priori*, but we have been less successful in blind identification of these corrupt regions for these signals.

Acknowledgements

During the course of this thesis I have accumulated a great debt of gratitude to a great many people:

My advisor, Richard Stern, who has been immensely supportive and has guided me through every problem I've encountered with patience, wisdom, and knowledge, and who has been the single greatest influence on this thesis,

My friends Pedro and Evandro, who have been there for me at various times of doubt, both materially and in spirit,

Rita, who I was forever turning to for her insights into my problem and for clarification of concepts,

Uday and Vipul, whose company and ready wit lightened many a heavy hour at work,

Ravi Mosur and Eric Thayer, formerly of the speech group in CMU, who developed the sphinx recognition system, and have always been forthcoming with their advice and help on matters related to the system,

My friends from an earlier life - Krishnan, who taught me all I know about speech recognition, Chakra, who gave me my first lessons on professionalism, Alok, who taught me to question the most mundane of statements, and Saravanan, who taught me to appreciate the meaning of mathematical symbols,

Mike, Carol, and Jon, who helped me greatly in arranging my defense,

Dr. Jordan Cohen, Dr. John Hampshire, and Dr. Raj Reddy, for being kind enough to serve on my thesis committee,

Many others who space, time, and memory will not permit me to list individually -

To all these people I owe the body of this thesis.

But my greatest debt of gratitude is to those pillars of my life, who have believed in me, supported me, encouraged me, and have been there for me at every bend of the road, every day of my life: my parents and my sister.

To them I owe the soul of this thesis, and the soul of every other thesis that might have been had I chosen a different walk in life.

Table of Contents

List of Figures

Glossary of terms

Bounded class-conditional imputation: Class-conditional imputation where estimates of missing components are found using bounded MAP estimation using the distributions of the classes.

Bounded cluster marginal reconstruction: Cluster marginal reconstruction where missing elements between $-\infty$ and an upper bound are integrated out of cluster distributions to estimate cluster membership of vectors. Missing elements are estimated using bounded MAP estimation based on the distribution of the estimated cluster.

Bounded covariance-based reconstruction: Covariance-based reconstruction, where missing elements are estimated using bounded MAP.

Bounded MAP estimation: MAP estimation where the estimated value of the variable is forced to lie within an upper bound.

Bounded marginalization: Marginalization where missing components of spectral vectors between $-\infty$ and a given upper bound are integrated out of the distributions of speech classes.

Class-conditional imputation: Missing-feature method where recognition is performed with incomplete spectrograms. In order to compute the likelihood of any sound class during recognition, missing components of spectral vectors are estimated based on the distribution of that class.

Classifier-compensation methods: Methods which modify the distributions of classes within the recognizer to compensate for the effect of noise.

Classifier-modification methods: Missing-feature methods where the classifier is modified to perform recognition directly using incomplete spectrograms.

Cluster-based reconstruction: Spectrogram reconstruction methods where spectral vectors are assumed to be segregated into clusters. Missing components of spectral vectors are estimated based on the distributions of these clusters.

Cluster marginal reconstruction: Cluster-based reconstruction method where the cluster membership of any vector is estimated by marginalizing the missing components of spectral vectors out of the distributions of the various clusters.

Cluster oracle reconstruction: Cluster-based reconstruction where the true cluster membership of incomplete vectors is known *a priori.*

Cluster time-interpolated reconstruction: Cluster-based reconstruction method where preliminary estimates of missing components of spectral vectors are obtained by linear interpolation along time, and the preliminary estimates are used in estimating cluster membership of vectors.

Cluster membership: The cluster that any spectral vector belongs to, in the cluster based representations used by cluster-based reconstruction methods.

Covariance-based reconstruction: Spectrogram reconstruction method where missing elements in spectrograms are estimated on the basis of their covariance with observed elements within the spectrogram.

Covariance individual reconstruction: Covariance-based reconstruction where missing elements in spectrograms are individually estimated.

Covariance joint reconstruction: Covariance-based reconstruction method where all missing elements within any spectral vector are jointly estimated.

Data-compensation methods: Methods which modify the incoming feature stream to compensate for the effect of noise on speech recognition systems.

Geometrical-reconstruction methods: Spectrogram reconstruction methods where missing elements of spectrograms are estimated by extrapolation of, or interpolation between, adjacent elements in the spectrogram.

Incomplete-spectrogram methods: Missing-feature methods where no information is assumed regarding the missing elements in the spectrogram.

Linear interpolation along frequency: Geometrical-reconstruction method where missing elements are estimated by linear interpolation between other observed elements within the same spectral vector.

Linear interpolation along time: Geometrical-reconstruction method where missing elements are estimated by linear interpolation between observed elements within the same frequency band in adjacent vectors.

Marginalization: Missing-feature method where recognition is performed with incomplete spectrograms. Missing components in spectral vectors are integrated out of the distributions of the various sound classes being considered by the recognizer.

MAP estimation: Maximum *a posteriori* estimation, where the value of a variable is estimated as the value at which the *a posteriori* distribution of the variable, conditioned on a set of observed variables, peaks.

Missing-feature methods: Methods which model the effect of noise as missing features in spectrograms and perform recognition based on the information in incomplete spectrograms.

Oracle masks: Spectrographic masks that have been obtained based on knowledge of the true SNR of the elements in the spectrograms of noisy speech.

Polynomial interpolation along frequency: Geometrical-reconstruction method where missing elements are estimated by polynomial interpolation between other observed elements within the same spectral vector.

Polynomial interpolation along time: Geometrical-reconstruction method where missing elements are estimated by polynomial interpolation between observed elements within the same frequency band in adjacent vectors.

Rational-function interpolation along frequency: Geometrical-reconstruction method where missing elements are estimated by fitting a rational function to other observed elements within the same spectral vector.

Rational-function interpolation along time: Geometrical-reconstruction method where missing elements are estimated by fitting a rational function to observed elements within the same frequency band in adjacent vectors.

Single cluster reconstruction: Cluster-based reconstruction where all spectral vectors are assumed to belong to a single cluster.

Spectrogram reconstruction methods: Missing-feature methods where missing regions of incomplete

spectrograms are estimated, to reconstruct complete spectrograms. Recognition is performed with features derived from the reconstructed spectrograms.

Spectral subtraction: Noise cancellation algorithm that maintains running estimates of the noise spectrum and subtracts them from the spectrum of noisy speech to estimate the spectrum of clean speech.

Spectral-subtraction-based mask estimation: Estimation of spectrographic masks based on the estimates of the noise spectrum computed for spectral subtraction.

Spectrographic mask: Information relating to an incomplete spectrogram that tags individual elements of the spectrogram as missing or observed.

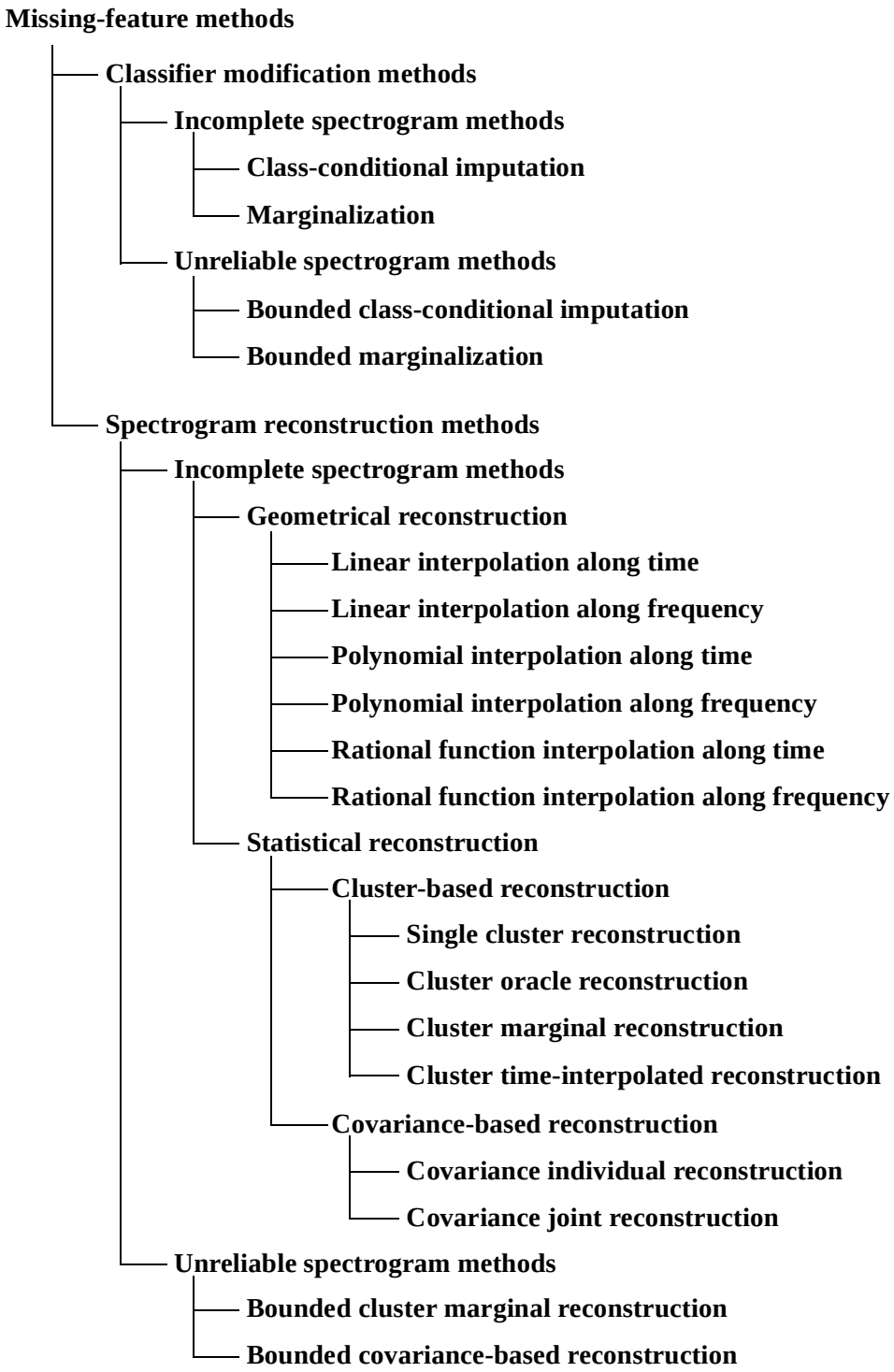
Statistical-reconstruction methods: Spectrogram reconstruction methods where missing elements of spectrograms are estimated based on their statistical relationships with the observed elements in the spectrogram.

Unreliable-spectrogram methods: Missing-feature methods where the upper bound on the missing elements of the spectrogram is assumed to be known.

Vector Taylor Series (VTS): Noise compensation algorithm that obtains maximum likelihood estimates of the noise spectrum and cancels this noise out of the parameters of noisy speech using an MMSE estimator.

VTS-based mask estimation: Estimation of spectrographic masks for noisy speech that is based on the estimate of the noise spectrum obtained by VTS.

Algorithm tree



Chapter 1
Introduction

The performance of automatic speech recognition (ASR) systems degrades greatly when the speech being recognized has been corrupted by noise [Acero 1993]. There are several reasons for this.

ASR systems are essentially statistical pattern classifiers that classify segments of sound as belonging to one of a set of sound classes. The classification is not performed using the speech signal itself, rather, the speech signal is parameterized into a sequence of *feature vectors*, and classification is performed using these feature vectors. The feature vectors themselves are variously derived from the power spectrum of short windowed segments, or *frames*, or speech. The ASR system learns the distribution of the feature vectors belonging to any sound, from a corpus of training speech. During recognition, a segment of speech is classified as belonging to the sound whose distribution is most likely to have generated the feature vectors belonging to that segment.

When speech is corrupted by stationary noise, one resulting effect is that the distribution of the feature vectors of the corrupted speech are no longer similar to the distributions learned from the training data. This mismatch results in mis-classification and poor recognition [Moreno 1996]. This effect can be minimized by training the recognition system with speech that has the same level of noise as the speech being recognized. But even in this situation, the addition of the noise results in increased error in the estimation of the spectrum of any frame of speech [Kay 1988], and therefore increases the inherent variability in the feature vectors corresponding to any sound. As a result, the variance of the distributions of the various sound classes increases, resulting in increased classification error, and increased mis-recognition over the situation where both training and test speech are noise free. Finally, when the corrupting noise itself is non-stationary, even training the system with speech corrupted to the same overall noise level as the test speech is not helpful. This is because, although the overall noise level in the training and test data are identical, this does not imply that the various examples of a sound in the training data are corrupted with the exact kind of noise that the test data has been corrupted by. Mismatches between distributions learned by the classifier and the distribution of the test data still persist.

The problem of reducing the mismatch between the distributions modeling the classes in the classifier and the distributions of the test data can be approached in two ways. In the first approach the test data are

"cleaned" in some manner in an attempt to make them similar to the training data whose distributions have been learned by the classifier. We refer to methods that attempt to compensate the test data for the effect of noise in this manner as *data-compensation methods*. In the second approach the distributions used by the classifier to models the various sound classes are modified to be similar to the distributions of the test data. We refer to methods that attempt to modify components of the classifier in this manner to compensate for the noise as *classifier-compensation methods*.

Several data compensation methods and classifier compensation methods have been proposed in the literature. Data compensation methods such as codeword dependent cepstral normalization (CDCN) [Acero 1993], vector Taylor series (VTS) [Moreno 1996], spectral subtraction [Boll 1979] and Wiener filtering [Porter 1984] attempt to compensate for the effect of the noise on the data based on estimates of the spectrum of the noise. Others such as multivariate Gaussian based cepstral compensation (RATZ) [Moreno 1996] and probabilistic optimal filtering (POF) [Neumeyer 1994] use explicit comparisons between data that have simultaneously been recorded in the training and test conditions to modify the test data. Classifier compensation methods such as parallel model combination (PMC) [Gales 1993] and model composition [Varga 1990] modify the distributions of the sound classes to account for the effect of additive noise. Others such as maximum likelihood linear regression (MLLR) [Leggetter 1994] on the other hand simply transform the parameters of the distributions to best fit the noisy test speech.

The drawback with all of these methods is that they assume, either explicitly or implicitly, that the underlying noise is stationary, and furthermore that the effect of the noise is representable by a linear transformation of the parameters of the distribution of the data. Thus, while all of these methods have been fairly successful against low to medium levels of stationary noise, *i.e.* noisy speech with signal-to-noise ratios (SNR) 10 dB or greater, they are less effective at higher levels of noise and completely ineffective in the presence of non-stationary noises [Raj 1997].

Two new approaches to robust speech recognition have been based on the observation that the human auditory system preferentially processes the high-energy components of the speech signal while suppressing the weaker components [Moore 1997]. These new approaches attempt to improve speech recognition performance by deweighting the contribution of the low SNR components of the speech to the recognition in some manner. *Multi-band* based approaches [Hermansky 1996] [Bourlard 1996] consider the fact that different frequency bands of the speech signal may be corrupted at different SNRs. They therefore decom-

pose the speech signal into separate frequency bands, and construct separate speech recognition systems for each band. The output of each of these recognition systems is then recombined to give the final output. The weight given to the output of the recognition system corresponding to each frequency band is ideally dependent on the SNR in that band, deweighting noisy bands with respect to the clean ones.

Missing-feature approaches [Cooke 1994] [Lippmann 1997], on the other hand, take into account the fact that SNR may be local not only to frequency but also in time. Speech is transformed into the time-frequency domain and represented as spectrographic images where the two axes of the image represent time and frequency respectively, and the pixel value of each element in the image represents the energy of the signal in that time-frequency location. Different regions of this spectrographic picture are corrupted to different degrees by the noise. In missing feature based approaches, the low SNR regions of this picture are selectively erased, and recognition performed on the basis of the remaining *incomplete spectrogram*. Since recognition is performed on the basis of incomplete spectrograms, we also refer to these methods as *incomplete-spectrogram methods*.

Incomplete-spectrogram methods have the advantage over other approaches that they make no assumptions, either explicit or implicit, about the stationarity of the corrupting noise. Also, they do not need to have a knowledge of the fine structure of the spectrum of the noise, needing only the coarse descriptions of the regions of the time-frequency plane as being either reliable or unreliable [Cooke 1994]. Incomplete spectrograms methods have been shown to result in recognition accuracies that are remarkably robust to high levels of noise corruption [Cooke 1999] [Cooke 2000].

All current incomplete spectrogram methods that have been reported in the literature so far are classifier-compensation methods [Cooke 1994][Lippmann 1997][Renevey 1999]. They model the effect of the incompleteness of the spectrographic data on the classifier and *the classifier is modified to compensate for the incompleteness of the data*. We refer to these missing-feature methods that modify the classifier as *classifier-modification methods* in this thesis. In order for such methods to be feasible, the classifier has to be trained with spectrographic features, *i.e.* spectra or log-spectra.

This is a serious drawback with these methods. It is well known that when recognition is performed with log spectra, the recognition accuracies obtained are much poorer than those obtained with other features, such as cepstra, that have been derived from the log spectra [Davis 1980]. As a result, even the base-

line recognition accuracy obtained with the cepstra of noisy speech with no compensation at all is frequently superior to the accuracy obtained with log spectra, even after missing-feature based compensation has been applied.

What this thesis is about

In this thesis we recast the missing-feature approach to noise robustness as a *data-compensation* problem. Instead of performing recognition directly with incomplete spectrograms we attempt to *estimate* the missing components of incomplete spectrograms and reconstruct complete spectrograms *prior to classification*. Estimation of missing regions of incomplete data has been much reported on in the fields of statistical analysis of data [Rubin 1987] [Quinlan 1989] [Ghahramani 1994]. However, to the best of our knowledge, this approach has not been applied to noise compensation for speech recognition prior to this work.

We refer to methods that estimate missing (noisy) regions of incomplete spectrograms to reconstruct complete spectrograms as *spectrogram reconstruction methods*.

The spectrogram reconstruction methods described in this thesis have several advantages over current incomplete spectrogram methods:

1) Since the reconstruction of the spectrogram is performed independently of the recognizer, the recognizer need not be modified in any manner.
2) They are more computationally efficient than classifier-compensation methods.
3) Since the reconstructed spectrograms can be transformed to cepstra, or other related features, and recognition performed with them, much better recognition accuracies can be obtained than with classifier compensation methods.

We approach the problem of estimating missing regions of spectrograms from two perspectives, one in which the missing regions of the spectrogram are treated as being completely unknown, and the second in which the noisy regions are assumed to be unknown, but bounded. We present methods which use simple statistical representations, other than that used by the speech recognition system, in order to reconstruct the missing regions. This gives us the freedom of using representations that are far simpler than that in the speech recognizer, while also permitting us to utilize information that is not represented by the recognizer to perform the reconstruction. We investigate several simple estimation techniques that reconstruct the

missing regions of spectrograms based both on purely geometrical constraints, as well as statistically motivated techniques that utilize the statistical correlations between various elements of a spectrogram. Experiments show that the recognition performance obtained with cepstral features derived from such reconstructed spectrograms is higher than that obtained with current missing feature methods that attempt to perform recognition directly with the incomplete spectrograms.

This thesis is organized in two parts. In the first part, consisting of Chapters 2, 3, 4, 5 and 6, we treat the noisy regions of speech spectrograms as completely unknown, or missing. In the second part consisting of Chapters 7 through 9 we treat them as missing but bounded.

In Chapter 2 we present a brief description of the speech recognition system, and also present a brief overview of missing data methods in statistical analysis. We especially describe the statistical methods that are applicable to techniques described later in the thesis. In Chapter 3 we describe the speech spectrogram, and how the effect of noise can be modelled as missing features on the spectrogram. In Chapter 4 we describe conventional missing-feature based recognition methods. In Chapter 5 we describe several inference methods that estimate the missing regions of incomplete spectrograms. In Chapter 6 we describe recognition experiments obtained with methods described in Chapter 5.

In Chapter 7 we present inference methods that assume that the unreliable regions of incomplete spectrograms are bounded, and describe experimental results with these methods.

One serious problem with missing-feature based methods is that in order for them to be applied effectively, the reliable and unreliable regions of the spectrogram have to be correctly identified. Although, missing feature methods only require very coarse information regarding the corrupted spectrogram, *i.e.* simple binary information about whether a particular element of the spectrogram is reliable or not, deriving such information, especially when the speech has been corrupted by non-stationary noise, is very difficult. In Chapter 8 we discuss this problem.

In Chapter 9 we summarize our findings and present our conclusions and ideas for future work.

Chapter 2
Background Information

2.1 Introduction

This thesis deals with the reconstruction of incomplete spectrograms to improve the performance of speech recognition systems on noisy speech. Reconstruction of incomplete spectrograms is an exercise in the inference of missing data in incomplete datasets, which is a well studied problem in statistics. The problem addressed in this thesis therefore involves three notions:

1) The manner in which automatic speech recognition systems function
2) The effect of noise on speech recognition systems
3) Inference of missing data in incomplete data sets

In this chapter we aim to clarify these fundamental notions to establish a basis for the work described in the following chapters. We first briefly outline the manner in which speech recognition systems function. We confine our discussion to recognition systems based on Hidden Markov Models (HMM) since the work described in this thesis has been evaluated using CMU Sphinx-III, an HMM based system. Nevertheless, the techniques described in the thesis are not specific to HMM based systems, and can be used with other statistical speech recognition systems as well. We then briefly describe the effect of noise on the performance of speech recognition systems. We also describe several current methods of compensating for the effect of the noise, and their drawbacks. Finally we present a brief review of existing literature on incomplete-data methods in other fields. Some methods which are explicitly used in the thesis are explained in greater detail.

2.2 Overview of Automatic Speech Recognition (ASR) systems

ASR systems are essentially pattern classification systems [Rabiner 1993]. Any utterance of speech is modeled as a sequence of sounds. These sounds may either be the phonemes in a language, words in that language, or larger units, depending on the vocabulary of the system and the task being performed by it. The complete set of sounds that the ASR system has to recognize forms the classes modeled by it. In this discussion we assume without loss of generality that the sound classes modeled by the system are words. The ASR system then classifies segments of speech as belonging to one of these classes.

Classification is not performed using the speech signal directly. Instead, the speech signal is parametrized into a sequence of *feature vectors*, or *parameter vectors*, and classification is performed using these feature vectors. The feature vectors used are usually cepstral coefficients [Davis 1980], or variants of cepstra [Hermansky 1990] derived from power spectra of short windowed segments, or *frames* of speech. Thus, a sequence of speech samples is transformed into a sequence of feature vectors each representing a single frame of speech, which is used to perform recognition.

Let $\boldsymbol{S}$ represent the sequence of parameter vectors derived from the utterance being recognized. Automatic speech recognition systems identify the sequence of words in that utterance using the optimal classifier equation

$$\hat{W} = \arg\max_W \{P(\boldsymbol{S}|W)P(W)\} \tag{2.1}$$

where $\hat{W}$ is the recognized sequence of words in that utterance. $P(W)$ is the *a priori* probability that the word sequence W was uttered and is usually specified by a *language model*. Further details of language models can be found in [Katz 1987]. $P(\boldsymbol{S}|W)$ is the likelihood of $\boldsymbol{S}$ given that the W was the sequence of words uttered. It is termed as the acoustic likelihood of the data and is obtained from the probability distribution of all parameter vectors that could represent the sequence of words W. In HMM-based speech recognition systems this probability distribution of sequences is modeled by an HMM. The following section describes the hidden Markov model in greater detail.

2.2.1 HMM-based modeling of the distributions of sequence of vectors

In HMM-based recognition systems the mechanism that generates the sequence of parameter vectors representing any word is modeled by an HMM [Rabiner 1993]. When generating the sequence, the generator is assumed to be in one of a finite set of states at any instant of time. A probability distribution function is associated with each of these states, which are referred to as the state probabilities. Thus, to generate the feature vector at any instant, the generator draws a vector from the distribution associated with the state it is in at that instant. The vectors that the generator draws from a state distribution are said to belong to that state. The HMM also has a set of *transition probabilities* associated with each state. The transition probabilities of a state refer to the probability distribution of the states that the generator can be in at the next

instant, given that it is in that state at the current instant. The generator draws from this distribution in order to determine which state it will be in at the next instant of time. If the generator is in state j at time instant $t+1$ after having been in state i at time t, it is said to *transit* from state i to state j at time instant t. The transition probabilities and the state distributions are all specific to the word being modeled by the HMM. Figure 2.1 shows an example of an HMM with 5 states. The HMM in this figure only permits transitions in one direction. All transitions with probability 0 are not shown. This HMM has a non-emitting initial state, and a non-emitting terminating state. Non-emitting states are states with which there are no probability distributions associated. Therefore no observations are generated when the generator is in these states. The non-emitting initial state in figure 2.1 implies that at $t = 0$, *i.e.* just *before* the generator begins generating vectors, it is in the initial state where it does not generate any observations. Similarly, if the generator enters the terminating state it can no longer transit to any of the other states in the HMM, nor can it generate any more observations.

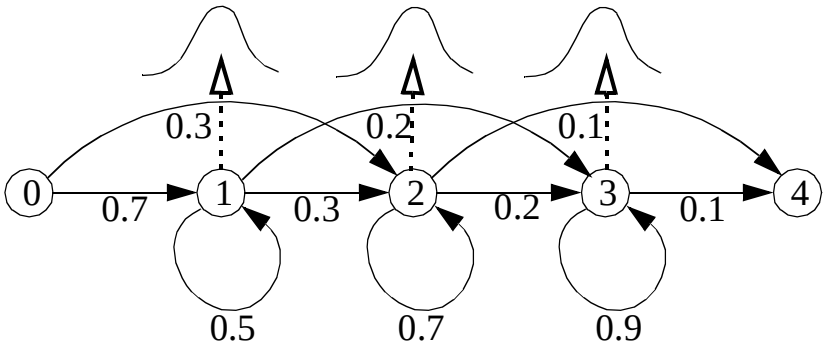


Figure 2.1 Example of a 5 state HMM with one non-emitting initial state, and a non-emitting terminating state. Each of the circles represents a state. The arrows represent valid transitions from the state, and the numbers below the arrows represent the probability of that transition. For example, the arrows from state 1 indicate that if the generator is in state 1 at time *t*, at time *t*+1 it can be in state 1 with probability 0.5, state 2 with probability 0.3 and state 3 with probability 0.2. The dotted arrows point to the state distributions associated with that state. An observation is drawn from this distribution every time the generator visits the state. The initial state (state 0) and the terminating state (state 4) have no state distributions associated with them, and no data are generated when the generator is in these states. Note that in this figure all transitions point left to right. In a more generic HMM, transitions may occur in any direction, from any state to any other state.

Thus, to generate a sequence of N vectors for the word, the generator transits through a sequence of $N+2$ states in the HMM, beginning with the non-emitting initial state and terminating in the final, absorbing state. At each time instant it draws observations from the state distribution of the state it is in at that time instant. The sequence of vectors so generated is said to be *generated by the HMM*.

The model for the generating mechanism for a *sequence* of words is also an HMM and easily constructed by concatenating HMMs for individual words. Figure 2.2 shows an example where the HMMs for three words have been concatenated to obtain an HMM modeling a sequence of three words.

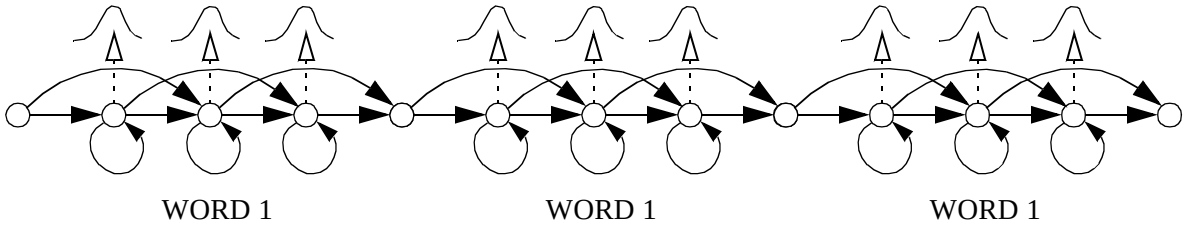


Figure 2.2 Example of constructing the HMM for a sequence of words from the HMMs of individual words. The non-emitting terminating state of any word is merged with the non-emitting initial state of the next word. The merged state is no longer an initial state or a terminating state. However, it remains non-emitting, and no state distribution is associated with it. The resulting HMM has a non-emitting initial state, a non-emitting terminating state and several intermediate non-emitting states as well.

The statistical parameters of the HMM representing a sequence of words W are the set of transition probabilities, represented as a matrix A_W, and the set of state probability distribution functions. The matrix A_W consists of elements $a_w(i,j)$, which represents the probability that the generator will be in state j in the next time instant, given that it is currently in state i. Thus, for an HMM with K states, we have

$$\sum_{j=1}^{K} a_w(i,j) = 1.0 \tag{2.2}$$

The state distribution of the k^{th} state is represented by $P_{W,k}(X)$, where X represents any parameter vector that belongs to the k^{th} state. In speech recognition systems the various state distributions are usually modeled as Gaussians or mixtures of Gaussians [Juang 1986]. Typically, for computational efficiency, these Gaussians are assumed to have diagonal covariance matrices, *i.e.* covariance matrices where the off-diagonal elements are all 0. For simplicity we represent the state distribution of the k^{th} state as

$$P_{W,k}(X) = MG(X;\phi_k^w) \tag{2.3}$$

where $MG(X;\phi_k^w)$ denotes a Gaussian mixture distribution corresponding to the k^{th} state of the HMM representing the word sequence W and ϕ_k^w represents the set of parameters associated with it. We denote the set of ϕ_k^w for all the states in the HMM for W as λ_W. A_W and λ_W represent the complete set of parameters needed to uniquely identify the HMM modeling W.

The probability of any vector sequence $\boldsymbol{S}$ that is generated by the HMM for W is now given by

$$P(\boldsymbol{S}|W) = \sum_{\boldsymbol{s} \in \Xi} P(\boldsymbol{S}, \boldsymbol{s}|W) = \sum_{\boldsymbol{s} \in \Xi, p} P(\boldsymbol{s}|W)P(\boldsymbol{S}|\boldsymbol{s}) \tag{2.4}$$

where $\boldsymbol{s}$ represents any state sequence that the generator can follow when generating $\boldsymbol{S}$, and Ξ represents the set of all possible state sequences. The state sequence $\boldsymbol{s}$ is, quite literally, a sequence of states, one for every feature vector in $\boldsymbol{S}$. That is,

$$\boldsymbol{s} = [s_1, s_2, s_3, \ldots, s_N] \tag{2.5}$$

where N is the total number of vectors in the sequence $\boldsymbol{S}$, and s_t is the state associated with the t^{th} vector in $\boldsymbol{S}$, $\boldsymbol{S}(t)$. The probability terms in the right hand side of Equation (2.4) can now be written as

$$\begin{aligned} P(\boldsymbol{S}|\boldsymbol{s}) &= \prod_t MG(\boldsymbol{S}(t);\phi_{s_t}^w) \\ P(\boldsymbol{s}|W) &= a(0, s_1)\prod_t a(s_t, s_{t+1}) \end{aligned} \tag{2.6}$$

where $a(0, s_1)$ represents the probability of transiting from the 0^{th} state (i.e the initial non-emitting state) of the HMM for W to the first state in the state sequence $\boldsymbol{s}$, and $a(s_t, s_{t+1})$ represents the probability of transiting from state s_t to state s_{t+1}. Equation (2.4) can now be rewritten as

$$P(\boldsymbol{S}|W) = \sum_{\boldsymbol{s} \in \Xi, p} \left(a(0, s_1)\prod_t a(s_t, s_{t+1}) \right) \left(\prod_t MG(\boldsymbol{S}(t);\phi_{s_t}^w) \right) \tag{2.7}$$

Ideally, recognition would be performed as

$$\hat{W} = \arg\max_{W}\left\{P(W)\sum_{\boldsymbol{s}\in\Xi}P(\boldsymbol{s}|W)P(\boldsymbol{S}|\boldsymbol{s})\right\} \tag{2.8}$$

However, for easy implementation, HMM based speech recognition systems usually estimate not just the best word sequence, but also the best state sequence associated with the word sequence. *i.e.* recognition is performed as

$$\hat{W} = \arg\max_{W,\boldsymbol{s}}\{P(W)P(\boldsymbol{S},\boldsymbol{s}|W)\} = \arg\max_{W,\boldsymbol{s}}\{P(W)P(\boldsymbol{s}|W)P(\boldsymbol{S}|\boldsymbol{s})\} \tag{2.9}$$

which can be further expanded into

$$\hat{W} = \arg\max_{W,\boldsymbol{s}}\left\{P(W)\left(a(0,s_1)\prod_t a(s_t,s_{t+1})\right)\left(\prod_t MG(\boldsymbol{S}(t);\phi^{w}_{s_t})\right)\right\} \tag{2.10}$$

In order to evaluate Equation (2.10) fully, the term within the braces would have to be computed for every possible word sequence in the language. This would be impractical. In practice, dynamic programming methods are used [Viterbi 1967] to obtain locally optimal estimates for $\hat{W}$.

The CMU Sphinx-III HMM based recognition system has been used exclusively to evaluate missing feature methods in this thesis. This is a phone-based recognition system. Words are further decomposed into sequences of phones and the HMMs for words are built by concatenating the phone HMMs. Further, in order to reduce the total number of parameters needed to construct HMMs for all the phonetic units modeled by the system, the state distributions of states of the HMMs of the various phonetic units are shared, *i.e.* the same distribution is used by the states of the HMMs of several phonetic units.

2.3 The effect of noise on speech recognition systems

Speech recognition systems function on the assumption that the distributions modeling the various sound classes in the recognizer are representative of the speech being recognized. In other words, it is assumed that the distributions of the feature vectors representing the various sound classes in the test data are very similar to the corresponding distributions in the recognizer. When the distributions in the recognizer have been trained from clean speech this is only true if the speech being recognized is clean as well. When the speech being recognized has been corrupted in any manner the two distributions are no longer

similar [Moreno 1996]. We can represent any noisy utterance as being a clean utterance that has been rendered noisy by some transformation. If we represent the t^{th} feature vector of an utterance of clean speech as $\boldsymbol{S}(t)$, and the corresponding feature vector of the corrupted utterance as $\boldsymbol{Y}(t)$, we could represent the relation between them as

$$\boldsymbol{Y}(t) = T(\boldsymbol{S}(t)) \tag{2.11}$$

where $T(\ )$ is the transformation that converts any clean speech feature vector to a noisy speech feature vector. If we represent the distribution of the feature vectors of clean speech representing a sound s as $P_s(\boldsymbol{S}(t))$, and the distribution of the corresponding vectors of the corrupted speech as $P_s(\boldsymbol{Y}(t))$, we have

$$P_s(\boldsymbol{Y}(t)) = P_s(T(\boldsymbol{S}(t))) \tag{2.12}$$

The recognizer models the sound s by the $P_s(\boldsymbol{S}(t))$, the distribution of clean speech vectors for that sound. However, the distribution of vectors in the test data for the sound s is $P_s(\boldsymbol{Y}(t))$. We see from Equation (2.12) that $P_s(\boldsymbol{Y}(t)) \neq P_s(\boldsymbol{S}(t))$ unless $T(\ )$ is an identity transformation. This mismatch between $P_s(\boldsymbol{S}(t))$ and $P_s(\boldsymbol{Y}(t))$ causes the performance of the recognition system to degrade greatly.

This mismatch can be eliminated if the distributions in the recognizer are learned using speech that has been subject to exactly the same kind of degradation as the test speech. However, even in this scenario, the effect of corrupting noise is to increase the inherent variability between different instances of any sound and the resulting recognition accuracy is significantly lower than when the data used to train the recognizer and the test data are both clean. Further, this requires precise control over the recording conditions of the test speech in order to keep them identical to that of the speech used to train the recognizer. In most practical situations mismatches between the distributions used by the recognizer and the distributions of the test data persist. Figure (2.3) shows the recognition performance of a speech recognizer on speech that has been corrupted by noise. As can be seen, the effect of noise is to degrade recognition accuracy greatly even when the distributions in the recognizer are perfectly matched to the distributions of the noisy speech data.

There are two possible approaches to reducing the mismatch between the distribution of test data and the recognizer distributions. In the first approach the test data is transformed in some manner such that the distributions of the transformed test data match the distributions in the recognizer. *i.e.*, $\boldsymbol{Y}(t)$ is transformed

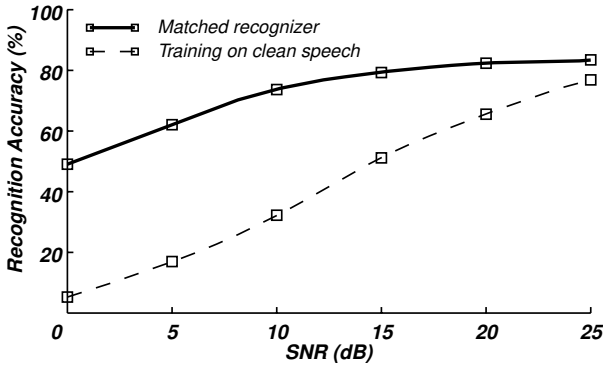


Figure 2.3 Recognition accuracy as a function of the signal-to-noise ratio of the speech being recognized. The lower curve represents a "mismatched" recognizer, where the recognition system has been trained on clean speech, but the test speech is noisy. The upper curve represents a "matched" recognizer, where the recognition system has been trained with speech that has been subject to the same level of noise as the test speech.

by a transformation $T_d(\ )$ such that

$$P(T_d(\boldsymbol{Y}(t))) \cong P(\boldsymbol{S}(t)) \tag{2.13}$$

Recognition is now performed with $T_d(\boldsymbol{Y}(t))$ instead of $\boldsymbol{Y}(t)$. This approach is referred to as the *data compensation* approach, since the noisy test data are being transformed to compensate for the corrupting noise.

The second approach to reducing the mismatch between the distributions in the recognizer and the distribution of the test data is to transform the recognizer distributions in some manner, such that they are now similar to the test data distribution. *i.e.*, the distributions $P_s(\boldsymbol{S}(t))$ are transformed by a transformation $T_m(\ )$ such that

$$T_m((P_s(\boldsymbol{S}(t))) \cong P_s(\boldsymbol{Y}(t))) \tag{2.14}$$

Recognition is now performed using $T_m(P_s(\boldsymbol{S}(t)))$ instead of $P_s(\boldsymbol{S}(t))$. Since components of the classifier are being modified to compensate for the noise, this approach is referred to as the *classifier-compensation* approach.

In HMM-based systems the recognizer distributions are transformed by transforming the parameters of the mixture Gaussian state distributions of the HMMs modeling the various speech sounds.

$$P_{W,k}(\boldsymbol{Y}(t)) \cong MG(\boldsymbol{S}(t); T_m(\phi_k^W)) \tag{2.15}$$

where $P_{W,k}(\boldsymbol{Y}(t))$ is the distribution that would have represented the k^{th} state of the HMM for the word sequence W, had the recognizer been trained with $\boldsymbol{Y}(t)$. Recognition is performed using $T_m(\phi_k^w)$ as the parameters of the state distributions.

Several data-compensation and classifier-compensation methods have been proposed in the literature. Among data compensation methods, methods such as CDCN [Acero 1993] and VTS [Moreno 1996] model the effect of noise on the feature vectors of clean speech using a parametric model and learn the parameters of this model based on samples of the noisy utterance being recognized and the *a priori* distributions of clean speech. They then attempt to transform the feature vectors of the noisy speech back to their clean counterparts using the learned parameters. Other methods such as RATZ [Moreno 1996] and POF [Neumeyer 1994] use "stereo data" - data that have been simultaneously recorded in clean and noisy environments - to learn the relations between the feature vectors of clean speech and those of noisy speech. This relationship is later used to estimate the clean speech feature vectors corresponding to the vectors of any noisy utterance. Still other methods such as spectral subtraction [Boll 1979] and Wiener filtering [Porter 1984] estimate the spectrum of the corrupting noise and use it to reduce the noise level in the noisy speech *signal*, rather than on its feature vectors.

Among classifier compensation methods, methods such as PMC [Gales 1993] and model composition [Varga 1990] use analytical models of the effect of noise on the feature vectors of clean speech and use these models to transform the parameters of the Gaussian mixture state distributions of the HMMs. Methods such as MLLR [Leggetter 1994], on the other hand, simply transform the parameters of the mixture Gaussian state distributions using an affine transform, to best fit the noisy speech. The parameters of the affine transform are learned from "adaptation data" - data that have been recorded under the same conditions as the noisy speech being recognized.

All of these methods assume, either explicitly or implicitly, that the noise that is corrupting the speech signal does not vary much over the course of the utterance. The noise is assumed to affect the feature vectors (or the distributions of the feature vectors) of any instance of a particular sound in exactly the same manner as it affects every other instance of the same sound. As a result, while these methods are fairly successful at compensating for stationary noises they are, in general, ineffective in the presence of non-stationary noises [Raj 1997].

The effect of non-stationary corrupting noises on speech recognition accuracy is different from that of stationary noises. Figure 2.4a compares recognition accuracy obtained on speech corrupted by stationary white noise, with that obtained on speech corrupted by music, which is a non-stationary signal. Figure 2.4b shows the improvement in recognition accuracy obtained when CDCN compensation is applied to both cases. We observe that at any given SNR the recognition accuracy obtained with speech that has been corrupted by music is greater than that obtained with speech corrupted by stationary noises. This is because at any given SNR the energy in music is much more localized in time due to its non-stationary nature than the energy in white noise. As a result, while some regions of the speech get corrupted to a greater degree by music than they do by white noise, other regions do not get corrupted much. The higher recognition performance of the recognizer in these less corrupt regions results in greater overall accuracy.

On the other hand CDCN compensation does not improve the recognition performance of speech corrupted with music, while it is quite effective on white noise. Similar results are obtained for speech corrupted by other non-stationary noises, and for other compensation methods. In general, while the effect of non-stationary noises is not as damaging to recognition accuracy, it is not possible to compensate for the effect of the noise effectively with current compensation techniques. Clearly, new approaches are required to handle non-stationary noises.

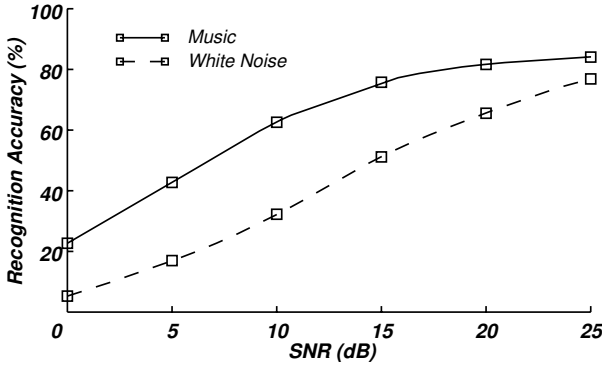


Figure 2.4a Recognition accuracy obtained with speech corrupted by white noise, and speech corrupted by a segment of music, at various SNRs.

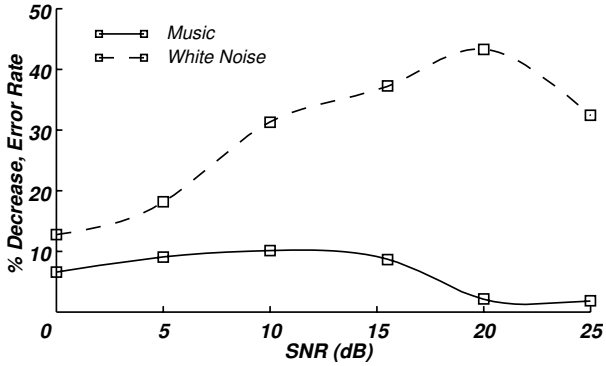


Figure 2.4b Relative improvement in recognition error rate obtained by applying CDCN compensation to speech corrupted by corrupted by white noise and music

It is well known that human beings are sell able to comprehend speech that has been heavily corrupted by both stationary and non-stationary noise [Lippmann 1997][Miller 1950]. It is also known that human listeners are able to comprehend speech which has undergone considerable spectral excisions. For example, normal conversation is possible with speech that has been either high-pass or low-pass filtered with a

cutoff frequency of 1800Hz [Fletcher 1953]. Similarly, speech that is occluded by interfering signals and corrupting noises is easily comprehended by humans. The human auditory system also exhibits the so-called *capture effect* [Moore 1997] by which locally more intense signal components dominate the neural response, suppressing the weaker components, sometimes completely. Phenomena such as excisions, occlusion, and the capture effect can be represented as occlusions of spectro-temporal regions in time-frequency representations of the speech signal. They therefore suggest that there is sufficient redundancy in the speech signal for it to be recognized based only on a fraction of spectro-temporal information present in it.

This observation has motivated two new approaches to robust recognition of noisy speech: the *multi-band recognition approach,* and the *missing-feature approach*. In these approaches the recognition system is modified to concentrate only on those portions of the speech signal that have been less corrupted by noise, rather than the entire signal. Let us represent the k^{th} component of the t^{th} feature vector of an utterance of clean speech as $S(t, k)$, and the corresponding component for the corrupted speech as $Y(t, k)$. Let $T_{t,k}(\ )$ be the transformation corrupting $S(t, k)$ such that

$$Y(t, k) = T_{t,k}(S(t, k)) \tag{2.16}$$

The new approaches attempt to improve recognition by concentrating only of those components of the noisy speech for which $Y(t, k) \cong T_{t,k}(S(t, k))$, *i.e.* the components for which the difference between the noisy speech and the clean speech, $E(t, k)$, is small, where $E(t, k)$ is defined as

$$E(t, k) = \left| Y(t, k) - T_{t,k}(S(t, k)) \right| \tag{2.17}$$

For these components the mismatch between the distributions in the recognizer and the distributions of the test data is also small. Components for which $E(t, k)$ is large are either deweighted or discarded completely.

Multi-band recognition approaches decompose speech into separate frequency bands and perform recognition independently on the various frequency bands. The recognition hypotheses of the individual bands are combined to obtain the final recognition hypothesis. During recognition the contributions of frequency bands where the error between the parameters of clean speech and those of noisy speech are expected to be large are given less weight with respect to the bands where the error is expected to be

smaller [Hermansky 1996] [Bourlard 1996].

Missing-feature approaches, on the other hand, do not decompose speech into frequency bands. Instead, they assume that those components of a spectrographic representation of speech for which $E(t, k)$ is large are in fact unknown or missing (hence the name "missing features"), and they perform recognition based only the remaining components [Cooke 1994][Lippmann 1997]. In other words, they model the effect of noise on speech as that of obscuring some of the components of the spectrographic representation, resulting in incomplete spectrographic data for that utterance. The problem of recognizing noisy speech then becomes one of *classification with incomplete data*. We describe how the effect of noise on speech can be modeled as missing spectrographic data in Chapter 3. Missing-feature approaches and the problem of classification with incomplete data are discussed in greater detail in Chapter 4.

Both multi-band and missing-feature approaches have the advantage that they do not make any explicit assumption about the characteristics of the noise corrupting the speech. The procedure for compensating for the noise is independent of whether the noise is stationary or non-stationary. They do however need to know beforehand which components of speech have been badly corrupted by the noise, and which have been less affected. The problem of estimating this information in the context of missing-feature approaches is discussed in greater detail in Chapter 8.

Missing-feature approaches have the advantage over multi-band approaches that they do not assume that different frequency bands are independent of each other. However, they have the disadvantage that they are restricted to performing recognition using spectral features. This restriction is discussed in greater detail in Chapter 4.

In this thesis we attempt to eliminate this restriction by *reconstructing* those components of a spectrographic representation of speech for which $E(t, k)$ is large *prior* to recognition. We are, in effect, re-formulating the missing-feature approach as one of *inference of missing data,* rather than that of classification with incomplete data. The problem of inference of missing data in incomplete data sets has been well studied in the field of statistics and other related fields. In the following section we briefly review the literature on the topic from these fields.

2.4 Incomplete Data Methods

Incomplete-data methods are methods that are applied for the analysis or study of data sets where some of the components are missing. Usually they deal with the estimation of the missing components of the data set. Alternately, they may deal with the estimation of the statistical properties of the data, based only on the incomplete data set.

Data sets could be incomplete for several reasons. For example, data could be missing due to the characteristics of the process that generated the samples. Incomplete data are frequently encountered in sample surveys [Madow 1983] where some respondents chose not to respond to certain queries in the survey, or prefer not to respond to the questionnaire at all. *Data occlusion* is another reason for incomplete data [Ahmed 1993]. This could happen, for example, where some of the regions of interest in a picture are occluded by irrelevant objects, or when some portions of a sound recording are occluded by noise. Incomplete data may also result from *loss* of data. Segments of sound recordings may be lost due to damage to the recording media. Portions of data may be lost during transmission over a communication channel. Data points that are obviously non representative can also give rise to missing data [Rubin 1987]. For example, in a survey where one of the queries is the age of a person, a response such as 937 is obviously erroneous and needs to be treated as unknown. Similarly, speech samples that are corrupted by very high levels of noise can be treated as unknown.

We note from the above examples that there are several mechanisms that render data unobservable to the observer. These mechanisms themselves, in turn, can have different characteristics. In the case of the incomplete or erroneously completed sample surveys, the non-response to a particular query may be related to the query itself (*e.g.* people who are unwilling to divulge their incomes), or to the actual response to the query (*e.g.* people belonging to a particular demographic group being unwilling to identify themselves as such). The non-response to a particular query may even be related to the response to other queries in the survey (*e.g.* people belonging to a particular demographic group being unwilling to divulge their incomes). Similarly, in the case of the picture with occluded regions also the missing data mechanism can vary. The mechanism can be completely random, as in the case of cars moving down a street occluding the objects on the other side of the street. The mechanism could be related to the content of the picture, *e.g.* bees occluding regions of flowers. In speech corrupted by noise, the mechanism causing incomplete data

depends on the content of the sound. The high energy and clearly enunciated regions of the signal are more likely to remain comprehensible than low energy regions.

In all the above cases we are able to distinguish between a truth and a missing data mechanism. By truth here we refer to the *true value* of the missing components of the data. In the case of the sample survey, for example, this would be the response that the respondent would have given to a query had he responded to it. For the picture this would be what the camera would have captured had the occluding object not been in place. When data is lost in transmission or storage this would be the value the data point had before it was lost. For all of these cases there exists a hypothetical data set corresponding to the incomplete data set, where all the components are present. The terminology we adopt [Little 1987] refers to this hypothetical data set, where no components are missing, as the *complete data*. Data sets that are missing some of their components are referred to as *incomplete data*. The missing data elements or components are referred to as *missing data*, or *missing features*. The mechanism that renders some of the complete data unobservable, thereby resulting in incomplete data, is referred to as the *missing data mechanism*.

Missing data mechanisms are usually categorized into three types [Ghahramani 1993]. The three categories are:

1) Missing Completely At Random (MCAR): The missing data mechanism in this case is completely random. As a result, the probability that any component of the complete data will be deleted by the mechanism is independent of both the component itself and the rest of the data set.
2) Missing At Random (MAR): Here the probability that any component of the complete data will be deleted depends on the value of the observed data.
3) Not Missing At Random (NMAR): Here the probability that any component of the complete data will be deleted depends both on the value of the observed data, as well as the value of the deleted data point itself.

Of these, MCAR missing data patterns are the most difficult to predict (based on the complete data), but the least problematic, because of the unsystematic nature of the deletions. MAR and NMAR missing data mechanisms on the other hand cause systematic deletions of data, and while the missing data patterns are more predictable, they can be very damaging to any analysis based on the data.

It is difficult, if not impossible, to perform any meaningful statistical analysis of the processes underlying any data if the data are incomplete. Similarly, classification of, or prediction on the bases of incomplete

data is not simple. Standard statistical procedures cannot be directly applied to such data sets since such procedures assume the existence of complete data.

Most incomplete data methods in the literature deal with the estimation of the missing data. The process of estimating the missing data components is referred to as *imputation*. The earliest used method of imputation was the so-called *mean imputation* [Ahmed 1993]. In mean imputation missing components of a vector are filled in by the average value of that component. This problem has the obvious disadvantages that it under represents the variability in the data, and also ignores the correlations between the various components of the data completely. The US Census Bureau attempts to handle missing data points in its sample surveys by a procedure known as *Hot Deck Imputation* [David 1983]. The hot deck procedure finds, for each incomplete data vector, a matching complete data vector, *i.e.* the data vector that is closest in terms of the components that are present in both vectors. The missing components of the incomplete data vector are then filled in with the corresponding components of the matching complete vector. Hot deck imputation, once again, has the shortcoming that the estimate of the missing data components are based on a single complete vector in the data set, ignoring any global properties of the data set. It also ignores the possibility that the matching vector itself may have been an outlier in the components of interest.

Several imputation methods have been proposed in the literature that use decision trees to impute the values of missing data points [Quinlan 1989]. Of these, methods based on Classification And Regression Trees have been most popular [Breiman 1984]. In these methods, the set of all complete data vectors is partitioned recursively into a tree based on a set of logical “questions”. Individual complete vectors form the leaves of this tree. Incomplete data vectors are passed down the tree based on their answers to the questions at each node in the tree, until they reach a leaf. The missing components of the vector are obtained from the vector at the leaf. While this procedure is simple and useful for multinomial data sets, their use becomes very complicated for data that can take values from a continuous range.

A more statistically motivated procedure that is frequently used is *Regression Imputation* [Mendenhall 1996]. The missing components of incomplete data vectors are imputed as a linear regression of the components of the vector that do exist. The regression coefficients are estimated from any existing set of complete data vectors. The drawback of this procedure is that all imputed values fall along a single regression line thereby under representing any variation inherent in the data. Also, there are implicit symmetry

assumptions about the distribution of the data that may not be valid.

Expectation Maximization, or EM, is a statistical technique that is highly suited to the estimation of the distributions of incomplete data sets [Dempster 1977]. The procedure iteratively finds the “expected value” of each of the missing elements in the data set, and uses these expected values to find the distribution of the data. EM can be further reinforced by the use of *a priori* statistics of the value of the missing components. This procedure is usually referred to as Bayesian EM [Ghahramani 1994].

Most missing data methods described in the literature are most suited to handle multinomial data sets and data such as incomplete sample surveys. Also, most of them assume that the incomplete data has occurred due to the deletion of elements from the complete data, and that no other data corrupting mechanisms are involved. Additionally, in most of the missing data methods described above, the missing components of an incomplete dataset are estimated based only the properties of observed portion of the incomplete data.

In the following section we describe three methods that assume *a priori* knowledge of the distribution of complete data, and use these in conjunction with the observed data to estimate the missing components of the data.

2.5 Statistical methods for estimating missing data

Statistical methods assume *a priori* knowledge of the distribution of the complete data, and use this knowledge to estimate the missing data. It is useful to introduce some mathematical notation here in order to simplify the explanations presented in the rest of the chapter. We represent the hypothetical complete data by the symbol $\mathbf{X}$. $\mathbf{X}$ in turn has two components, the observed data $\mathbf{X}_o$ and the missing data $\mathbf{X}_m$. The complete data is the combination of the two, a relation that we denote by $\mathbf{X} = \mathbf{X}_o, \mathbf{X}_m$.

Statistical estimation methods assume that either the probability distribution of the complete data, or that some of the statistical properties of the data that are derivable from this distribution, are known. Let $P(\boldsymbol{X};\phi)$ represent a parametric model for the distribution of the complete data, where ϕ represents the parameters of the parametric model. If $P(\boldsymbol{X};\phi)$ were Gaussian, ϕ would refer to the mean and the variance of the distribution. From this distribution the conditional probability distribution $P(\mathbf{X}_m \mid \mathbf{X}_o;\phi)$ and

the conditional distribution $P(\mathbf{X}_o|\mathbf{X}_m;\phi)$, can be derived. Statistical estimation methods can be employed where any of these distributions are known.

2.5.2 Minimum Mean Squared Error (MMSE) estimation

In MMSE estimation the missing data are estimated to minimize the expected mean squared error between the estimates and the true value of the elements, conditioned on the observed data [Therrien 1992].

$$\hat{\boldsymbol{X}}_m = \arg\min_{\boldsymbol{X}_m}\left\{E[\left\|\boldsymbol{X}_m - \boldsymbol{X}_m^t\right\|^2 \middle| \boldsymbol{X}_o]\right\} \tag{2.18}$$

where $\hat{\boldsymbol{X}}_m$ is the estimate for the missing data, and $\boldsymbol{X}_m^t$ is the true value of $\boldsymbol{X}_m$. Only the first and second moments of the conditional distribution of the missing data, $P(\mathbf{X}_m|\mathbf{X}_o;\phi)$ are needed for MMSE estimation.

2.5.3 Maximum Likelihood (ML) estimation

In ML estimation the missing data are estimated so as to maximize the conditional likelihood of the values of the observed data $\boldsymbol{X}_o$ [Therrien 1992].

$$\hat{\boldsymbol{X}}_m = \arg\max_{\boldsymbol{X}_m}\{P(\boldsymbol{X}_o|\boldsymbol{X}_m, \phi)\} \tag{2.19}$$

This method bases the estimates of the missing values entirely on the observed data $\boldsymbol{X}_o$, with no reference to the inherent statistical distribution of the missing data. It has the advantage however, that the *a priori* distribution of the missing data, $P(\boldsymbol{X}_m;\phi)$ need not be known.

2.5.4 Maximum A-Posteriori (MAP) estimation

MAP estimation has been extensively used in the methods described in this thesis. We therefore describe the MAP estimation procedure in somewhat greater detail than the previous methods. In MAP estimation the missing data are estimated to maximize their likelihood, conditioned on the values of the

observed data [Therrien 1992].

$$\hat{\boldsymbol{X}}_m = \arg\max_{\boldsymbol{X}_m}\{P(\boldsymbol{X}_m|\boldsymbol{X}_o;\phi)\} \tag{2.20}$$

We note that when $P(\boldsymbol{X}_m|\boldsymbol{X}_o;\phi)$ is Gaussian, MMSE and MAP estimates are identical since a Gaussian distribution is completely described by its first and second moments.

MAP estimation can be simplified to a linear regression when the distribution of the complete data is Gaussian. For simplicity, we assume that the elements of the complete data $\boldsymbol{X}$ have been arranged into a vector, such that the observed components of the complete data form the initial portion of this vector, and the missing components form the training portion. The observed and missing data, $\boldsymbol{X}_o$ and $\boldsymbol{X}_m$, are also arranged into vectors, such that

$$\boldsymbol{X} = [\boldsymbol{X}_o, \boldsymbol{X}_m] \tag{2.21}$$

This assumption does not lead to any loss of generality since any set of M elements can be represented as a vector in an M dimensional space, and the order in which the components are listed in this vector merely denotes the order in which the various dimensions are arranged.

Let $P(\boldsymbol{X};\mu, \Theta)$ be a Gaussian distribution with mean vector μ and covariance matrix Θ. The distributions of $\boldsymbol{X}_o$ and $\boldsymbol{X}_m$, $P(\boldsymbol{X}_o;\mu, \Theta)$ and $P(\boldsymbol{X}_m;\mu, \Theta)$ would therefore also be Gaussian [Papoulis 1991]. If the mean vectors of $P(\boldsymbol{X}_o;\mu, \Theta)$ and $P(\boldsymbol{X}_m;\mu, \Theta)$ are given by μ_o and μ_m respectively, and their covariance matrices by Θ_{oo} and Θ_{mm} respectively, we have

$$\mu = [\mu_o, \mu_m] \tag{2.22}$$

and

$$\Theta = \begin{bmatrix} \Theta_{oo} & \Theta_{om} \\ \Theta_{mo} & \Theta_{mm} \end{bmatrix} \tag{2.23}$$

where Θ_{om} is the cross covariance between $\boldsymbol{X}_o$ and $\boldsymbol{X}_m$, and $\Theta_{mo} = \Theta^T_{om}$.

It can now be shown that $P(\boldsymbol{X}_m|\boldsymbol{X}_o, \mu, \Theta)$ is given by

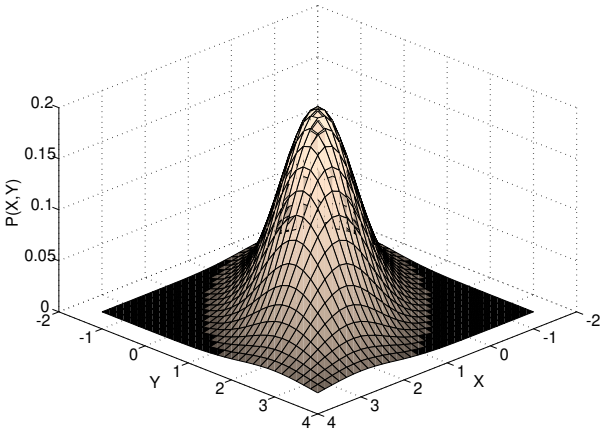


Figure 2.5a Gaussian distribution of a 2 dimensional random vector. The mean of the Gaussian is at [1,1]. The *X* and *Y* components have covariance 1.0, and the covariance between *X* and *Y* is 0.5.

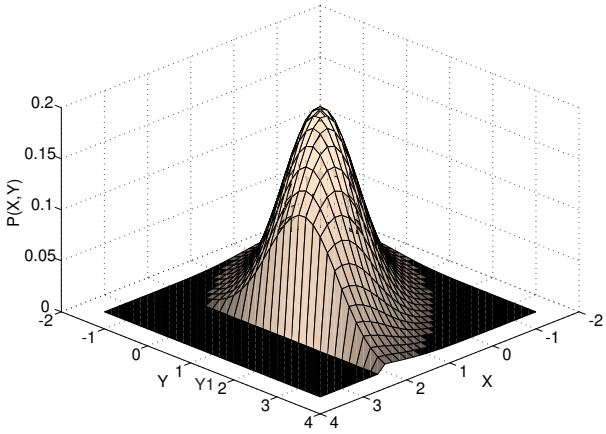


Figure 2.5b The same Gaussian sliced at $X = 2$. The flat surface in the figure represents the distribution of all vectors whose *X* component is 2. This distribution peaks at $Y = Y1$. Thus $Y1$ is the MAP estimate of *Y* when *X* is 2

$$P(\boldsymbol{X}_m|\boldsymbol{X}_o, \mu, \Theta) = C\exp(-0.5(\boldsymbol{X}_m - \mu_m - \Theta_{mo}\Theta^{-1}{}_{oo}(\boldsymbol{X}_o - \mu_o))^T(\Theta_{mm} - \Theta_{mo}\Theta_{oo}^{-1}\Theta_{om})^{-1} \quad (2.24)$$
$$(\boldsymbol{X}_m - \mu_m - \Theta_{mo}\Theta^{-1}{}_{oo}(\boldsymbol{X}_o - \mu_o)))$$

where C is a normalizing constant. Combining Equation (2.20) and Equation (2.24), we get

$$\hat{\boldsymbol{X}}_m = \text{argmax}_{\boldsymbol{X}_m}\{P(\boldsymbol{X}_m|\boldsymbol{X}_o, \mu, \Theta)\} = \boldsymbol{X}_m + \Theta_{mo}\Theta^{-1}{}_{oo}(\boldsymbol{X}_o - \mu_o) \quad (2.25)$$

The MAP solution given in Equation (2.25) is best visualized using the two dimensional example shown in figures 2.5a and 2.5b. In this example $\boldsymbol{X}$ is a two dimensional vector, with components X, and Y, of which X has been observed and Y is missing. In the example the observed value of X is 2. The distribution of $\boldsymbol{X}$, $P(\boldsymbol{X}|\mu, \Theta)$, is a Gaussian, and is shown in figure 2.5a.

Since X is known to be 2, we are only interested in the distribution of vectors with $X = 2$. Figure 2.5b shows the slice of the distribution $P(\boldsymbol{X}|\mu, \Theta)$ at $X = 2$. The vertical face in figure 2.5b represents $\alpha P(Y|X = X_0, \mu, \Theta)$, where α is a scaling constant. This distribution is observed to peak at $Y = Y_1$, which indicates that the most frequently occurring values of Y lie in an small region around Y_1. The MAP estimate for Y is therefore Y_1.

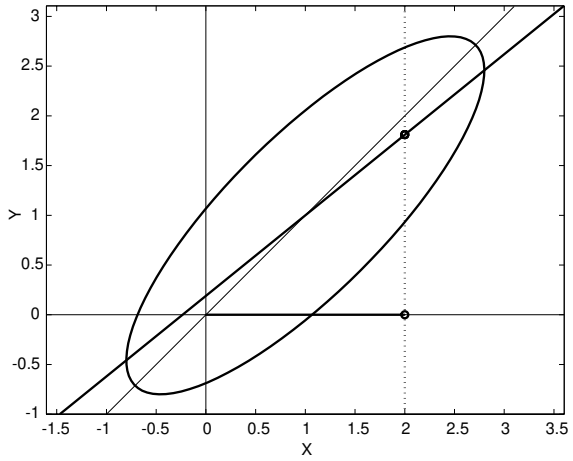


Figure 2.6 Cross section of Gaussian in figure 2.5a. The solid horizontal line shows the observed value of *X*. The circle on the intersection of the solid diagonal line, and the dotted line, shows where the distribution of vectors with *X*=2 peaks. This is the MAP estimate of *Y* when *X*=2. The solid diagonal line shows how the position of this peak varies at each value of X.

Figure 2.6 shows a projection of the Gaussian shown in figure 2.5a the X-Y plane. The solid line in the figure traces value of Y at which a slice of the Gaussian peaks at any value of X. *i.e.* the line traces the MAP estimate for Y as a function of the observed value of X. As can be seen, the relationship between the two is a line, the equation for which is given by Equation (2.25).

2.6 Summary

In this chapter we have presented a brief overview of automatic speech recognition systems and a brief survey of current literature on missing data methods. We have also explained some statistical missing data inference methods in relatively greater detail.

In the next chapter we describe time-frequency representations of speech, and how the effect of noise on speech can be modeled as missing features in these representations.

Chapter 3
Modeling the effect of noise as missing features

3.1 Introduction

Missing-feature methods model the effect of noise on speech as the deletion of regions of time-frequency representations of the speech signal. While there are several time-frequency representations where the effect of noise can be so modeled, the time-frequency representation most commonly used is the spectrogram [Rabiner 1978]. This is the representation that has been used in this thesis.

In this chapter we describe the spectrogram, and its mel-spectral variant, which we specifically use. We also describe how noise affects the spectrogram, and how the effects can be modeled as missing features.

3.2 The Spectrogram

The spectrogram is a commonly used two dimensional representation of the speech signal. It is a pictorial representation of the short-time periodogram [Therrien 1992] of the signal. The short-time periodogram of a signal is given by

$$P_x(l, \omega) = \frac{1}{2L+1}\left|X(l, \omega)\right|^2 \tag{3.1}$$

where

$$X(l, \omega) = \sum_{k=-\infty}^{\infty} x[k]w[l-k]e^{-j\omega k} \tag{3.2}$$

where $w[l]$ is a window of length $2L+1$. Each windowed segment of the signal is referred to as a *frame* of the signal. $X(l, \omega)$ is the value at ω of the Fourier transform of a frame of speech centered around l. $X(l, \omega)$ is also called the short-time Fourier transform [Rabiner 1978] of the signal.

The short-time periodogram of a speech signal therefore consists of a sequence of power spectra, one for each sample in the signal. $P_x(l, \omega)$ represents the power in frequency ω at time instant l in the signal. In practice, the short-time periodogram is not computed for every frequency, or at every time instant. It is

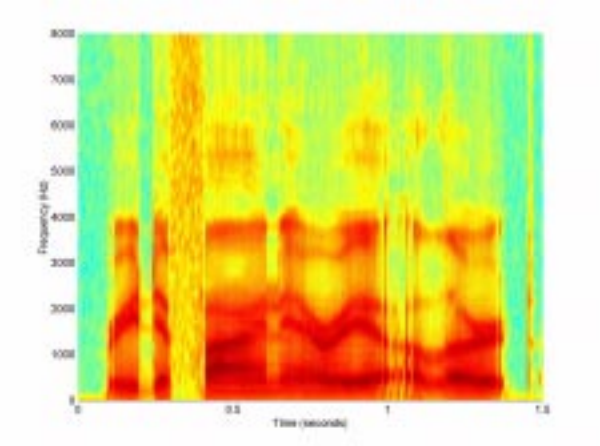


Figure 3.1 This figure shows the wideband spectrogram of the utterance "Redefine Area Alert". The length of the analysis windows was 10ms. Adjacent windows were overlapped by 5ms. The dark bands represent peaks in the spectral envelope. These peaks are called "formants" and their trajectories are characteristic of the sounds in the speech signal.

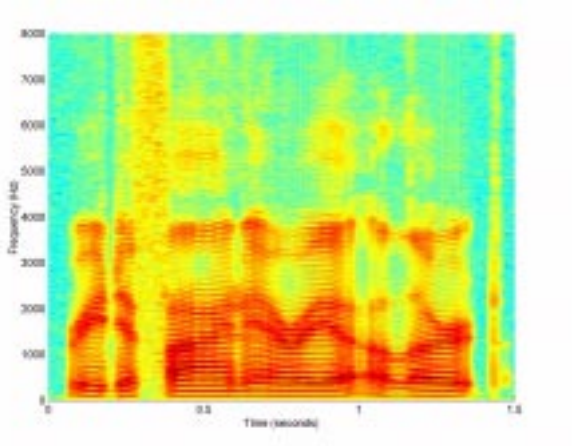


Figure 3.2 This figure shows the narrowband spectrogram of the same utterance. The length of the analysis window was 30ms. Adjacent windows were overlapped by 5ms. The harmonic nature of speech is evident in the figure due to the length of the analysis windows. However the formants are not so clearly visible in this figure.

sufficient to compute $X(l, \omega)$ at only $2L + 1$ points along the frequency axis, and for every $L/2^{\text{th}}$ point in the sequence, for it to be completely invertible. In practice, for speech recognition systems, the time axis is sampled less frequently: typically the short-time Fourier transform is computed for every L^{th} sample in the sequence. The short-time periodogram derived from it therefore consists of a sequence of power spectral vectors, each of which has $2L + 1$ components and represents a short segment of the speech signal.

The spectrogram represents the short-time periodogram as a picture as in Figures 3.1 and 3.2. In these figures the abscissa represents the time (l) axis, the ordinate represents the frequency (ω) axis, and the color, or the intensity of the picture at any location (l, k) in the picture encodes the value of $\log(P_x(l, \omega_k))$, where $P_x(l, \omega_k)$ is the k^{th} component of the l^{th} power spectral vector in the short term periodogram. Although the term *spectrogram* usually refers to these pictorial representations, we also use it to refer to the logarithm of the short-time periodogram. Thus the spectrogram consists of a sequence of log-spectral vectors where $S(l, k)$, the k^{th} component of the l^{th} log-spectral vector in the spectrogram is given by

$$S(l, k) = \log(P_w(l, \omega_k)) \tag{3.3}$$

The difference between Figures 3.1 and 3.2 is in the length of the analysis window, $w[l]$. Longer windows result in greater frequency resolution, but lower resolution of quick changes in the spectrum with time.

The MEL spectrogram

A variant of the short-time Fourier transform of the speech signal that is used commonly by speech recognition systems is the mel-spectral representation [O'Shaughnessy 1987]. The mel spectrum, in principle, tracks the power at the output of a band of filters, called the mel filters. In practice, the mel spectrum of a frame of speech is approximated by integrating over the DFT of the windowed speech signal as follows:

$$P_x(l, k) = \sum_{j=0}^{2L} m_k(j)|X(l, j)|^2 \tag{3.4}$$

where $P_x(l, k)$ is the k^{th} component of the mel spectrum in the l^{th} analysis window and $m_k(j)$ is the j^{th} DFT coefficient of the impulse response of the k^{th} mel filter. $X(l, j)$ is the j^{th} frequency component of the DFT of the l^{th} analysis window of the speech signal $x[n]$. The mel spectrum can be viewed as an spectrally smeared version of the short-time periodogram. Frames are typically 25 ms long, and overlap by 15 ms for the mel-spectral representation.

The *mel spectrogram* is simply obtained from the mel spectrum as

$$S_x(l, k) = \log(P_x(l, k)) \tag{3.5}$$

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 a variant of the spectrogram that uses an spectrally smeared version of the short-time periodogram. In subsequent chapters of this thesis we therefore use the term "spectrogram" to refer to both the spectrogram described in Section 3.2 as well as the mel spectrogram. We use the term "spectral vector" to represent both the log-spectral vectors of the spectrogram, and the log-mel-spectral vectors of the mel spectrogram. We generically refer to the components of a spectral vector as *frequency components* of that vector, irrespective of whether the underlying spectrogram is a true spectrogram or a mel spectrogram. This should not cause any confusion, however, since all spectrogram-

related methods described in this thesis apply equally to both kinds of spectrograms.

All experiments in this thesis were conducted with mel spectrograms. Typically, speech recognition systems use 40 mel filters to parametrize broadband speech. However, for all the experiments in this thesis we have used only 20 filters covering the frequency range. Figure 3.3 shows the mel spectrogram of a speech utterance. The abscissa represents the frame index, the ordinate represents the mel filter index. The color/shade of the picture at any (l, k) encodes the value of the corresponding $S_x(l, k)$. Since the normal spectrogram described in Section 3.2 is obtained from the short-time Fourier transform of the signal, it can, in principle, be inverted to retrieve the speech signal (provided the phase information in the short-time Fourier transform is available). The mel spectrogram, on the other hand, cannot be inverted to retrieve the speech signal, except to a very crude approximation.

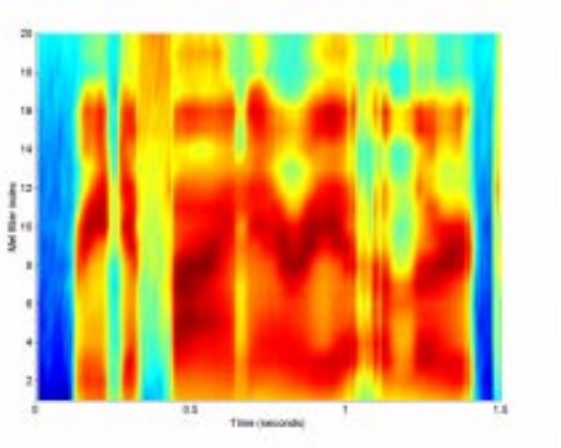


Figure 3.3 Mel spectrogram of the utterance “Redefine Area Alert”. 20 mel filters covering the frequency range 150 Hz to 8 KHz have been used for this representation. The vertical axis represents the index of the mel filter. The horizontal axis represents the index of the mel-spectral vectors in the spectrogram. The analysis windows were 25 ms long. Adjacent windows are overlapped by 15 ms.

3.3 Effect of noise on the spectrogram

When the speech signal is corrupted by additive noise, we have

$$y[l] = x[l] + n[l] \tag{3.6}$$

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. The short-time Fourier transform of the noisy signal is given by

$$Y(l, \omega) = \sum_{k=-\infty}^{\infty} (x[k] + n[k])w[l-k]e^{-j\omega k} \tag{3.7}$$

$$Y(l, \omega) = X(l, \omega) + N(l, \omega) \tag{3.8}$$

where $N(l, \omega)$ is the short-time Fourier transform of the noise. If we assume that the noise is uncorrelated to the speech signal, the short-time periodogram of the noisy signal is given by [Oppenheim 1989]

$$P_y(l, \omega) = P_x(l, \omega) + P_n(l, \omega) \tag{3.9}$$

where $P_n(l, \omega)$ is the short-time periodogram of the noise. The signal-to-noise ratio (SNR) in the spectrogram of the noisy signal at any (l, k) is given by

$$SNR(l, k) = 10\log\left(\frac{P_x(l, \omega_k)}{P_n(l, \omega_k)}\right) \tag{3.10}$$

As can be seen from Equations (3.9) and (3.10) above, the SNR of the elements of the spectrogram is a function of both time and frequency. Typically, for any level of noise, the spectrogram would have regions of very high SNR, as well as regions of very low SNR. As the global SNR of the noisy utterance decreases the proportion of high-SNR regions decreases, while the proportion of low-SNR regions increases. Figure 3.4 shows a quantized version of the wideband spectrogram of an utterance of speech corrupted to 20 dB by white noise. All regions of the spectrogram where the local SNR is less than 0 dB are colored white and all regions where the SNR is greater than 0 dB are colored black. Figure 3.5 shows a similar quantized spectrogram of speech corrupted to a global SNR of 0dB. As is apparent from the two pictures, the fraction of the pictured covered by the black regions is considerably lesser in Figure 3.5 than in Figure 3.4. Conversely, the fraction of the picture colored white, *i.e.* low SNR regions, is considerably higher in Figure 3.5.

The same logic can be applied to the mel spectrogram to show that the mel spectrogram of the noisy speech signal given in Equation (3.6) can be expressed as the sum of the mel spectrum of the clean speech signal and the mel spectrum of the noise

$$P_y(l, k) = P_x(l, k) + P_n(l, k) \tag{3.11}$$

Similarly to the spectrogram, the local signal to noise ratio of the mel spectrogram of the noisy speech sig-

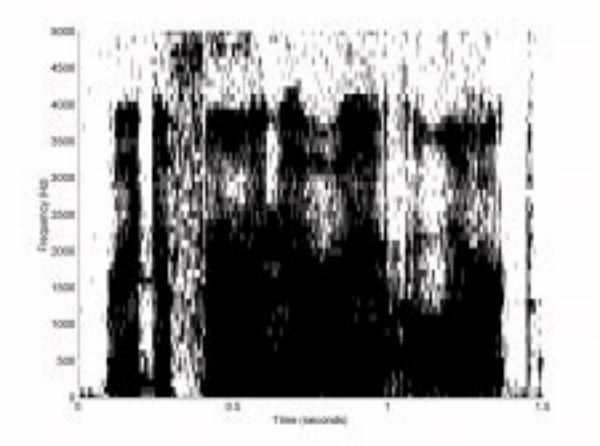


Figure 3.4 Quantized spectrogram of an utterance of speech that has been corrupted to 20 dB by additive white noise. All regions of the spectrogram where the local SNR is greater than 0dB (*i.e.* where the speech energy was greater than the noise energy) are colored black. All regions with local SNR less than 0 dB are colored white. Only frequencies up to 5 KHz have been shown in the figure.

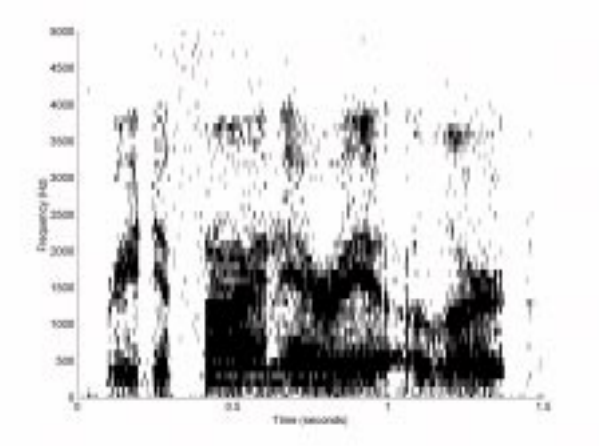


Figure 3.5 Quantized spectrogram of the same utterance, when corrupted to 0 dB by additive white noise. Once again, all regions of the spectrogram with local SNR greater than 0 dB have been colored black, and all regions with local SNR less than 0 dB have been colored white. Once again, only frequencies up to 5 KHz have been shown. The fraction of white regions here is clearly much greater here than in figure 3.4.

nal, given by

$$SNR(l, k) = 10\log\left(\frac{P_x(l, k)}{P_n(l, k)}\right) \tag{3.12}$$

varies with both frame index l and filter index k. The mel spectrogram of noisy speech also exhibits both regions of high SNR and regions of low SNR. Figure 3.6 shows the SNR of the mel spectrogram of an utterance of speech corrupted to 10 dB by white noise. The abscissa represents the frame index and the ordinate the mel filter index. The SNR is coded by gray shade - the darker the color, the greater the SNR. As can be seen from the figure, there are several regions of high SNR, and several other regions of very low SNR. In this figure the lowest SNR regions correspond to segments where there is no speech at all and the signal consists entirely of noise.

3.4 Modeling the effect of noise as missing features in the spectrogram

As mentioned in Chapter 2, there is empirical evidence to the effect that human listeners concentrate on the high energy regions of the speech [Moore 1997], effectively ignoring the low energy regions in dealing with noise. In other words, the evidence is taken from the so-called *reliable* spectro-temporal regions of

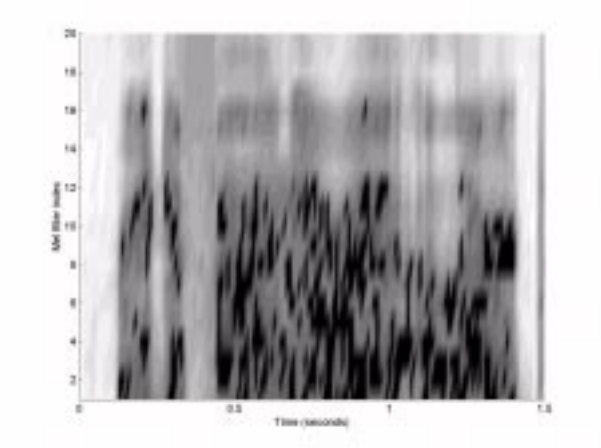


Figure 3.6 Local SNR of the elements of the mel-spectrogram of an utterance corrupted to 10dB by additive wh noise. The SNR is gray coded - the darker the color the higher the SNR of the element.

speech, while ignoring or deweighting the so-called *unreliable* regions. It stands to reason that an identical concept could be applied to automatic speech recognition systems as well. If the evidence used for recognition were derived only from the reliable regions of the spectrogram, eliminating the unreliable regions, the recognition performance of the system would be expected to become much more robust to noise.

One way of measuring the “reliability” of any region in the time-frequency plane is to measure the SNR of the signal component in that region. The higher the SNR the greater the reliability of the signal components in that region, and the lower the SNR the lower the reliability. In order to eliminate low-reliability regions from the spectrogram we would therefore *erase* all low-SNR regions of the spectrograms, retaining the high-SNR regions of the spectrogram alone.

Figure 3.7 shows the spectrogram of a noisy utterance. Figure 3.8 shows the same spectrogram, where all those portions of the spectrogram where the noise energy was greater than the speech energy, *i.e.* where the local SNR was less than 0 dB, have been termed unreliable and erased. The resultant picture has several elements missing. We refer to the pattern of present and deleted regions in the spectrogram as the *spectrographic mask* for the spectrogram. We would now have to perform the task of recognizing what has been said in the utterance, a statistical inference task, based on this incomplete picture.

In the Figures 3.7 and 3.8 all regions of the spectrogram where the local SNR was less than 0 dB have been deemed unreliable. The threshold of 0 dB used here was arbitrarily chosen. Cooke et. al. [Cooke 1999] report that regions of the spectrogram where the SNR is lower than 15 dB are unreliable, and con-

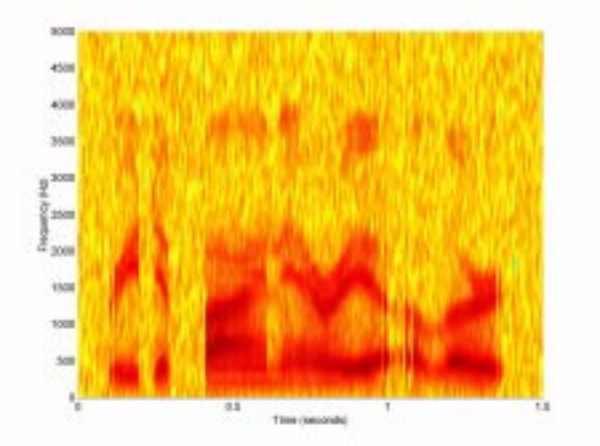


Figure 3.7 Wideband spectrogram of an utterance of speech that has been corrupted to 15 dB by additive white noise. The utterance is “Redefine Area Alert”.

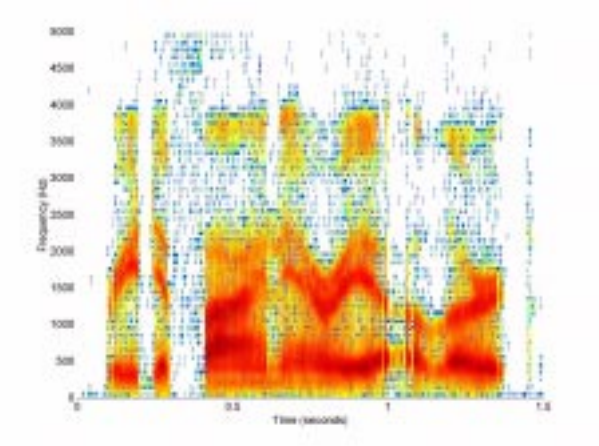


Figure 3.8 Wideband spectrogram of the same utterance when all regions with a local SNR less than 0 dB have been deleted. The white regions in the figure represent the deleted regions of the spectrogram.

tribute negatively to recognition performance. Using their definition, all regions of the spectrogram where the local SNR is less than 15dB would be deemed unreliable and erased. Our experiments (reported in Chapter 6) indicate that the optimal SNR threshold below which the spectrogram regions begin to affect recognition performance poorly lies between 5 dB and -5 dB.

All statements in the above discussion apply to mel spectrogram as well. All regions with a local SNR below a preset threshold could be deemed unreliable and erased. Recognition would have to be performed on the remaining figure. Figure 3.9 and Figure 3.10 show the mel spectrogram of an utterance of noisy

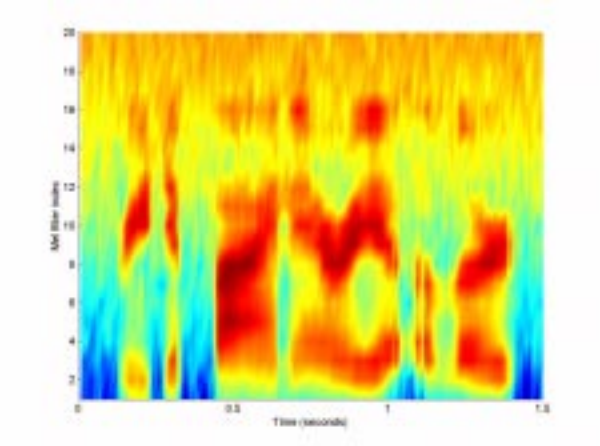


Figure 3.9 Mel spectrogram of an utterance of speech that has been corrupted to 10 dB by additive white noise. The utterance is “Redefine Area Alert”.

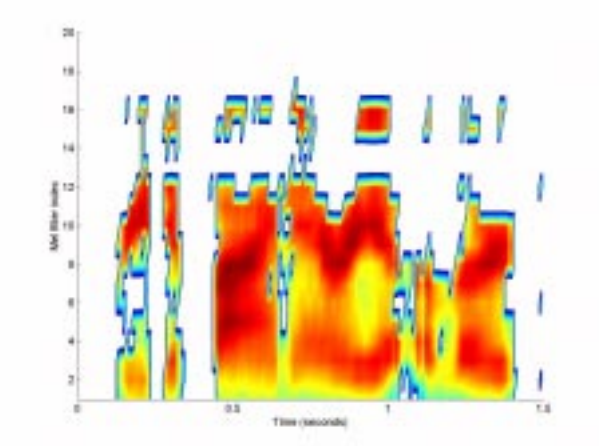


Figure 3.10 Mel spectrogram of the same utterance when all regions with a local SNR less than 0 dB have been deleted. The white regions in the figure represent the deleted regions of the spectrogram.

speech, and the same spectrogram where unreliable elements have been erased, respectively.

3.5 Summary

In this chapter we have described the spectrographic and mel-spectrographic representations of the speech signal. We have also described how the effect of noise corruption can be modeled as deletions of regions of the spectrogram. The result of such deletions are incomplete spectrograms. In the next chapter we will describe conventional methods of recognizing speech with incomplete spectrograms.

Chapter 4
Recognizing speech with incomplete spectrograms

4.1 Introduction

As explained in Chapter 2, a speech recognition system is a statistical pattern classifier. Statistical pattern classification is the problem of identifying which of a set of L classes a given data set X belongs to [Duda 1973]. Given a set of classes C_i, and the distribution of the data belonging to each of the classes $P(X|C_i)$, the optimal statistical pattern classifier estimates the class $class(X)$ that a data set X belongs to as [Duda 1973]

$$class(X) = \arg\max_k \{P(X|C_k)P(C_k)\} \tag{4.1}$$

where $P(X|C_k)$ is the likelihood of X given that it belongs to the k^{th} class C_k, and $P(C_k)$ is the prior probability of the k^{th} class. $P(X|C_k)P(C_k)$ is the *a posteriori* probability of the class C_k

Consider a situation where some components of the data set are missing or occluded (*i.e*. they cannot be observed or measured due to some reason). Let X_o represent the observed portion of X, and X_m the missing portion. The complete data is the combination of the observed and missing data, *i.e.* $X = (X_o, X_m)$. In this case Equation (4.1) becomes

$$class(X) = \arg\max_k \{P(X_o, X_m|C_k)P(C_k)\} \tag{4.2}$$

Clearly, this cannot be directly evaluated since X_m is not known and therefore its likelihood cannot be computed. We are therefore faced with the problem of *classification with incomplete data.*

In the context of speech recognition systems the problem would be stated in the following manner. Let $\boldsymbol{S}$ represent the sequence of parameter vectors derived from the utterance being recognized. The optimal classifier given by Equation (4.1) becomes

$$\hat{W} = \arg\max_W \{P(\boldsymbol{S}|W)P(W)\} \tag{4.3}$$

where $\hat{W}$ is the recognized sequence of words in that utterance, $P(\boldsymbol{S}|W)$ is the likelihood of $\boldsymbol{S}$ given that

W was the sequence of words uttered, and $P(W)$ is the prior probability that W was uttered. Let $\boldsymbol{S}$ be a spectrogram with some components missing. $\boldsymbol{S}$ can be decomposed into its observed and missing components as $\boldsymbol{S} = \{\boldsymbol{S}_o, \boldsymbol{S}_m\}$, where $\boldsymbol{S}_o$ is the observed portion of the spectrogram and $\boldsymbol{S}_{\boldsymbol{m}}$ is the missing portion. Equation (4.3) then becomes

$$\hat{W} = \arg\max_W \{P(\boldsymbol{S}_o, \boldsymbol{S}_m | W) P(W)\} \tag{4.4}$$

Once again, this cannot be evaluated directly since the value of $\boldsymbol{S}_{\boldsymbol{m}}$ is not known. Thus the problem of recognizing speech with incomplete spectrograms is also one of classification with incomplete data.

In order to perform classification (or recognition) with incomplete data (or spectrograms) it becomes necessary to develop procedures that can compensate for the missing data in some manner. When applied to speech recognition systems, we refer to these procedures as *incomplete-spectrogram methods* of recognition. Traditional incomplete-data methods such as those described in section 2.4 deal with *analysis* of data with components missing, or with *inference* of missing data. However, the final goal here is not to analyze spectrograms with missing regions, or even to infer the values of the missing regions, but rather to perform *classification* or *recognition* when some of the data are missing. The solution to the incomplete data problem in this situation has to keep the final goal (of classification or recognition) in mind, and in this respect it varies significantly from missing data inference methods.

There are two possible approaches to handling the problem of classification with incomplete data. The first approach is the so called *data imputation* approach where the missing portion of the data, $X_{\boldsymbol{m}}$, are estimated somehow. Classification is then performed using the estimated value, $\hat{X}_{\boldsymbol{m}}$

$$\begin{aligned} \hat{X}_m &= estimate(X_m) \\ class(X) &= \arg\max_k \{P(X_o, \hat{X}_m | C_k) P(C_k)\} \end{aligned} \tag{4.5}$$

One specific imputation based solution that is well suited to estimates used for classification is the so-called *class-conditional imputation*. Class-conditional imputation utilizes the distributions of the data as modeled by the classifier, in order to obtain statistical estimates for the missing components. This has been the imputation method of choice for speech recognition researchers and has been extensively investigated

and reported on [Cooke 1994][Lippmann 1997].

The other approach to recognizing speech with incomplete spectrograms is to reformulate the classification so as to perform the classification based on the observed components alone. This approach is referred to as the *marginalization* method since the unobserved components are marginalized out of the classification procedure. This has been the most successful method for recognition with incomplete spectrograms, and has also been extensively reported [Cooke 1999][Lippmann 1997][El-Maliki 1999].

Both class-conditional imputation and marginalization modify the manner in which the classifier, or recognizer, computes the *a posteriori* probabilities of the various classes in order to facilitate classification or recognition with incomplete data. We therefore refer to them as *classifier-modification methods*.

The following sections describe class-conditional imputation and marginalization in greater detail.

4.2 Class-conditional imputation

In class-conditional imputation a separate estimate of missing data X_m is obtained for each of the classes C_k, using the distribution of that class, conditioned on the observed data X_o. The Maximum A Posteriori (MAP) estimation procedure described in section 2.5.4 is used for the estimation. The *a posteriori* probability of any of the classes computed using the estimates of the missing data obtained using the distribution of that class. The classification procedure is therefore given by

$$\begin{aligned} \hat{X}_{m,k} &= \arg\max_X \{P(X|X_o, C_k)\} \\ class(X) &= \arg\max_k \{P(X_o, \hat{X}_{m,k}|C_k)P(C_k)\} \end{aligned} \tag{4.6}$$

where $\hat{X}_{m,k}$ is the MAP estimate of the missing data obtained assuming that the complete data X belonged to the k^{th} class C_k, conditioned on the observed data X_o. This procedure gets its name because the estimates of the missing data are conditional to the class being considered and are specific to that class. Figure 4.1 shows a schematic representation of the class-conditional imputation procedure.

When applied to a speech recognition system class-conditional imputation performs recognition as

$$\hat{W} = \arg\max_W \{P(\boldsymbol{S}_o, \hat{\boldsymbol{S}}_{m,W}|W)P(W)\} \tag{4.7}$$

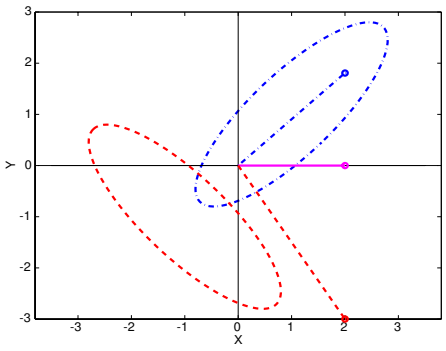


Figure 4.1 Schematic example for class-conditional imputation. The two ellipses represent the cross sections of the Gaussian distributions of the two classes in a two-class classification problem. An incomplete vector is to be classified as belonging to one of these classes. The solid line shows the X component of the vector whose Y component is missing. The MAP estimate for the complete vector obtained using the distribution of the class represented by the dashed ellipse, is given by the dashed line. Similarly, the MAP estimate obtained using the distribution of the dash-dotted ellipse is shown by the dash-dotted line. In class-conditional imputation, the *a posteriori* probability of the dashed class is computed using the dashed line, and the *a posteriori* probability of the dash-dotted class is computed using the dash-dotted line. The class with the higher likelihood is chosen as the estimate of the class that the complete vector belongs to.

where $\hat{\boldsymbol{S}}_{m,W}$, the MAP estimate of $\boldsymbol{S}_m$, is dependent on the particular word hypothesis being considered.

$$\hat{\boldsymbol{S}}_{m,W} = \arg\max_S \{P(S|\boldsymbol{S}_o, W)\} \tag{4.8}$$

As explained in section 2.2.1, in HMM-based speech recognition systems the recognizer estimates not just the best word sequence, but also the best state sequence associated with the word sequence. Equation (4.7) and Equation (4.8) therefore get modified to

$$\begin{aligned} \hat{W} &= \arg\max_{W,s} \{P(\boldsymbol{S}_o, \hat{\boldsymbol{S}}_{m,W,s}, \boldsymbol{s}|W)P(W)\} \\ \hat{W} &= \arg\max_{W,s} \{P(\boldsymbol{S}_o, \hat{\boldsymbol{S}}_{m,W,\boldsymbol{s}}|\boldsymbol{s}, W)P(\boldsymbol{s}|W)P(W)\} \end{aligned} \tag{4.9}$$

where $\boldsymbol{s}$ represents any valid state sequence that can be generated by the HMM for W. The estimate for the missing data is given by

$$\hat{\boldsymbol{S}}_{m,W,\boldsymbol{s}} = \arg\max_S \{P(S|\boldsymbol{S}_o, \boldsymbol{s}, W)\} = \arg\max_S \{P(S|\boldsymbol{S}_o, \boldsymbol{s})\} \tag{4.10}$$

where the second term is dependent only on $\boldsymbol{s}$ since W is redundant once $\boldsymbol{s}$ is known. We recall that the state sequence $\boldsymbol{s}$ is simply a sequence of states, one for every spectral vector in $\boldsymbol{S}$.

$$\boldsymbol{s} = [s_1, s_2, s_3, \ldots, s_N] \tag{4.11}$$

where N is the total number of spectral vectors in $\boldsymbol{S}$ and s_k is the state associated with the k^{th} vector in $\boldsymbol{S}$. The estimate for the missing components is therefore given by

$$\hat{\boldsymbol{S}}_{m,W,\boldsymbol{s}} = \text{argmax}_S\{P(S|\boldsymbol{S}_o, s_1, s_2, s_3, \ldots, s_N)\} \tag{4.12}$$

If we refer to the individual spectral vectors in $\boldsymbol{S}$ as $\boldsymbol{S}(t)$ and separate the missing and observed components of $\boldsymbol{S}(t)$ into $\boldsymbol{S}_m(t)$ and $\boldsymbol{S}_o(t)$ respectively, we get

$$\hat{\boldsymbol{S}}_{m,W,s} = [\hat{\boldsymbol{S}}_{m,W,s}(1), \hat{\boldsymbol{S}}_{m,W,s}(2), \hat{\boldsymbol{S}}_{m,W,s}(3), \ldots, \hat{\boldsymbol{S}}_{m,W,s}(N)] \tag{4.13}$$

where $\hat{\boldsymbol{S}}_{m,W,s}(t)$ refers to the estimate of $\boldsymbol{S}_m(t)$, the vector of missing components in $\boldsymbol{S}(t)$, the t^{th} vector of $\boldsymbol{S}$, when the word hypothesis being considered is W and the state sequence being considered is $\boldsymbol{s}$. Since HMMs assume that the individual vectors of the spectrogram are independent, Equation (4.12) leads us to

$$\hat{\boldsymbol{S}}_{m,W,s}(t) = \text{argmax}_S\{P(S|\boldsymbol{S}_o(t), s_t)\} \tag{4.14}$$

The right hand side of Equation (4.14) is independent of both the word sequence W and the complete state sequence $\boldsymbol{s}$, and is only dependent on the particular state s_t whose likelihood is being considered. Therefore in computing the likelihood of any state sequence that includes s_t we would use $\hat{\boldsymbol{S}}_{m,W,s}(t)$ as the estimate of the missing components of $\boldsymbol{S}(t)$. The implication of Equation (4.9), Equation (4.13) and Equation (4.14) is that the missing components of a vector are estimated separately for every state considered during recognition, conditioned on the observed components of that vector, and based on the distribution associated with that state. In the computation of the likelihood of any state for any vector, the estimates for the missing components of that vector that were obtained using the distribution of that state are used.

4.3 Marginalization

Another method of solving the problem of classification with incomplete data is to perform the classifi-

cation based on the observed data alone. The optimal classifier in this case is given by

$$class(X) = \arg\max_k \{P(X_o|C_k)P(C_k)\} \tag{4.15}$$

where $P(X_o|C_k)$ is the likelihood of the observed data, given that the data belongs to the k^{th} class. Distribution of the classes are usually defined on the complete data X and not on the observed components X_o alone. *i.e.* the defined distribution for any class C_k is $P(X|C_k) = P(X_o, X_m|C_k)$ and not $P(X_o|C_k)$. Indeed, it may be difficult to specify the distributions of the observed components alone since the precise set of observed components may vary from data set to data set. As a result it becomes necessary to obtain the distribution of the observed components by integrating the distribution of the complete data over all the missing components:

$$P(X_o|C_k) = \int_{-\infty}^{\infty} P(X_o, X_m|C_k)dX_m = \int_{-\infty}^{\infty} P(X|C_k)dX_m \tag{4.16}$$

$P(X_o|C_k)$ is traditionally referred to as the *marginal* distribution of X_o and the process of obtaining $P(X_o|C_k)$ from $P(X|C_k)$ is referred to as *marginalization*. Optimal classification is performed using the marginal distributions obtained using Equation (4.16).

$$class(X) = \arg\max_k \left\{ P(C_k) \int_{-\infty}^{\infty} P(X|C_k)dX_m \right\} \tag{4.17}$$

Since the classification is being performed using distributions that have been obtained by marginalization, the procedure of classifying with marginal distributions is also referred to as *marginalization.* Figure 4.2 shows a schematic representation of marginalization based classification with incomplete data.

When applied to a speech recognition system marginalization based recognition with incomplete spectrograms is performed as

$$\hat{W} = \arg\max_W \{P(\boldsymbol{S}_o|W)P(W)\} \tag{4.18}$$

Once again, since the distribution of the data associated with any word sequence W is defined on the

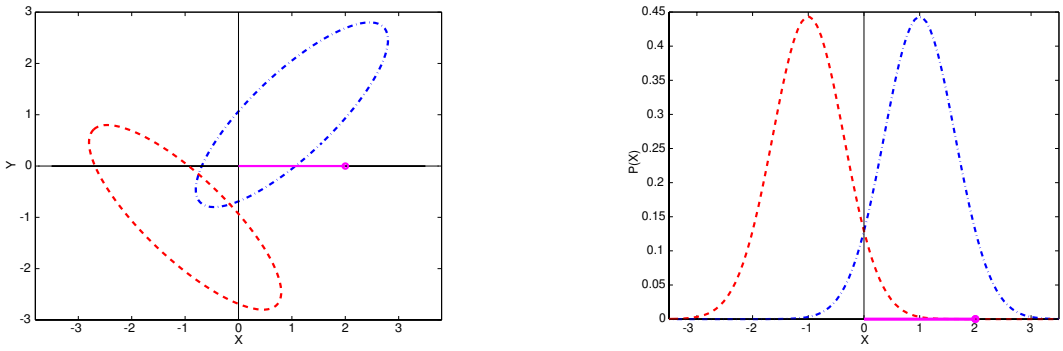


Figure 4.2 Schematic example for marginalization. In the left panel the two ellipses show the cross section of the Gaussian distribution of each of the classes. The sold line shows the X component of the vector whose Y component is missing. In marginalization the Y component of the two class distributions is eliminated by integrating it out of the distributions. The resulting distributions give only the distribution of the X components of the classes. The right panel shows the distribution of the X components of the two classes. Since the original distribution was Gaussian, these are also Gaussian. The Y component no longer figures in the problem. In this reduced situation, the *a posteriori* probability of the classes is computed based on the likelihood of the X component of the incomplete vector (given by the solid line) is computed on the Gaussians shown and the class with the higher *a posteriori* probability is chosen as the estimate of the class that the complete vector belongs to.

complete spectrogram rather than on the observed components alone, the marginal distributions of the observed components of the spectrogram would have to be obtained by integrating out the missing components from the distribution. The optimal recognition would now be defined over the marginal distributions so obtained as

$$\hat{W} = \arg\max_W \left\{ P(W) \int_{-\infty}^{\infty} P(\boldsymbol{S}_o, \boldsymbol{S}_m | W) d\boldsymbol{S}_m \right\} \tag{4.19}$$

HMM-based speech recognition systems jointly estimate the best state sequence along with the word sequence. Equation (4.19) therefore gets modified to

$$\begin{aligned} \hat{W} &= \arg\max_W \arg\max_{\boldsymbol{s}} \left\{ P(W) \int_{-\infty}^{\infty} P(\boldsymbol{S_o}, \boldsymbol{S}_m | \boldsymbol{s}, W) P(\boldsymbol{s} | W) d\boldsymbol{S}_m \right\} \\ \hat{W} &= \arg\max_W \arg\max_{\boldsymbol{s}} \left\{ P(\boldsymbol{s} | W) P(W) \int_{-\infty}^{\infty} P(\boldsymbol{S_o}, \boldsymbol{S}_m | \boldsymbol{s}) d\boldsymbol{S}_m \right\} \end{aligned} \tag{4.20}$$

where $\boldsymbol{s}$ represents any valid state sequence that can be generated by the HMM for W. As mentioned in Equation (4.11), $\boldsymbol{s} = [s_1, s_2, s_3, \ldots, s_N]$, where s_k is the state associated with the k^{th} vector in $\boldsymbol{S}$.

Referring to the individual vectors of $\boldsymbol{S}$ as $\boldsymbol{S}(t)$ and the missing and observed components of $\boldsymbol{S}(t)$ as $S_m(t)$ and $S_o(t)$ respectively as before, we get

$$P(\boldsymbol{S_o}, \boldsymbol{S_m}|\boldsymbol{s}) = P(S_o(1), S_m(1), S_o(2), S_m(2), \ldots, S_o(N), S_m(N)|s_1, s_2, \ldots, s_N)$$

$$P(\boldsymbol{S_o}, \boldsymbol{S_m}|\boldsymbol{s}) = \prod_{n=1}^{N} P(S_o(n), S_m(n)|s_n) \tag{4.21}$$

Combining Equation (4.20) and Equation (4.21), the HMM assumption of independence of individual vectors in the spectrogram leads us to

$$\hat{W} = \text{argmax}_W \text{argmax}_{\boldsymbol{s}} \left\{ P(\boldsymbol{s}|W)P(W) \prod_{n=1}^{N} \int_{-\infty}^{\infty} P(S_{\boldsymbol{o}}(n), S_m(n)|s_n) dS_m(n) \right\} \tag{4.22}$$

An alternate way of viewing Equation (4.22) is that the missing components in each vector of the spectrogram are integrated out of the distributions of all the states in the recognizer, in order to compute the likelihood of that vector. Since the set of missing components can vary from vector to vector this integration would have to be performed for every vector in the spectrogram.

4.4 Experimental results

The effectiveness of class-conditional imputation and marginalization in recognizing speech based on incomplete spectrograms was evaluated on incomplete spectrograms with simulated patterns of missing elements. Incomplete spectrograms were generated by erasing random elements of a mel-spectrographic representation of speech so as to obtain the desired fraction of missing elements. No noise was added to the observed regions in the spectrogram. We refer to this procedure of generating incomplete spectrograms as the *random-drop* mechanism, and the paradigm of evaluating incomplete-spectrogram methods on such spectrograms as the *random-drop paradigm*.

It is important to note here that the random-drop mechanism is not a realistic model for the effect of noise on the spectrograms of speech by any means. It is merely a useful paradigm for the quick evaluation of missing-data techniques, and is used only as a preliminary test for the techniques developed in this thesis. The true performance of these techniques can only be evaluated on speech corrupted by noise. The

deletions induced by noise tend to be much more systematic and occur in blocks. We describe the true nature of deletions induced by noise in Chapter 6 in greater detail. Nevertheless, the random-drop paradigm remains a very useful paradigm for evaluating the efficacy of missing-feature methods, since the patterns of missing regions are not biased by the systematic behavior of any corrupting noise. Furthermore, the additional effect of noise on the observed regions of the spectrogram need not be considered.

Figure 4.3 shows a typical mel spectrogram when different fractions of the spectrogram have been randomly erased. In all of our experiments the mel-spectral representation with 20 mel filters, *i.e.* a mel spectrogram where the individual vectors have 20 components, has been used.

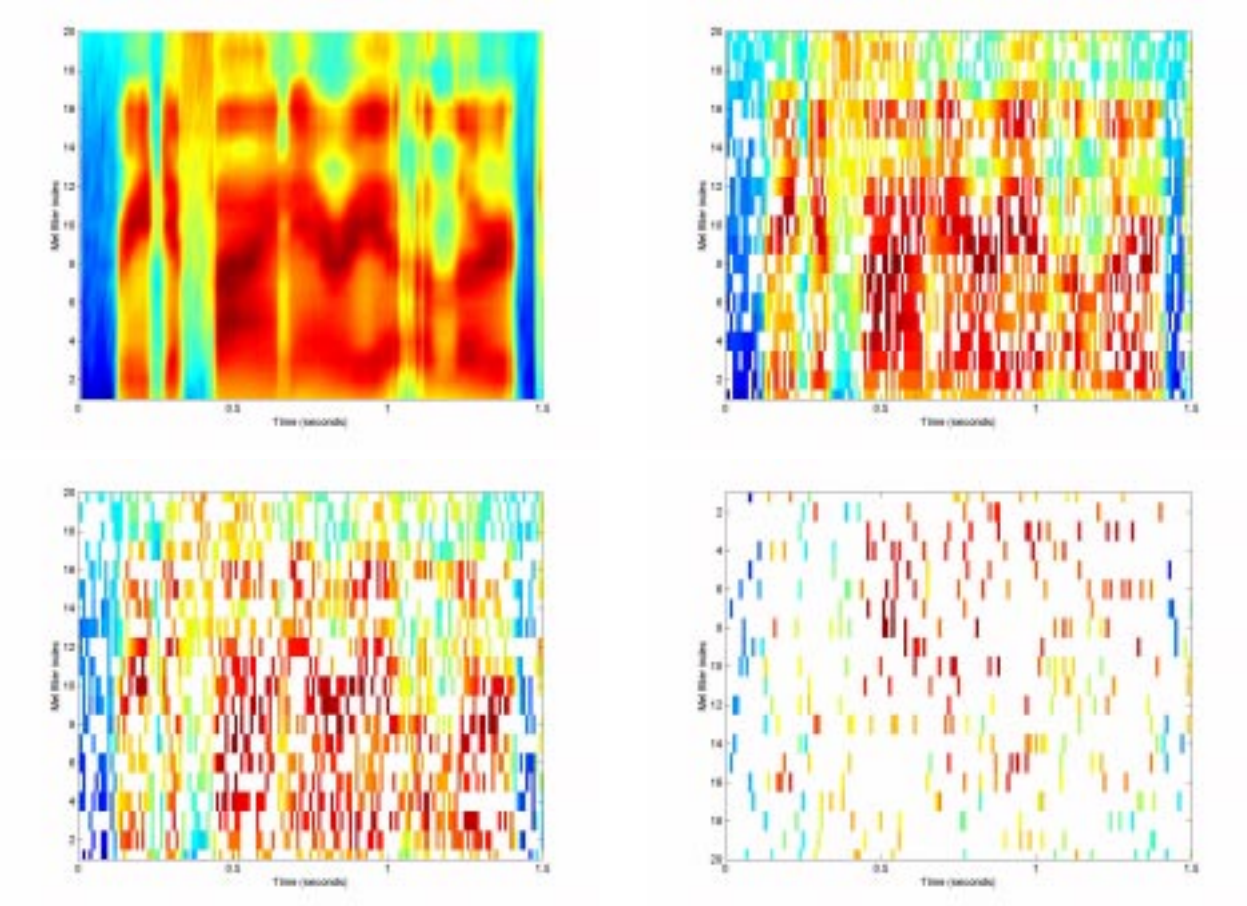


Figure 4.3 Examples of a mel spectrogram with randomly missing regions. The top left panel shows the original mel spectrogram for the utterance “Redefine Area Alert”. The top right panel shows the same spectrogram when 40% of its elements have been randomly deleted. The white portions of the picture represent the deleted regions. The bottom left panel shows the spectrogram when 60% of its elements have been randomly deleted. The bottom right panel shows it with 90% of its elements deleted.

Experiments were run using the DARPA Resource Management (RM) database [Price 1988] on the

CMU Sphinx-III HMM based recognition system. Continuous HMMs with single Gaussian state distributions were trained. The 20 dimensional log-mel-spectral vectors were used as the features to train the recognition system. The system was trained with 2880 utterances of uncorrupted spectrograms. The test set consisted of 1600 utterances from the RM database. Random elements were dropped from the spectrograms of the test data as described above.

Figure 4.4 shows the recognition accuracy obtained using class-conditional imputation and marginalization as a function of the fraction of elements missing in the test spectrograms. As can be seen, these

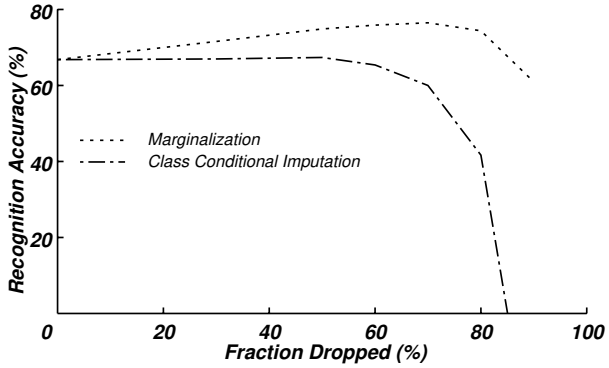


Figure 4.4 Recognition accuracy as a function of drop fraction for class-conditional imputation and marginalization. The horizontal axis show the drop fraction, *i.e.* the fraction of elements deleted from the spectrogram. The vertical axis shows the recognition accuracy obtained using the incomplete spectrograms.

methods are highly effective at handling missing data in spectrograms. Class-conditional imputation results in recognition accuracies comparable to those obtained with uncorrupted spectrograms when 70% of the elements in the spectrogram are missing. Marginalization performs recognition using the optimal classifier, given only the observed elements of the spectrogram. It is therefore expected to perform better than class-conditional imputation. We observe from Figure 4.4 that marginalization is indeed far more effective than class-conditional imputation. The recognition accuracy obtained when 90% of the spectrogram is missing is only slightly worse than that obtained with the uncorrupted spectrogram. While these results speak highly of the these methods, they also seem indicative of the high degree of redundancy in the speech signal. This is in agreement with human performance which is very robust to high degrees of degradation or spectro-temporal excision of the speech signal.

We would like to point out the anomalous results seen in Figure 4.4 whereby the recognition accuracy obtained with spectrograms where 80% of the elements have been deleted is actually superior to the perfor-

mance obtained with complete spectrograms. We do not have a satisfactory explanation for this behavior. We hypothesize that this behavior is characteristic to the Resource Management database used in these experiments. Other researchers who have used this database have obtained similar results with the random drop paradigm [Cooke 1994]. However, this behavior has not been seen with other databases.

4.5 Drawbacks with classifier modification methods

While class-conditional imputation and marginalization are very effective at recognizing speech based on spectrograms with random elements missing, they suffer from several drawbacks.

Both class-conditional imputation and marginalization are classifier compensation methods. They attempt to compensate for the missing data either by estimation on the basis of, or modification of, the distributions of the classes. In order to be able to either estimate the missing components based on the distributions of the classes, as in class-conditional imputation, or to be able to marginalize out the missing components, it becomes essential that the distributions be defined on the same parameters where the missing components are identified. *Since components of the spectrogram are missing, it becomes necessary to train the recognizer using spectrographic features. As a result recognition can only be performed using log spectral vectors*. Figure 4.5 explains this limitation schematically.

This limitation gives rise to several problems:

1) It is known that, with uncorrupted vectors, the performance of HMM based recognition systems is

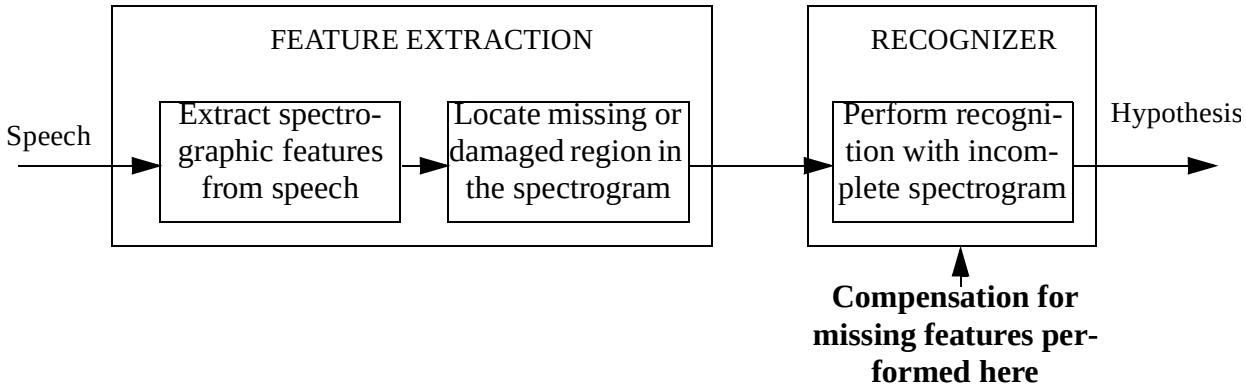


Figure 4.5 Block diagram explaining classifier compensation methods of recognition with incomplete spectrograms. The speech recognition system has the two modules. The feature extraction module extracts features from the speech signal. The recognition module performs recognition with the features. In classifier compensation techniques, the feature extraction module generates incomplete spectrograms. The recognizer recognizes speech based on these incomplete spectrograms. Thus, the recognizer has to be trained on spectrographic features.

much better when the recognizer is trained using cepstra, rather than log spectra [Davis 1980]. Table 4.1 compares the recognition accuracy obtained on clean speech using cepstra with that obtained using log spectra. Clearly, the accuracy obtained with cepstra is much higher. Similar situations arise with other kinds of classifiers which may perform better with other features than with spectral vectors.

Recognition accuracy with log spectral vectors	Recognition accuracy with cepstral vectors
63%	82%

Table 4.1 Comparison of the recognition accuracy obtained with log-spectral vectors with the recognition accuracy obtained with cepstral vectors on the RM database. In both cases an HMM-based recognizer with 2000 tied states, each modeled by a single Gaussian, was used.

2) Difference and double-difference parameters are commonly used to improve recognition performance. Difference parameters are computed as the difference between vectors.

$$\begin{aligned} d\boldsymbol{S}(t) &= \boldsymbol{S}(t+\tau) - \boldsymbol{S}(t-\tau) \\ dd\boldsymbol{S}(t) &= d\boldsymbol{S}(t+\tau) - d\boldsymbol{S}(t-\tau) \end{aligned} \tag{4.23}$$

where τ typically takes values between 1 and 4. $dd\boldsymbol{S}(t)$ is the double-difference parameter vector at time t, $d\boldsymbol{S}(t)$ is the difference parameter vector at time t, and $\boldsymbol{S}(t)$ is the spectral vector at time t. If an element of either $\boldsymbol{S}(t+\tau)$ or $\boldsymbol{S}(t-\tau)$ is missing the corresponding element in $d\boldsymbol{S}(t)$ cannot be computed, and would therefore also be missing. It is easy to see that the fraction of missing elements can be up to twice as high in the difference parameters as in the spectral vectors. Similarly, the fraction of missing elements in the double-difference parameters can be up to four times as high as the spectral vectors themselves. Thus, the missing feature methods described in this chapter would have to compensate for the much higher fractions of missing elements in the difference and double-difference parameters, reducing the contributions of these parameters to recognition performance greatly.

3) Mean normalization is a procedure by which the mean of the spectral vectors in any utterance is subtracted from all the vectors in the utterance. Variance normalization similarly normalizes the vectors in the utterance by their variance. Both procedures have been shown to improve the recognition performance of speech recognition systems. However, when the spectrographic parameters that are used for recognition have missing elements, the estimates of the means and the variance of the spec-

tral vectors can be biased by the patterns of the missing elements. This can render both mean normalization and variance normalization ineffective.

The main reason for all of the problems above is that class-conditional imputation and marginalization attempt to perform classification with incomplete spectrograms, directly.

The data-compensation approach to the missing feature paradigm

In this thesis we recast the problem of recognition with incomplete spectrograms as a *data compensation* problem. Instead of performing recognition directly with incomplete spectrograms, we *reconstruct* all the missing regions of the spectrograms in a preliminary pre-processing step. We call this the *spectrogram reconstruction approach*. Cepstral features can now be derived from the fully reconstructed spectrogram and used to perform recognition. Since the reconstruction of spectrograms is done independently of the recognizer, the recognizer need not be modified in any manner. Figure 4.6 represents the proposed approach as a block diagram.

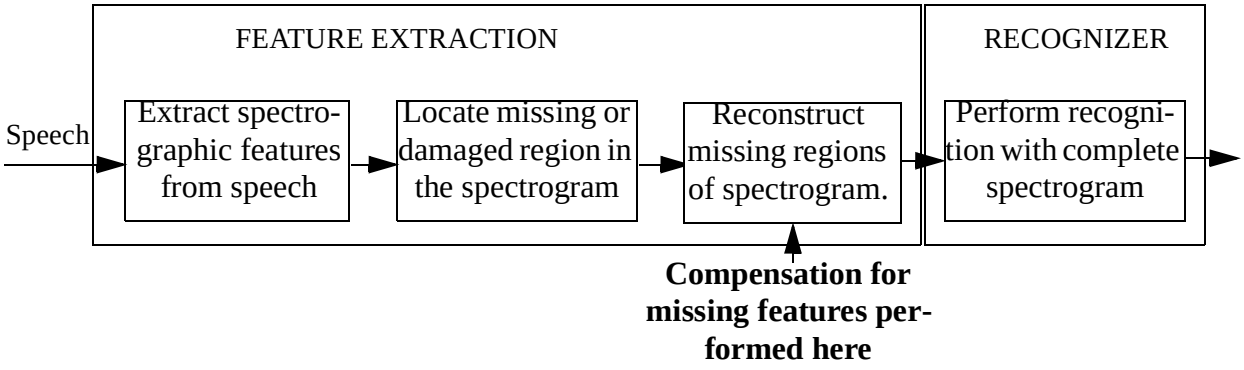


Figure 4.6 Block diagram explaining the data-compensation approach to recognition with incomplete spectrograms. The missing regions of the incomplete spectrograms are reconstructed in the feature extraction module itself. Thus, the output of the feature extraction module is a complete, reconstructed spectrogram. This reconstructed spectrogram can then be transformed to any feature of choice, if desired, before being passed on to the recognizer. The recognizer works on complete features, and can work with any feature extracted from the complete spectrogram.

If recognition is to be performed using spectrographic features, the data-compensation approach is sub-optimal to classifier-modification methods. This is because the reconstruction approach uses *estimates* of the missing data, and these are bound to be erroneous to varying degrees, depending on the manner in which they are obtained. In such a situation it can be argued that classification based on the observed data alone is more optimal than classification that uses estimates for the missing data [Moreno 1996]. This argument is also borne out by the fact that recognition accuracies obtained using marginalization, which

uses only the observed data, are higher than that obtained using class-conditional imputation which uses estimates of the missing data.

There are, however, many advantages to the data-compensation approach. The primary advantage is that, since the complete reconstructed spectrogram is now available, the recognizer is no longer constrained to perform recognition using spectrographic features. We can derive a more optimal set of parameters from the reconstructed spectrogram and use these features to perform the recognition. It is expected that the improvement in classification accuracy obtained due to the use of the more optimal feature set more than offsets the reduction in accuracy occurring due to the use of estimated values for the missing data in classification. Furthermore, since the complete spectrogram, or the set of cepstral or other features derived from the complete spectrogram, are now available computation of difference parameters and variance and mean normalization can be performed in the usual fashion. Another advantage is that since a complete spectrogram is now available for recognition, the recognizer itself need not be modified in any manner to account for the missing data. The missing feature estimation procedure can be performed independently of the recognizer, permitting any standard recognizer to be used. Finally, the proposed procedure permits reconstruction methods that use different models for speech than that used by the recognizer. The spectrogram reconstruction procedure can be performed using very simple statistical and parametric models of speech spectrograms. The resulting methods can be much simpler and much more computationally efficient than classifier compensation methods such as class-conditional imputation and marginalization.

We investigate spectrogram reconstruction methods for recognition with incomplete spectrograms in the following chapters.

Chapter 5
Spectrogram reconstruction methods for missing data

5.1 Introduction

In this chapter we address the problem of estimating missing regions of incomplete spectrograms to reconstruct complete spectrograms. We investigate several simple *spectrogram reconstruction methods* that estimate missing elements based on the geometric structure of speech spectra and on simple statistical information culled from available corpora of uncorrupted speech. Since this thesis is primarily concerned with speech recognition, our goal is not simply good reconstruction or analysis of spectrograms but also that of achieving good recognition performance with the reconstructed spectrograms. The developed techniques are therefore evaluated based on the recognition performance achieved with the reconstructed spectrograms.

The simplest manner of reconstructing missing regions in spectrograms would be to do it based only on the geometrical placement of the observed regions of the spectrogram. We refer to these as *geometrical reconstruction methods* since all the information used to reconstruct the missing regions is present within the spectrogram, *i.e.* it is local to the spectrogram. No additional sources of information are used.

The features of a spectrogram show continuity across both frequency and time. Therefore, it can be expected that the frequency components of the spectral vectors in the spectrogram show statistical dependencies both with other components within the same vector as well as with the components of the other vectors in the spectrogram. Where additional corpora of uncorrupted speech are available, the statistical relations between the various components of the spectrogram can be learned from these corpora. The statistical relations learned can be “vector statistics”, *i.e.* the distribution of spectral vectors and the statistical relationship between the various frequency components within spectral vectors, or “covariance statistics”, *i.e.* the statistical relationship between the components of different vectors in the spectrogram. These statistical relations can then be used to condition the reconstruction of missing features. We refer to these spectrogram reconstruction methods as *statistical reconstruction methods*

In the following sections we investigate three types of reconstruction algorithms:

1) Geometrical reconstruction methods based on linear and non-linear interpolation

2) Statistical reconstruction methods that utilize *vector statistics* learnt from uncorrupted spectrograms of clean speech to reconstruct incomplete spectrograms
3) Statistical reconstruction methods that use *covariance statistics* learnt from uncorrupted spectrograms to perform reconstruction.

We evaluate all the spectrogram reconstruction methods described in this chapter both on the basis of the accuracy of the reconstruction and on the recognition accuracy of a speech recognition system which uses the estimated spectrograms. The random-drop paradigm described in Section 4.4, wherein randomly chosen elements of the spectrogram are deleted, has been used to evaluate all the reconstruction methods. We would like to reiterate here that the random-drop paradigm is not a realistic model for the effect of noise on speech spectrograms. When deletions in spectrograms are noise induced the missing regions in the spectrogram do not occur at random. Instead they occur in blocks and are systematic. Another difference between deletions generated by the random-drop paradigm and noise-induced deletions is that in the random-drop paradigm it is assumed that the locations of the missing elements are known *a priori*. When deletions in the spectrogram are noise induced, the locations of the deleted regions would not be known *a priori* and would have to be estimated. Thus, it should not be expected that recognition results obtained with deletion patterns generated by the random-drop paradigm would carry over to spectrograms with noise-induced deletions. However, the random-drop paradigm is a useful tool for preliminary evaluation of the spectrogram reconstruction methods, and has been used only to that end in this chapter. We evaluate the efficacy of the spectrogram reconstruction methods developed in this chapter on noise-induced deletions in Chapter 6.

In the rest of this chapter we follow the notation introduced in earlier chapters to denote a spectrogram by $\boldsymbol{S}$. The observed portion of the spectrogram is denoted by $\boldsymbol{S}_o$ and the missing portion by $\boldsymbol{S}_{\boldsymbol{m}}$. We represent an arbitrary spectral vector as S and the t^{th} spectral vector in the spectrogram $\boldsymbol{S}$ by $\boldsymbol{S}(t)$. The entire spectrogram consists of the sequence of spectral vectors $\boldsymbol{S}(1), \boldsymbol{S}(2), \boldsymbol{S}(3), \ldots, \boldsymbol{S}(N)$, represented more compactly as, $\boldsymbol{S}(t), 1 \leq t \leq N$, where N represents the total number of spectral vectors in the spectrogram. The missing components of the t^{th} spectral vector, $\boldsymbol{S}(t)$ are represented by $\boldsymbol{S}_m(t)$ and the observed components by $\boldsymbol{S}_o(t)$. The k^{th} frequency component of the t^{th} spectral vector, $\boldsymbol{S}(t)$, is represented by $S(t, k)$. The sequence of components $S(t, 1), S(t, 2), S(t, 3), \ldots, S(t, K)$, represented more

compactly as $S(t, k), 1 \le k \le K$, comprises the entire spectral vector $\boldsymbol{S}(t)$, where K is the total number of frequency components in the vector. In a mel spectrogram K would refer to the total number of mel filters being used (Section 3.2).

The following section deals with geometrical reconstruction methods.

5.2 Geometrical reconstruction methods

The simplest method of reconstructing a missing element in a spectrogram is by interpolating between adjacent observed elements in the spectrogram. Since the spectrogram has a two-dimensional support (frequency and time) these elements could be adjacent along either of the axes, frequency or time. When the elements used for interpolation are adjacent along the frequency axis, we refer to it as interpolation along frequency. When the elements are adjacent in time we refer to it as interpolation along time.

The interpolation used could be simple linear interpolation, or it could use other higher-order functional forms such as polynomials, rational functions, or splines. We will now describe and evaluate missing-feature reconstruction by linear and non-linear interpolation, both along frequency and along time.

5.2.1 Linear interpolation

The simplest form of interpolation is linear interpolation. Consider any sequence of numbers $s[1], s[2], \ldots, s[M]$, where the samples in the interval $[l_1, l_2]$ are unknown or missing, *i.e.* the values $s[l], l_1 \le l \le l_2$ are missing. Linear interpolation based estimates of the missing values are obtained by drawing a straight line between the nearest known neighbors, $s[l_1 - 1]$ and $s[l_2 + 1]$, and reading the estimated values of $s[l_1]$ through $s[l_2]$ off this line. Mathematically, the estimated value $\hat{s}[l]$ for any missing element $s[l]$ in the range $l_1 \le l \le l_2$ is given by [Press 1992]

$$\hat{s}[l] = s[l_1 - 1] + \frac{(s[l_2 + 1] - s[l_1 - 1])(l - l_1 + 1)}{l_2 - l_1 + 2} \qquad (l_1 \le l \le l_2) \tag{5.1}$$

Linear interpolation along frequency: Linear interpolation can be used to estimate the missing components of a spectral vector based on the observed components within the same vector. In this case, the

sequence considered in Equation (5.1) would be the components of the spectral vector, $S(t, k), 1 \le k \le K$. Here if frequency components $[k_1, k_2]$ in the l^{th} vector, *i.e*, $S(l, k), k_1 \le k \le k_2$, are missing the estimate for the missing values would be given by

$$\hat{S}(l, k) = S(l, k_1 - 1) + \frac{S(l, k_2 + 1) - S(l, k_1 - 1)}{k_1 - k_1 + 2}(k - k_1 + 1) \tag{5.2}$$

Since the estimates for the missing components are obtained by interpolation between other frequency components within the same vector, we refer to this method as *linear interpolation along frequency.*

Linear interpolation along time: Missing components of the spectrogram can also be estimated by linear interpolation between the same frequency components in adjacent spectral vectors. In this case, the sequence of points considered for interpolation would be a single slice of the spectrogram, parallel to the time axis, *i.e.* $S(l, k), 1 \le l \le N$. For brevity we refer to such a slice of the spectrogram as a *time slice* of the spectrogram. Here if the k^{th} frequency component in vector numbers $[l_1, l_2]$, *i.e.* $S(l, k), l_1 \le l \le l_2$, were missing, the estimate for these missing values would be given by

$$\hat{S}(l, k) = S(l_1 - 1, k) + \frac{S(l_2 + 1, k) - S(l_1 - 1, k)}{l_2 - l_1 + 2}(l - l_1 + 1) \tag{5.3}$$

Since the estimates are now obtained by interpolation between the same frequency components at other time instants, we refer to this method as *linear interpolation along time.*

For both interpolation along frequency and interpolation along time, if the missing elements being estimated lie at the boundaries of the spectrogram, they cannot estimated by interpolation. For example, if $S(l, k), k_1 \le k \le k_2$ are missing and $k_1 = 1$ or $k_2 = K$, these elements cannot be estimated by interpolation along frequency since the spectral vector has observed components on only one side of the missing elements. Similarly, if $S(l, k), l_1 \le l \le l_2$ are missing and $l_1 = 1$ or $l_2 = N$, they cannot estimated by interpolation along time since all the observed values of frequency component k are to one side of the missing segment. In both these cases the missing elements have to be estimated by linear extrapolation of the two closest observed elements instead of interpolation. For the case of estimation by linear extrapolation along frequency, if the closest observed components of the vector are $S(l, k_3)$ and $S(l, k_4)$, the miss-

ing boundary elements would be given by

$$\hat{S}(l, k) \;=\; S(l, k_3) + \frac{S(l, k_4) - S(l, k_3)}{k_4 - k_3}(k - k_3) \tag{5.4}$$

Similarly, where elements are being estimated by extrapolation along time, if the closest observed elements to the missing components in the time slice are $S(l_3, k)$ and $S(l_4, k)$ the missing boundary elements are obtained as

$$\hat{S}(l, k) \;=\; S(l_3, k) + \frac{S(l_4, k) - S(l_3, k)}{l_4 - l_3}(l - l_3) \tag{5.5}$$

Alternately, missing boundary points could be filled in by simple replication of the last observed element.

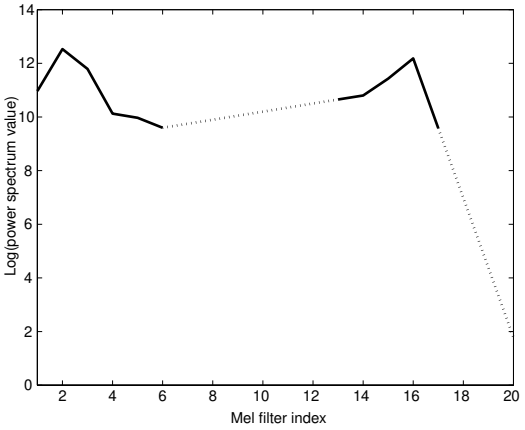


Figure 5.1 Plot of a single spectral vector. The dotted regions are linear interpolation/extrapolation estimates of missing values.

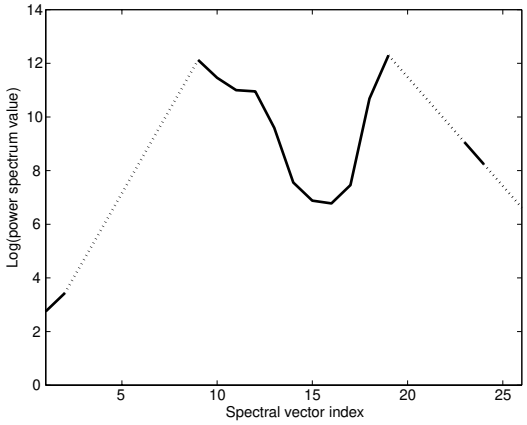


Figure 5.2 Plot of the trajectory of a single frequency component with time. The dotted regions are linear interpolation/extrapolation estimates of missing values

Figure 5.1 shows an example of estimation by interpolation along frequency. The figure plots the values of the frequency components of a single spectral vector against the index of the frequency component. Elements that are missing in the middle of the vector have been estimated using interpolation while those missing towards the edges have been estimated by extrapolation. Figure 5.2 similarly illustrates estimation by interpolation along time. The trajectory of a single frequency component is traced (*i.e.* a time slice of the spectrogram). Data points missing in the middle of the plot have been estimated by interpolation and those missing towards the edges have been estimated by extrapolation.

5.2.2 Nonlinear interpolation with polynomial functions

A polynomial of order M relating two variables x and y is a function of the form

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_M x^M \tag{5.6}$$

A line is a polynomial of order one. There is a unique line through any two points. Extrapolating that statement it can be shown that through any N points there is a unique polynomial of order $N-1$. Given a set of L points on a plane, $\{(x_1, y_1), (x_2, y_2), \ldots, (x_L, y_L)\}$, the unique $L-1^{\text{th}}$ order polynomial that passes through the L points can be determined using Lagrange's formula [Press 1992]

$$y = f_{L-1}(x) = \frac{(x-x_2)(x-x_3)\ldots(x-x_L)}{(x_1-x_2)(x_1-x_3)\ldots(x_1-x_L)} y_1 + \frac{(x-x_1)(x-x_3)\ldots(x-x_L)}{(x_2-x_1)(x_2-x_3)\ldots(x_2-x_L)} y_2 + \ldots + \frac{(x-x_1)(x-x_2)\ldots(x-x_{L-1})}{(x_L-x_1)(x_L-x_2)\ldots(x_L-x_{L-1})} y_L \tag{5.7}$$

While Lagrange's formula gives us a direct polynomial formulaic relation between an arbitrary x and the corresponding y, a procedurally and computationally simpler method to obtain y for a given x is to use Neville's algorithm. Neville's algorithm is a recursive procedure that begins by computing L zeroth order polynomials (constants) and recursively computes the M^{th}-order polynomial as a linear interpolation between two polynomials of order $M-1$. The details of the algorithm can be found in [Press 1992].

Polynomial functions can be used to estimate missing values in a sequence. Consider any sequence of numbers $s[1], s[2], \ldots$ where the values of the sequence in the interval $[l, r]$ are unknown or missing, *i.e* the values $s[n], l \le n \le r$, are missing. We can denote any element $s[l]$ in the sequence as a point $(l, s[l])$ on a plane. Let $(l_1, s[l_1]), (l_2, s[l_2]), \ldots, (l_P, s[l_P])$ be the set of P observed points in the sequence immediately preceding the point $(l, s[l])$ (*i.e.* $l_1, l_2, \ldots, l_P < l$). Similarly, let the set of points $(r_1, s[r_1]), (r_2, s[r_2]), \ldots, (r_Q, s[r_Q])$ be the set of Q observed values immediately following the point $(r, s[r])$ (*i.e.* $r_1, r_2, \ldots, r_Q > r$). A polynomial $f_{P+Q-1}(n)$ of order $P+Q-1$ can be fitted to these $P+Q$ points using Equation (5.7). The estimates for values of the points in the missing interval can now be derived from the polynomial as

$$\hat{s}[n] = f_{P+Q-1}(n), \quad l \le n \le r \tag{5.8}$$

This procedure is referred to as polynomial interpolation. P and Q can be chosen according to the kind of polynomial fit desired. Typically when polynomial interpolation with a polynomial of order $L-1$ is desired, the $P = L/2$ points immediately preceding the missing points and the $Q = L/2$ points immediately following them are used to determine the polynomial.

Missing regions in spectrograms can be estimated using polynomial interpolation. Once again, the interpolation can be performed either across frequency or across time. As before, when estimates are obtained by interpolating between the frequency components of the same vector we refer to the procedure as *polynomial interpolation along frequency.* To interpolate across frequency the sequence of points considered in Equation (5.7) consists of the components of a single spectral vector, $S(t, k), 1 \le k \le K$. Here, if the frequency components $[l, r]$ in the t^{th} vector, *i.e.* $S(t, k), l \le k \le r$, are missing, we would locate the $L/2$ closest observed frequency components of the vector preceding the missing region and the $L/2$ closest frequency components following it and use these in Equation (5.7) to obtain an $L-1^{\text{th}}$ order polynomial, $f_{L-1}(l)$. The estimates of the missing components are obtained as

$$\hat{S}(t, k) = f_{L-1}(k), \quad l \le k \le r \tag{5.9}$$

If the missing points are estimated by interpolating between the same frequency components of adjacent spectral vectors we refer to the procedure as *polynomial interpolation along time*. In this case the sequence of points considered in Equation (5.7) consists of a single time slice of the spectrogram, *i.e.* $S(t, k), 1 \le t \le N$. Here, if the k^{th} frequency component in vector numbers $[l, r]$, *i.e.* $S(t, k), l \le t \le r$, were missing, we would locate the $L/2$ vectors immediately preceding the missing region whose k^{th} components are present and similarly the $L/2$ vectors immediately following the missing regions and use these in Equation (5.7) to obtain an $L-1^{\text{th}}$ order polynomial, $f_{L-1}(t)$. The estimates of the missing components are obtained from this polynomial as

$$\hat{S}(t, k) = f_{L-1}(t), l \le t \le r \tag{5.10}$$

In both interpolation along frequency and interpolation along time, some missing regions may not have

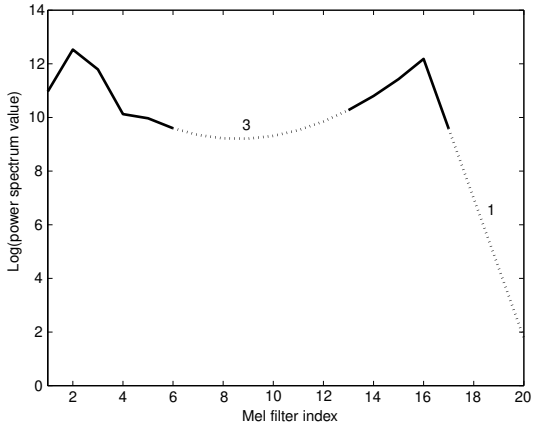


Figure 5.3 Plot of a single spectral vector. The dotted regions are polynomial-interpolation estimates of missing values. The order of the polynomial used is given above the dotted lines. Missing boundary elements are obtained by extrapolation.

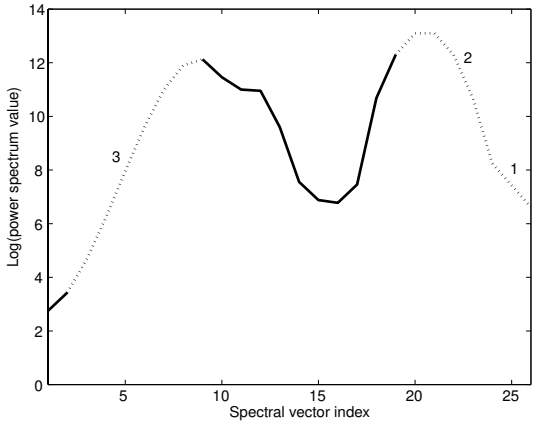


Figure 5.4 Plot of the trajectory of a single frequency component with time. The dotted regions are polynomial-interpolation estimates of missing values. The order of the polynomial used is shown. Missing boundary elements are obtained by extrapolation

$L/2$ observed components preceding or following them, *i.e.* either P or Q (or both) in Equation (5.9) would be less than $L/2$. Here a polynomial of lower order is fitted to the available points and the estimates for the missing values are obtained from this lower order polynomial. Also, interpolation is not possible for missing elements on the boundaries of the spectrogram. These regions are estimated by linear extrapolation as described in Section 5.2.1.

Figure 5.1 illustrates estimation by polynomial interpolation along frequency pictorially. The values of the frequency components of a single spectral vector are plotted against the index of the frequency component. Figure 5.2 similarly illustrates estimation by polynomial interpolation along time pictorially. The trajectory of a single frequency component is traced (*i.e.* a time slice of the spectrogram). In both figures, polynomial interpolation with a polynomials of order 3 has been performed. Where the number of points available for interpolation was insufficient a polynomial of lower order has been used. Missing regions at the boundaries have been estimated by linear extrapolation.

5.2.3 Nonlinear interpolation with rational functions

A rational function is defined as a quotient of polynomials. For example, the function

$$R_{N,M}(x) = \frac{P_N(x)}{Q_M(x)} = \frac{1 + a_1 x + a_2 x^2 + \ldots + a_N x^N}{b_0 + b_1 x + b_2 x^2 + \ldots + b_M x^M} \tag{5.11}$$

is a rational function, with a "numerator polynomial" of order N and a "denominator polynomial' of order M. We refer to such a function as a rational function of order (N, M). A rational function of order (N, M) such as the one given in Equation (5.11) has $N + M + 1$ parameters and is therefore uniquely described by $N + M + 1$ points. Given $N + M + 1$ points one can therefore construct the order (N, M) rational function.

An efficient algorithm to construct rational functions for the special cases when $M = N$ or $M = N + 1$ is the *Bulirsch-Stoer* algorithm [Press 1992]. The Bulirsch-Stoer algorithm is a recursive procedure that constructs increasing orders of rational functions from rational functions of lower order. The constraint, however, is that the order of the denominator polynomial has to be the same as, or one more than the order of the numerator polynomial, *i.e.* $N \leq M \leq N + 1$.

Rational-function interpolation is performed very similarly to polynomial interpolation. A rational function of the desired order is fitted to the points immediately adjacent to the missing points in a sequence, and the estimates of the missing points are derived from the rational function. Rational-function interpolation can be used to estimate missing points in a spectrogram. Interpolation along frequency and interpolation along time are both possible. In order to use an order (N, M) function for estimation we would need $N + M + 1$ observed points to compute the function. Of these, ideally, $(N + M + 1)/2$ of the observed points would precede the points to be estimated and $(N + M + 1)/2$ would follow them. Once the rational function has been obtained from these points, the estimates for the missing points can be obtained as the value of the rational function at the appropriate indices.

Once again, if $N + M + 1$ points are not available for the estimation the order of the rational function would have to be reduced to accommodate the available points. Also, as in the case of linear and polynomial interpolation, missing points near the boundaries of the spectrogram would have neighbors available on only one side and would therefore have to be estimated by extrapolation instead of interpolation.

There are several other interpolation techniques such as cubic spline interpolation etc. that can be used

to estimate the values of missing points. However, they have not been attempted in this thesis since we expect their performance to not be greatly different from those obtained with the interpolation methods described in this section.

5.2.4 Experimental results with interpolation based estimation of missing points

The principal goal of reconstruction (*i.e.* estimation of missing elements of the spectrogram) is not so much to effect an accurate, error-free reconstruction of the missing points as to reconstruct a complete spectrogram that can be used for recognition without much degradation in recognition accuracy. These goals are not unrelated to each other to the extent that error-free reconstruction of missing regions would result in high recognition accuracy. However, the converse is not necessarily true - it is not necessary that reconstructed spectrograms that result in high recognition accuracies would be very similar to the original, uncorrupted, spectrogram. Thus, while the *accuracy* of the reconstruction methods is evaluated by the error in the reconstruction, the *effectiveness* of the reconstruction methods in achieving the primary goal of the reconstruction is measured by the recognition performance obtained with the reconstructed spectrograms.

Experiments were conducted to evaluate the effectiveness of the interpolation-based reconstruction methods described above. The spectrogram reconstruction methods were evaluated on spectrograms with elements randomly deleted following the paradigm explained in Section 4.4. The experimental setup used was also identical to the one used to evaluate marginalization and class-conditional imputation in Section 4.4. A 20 mel-filter based mel-spectral representation was used to parametrize speech. The recognition system was trained directly with the log-mel-spectral parameters. The fully-continuous HMM-based SPHINX-III system was used for all experiments with the DARPA resource management database. Random elements of the mel spectrogram were deleted and reconstructed using linear interpolation, polynomial interpolation with polynomials of order 3, and rational-function interpolation with rational functions of order (1,2). Both interpolation across frequency and interpolation across time were evaluated. The orders for the polynomial and rational-function interpolation were chosen such that an even number of elements would be needed to determine the functions. This permits the number of observed elements used from either side of the missing elements to be the same, thereby giving us a symmetric estimator. Where the requisite number of points to determine the functions were not available lower order polynomials and

rational functions were used. All missing points at boundaries were estimated by linear extrapolation.

The mean squared error (MSE) between the estimated portions of the reconstructed spectrogram and the corresponding regions of the original spectrogram is a measure of the accuracy of the reconstructed spectrogram. Representing the elements of the true uncorrupted spectrogram from which the incomplete spectrogram was derived as $S_u(t, k)$, the elements of the reconstructed spectrogram as $\hat{S}(t, k)$, and the number of missing elements in the spectrogram as N_{miss}, we define the MSE of reconstruction as

$$MSE(\hat{\boldsymbol{S}}) = \frac{\sum_{i=1}^{N}\sum_{k=1}^{K}\left|(\hat{S}(t, k) - \boldsymbol{S}_u(t, k))\right|^2}{N_{miss}} \tag{5.12}$$

Clearly, the greater the MSE, the greater the divergence between the reconstructed and uncorrupted spectrograms, and the lower the accuracy of the reconstruction.

The accuracy of the reconstructed spectrograms was measured in terms of the mean squared error (MSE) between the reconstructed spectrogram and the original uncorrupted spectrogram. The recognition accuracy obtained with the reconstructed spectrograms was measured to evaluate the effectiveness of the reconstruction procedures.

Figures 5.5 and 5.6 show the uncorrupted mel spectrogram of an utterance and the mel spectrogram when 50% of the elements in the picture are missing, respectively. Figure 5.7 shows the mel spectrogram when all the missing elements have been reconstructed using linear interpolation along frequency. Figure 5.8 shows a similar figure where all the missing elements have been reconstructed using linear interpolation along time. Figures 5.9 and 5.10 show the reconstructions obtained using polynomial interpolation along frequency using cubic polynomials (*i.e.* polynomials of order 3) and cubic polynomial interpolation along time respectively. Similarly Figures 5.11 and 5.12 show the reconstruction obtained using rational functions of order (1,2) for interpolation along frequency and time respectively.

We observe from Figures 5.5 through 5.12 that, in general, reconstructed spectrograms obtained by interpolation along time match the original spectrogram more closely that reconstructions obtained by interpolation along frequency. Furthermore, reconstruction by linear interpolation is seen to be better than

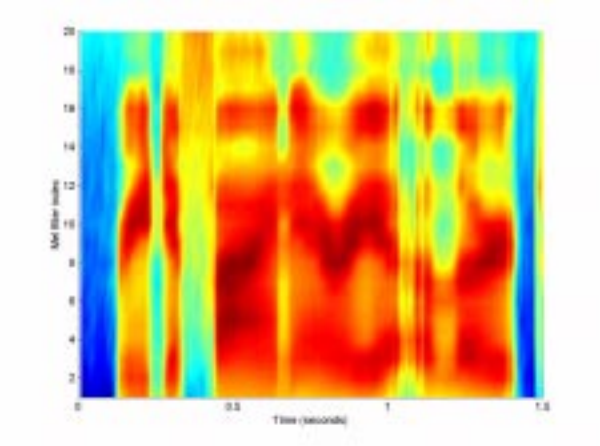


Figure 5.5 Mel spectrogram of an utterance.

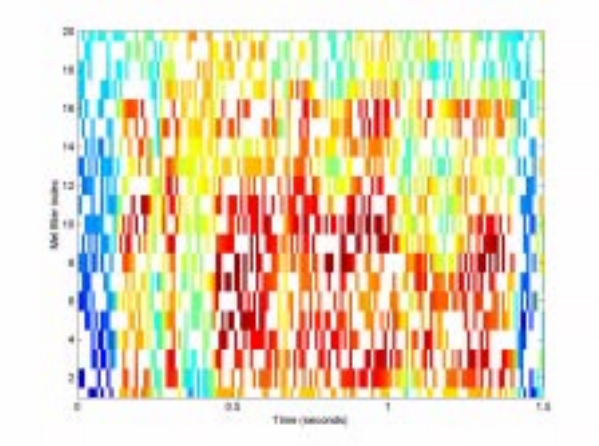


Figure 5.6 The same spectrogram when a randomly selected 50% of its elements have been deleted.

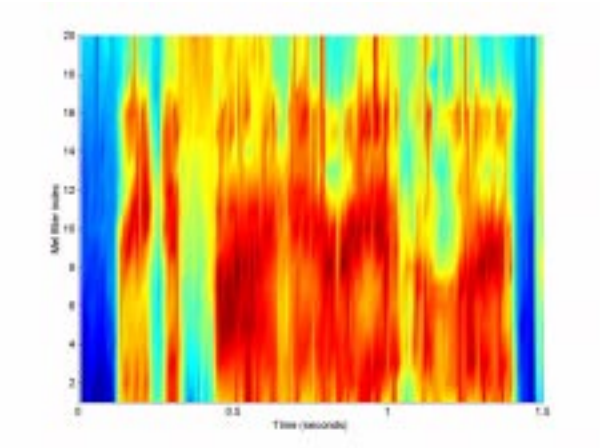


Figure 5.7 Spectrogram obtained by estimating the missing regions by linear interpolation across frequency

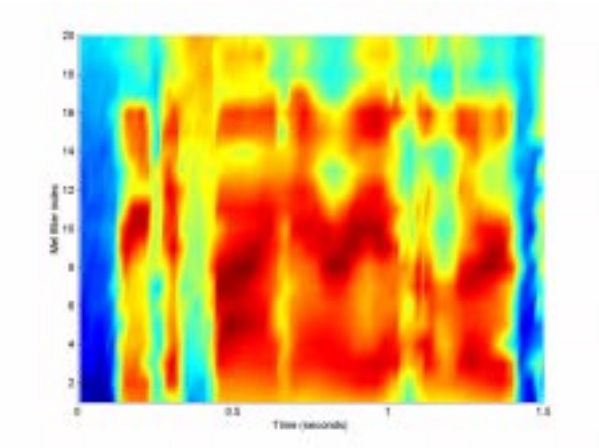


Figure 5.8 Spectrogram obtained by estimating the missing regions by linear interpolation across time.

polynomial or rational function interpolation in general and reconstruction obtained by linear interpolation along time matches the original spectrogram most closely, overall. Figure 5.13 below plots the mean square error between the reconstructed elements of the spectrograms and their actual values as a function of the fraction of elements that were missing from the spectrograms. We refer to this fraction as the *drop fraction* in the spectrogram. The MSE obtained using each of the reconstruction methods represented in Figures 5.7 through 5.12 is shown. This figure confirms the visual observation from the earlier set of figures that the lowest mean squared error overall is obtained with reconstruction by linear interpolation across time. We do note, however, that when the fraction of missing points is small, reconstruction using cubic polynomial interpolation along time results in the best mean squared error. However, as the fraction

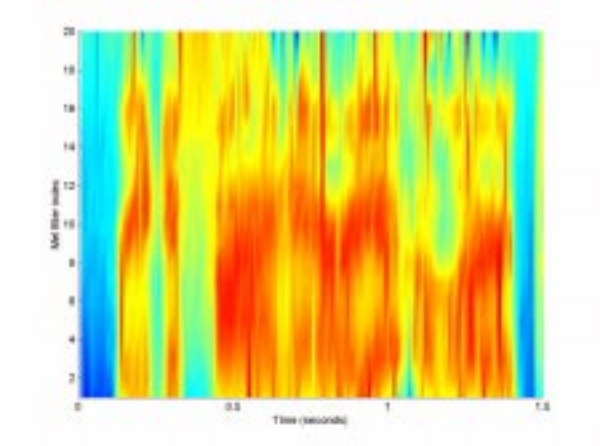


Figure 5.9 Spectrogram obtained by reconstructing missing regions by polynomial interpolation along frequency. Polynomials of order 3 were used when at least two observed elements were present on either side of the missing elements. When the number of available observed neighbors was lesser, lower order polynomials were used.

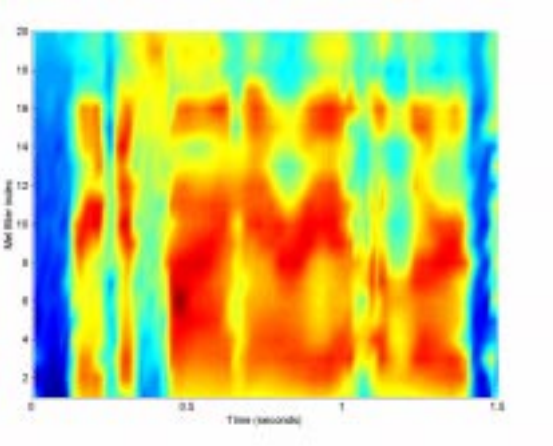


Figure 5.10 Spectrogram obtained by reconstruction missing regions by polynomial interpolation along time. Polynomials of order 3 were used where at least two observed elements were present on either side of the missing elements. Otherwise lower order polynomials were used.

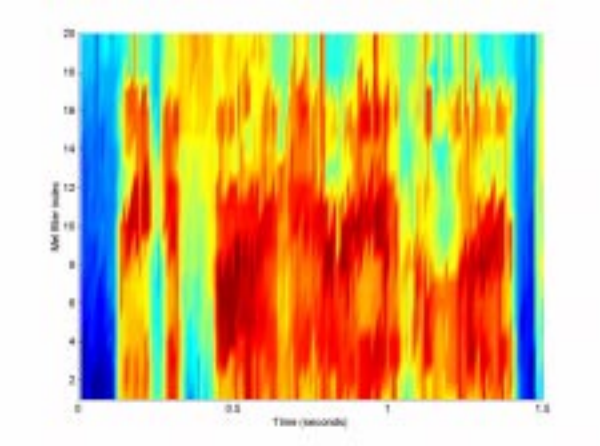


Figure 5.11 Spectrogram obtained by estimating missing regions by rational function interpolation along frequency. Rational functions of order (1,2) were used where at least two observed elements were present on either side of the missing elements. Otherwise rational functions of a lower order were used.

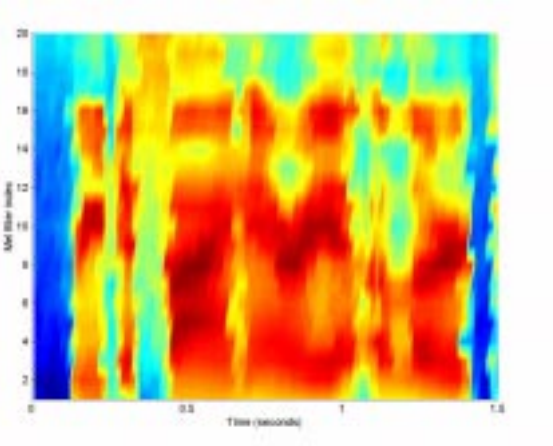


Figure 5.12 Spectrogram obtained by estimating missing regions by rational function interpolation along time. Rational functions of order (1,2) were used where possible. Otherwise, lower order rational functions were used.

of missing points increases and the mean distance from any missing point to the closest observed point increases, the reconstruction error with cubic polynomial interpolation increases faster than that of linear interpolation

Figure 5.14 plots the recognition accuracies obtained using the reconstructed mel spectrograms obtained using all the methods represented in Figure 5.13 against the fraction of elements missing in the

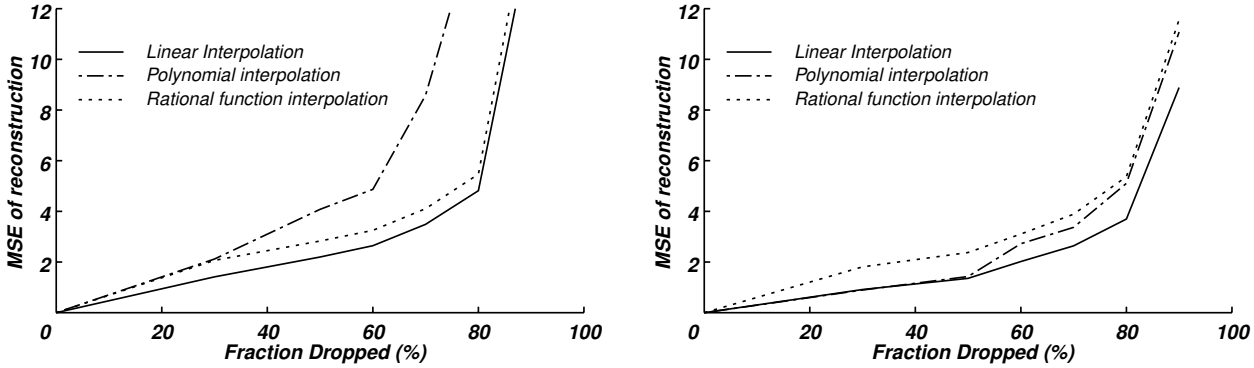


Figure 5.13 Mean Squared Error (MSE) of reconstruction for linear and non-linear interpolation, along frequency and time vs. fraction of elements missing in the incomplete spectrogram

Left Panel: MSE obtained with interpolation along frequency. Linear interpolation, polynomial interpolation with polynomials of order 3, and rational-function interpolation with rational functions of order (1,2) are represented

Right Panel: MSE obtained with interpolation along time. Linear interpolation, polynomial interpolation with polynomials of order 3, and rational function interpolation with rational functions of order (1,2) are represented

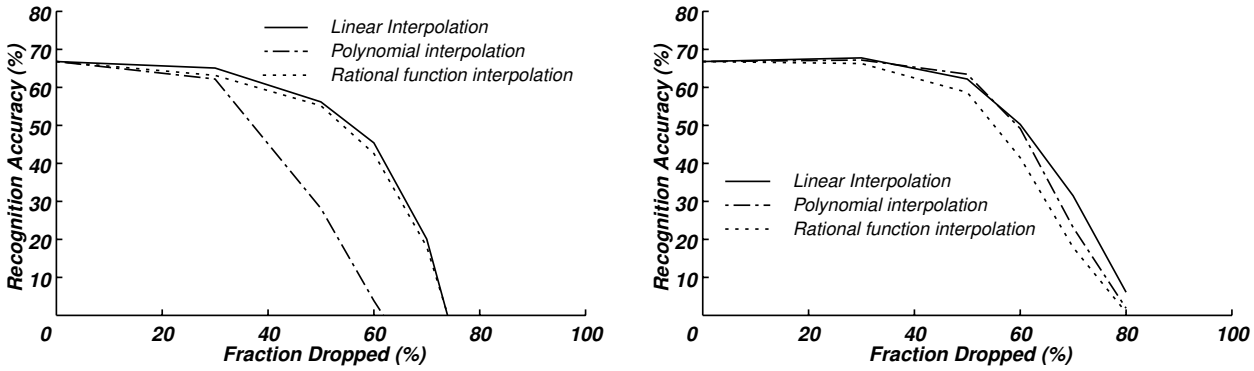


Figure 5.14 Recognition accuracy vs. drop fraction for spectrograms reconstructed by linear and non-linear interpolation along frequency and time.

Left Panel: Recognition accuracy obtained with reconstructed spectrograms where missing elements were estimated by interpolation along frequency. Linear interpolation, polynomial interpolation with polynomials of order 3, and rational-function interpolation with rational functions of order (1,2) are represented

Right Panel: Recognition accuracy obtained with reconstructed spectrograms where missing elements were estimated by interpolation along time. Linear interpolation, polynomial interpolation with polynomials of order 3, and rational-function interpolation with rational functions of order (1,2) are represented

picture. We observe that the trends are similar to those observed in Figure 5.13. Non-linear interpolation techniques result in poorer recognition accuracies than linear interpolation in general. Also, interpolation along frequency generally results in lower accuracies than interpolation along time. The best performance overall is achieved with linear interpolation along time.

5.2.5 Geometrical reconstruction methods: summary and conclusion

The recognition accuracies in Figure 5.14 show that even simple geometrical reconstruction methods such as linear interpolation based estimation of missing points can be quite effective in reconstructing spectrograms when random elements of the spectrogram are missing. Spectrograms reconstructed by linear interpolation along time show minimal loss in recognition accuracy when fully half the picture is missing. The best reconstruction, both in terms of MSE and recognition accuracy, is obtained by simple linear interpolation, and increasing the complexity or order of the functions used to estimate the missing regions results in no improvement in the reconstruction. One likely conclusion drawn from this is that the values of the elements in the spectrogram do not follow any specific pattern that can be captured by any single functional form. As a result the estimates obtained with more detailed models such as polynomials and rational functions are more likely to be erroneous than estimates obtained with simple first order functions.

Another noteworthy fact is that interpolation along time is generally more effective than interpolation along frequency. One of the reasons for this is that spectral vectors in the mel spectrograms used in these experiments have only 20 components. Consequently, observed elements frequently cannot be found on one side of missing elements, especially at high drop fractions, and these elements have to be reconstructed by extrapolation, rather than by interpolation. Extrapolation is known to result in poorer estimates than interpolation. Interpolation along time, on the other hand, does not face this problem since time slices of spectrograms have as many elements as the number of spectral vectors in the spectrogram. Another possible reason for the greater effectiveness of interpolation along time could be that spectrograms exhibit greater continuity along time, than along frequency.

All the methods mentioned in this section, are *local* reconstruction methods in that they reconstruct missing elements solely on the basis of the elements remaining in the picture. All the information used to reconstruct the missing points is obtained from the spectrogram itself, with no reference to any external sources of information. Such reconstruction methods have several drawbacks. First, when the fraction of missing elements is very high there might not be sufficient information remaining in the picture to reconstruct the missing elements properly. Second, if the observed elements in the spectrogram were to be distorted due to any reason such as due to noise, all missing elements reconstructed on the basis of these points alone would also be distorted similarly.

These shortcomings could be avoided if the reconstruction process were directed by other external information about the structure of speech spectrograms. This has the advantages of permitting better reconstruction when there is insufficient information in the damaged spectrogram as well as ensuring that the reconstructed spectrogram conforms to the notion of a *clean* spectrogram as represented by these external sources of knowledge. Some easily accessible sources of information are the large corpora of speech data that are readily available to train a speech recognition system. The distribution of the elements of spectrograms and the statistical relations between them can be learned from these corpora and used in the reconstruction.

In the following section we discuss reconstruction methods that utilize *vector statistics*, *i.e.* the distribution of the spectral vectors of clean speech.

5.3 Cluster-based reconstruction: statistical reconstruction using distributions of uncorrupted spectral vectors

In the methods described in this section we use the *vector statistics* of the spectral vectors for reconstruction of the complete spectrogram. These methods treat each spectral vector independently of every other vector in the spectrogram, *i.e.* they model the sequence of spectral vectors in the spectrogram as the output of an independent identically distributed (IID) random process. The statistical relations between components of different vectors are not modeled. The distribution of spectral vectors obtained under the IID assumption is used to condition the estimates of missing components.

The distribution of the spectral vectors of clean, uncorrupted speech is not known beforehand and has to be learned from a training corpus of uncorrupted speech. Since the precise form of the distribution of the spectral vectors is not known, a parametric form for the distribution must be assumed. The simplest and possibly the most commonly used representation for the distribution of speech vectors is the cluster-based representation. In a cluster-based representation spectral vectors are assumed to be segregated into a set of clusters. All vectors belonging to any cluster are further assumed to have a parametric distribution, which we refer to as the *cluster distribution*. Cluster-based representations therefore have two types of parameters:

1) The *a priori* probability that a random vector belongs to any of the clusters

2) The parameters of the cluster distributions

Let S represent an arbitrary spectral vector. Let c_k represent the *a priori* probability that a vector belongs to the k^{th} cluster, and let $g_k(S;\phi_k)$ represent the parametric distribution of vectors belonging to the k^{th} cluster. In a cluster-based representation the distribution of S is therefore modeled as

$$P(S) = \sum_{k=0}^{K} c_k g_k(S;\phi_k) \tag{5.13}$$

where K is the total number of clusters and ϕ_k represents the set of parameters of $g_k(S;\phi_k)$.

If we assume that within any cluster vectors are distributed according to a Gaussian distribution, then the overall distribution of the data set can be represented as a mixture of multivariate Gaussians

$$P(S) = \sum_{k=1}^{K} c_k (2\pi|\Theta_k|)^{-\frac{d}{2}} e^{-\frac{1}{2}(S-\mu_k)^T \Theta_k^{-1}(S-\mu_k)} \tag{5.14}$$

where d is the dimensionality of the vectors. μ_k and Θ_k are the mean vector and covariance matrix, respectively, of the Gaussian distribution of the vectors belonging to the k^{th} cluster. The parameters of the distribution represented by Equation (5.14), namely the values of c_k, μ_k, and Θ_k for all the clusters must learned from the training corpus. In order to learn these parameters the vectors of the training corpus can be clustered into the desired number of clusters using techniques such as k-means clustering [McQueen 1967], the LBG algorithm [Linde 1980] etc., and the distributions of the individual clusters can be obtained once the clusters are obtained. More consistent parameter estimates are obtained using maximum likelihood (ML) methods [Mclachlan 1988].

While the distribution represented by Equation (5.13) is more generic and therefore better able to model a wider class of distributions, the Gaussian mixture distribution given by Equation (5.14) has several advantages:

- Most distributions of infinite extent (*i.e.* distributions which are non-zero everywhere except at infinity) can be modeled by mixtures of Gaussians with arbitrary precision [Mclachlan 1988].
- Gaussian densities are completely defined by their first and second order moments. As a result, we only need to know the first and second order moments of the individual Gaussians comprising the

mixture to completely describe the density. The estimation errors inherent in the estimation of higher order moments needed by other density functions are thereby avoided.

- The parameters of a mixture Gaussian distribution can be easily estimated using the EM algorithm [Dempster 1977]. It is also very easy to derive EM type solutions for most other ML estimation problems where the random variables involved have mixture Gaussian distributions.
- Most methods of estimating missing elements in a spectrogram that are discussed in this thesis involve *maximum a posteriori* (MAP) estimation of the missing elements. MAP estimation is very simple when the underlying distribution of the data is Gaussian.

In light of these advantages, we model the distribution of spectral vectors as a mixture Gaussian for the missing feature methods described in this section.

The cluster-based representation leads to a very simple solution for the estimation of missing elements of the spectrogram. Given any spectral vector $\boldsymbol{S}(t)$ with missing components $\boldsymbol{S}_m(t)$ we only have to identify the cluster that the vector $\boldsymbol{S}(t)$ belongs to and use the distribution of the vectors belonging to that cluster to obtain an estimate for $\boldsymbol{S}_m(t)$. We refer to the cluster that any vector belongs to as the *cluster membership* of that vector. This cluster membership localizes the region that the vector $\boldsymbol{S}(t)$ can lie in, and thereby the range of values that the missing components of the vector can take. Thereafter, the distribution of the cluster can be used to obtain a statistical best guess for the missing components of vector within the localized region.

As discussed in Chapter 2, several statistical methods exist to estimate the missing components of a data set given the distribution of the complete data. While all of these methods can be used to estimate the missing components of a vector, MAP estimation is arguably the best motivated procedure among them. MAP estimation also provides a tractable framework for incorporating additional constraints where available, in the estimation.

We therefore use MAP estimation to estimate the missing components of vectors. Once the cluster membership of a vector is identified the missing components of the vector are obtained as MAP estimates based on the distribution of the identified cluster, conditioned on the observed components of the vector.

Figure 5.15 shows a schematic representation of cluster-based estimation of missing elements of a spec-

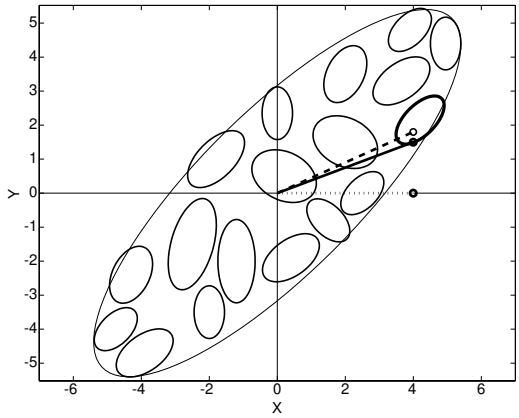


Figure 5.15 Schematic representation of cluster-based reconstruction. The big ellipse represents the outline of the distribution of a set of two dimensional vectors. The data has been segregated into a number of small clusters. The smaller ellipses represent the cross section of the Gaussian distributions of these individual clusters. The solid line represents a complete vector. In the observed data, the Y component of this vector is missing, and only the X component, represented by the dotted line along the X axis, is observed. The cluster-based reconstruction method identifies the thick ellipse as the cluster that the complete vector belongs to, and uses the distribution of that cluster to obtain an MAP estimate for the missing Y component, and thereby the complete vector. The estimate complete vector is represented by the dashed line.

tral vector.

In order to obtain a complete cluster-based reconstruction method for incomplete spectrograms the following issues also have to be resolved:

- The number of clusters to use in the cluster-based representation in order to obtain optimal reconstruction is not known.
- The manner in which the cluster membership of any vector is determined is not known. The fact that some components of the vector may be missing makes cluster membership identification a difficult problem.

Depending on the particular solution for each of the above problems the reconstructed spectrogram and the recognition accuracy obtained using the reconstructed spectrogram can vary. In the following sections we address these issues and describe three cluster-based reconstruction methods:

- single-cluster-based reconstruction,
- multiple-cluster-based reconstruction with marginalization-based cluster identification,
- multiple-cluster-based reconstruction with time-interpolation-based cluster identification, and

These methods vary only in the number of clusters used to represent the distribution and the manner in which clusters are identified.

5.3.1 Single cluster based reconstruction: modeling the distribution with a single cluster

The simplest cluster-based representation of the distribution of a data set is where all data are assumed to belong to a single cluster. The distribution of the cluster is simply the global distribution of the data. In single cluster-based estimation therefore, all spectral vectors are assumed to belong to a single cluster. We assume the distribution of the cluster to be a Gaussian. The mean vector and covariance matrix of the Gaussian can be directly obtained from a training corpus of clean speech spectrograms.

Since there is only one cluster that any vector can belong to no further cluster membership identification is necessary during the estimation. The MAP estimate of the missing components of any vector is based on the cluster distribution of this single cluster, *i.e.* the global distribution of the data, and conditioned on the observed elements in that vector. We denote the missing components of the t^{th} spectral vector $\boldsymbol{S}(t)$, by the vector $\boldsymbol{S}_m(t)$ and its observed components by the vector $\boldsymbol{S}_o(t)$ such that $\boldsymbol{S}(t) = A_t[\boldsymbol{S}_o(t), \boldsymbol{S}_m(t)]$, where A_t is the permutation matrix that rearranges the components of $\boldsymbol{S}_o(t)$ and $\boldsymbol{S}_m(t)$ to obtain $\boldsymbol{S}_o(t)$. Note that A_t is specific to the t^{th} vector since the precise set of components that are missing from any spectral vector can vary from vector to vector. The estimated value of the vector of missing components, $\hat{\boldsymbol{S}}_m(t)$ and the corresponding estimate of the complete vector $\hat{\boldsymbol{S}}(t)$ are now obtained in the manner described in Section 2.5.4 as

$$\begin{aligned} \hat{\boldsymbol{S}}_m(t) &= \mu_m + \Theta_{mo}\Theta_{oo}^{-1}(\boldsymbol{S}_o(t) - \mu_o) \\ \hat{\boldsymbol{S}}(t) &= A_t[\boldsymbol{S}_o(t), \hat{\boldsymbol{S}}_m(t)] \end{aligned} \tag{5.15}$$

where μ_o and Θ_{oo} are the mean and covariance of the observed components (*i.e.* the mean vector and the covariance matrix of the marginal distribution of $\boldsymbol{S}_o(t)$), μ_m is the mean of the missing components (*i.e.* the mean vector of the marginal distribution of $\boldsymbol{S}_m(t)$), and Θ_{mo} is the cross covariance between the observed components and the missing components, *i.e.* $\Theta_{mo} = E[(\boldsymbol{S}_m(t) - \mu_m)(\boldsymbol{S}_o(t) - \mu_o)^T]$, where $E[\ ]$ refers to the expectation operator [Papoulis 1991]. μ_m, μ_o, Θ_{oo} and Θ_{mo} are all easily obtained

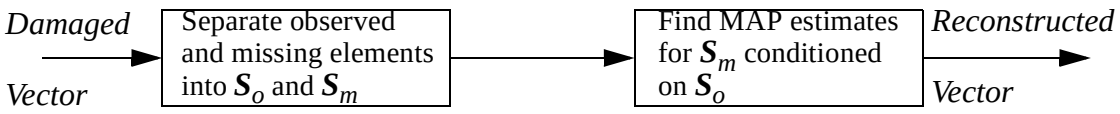


Figure 5.16 Block diagram explaining the procedure for estimating the missing components of a vector. The complete spectrogram is obtained by reconstructing the missing elements of each vector in the spectrogram using this procedure.

from the parameters of the cluster distribution as explained in Section 2.5.4. Figure 5.16 represents the procedure of reconstructing the damaged components of a vector as a block diagram.

The complete spectrogram $\hat{\boldsymbol{S}}$ is reconstructed by reconstructing each incomplete spectral vector in the spectrogram using Equation (5.15) as

$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{S}}(1), \hat{\boldsymbol{S}}(2), \hat{\boldsymbol{S}}(3), \ldots, \hat{\boldsymbol{S}}(N) \tag{5.16}$$

Recognition is now performed using the estimated complete spectrogram.

For brevity we refer to single-cluster-based reconstruction as *single cluster reconstruction* in future references to the method.

5.3.1.1 Experimental results with a single cluster based reconstruction

Single cluster reconstruction was evaluated with the random-drop paradigm. Experiments were run using the RM database and the same setup used to evaluate geometrical reconstruction methods in Section 5.2.4.

Figure 5.17 shows the same incomplete spectrogram shown in Figure 5.6. Figure 5.18 shows the reconstructed spectrogram obtained when the missing regions of this spectrogram have been reconstructed using single-cluster-based estimation.

Figure 5.19 shows the mean squared error of spectrograms reconstructed with single cluster reconstruction, as a function of the fraction of elements missing in the spectrogram (*i.e.* the drop fraction). As seen from this figure, the MSE of the reconstructed spectrogram increases as the drop fraction increases, *i.e.* the accuracy of the reconstruction decreases as the drop fraction increases.

Figure 5.20 shows the recognition accuracy obtained using spectrograms that have been reconstructed

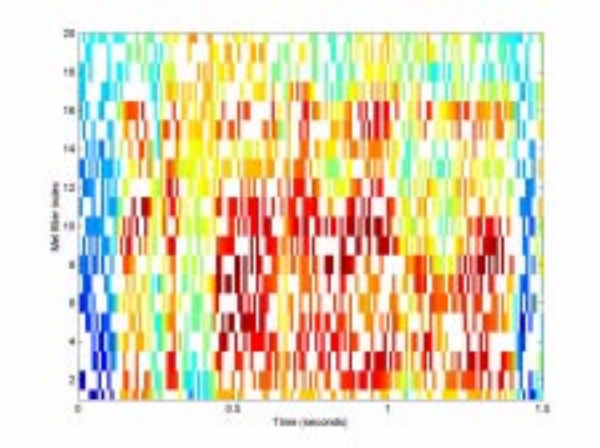


Figure 5.17 Spectrogram of an utterance of speech, where 50% of the elements have been randomly deleted

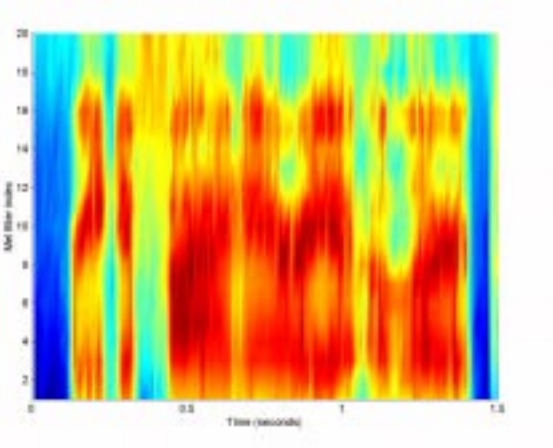


Figure 5.18 The same spectrogram where the missing elements have been reconstructed by single cluster reconstruction.

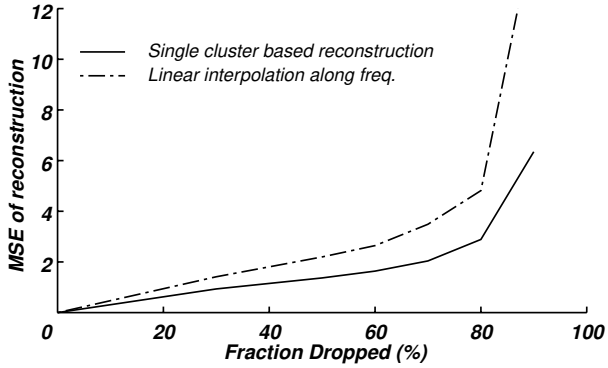


Figure 5.19 Mean squared error between the estimated regions of the reconstructed spectrogram obtained using single cluster reconstruction and the corresponding regions of the original uncorrupted spectrogram, as a function of the drop fraction. The MSE obtained with linear interpolation along frequency is also shown for comparison.

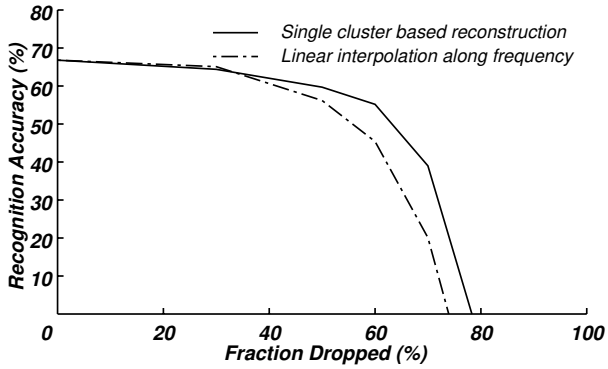


Figure 5.20 Word recognition accuracy obtained with reconstructed spectrogram as a function of the drop fraction. The recognition accuracy obtained with linear interpolation along frequency is also shown for comparison.

using single cluster reconstruction, as a function of the drop fraction.

The recognition accuracy is seen to decrease as the drop fraction increases. This correlates well with the fact that the mean squared error of the reconstructed spectrograms increases with increasing drop fraction.

5.3.1.2 Discussion and analysis of experimental results

We note from the earlier Section that the MSE of reconstruction, and consequently the recognition accuracy, degrade as the drop fraction increases. This happens due to several reasons.

It can be shown the expected MSE of reconstruction for $\boldsymbol{S}_m(t)$, the missing components of the t^{th} vector $\boldsymbol{S}(t)$ is given by [Appendix A]

$$MSE(\hat{\boldsymbol{S}}_m(t)) = trace(\Theta_{mm}) - trace(\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om}) \tag{5.17}$$

where Θ_{mm} is the covariance matrix of $\boldsymbol{S}_m(t)$, Θ_{oo} the covariance matrix $\boldsymbol{S}_o(t)$ and Θ_{mo} is the cross covariance between $\boldsymbol{S}_m(t)$ and $\boldsymbol{S}_o(t)$. It can also be shown that the MSE of reconstruction increases as the number of missing elements in $\boldsymbol{S}(t)$, *i.e.* the number of elements in $\boldsymbol{S}_m(t)$, increases [Appendix A]. As the drop fraction increases, the average number of elements in $\boldsymbol{S}_m(t)$ does increase. As a result, the MSE of reconstruction increases with increasing drop rate.

Another factor that affects the estimation of the missing regions is the actual covariance between the missing components and the observed components of the vector. As the drop fraction increases, the average distance between a frequency component and the nearest observed frequency component increases [Appendix A]. Figure 5.21 plots the mean distance between any missing frequency component and the closest observed frequency component as a function of the drop fraction. Figure 5.22 shows how the average relative covariance between two frequency components varies as the distance between the components increases. We observe from these figures that as the drop fraction increases, the cross-covariance between the missing component and the observed components decreases. It is easy to see in Equation (5.17) that as the individual elements in the cross-covariance matrix Θ_{mo} decrease, $trace(\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om})$ decreases [Appendix A] and the MSE of reconstruction increases. Thus, another reason for the increase in MSE with increasing drop fraction is the corresponding decrease in the covariances between the missing components

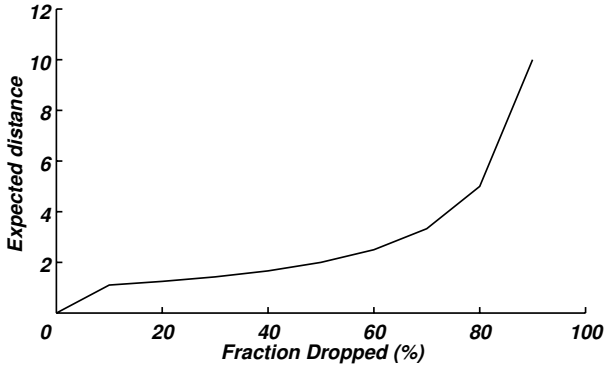


Figure 5.21 Mean distance between a missing component and its closest observed neighbor as a function of drop rate.

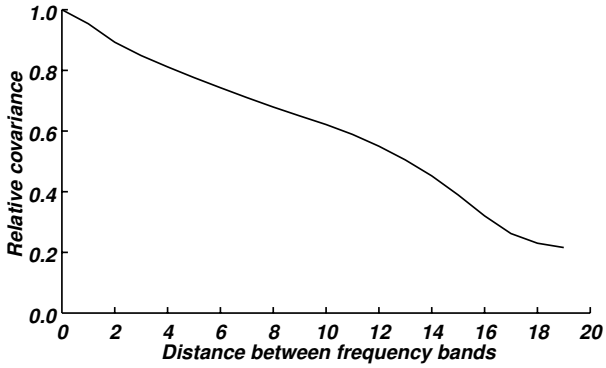


Figure 5.22 Relative covariance between two frequency components as a function of the distance between them.

and the observed components of the spectral vector.

The recognition accuracy obtained with the reconstructed spectrogram is clearly related to the accuracy of the reconstruction. The error in reconstruction can be viewed as noise added to the corrected spectral values. The greater the error, the greater the noise. At higher drop fractions the higher MSE of reconstruction corresponds to noisier spectrograms resulting in poorer recognition accuracy.

In general we note that even this very simple clustering method using only a single cluster results in reasonably good reconstructions of the spectrogram when the fraction of missing elements is less than 50%. The difference in recognition performance between the reconstructed spectrograms and the original spectrograms is not appreciable at these drop fractions (*i.e.* fractions of missing data). Comparison of reconstructed spectrograms obtained by linear interpolation across frequency (Figure 5.7), and by single cluster reconstruction (Figure 5.18), show both reconstructed spectrograms to be similar in nature. This is because both reconstruction by interpolation across frequency and singe cluster reconstruction are based on the assumption that the energy in adjacent frequency bands varies continuously and smoothly. However, cluster reconstruction uses additional statistical information about the statistical correlations between frequency bands. Comparison of the MSE of reconstruction and the recognition accuracy obtained using reconstructed spectrograms for the two reconstruction methods (Figures 5.13, 5.14, 5.19, and 5.20) shows us that the additional statistical information used in the cluster-based reconstruction methods does indeed result in better reconstruction.

5.3.2 Multiple cluster based reconstruction

So far we have discussed a spectrogram reconstruction method in which we modeled the distribution of spectral vectors with a single cluster. A more detailed representation would use multiple clusters to model the distribution of spectral vectors. The means and variances of the distributions of the individual clusters, and the proportion of vectors belonging to each of the clusters, *i.e.* the *a priori* probability of the individual clusters, can be learned from a training corpus of clean spectrograms using the EM algorithm [Dempster 1977].

When the distribution is represented by multiple clusters the procedure for estimating the missing portions of an incomplete vector has two steps. In the first step the *cluster membershi*p of the vector, *i.e.* the cluster that the vector belongs to, is identified. Once the cluster membership of the vector is established the distribution of that cluster is used to obtain MAP estimates for the missing components of that vector. Figure 5.23 represents the entire procedure for estimating the missing regions of an incomplete vector as a block diagram.

A vector is said to belong to the cluster that is most likely to have generated it. Since all cluster distributions are assumed to be Gaussian, the cluster membership $k_{S(t)}$ of the vector $\boldsymbol{S}(t)$ is defined as

$$k_{S(t)} = \arg\max_k\{P(k|\boldsymbol{S}(t))\} = \arg\max_k\{P(\boldsymbol{S}(t)|k)P(k)\}$$

$$k_{S(t)} = \arg\max_k\left\{c_k|\Theta_k|^{-\frac{1}{2}}\exp(-0.5(\boldsymbol{S}(t)-\mu_k)^T\Theta_k^{-1}(\boldsymbol{S}(t)-\mu_k))\right\} \tag{5.18}$$

where μ_k and Θ_k are the mean vector and the covariance matrix respectively of the k^{th} cluster, and c_k is the *a priori* probability that any vector belongs to the k^{th} cluster. This treats the identification of cluster

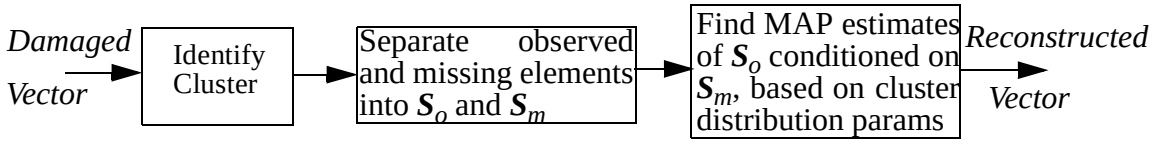


Figure 5.23 Block diagram showing estimation of missing elements in a spectral vector using a multiple-cluster based representation of the distribution of spectral vectors.

membership as a classification problem, where the clusters are the classes. Equation (5.19) defines the optimal bayesian classifier which determines which cluster a vector belongs to. The definition of cluster membership also has a geometrical interpretation. It can be restated in terms of distance if we define distance as the negative of the log-likelihood of the vector:

$$k_{S(t)} = \arg\min_k \{0.5(\boldsymbol{S}(t) - \mu_k)^T \Theta_k^{-1} (\boldsymbol{S}(t) - \mu_k) - \log(c_k) + 0.5\log(|\Theta_k|)\} \tag{5.19}$$

Using this definition of distance, the cluster membership is defined as the cluster that the vector is closest to. Equations (5.18) and (5.19) implicitly define the boundaries between the various clusters. Thus, any vector that falls within the boundaries of a particular cluster is said to belong to that cluster.

Once the cluster membership of the vector is identified, the distribution of that cluster is used to obtain MAP estimates for the missing components of that vector. As before, separating the missing and observed components of $\boldsymbol{S}(t)$ into $\boldsymbol{S}_m(t)$ and $\boldsymbol{S}_o(t)$ such that $\boldsymbol{S}(t) = A_t[\boldsymbol{S}_o(t), \boldsymbol{S}_m(t)]$, the estimated value of $\hat{\boldsymbol{S}}_m(t)$ and the corresponding complete vector $\hat{\boldsymbol{S}}(t)$ are now obtained as

$$\begin{aligned} \hat{\boldsymbol{S}}_m(t) &= \mu_{k,m} + \Theta_{k,mo}\Theta_{k,oo}^{-1}(\boldsymbol{S}_o(t) - \mu_{k,o}) \\ \hat{\boldsymbol{S}}(t) &= A_t[\boldsymbol{S}_o(t), \hat{\boldsymbol{S}}_m(t)] \end{aligned} \tag{5.20}$$

where k is the cluster membership of $\boldsymbol{S}(t)$ and $\mu_{k,o}$ and $\Theta_{k,oo}$ are the mean and covariance of the observed components, $\mu_{k,m}$ is the mean of the missing components given that $\boldsymbol{S}(t)$ belongs to the k^{th} cluster. $\Theta_{k,mo}$ is the cross covariance between $\boldsymbol{S}_m(t)$ and $\boldsymbol{S}_o(t)$, given that $\boldsymbol{S}(t)$ belongs to the k^{th} cluster. *i.e.*,

$$\Theta_{k,mo} = E[(\boldsymbol{S}_m(t) - \mu_{k,m})(\boldsymbol{S}_o(t) - \mu_{k,o})^T \,|\, cluster = k] \tag{5.21}$$

The means and covariances, $\mu_{k,m}$, $\mu_{k,o}$, $\Theta_{k,oo}$ and $\Theta_{k,mo}$ are all obtained from the parameters of the cluster distribution of the k^{th} cluster. The estimated complete spectrogram $\hat{\boldsymbol{S}}$ is obtained by reconstructing the missing components of each spectral vector in the spectrogram using Equation (5.20) as

$$\hat{\boldsymbol{S}} = \hat{\boldsymbol{S}}(1), \hat{\boldsymbol{S}}(2), \hat{\boldsymbol{S}}(3), \ldots, \hat{\boldsymbol{S}}(N) \tag{5.22}$$

Recognition is now performed using the estimated complete spectrogram.

An important parameter in a multiple-cluster-based representation of the distribution of a data set is the number of clusters used in the representation. We refer to this number as the *codebook size* of the representation. As the codebook size increases the representation becomes more detailed and the size of the individual clusters decreases. Thus, as the codebook size increases the cluster membership of a vector increasingly localizes its position. Therefore the error in the estimates of the missing components can be expected to decrease with increasing codebook size if the cluster membership of every vector is always correctly known.

When complete spectral vectors are available, cluster membership of vectors can be directly obtained by evaluating Equation (5.19). However, when dealing with incomplete spectrograms, several components of the spectral vector could be missing. Direct computation of Equation (5.19) is not possible with incomplete vectors. From a geometrical perspective, it is not possible to determine whether a vector lies within the boundaries of the cluster when some of the components of the vector are not known. In this situation cluster membership has to be estimated using one of the following solutions:

- Identify cluster membership based only on the observed components
- Pre-estimate the missing components in some manner, and then use the complete vector to identify cluster membership

Since the cluster membership found by these methods is only an estimate of the true cluster membership it is likely to be erroneous. Error in cluster-membership identification results in the distribution of the wrong cluster being used for estimating the missing components of the spectral vector resulting in increased MSE in the reconstruction. Furthermore, the error in estimating cluster membership with incomplete vectors can be expected to increase as the codebook size increases and the clusters become more localized. This resulting increase in MSE due to the increased error in cluster membership identification is likely to compensate for some or all of the improvement in the reconstruction accuracy that would be expected with increasing codebook size had cluster membership been perfectly known.

In the following section we investigate the ideal situation where the correct cluster membership of all vectors in the spectrogram is known *a priori* for the estimation of missing elements, and we also evaluate the effect of codebook size in this ideal situation. In subsequent sections we address the problem of estimating cluster membership of incomplete vectors. We present two multiple-cluster-based reconstruction

methods where we estimate cluster membership of vectors with incomplete vectors and reconstruct vectors based on the estimated cluster membership.

5.3.3 Oracle experiments with perfect knowledge of cluster membership

In an ideal situation the cluster membership of every vector in the incomplete spectrogram would be known *a-priori*. The reconstruction obtained under this ideal condition can be considered to be an upper bound to the performance of cluster-based reconstruction methods. We attempt to estimate this upper bound experimentally with an “oracle” experiment where the correct cluster membership of the vectors is given beforehand.

For the oracle experiment random elements of the spectrogram were dropped using the random-drop paradigm and reconstructed as described in Section 5.3.2. The cluster membership of each of the vectors was determined using the corresponding vector from the original, complete, spectrogram. MAP estimates of the missing components of vectors were estimated using the distribution of the correct cluster.

We refer to this procedure where incomplete spectral vectors are reconstructed using oracle knowledge of cluster membership as *cluster oracle reconstruction*.

Figure 5.24 shows the mean squared error of spectrograms reconstructed by cluster oracle reconstruction with cluster-based representations of different codebook sizes, as a function of the drop fraction obtained. Each line in the figure plots the MSE of reconstruction obtained using a cluster-based representation of a particular size. The MSE of reconstruction when the codebook size is one is identical to that obtained with single cluster reconstruction (Figure 5.19) since with only a single cluster there is no identification of cluster membership necessary. As the codebook size increases the MSE of reconstruction at any drop fraction is observed to decrease monotonically as predicted in Section 5.3.2. Also, for any codebook size the MSE increases with increasing drop fraction as was observed in single cluster reconstruction.

Figures 5.25 shows an example spectrogram with 70% of its elements missing and the reconstructed spectrogram obtained with oracle knowledge of cluster membership with cluster-based representations of increasing codebook sizes. We see from the pictures that the reconstruction follows the same pattern as the MSE - the reconstructed spectrogram resembles the original increasingly with increasing codebook size.

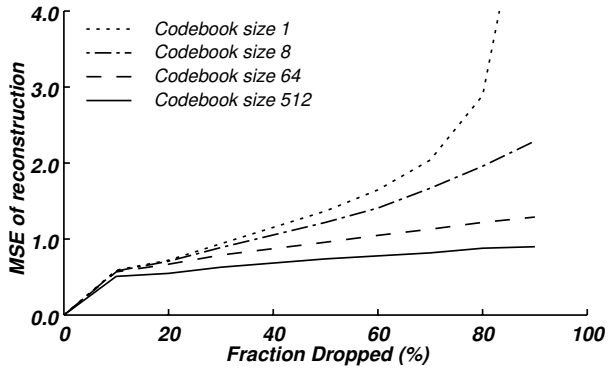


Figure 5.24 Mean squared error of the reconstructed spectrogram as a function of drop rate for various codebook sizes. Each line in the figure plots the MSE of reconstruction for a particular codebook size.

As explained in Section 5.3.1.2, spectrograms with larger MSE of reconstruction can be expected to result in lower recognition accuracy than spectrograms with lower MSE. Therefore, the recognition accuracy obtained with the reconstructed spectrograms can be expected to reflect the trends of the MSE of reconstruction. It can be expected that the recognition accuracy obtained with cluster oracle reconstructed spectrograms will increase with increasing codebook size at any drop fraction, and that it will decrease with increasing drop fraction for any codebook size. Figure 5.26 shows the recognition accuracy obtained with cluster oracle reconstructed spectrograms as a function of drop fraction, for cluster-based representations of various codebook sizes. The trends seen in this figure are exactly as expected. The recognition accuracy obtained with a codebook size of one is identical to that obtained with single cluster reconstruction (Figure 5.20). As the codebook size increases the recognition accuracy at any drop fraction improves monotonically. For codebook size 512 the reconstructed spectrogram obtained when 90% of the original spectrogram is missing results in almost the same recognition accuracy as the uncorrupted spectrogram (0% drop fraction). Also, for all codebook sizes recognition accuracy degrades with increasing drop fraction.

Figures 5.24 through 5.26 above indicate that good reconstruction and very high recognition accuracies are possible, in principle, using cluster-based reconstruction. However, the actual performances seen in these figures are only upper bounds and, indeed, may be unachieveable. In a real situation the cluster membership of any vector would not be known *a priori* and would have to estimated. As the codebook size increases, the error in cluster membership identification can also be expected to increase. These would

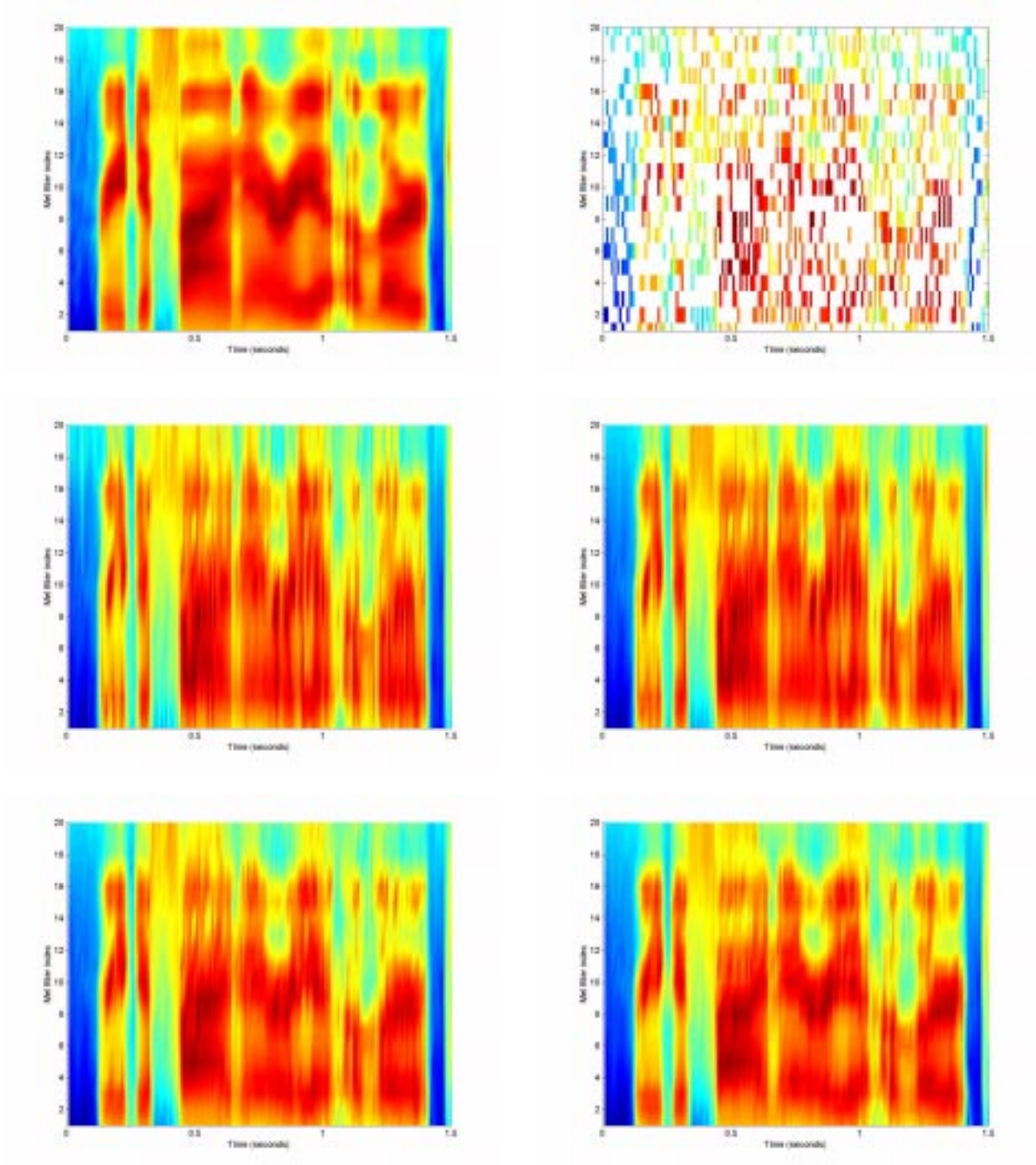


Figure 5.25 Examples of reconstructed spectrogram with oracle knowledge of cluster membership

Panel 1: Original spectrogram

Panel 2: Spectrogram with 70% of its elements randomly deleted

Panel 3: Spectrogram reconstructed with cluster based representation of codebook size 1

Panel 4: Spectrogram reconstructed with codebook size 8

Panel 5: Spectrogram reconstructed with codebook size 64

Panel 6: Spectrogram reconstructed with codebook size 512

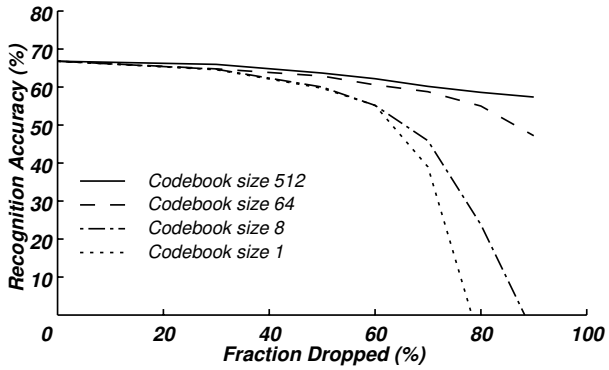


Figure 5.26 Recognition accuracy obtained with spectrograms reconstructed with oracle knowledge of cluster membership, as a function of drop fraction. Recognition accuracies are plotted for the reconstructed spectrograms obtained for several codebook sizes

increase the MSE in reconstruction and reduce the recognition accuracy.

5.3.4 Cluster Marginal Reconstruction: Identifying cluster membership based on observed components alone

Consider an incomplete spectral vector $\boldsymbol{S}(t)$ with missing components $\boldsymbol{S}_m(t)$ and observed components $\boldsymbol{S}_o(t)$. Equation (5.18) can now be restated for $\boldsymbol{S}(t)$ as

$$k_{S(t)} = \arg\max_k \{P(k|\boldsymbol{S}_o(t), \boldsymbol{S}_m(t))\} = \arg\max_k \{P(\boldsymbol{S}_o(t), \boldsymbol{S}_m(t)|k)P(k)\} \tag{5.23}$$

Since the value of $\boldsymbol{S}_m(t)$ is unknown this cannot be evaluated and the correct cluster membership of $\boldsymbol{S}(t)$ cannot be obtained directly. One solution to this problem is to attempt to identify the cluster membership of the vector based on the observed components of the vector alone.

$$\hat{k}_{S(t)} = \arg\max_k \{P(k|\boldsymbol{S}_o(t))\} = \arg\max_k \{P(\boldsymbol{S}_o(t)|k)P(k)\} \tag{5.24}$$

The cluster distributions are defined on the entire spectral vector $\boldsymbol{S}(t)$. Therefore, to obtain the distribution of the observed parameters we would have to integrate the missing components out of the distributions, *i.e.* by marginalization.

$$P(\boldsymbol{S}_o(t)|k) = \int_{-\infty}^{\infty} P(\boldsymbol{S}_o(t), \boldsymbol{S}_m(t)|k) d\boldsymbol{S}_m(t) = \int_{-\infty}^{\infty} P(\boldsymbol{S}(t)|k) d\boldsymbol{S}_m(t) \tag{5.25}$$

Cluster membership therefore would be estimated as

$$\hat{k}_{S(t)} = \arg\max_k \left\{ P(k) \int_{-\infty}^{\infty} P(\boldsymbol{S}(t)|k) d\boldsymbol{S}_m(t) \right\} \quad (5.26)$$

This method of estimating cluster membership with incomplete data is very similar in principle to marginalization based classification with incomplete data (Section 4.3). As in the case of cluster-membership identification with complete vectors, Equation (5.24) can be expressed in terms of distance, which is defined as the negative of the log likelihood of the observed components of the vector

$$\hat{k}_{S(t)} = \arg\min_k \{0.5(\boldsymbol{S}_o(t) - \mu_{k,o})^T \Theta_{k,oo}^{-1} (\boldsymbol{S}_o(t) - \mu_{k,o}) - \log(c_k) + 0.5\log(|\Theta_{k,oo}|)\} \quad (5.27)$$

where $\mu_{k,o}$ and $\Theta_{k,oo}$ are the mean and covariance of the observed components, given that the vector belongs to the k^{th} cluster.

The cluster membership estimated by Equation (5.27) is likely to be erroneous since the contribution of the missing components to the likelihoods of clusters is not being considered. As the fraction of elements missing from the vector increases Equation (5.27) is computed on fewer and fewer components and the estimated cluster membership becomes increasingly erroneous. In the limit where the entire vector is missing it is not possible to identify the cluster at all. In this situation we arbitrarily select, for the totally corrupt vector, the estimated cluster identity of the closest vector that is not completely corrupted.

Once the cluster membership of a vector is estimated, the distribution of the estimated cluster is used to estimate the missing components in the vector, and thereby the complete vector, using Equation (5.20).

We refer to this procedure of cluster membership estimation by marginalization and reconstruction of vectors with clusters so identified as *cluster marginal reconstruction.* The nomenclature is indicative of the fact that cluster-based reconstruction is being used, and that cluster membership has been identified by marginalization.

5.3.4.1 Experimental evaluation

Cluster marginal reconstruction was evaluated using the random-drop paradigm and the same experimental setup used in Section 5.3.3. In multiple-cluster-based reconstruction an additional factor affecting

reconstruction are the errors in cluster membership identification. Figure 5.27 plots the percentage of clusters that are wrongly identified as a function of the fraction of the elements missing in the incomplete spectrogram, for various codebook sizes. The percentage of wrongly identified clusters varies approximately linearly with the fraction of missing elements. We also observe that the fraction of wrongly identified clusters also increases as the codebook size increases, as expected. We noted in Section 5.3.3 that when cluster membership of vectors is perfectly known the MSE of estimation decreases monotonically with increasing codebook size. However, when cluster member is not known the increased error in cluster-membership identification with increasing codebook size introduces errors in the estimation that are likely to compensate for some, or all of the improvement obtained due to increased codebook size. Figure 5.28 shows the MSE in reconstruction as a function of the fraction of missing elements for various codebooks sizes. Note that the MSE obtained with codebook size 1 is the same as that obtained with cluster oracle reconstruction since there is no identification of cluster membership needed. Increasing the codebook size not improve the MSE significantly with increasing codebook size confirming our hypothesis that the increased error in cluster identification compensates for the improved reconstruction with increasing codebook size.

Figure 5.29 shows the reconstructed spectrogram obtained with different codebook sizes when the incomplete spectrogram has 70% of its elements missing. We observe that there is no appreciable visual difference between the reconstructed spectrograms that could be attributed to codebook size.

It is logical to conclude that since increasing codebook size does not improve the MSE of reconstruction it will not improve the recognition accuracy either. This hypothesis in confirmed by Figure 5.30, which shows the recognition performance obtained with the reconstructed spectrograms for various code-

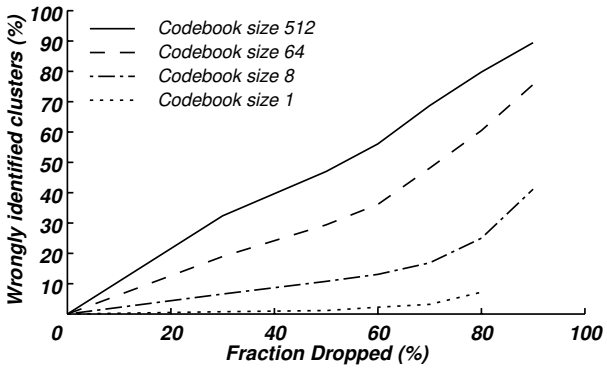


Figure 5.27 Percentage of clusters wrongly identified as a function of drop fraction for cluster-based representations of various codebook sizes.

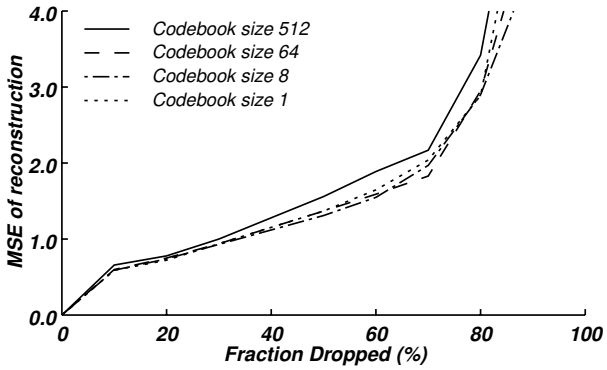


Figure 5.28 MSE of reconstruction as function of drop rate, for cluster-based representations of various codebook sizes.

books sizes. Increasing codebook size does not improve recognition accuracy significantly. In fact, the performance with codebook size 512 is worse than that obtained with codebook size 1

The lack of improvement in reconstruction accuracy and recognition performance with increasing codebook size is attributable entirely to errors in cluster membership identification. It stands to reason that substantially better reconstruction could be achieved at higher drop rates if the cluster identification accuracy could be improved.

5.3.5 Cluster membership estimation with preliminary estimates

We can avoid the problem of having to ignore the missing components entirely in cluster identification by using a *preliminary* estimate for the missing components in the vector, for cluster membership identification. If we represent the preliminary estimate for the vector of missing elements $\boldsymbol{S}_m(t)$ as $\bar{\boldsymbol{S}}_m(t)$, then

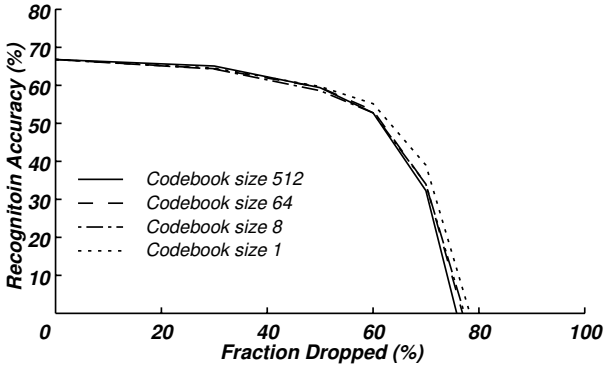


Figure 5.30 Recognition accuracy vs. drop fraction using spectrograms reconstructed by cluster marginal reconstruction, for various codebook sizes.

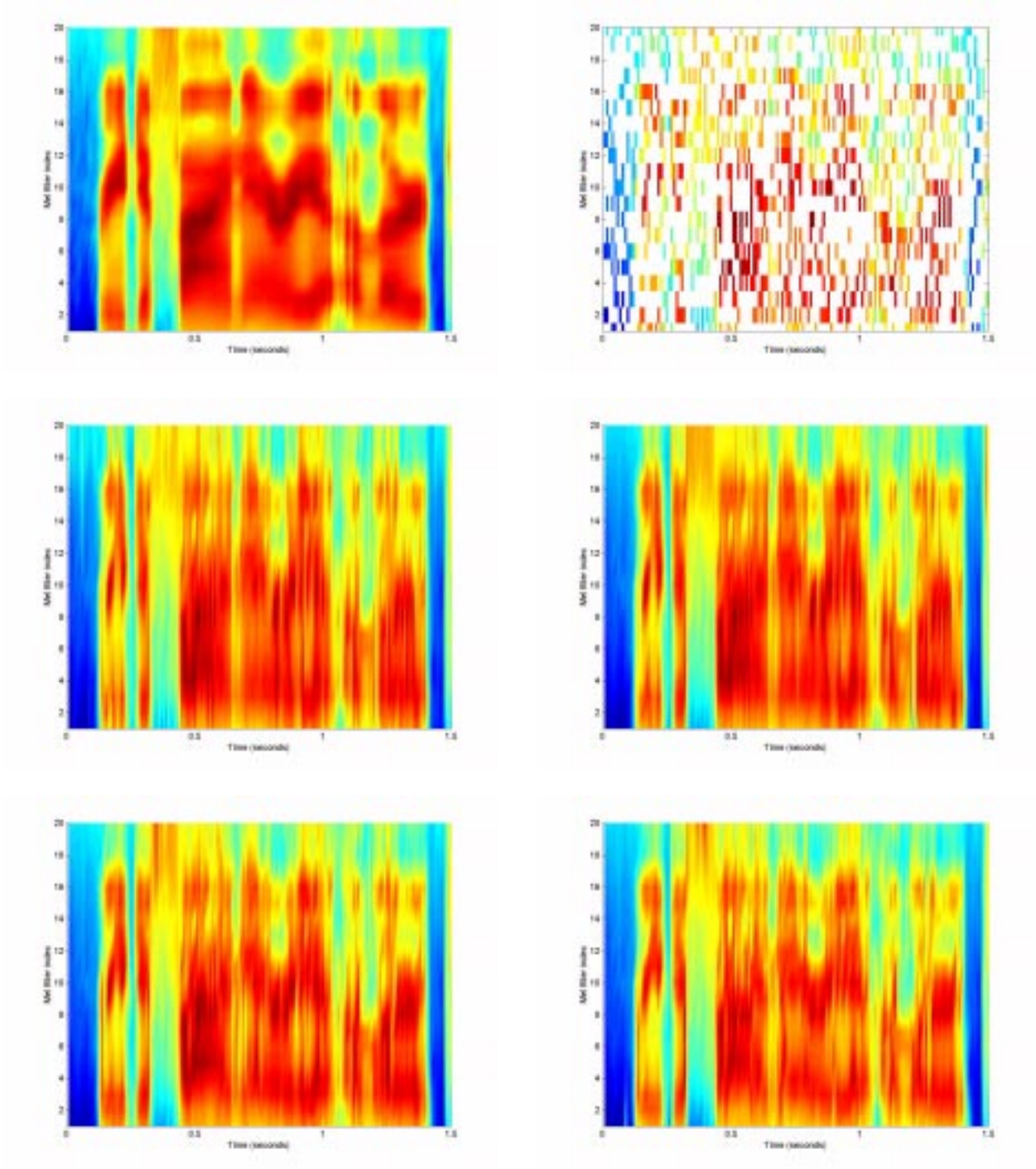


Figure 5.29 Reconstructed spectrogram obtained by marginalization based estimation, for several codebook sizes

Panel 1: Original spectrogram

Panel 2: Spectrogram with 70% of its elements randomly deleted

Panel 3: Spectrogram reconstructed with cluster based representation of codebook size 1

Panel 4: Spectrogram reconstructed with codebook size 8

Panel 5: Spectrogram reconstructed with codebook size 64

Panel 6: Spectrogram reconstructed with codebook size 512

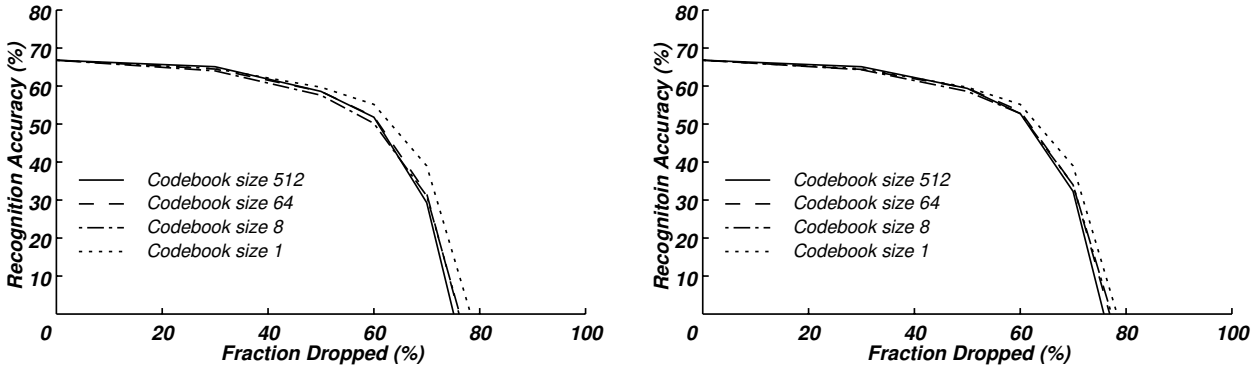


Figure 5.32 The left frame shows recognition accuracy obtained spectrograms reconstructed by frequency interpolation based estimation of cluster membership, for codebook sizes 1, 8 64 and 512. The right panel shows the same for cluster marginal reconstruction.

we would obtain the preliminary estimate for the complete vector as

$$\bar{\boldsymbol{S}}(t) \;=\; A_t[\boldsymbol{S}_o(t), \bar{\boldsymbol{S}}_m(t)] \tag{5.28}$$

The cluster membership of the vector can then be estimated using the preliminary estimate as

$$\hat{k}_m \;=\; \arg\min_k\{0.5(\bar{\boldsymbol{S}}(t)-\mu_k)^T\Theta_k^{-1}(\bar{\boldsymbol{S}}(t)-\mu_k)-\log(c_k)+0.5\log(\left|\Theta_k\right|)\} \tag{5.29}$$

It is important to distinguish between the preliminary estimate of the complete vector $\bar{\boldsymbol{S}}(t)$ and the final estimate of the complete vector $\hat{\boldsymbol{S}}(t)$ that is used to reconstruct the complete spectrogram in Equation (5.20). The density of the cluster identified in Equation (5.29) is used to obtain the final estimate of the missing elements of the vector $\boldsymbol{S}(t)$. The complete vector $\hat{\boldsymbol{S}}(t)$ is obtained as in Equation (5.20).

The preliminary estimate for the missing components, $\bar{\boldsymbol{S}}_m(t)$ can be obtained by any of the geometrical reconstruction method described in Section 5.2. Of these, it was seen that simple linear interpolation was superior to non-linear interpolation methods. Therefore, linear interpolation based estimation methods are good candidate methods for obtaining the preliminary estimate of missing elements.

5.3.5.1 Preliminary estimate by frequency interpolation

The preliminary estimate $\bar{\boldsymbol{S}}_m(t)$ of the missing components in the vector can be obtained by linear interpolation across frequency, as described in Section 5.2.1. This preliminary estimate can then be used for estimating cluster membership.

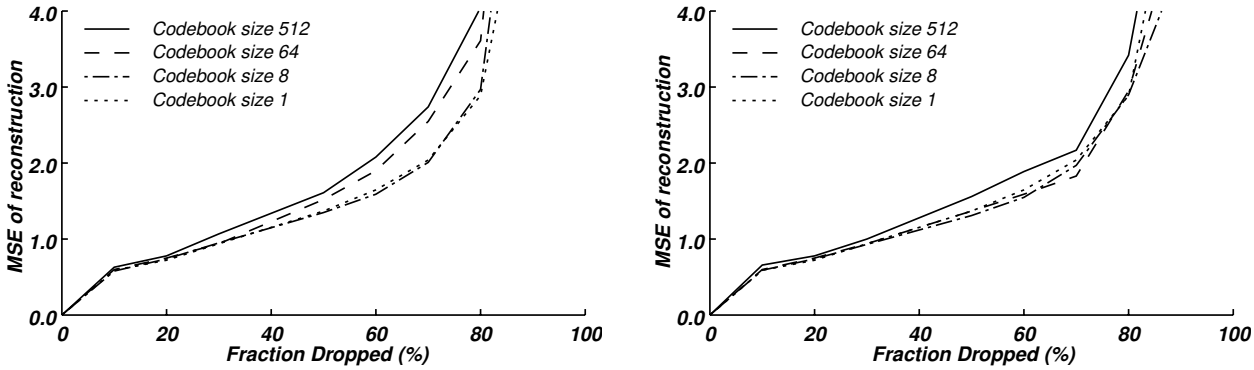


Figure 5.31 The left frame shows MSE of reconstruction for frequency interpolation based estimation of cluster membership, for codebook sizes 1, 8, 64 and 512. The right panel shows the same for cluster marginal reconstruction.

Figure 5.31 shows the MSE of reconstruction for frequency interpolation based estimation of cluster membership, as a function of the drop fraction, for several codebook sizes, and compares it with the MSE for cluster marginal reconstruction (Section 5.3.4). The random-drop paradigm and the DARPA RM database were used, as in the other experiments reported in this chapter. Figure 5.32 compares the recognition accuracy obtained with reconstructed spectrogram for the two cases. As can be seen, there is no appreciable improvement to be obtained by interpolating across frequency. In fact, there seems to be a slight degradation of performance at higher codebook sizes. This is only to be expected since interpolation across frequency depends on the continuity across frequency bands to obtain estimates of missing components. Cluster-based representations already model the correlations between frequency bands explicitly. Since no information other than this is used in the preliminary estimate, improvement in reconstruction due to using the preliminary estimate can only be expected to be marginal, if any.

Since no improvement is to be gained by this procedure we make no further reference to it in this thesis.

5.3.5.2 Preliminary estimate by time interpolation

The preliminary estimate of the missing components, $\bar{\boldsymbol{S}}_m(t)$ used in the estimation of cluster membership can be obtained by linear interpolation across time, as described in Section 5.2.1. Linear interpolation along time takes advantage of the temporal continuity of frequency components to estimate missing components of the spectrogram. A cluster-based representation models the distribution of each vector independently of any other vector and does not model temporal continuity in any manner. Therefore, the temporal

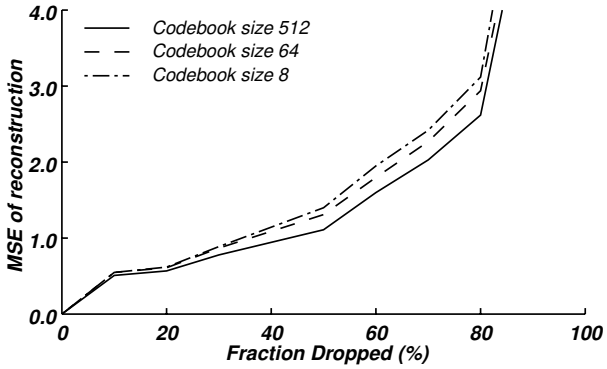


Figure 5.34 MSE for spectrogram reconstructed by cluster time-interpolated reconstruction, as a function of drop fraction, for various codebook sizes

constraints imposed by linear interpolation across time represent an additional source of information and are expected to improve cluster membership identification and reconstruction accuracy.

We refer to this reconstruction procedure where cluster membership is identified based on preliminary estimates given by linear interpolation along time as *cluster time-interpolated reconstruction*.

Figure 5.33 shows the percentage of vectors whose cluster membership was wrongly identified when time-interpolation based preliminary estimates are used, as a function of the drop fraction, for various codebook sizes. Comparison with Figure 5.27 shows that the cluster-membership-identification error is significantly less than that seen when clusters memberships are identified based on observed elements alone. The temporal continuity imposed by the preliminary estimates improves the cluster membership identification greatly. Figure 5.34 shows the MSE in reconstruction as a function of the fraction of elements missing, for various codebook sizes. We observe, in this case, that unlike the case of cluster mar-

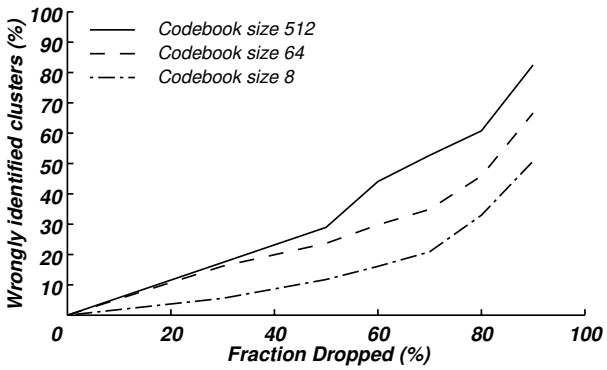


Figure 5.33 Percentage of vectors whose cluster membership was wrongly identified, as a function of drop fraction, for various codebook sizes.

ginal reconstruction (Figure 5.28), the MSE actually improves with increasing codebook size. This is due to the improved estimation of cluster membership.

Figure 5.35 shows an example of the estimated complete spectrogram obtained by cluster time-interpolated reconstruction, with different codebook sizes, when 70% of the elements in the incomplete spectrogram are missing. Predictably, the reconstructed spectrogram visually resembles the original spectrogram as codebook size increases.

Figure 5.36 shows recognition accuracy obtained using reconstructed spectrograms for various codebook sizes. Recognition accuracy is seen to improve with every increase in codebook size, following the trend in the MSE. Recognition accuracies for codebook size 512 are significantly greater than those obtained by cluster marginal reconstruction. Comparison with Figure 5.14 also establishes that the recognition accuracy obtained here is higher than that using purely geometrical reconstruction using linear interpolation across time (Section 5.2.4).

5.3.6 Cluster-based reconstruction methods summary

Cluster-based reconstruction methods can be very effective in reconstructing missing regions of spectrograms. The introduction of the vector statistics of spectral vectors improves the reconstruction significantly over methods that use purely local information, such as linear and non-linear interpolation.

When clusters memberships are identified based only on the observed components of spectral vectors, the performance obtained with multiple-cluster-based representations is similar to that obtained when the distribution of spectral vectors is modeled as a single Gaussian. This seems to indicate that the single

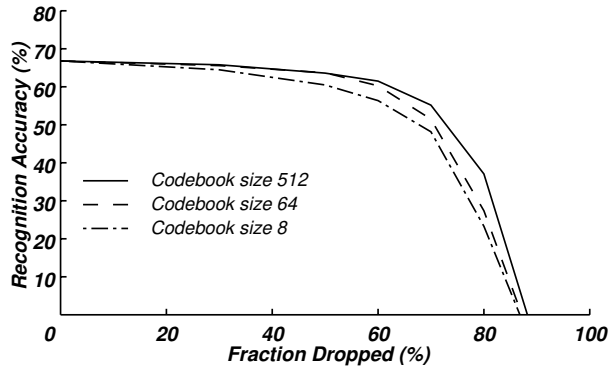


Figure 5.36 Recognition accuracy with reconstructed spectrograms as a function of drop fraction, for various codebook sizes.

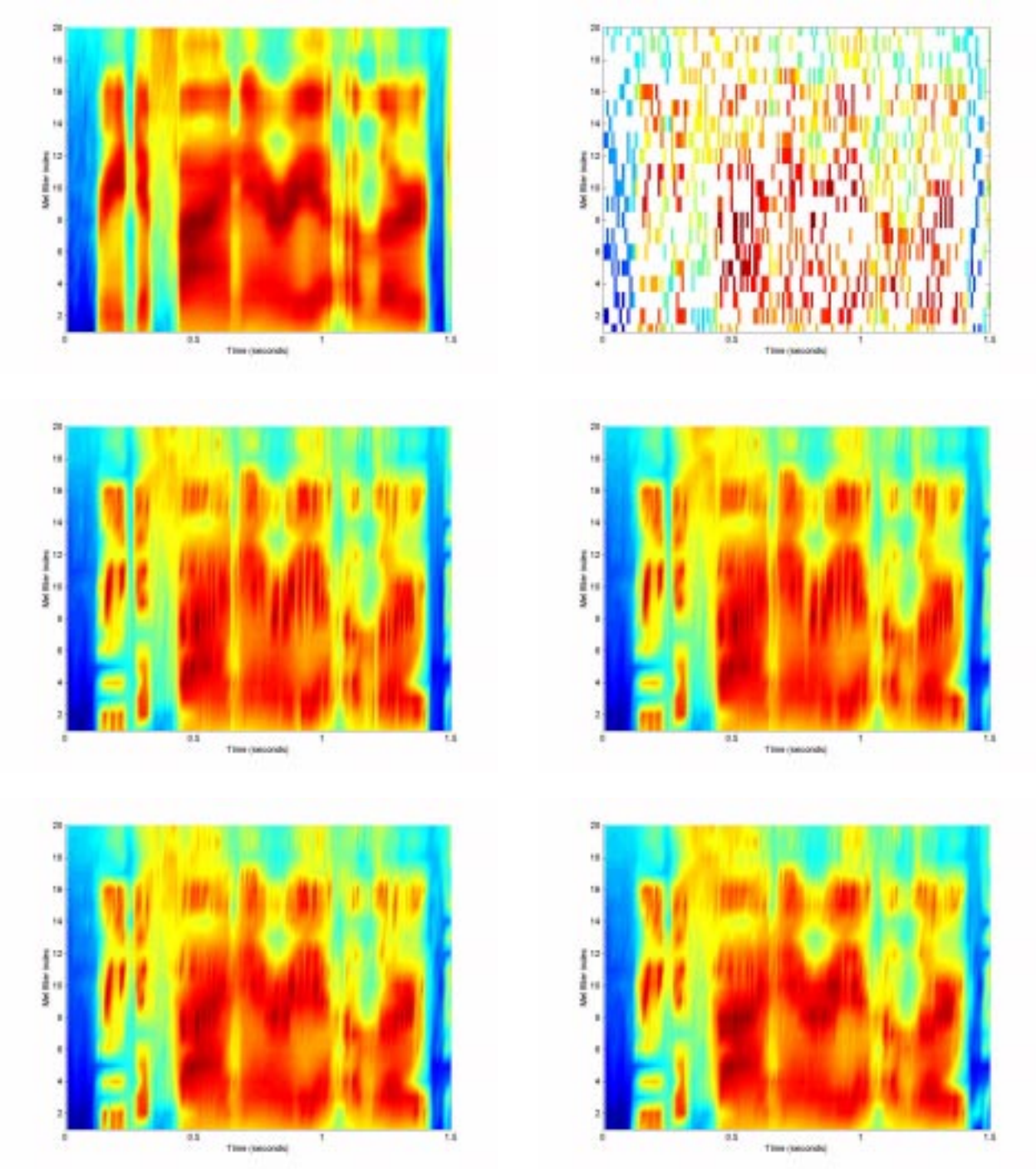


Figure 5.35 Reconstructed spectrograms when cluster membership was identified based on a preliminary estimate by linear interpolation along time

Panel 1: Original spectrogram

Panel 2: Spectrogram with 70% of its elements randomly deleted

Panel 3: Spectrogram reconstructed with cluster based representation of codebook size 2

Panel 4: Spectrogram reconstructed with codebook size 8

Panel 5: Spectrogram reconstructed with codebook size 64

Panel 6: Spectrogram reconstructed with codebook size 512

Gaussian model for the distribution of vectors is as good as, or better than the Gaussian mixture model implied by multiple-cluster-based representations, for the purpose of reconstruction. In the absence of any additional criterion of localizing the position of the complete vector, the two models perform similarly.

Preliminary estimates of missing elements given by linear interpolation along time provide the localization of the vector needed to obtain better performance with multiple-cluster-based representations. The inclusion of temporal information in the reconstruction procedure in the form of the temporal continuity enforced by the preliminary estimate improves the quality of the reconstruction significantly. However, the information used in the preliminary estimate is purely local. It stands to reason that if prior information regarding the *statistical relationships* between the components of different vectors could be used the quality of the reconstruction can be further improved.

There are several ways of statistically modeling the temporal continuity between elements in the spectrogram. One method would be to model the sequence of spectral vectors as the output of a hidden Markov model (HMM) [Therrien 1992], or a higher-order HMM [Therrien 1992], rather than as a sequence of IID vectors. However HMMs and higher order HMMs are complicated models requiring many parameters. A much simpler model would be to simply model the statistical correlations between the various elements in the spectrogram explicitly. The following section deals with such a method.

5.4 Covariance-based reconstruction

A very simple statistical model for the spectrogram is to consider the sequence of spectral vectors that constitute a spectrogram to be the output of a Gaussian wide-sense stationary (WSS) random process [Papoulis 1991]. All possible spectrograms are assumed to be individual observations from a single process. The statistical parameters of this process are then used to obtain estimates for the missing components of incomplete spectrograms.

We refer to spectrogram reconstruction methods based on this model as *covariance-based reconstruction* methods.

The assumption of wide-sense stationarity leads to the assumption that the means of the spectral vectors, and the covariances between elements in the spectrogram are independent of their position in the spectrogram. If we define the mean of the k^{th} element of the t^{th} spectral vector $S(t, k)$, $\mu(t, k)$, and the

covariance between the k_1 th element of the t_1 th spectral vector $S(t_1, k_1)$ and the k_2 th element of the t_2 th spectral vector $S(t_2, k_2)$, $c(t_1, t_2, k_1, k_2)$, as

$$\begin{aligned} \mu(t, k) &= E[S(t, k)] \\ c(t_1, t_2, k_1, k_2) &= E[(S(t, k_1) - \mu(t, k_1))(S(t_2, k_2) - \mu(t_2, k_2))] \end{aligned} \tag{5.30}$$

where $E[\ ]$ stands for the expectation operator, the assumption of wide-sense stationarity gives us the following properties for these parameters [Papoulis 1991]

$$\mu(t, k) = \mu(t_1, k) = \mu(k) \tag{5.31}$$

$$c(t, t + \tau, k_1, k_2) = c(t_1, t_1 + \tau, k_1, k_2) = c(\tau, k_1, k_2) \tag{5.32}$$

In other words, the expected value $\mu(k)$ of the k th component of a spectral vector is not dependent on where the vector occurs in the spectrogram. Similarly, the covariance between the components of two spectral vectors depends only on the distance τ between the vectors (along the time axis) and not on where they occur in the spectrogram. The means of the components of the spectral vectors $\mu(k)$ and the various covariance parameters $c(\tau, k_1, k_2)$ can now be learned from a training corpus of uncorrupted spectrograms. The implication of the assumption of a Gaussian process is that the joint distribution the components of all the spectral vectors in a sequence of vectors is Gaussian. Additionally, the distribution of any subset of the components in a sequence of vectors is also Gaussian [Papoulis 1991]. Therefore these means and covariances describe the process completely and are all that are needed to estimate missing components of spectrograms.

The $\mu(k)$ values define the expected value of every component in a spectrogram and the $c(\tau, k_1, k_2)$ values define the covariance between any component in the spectrogram and any other component in it.

$$\begin{aligned} E[S(t, k)] &= \mu(k) \\ E[(S(t_1, k_1) - \mu(t_1, k_1))(S(t_2, k_2) - \mu(t_2, k_2))] &= c(t_1 - t_2, k_1, k_2) \end{aligned} \tag{5.33}$$

To reconstruct an incomplete spectrogram $\boldsymbol{S}$ the observed components of the spectrogram are arranged into a vector $\boldsymbol{S}_o$. The missing components are arranged into a vector $\boldsymbol{S}_m$. Since we know the mean values

S(1,1)	S(2,1)	S(3,1)	S(4,1)
S(1,2)	S(2,2)	S(3,2)	S(4,2)
S(1,3)	S(2,3)	S(3,3)	S(4,3)
S(1,4)	S(2,4)	S(3,4)	S(4,4)

Figure 5.37 Example showing how the missing and observed components of a spectrogram can be separated into a vector of missing components and a vector of observed components, and the corresponding mean and covariance values. The figure represents a spectrogram with 4 spectral vectors, each with 4 elements. Each column of elements represents a single spectral vector. The grey elements are missing.

of all the components in the spectrogram and the covariance between any two components in it, the means of the individual components of $\boldsymbol{S}_o$ and $\boldsymbol{S}_m$ and the covariances between their various components are all known. These can be used to construct μ_m^S and μ_o^S, the mean vectors of $\boldsymbol{S}_m$ and $\boldsymbol{S}_o$ respectively, $\boldsymbol{C}_{oo}$, the autocovariance matrix of $\boldsymbol{S}_o$, and $\boldsymbol{C}_{mo}$, the cross covariance between $\boldsymbol{S}_m$ and $\boldsymbol{S}_o$.

Explaining the construction of $\boldsymbol{S}_m$ and $\boldsymbol{S}_o$ with an example

We illustrate the construction of $\boldsymbol{S}_m$ and $\boldsymbol{S}_o$ with a simple example. Figure 5.37 shows an example of a small spectrogram consisting of only four spectral vectors, each of which has only four components. Each of the elements in the spectrogram has been identified by a tag for convenience. All grey boxes in the figure represent missing elements.

The vector of observed elements $\boldsymbol{S}_o$, and the vector of missing elements $\boldsymbol{S}_m$ are constructed for this example as

$$\boldsymbol{S}_o = [S(1,2), S(1,4), S(2,1), S(2,3), S(3,2), S(3,3), S(3,4), S(4,1), S(4,4)]^T$$

$$\boldsymbol{S}_m = [S(1,1), S(1,3), S(2,2), S(2,4), S(3,1), S(4,2), S(4,3)]^T$$

The expected value of all elements in any row is assumed to be the same (since the vectors are assumed to be the output of a WSS process). The mean vectors for $\boldsymbol{S}_o$ and $\boldsymbol{S}_m$ are therefore constructed as

$$\mu_o^S = [\mu(2), \mu(4), \mu(1), \mu(3), \mu(2), \mu(3), \mu(4), \mu(1), \mu(4)]^T$$

$$\mu_m^S = [\mu(1), \mu(3), \mu(2), \mu(4), \mu(1), \mu(2), \mu(3)]^T$$

The autocovariance matrix of $\boldsymbol{S}_o$ is a 9x9 matrix constructed as

$$\boldsymbol{C}_{oo} = \begin{bmatrix} c(0,2,2) & c(0,2,4) & c(1,2,1) & \dots & c(3,2,4) \\ c(0,4,2) & c(0,4,4) & c(1,4,1) & \dots & c(3,4,4) \\ c(-1,1,2) & c(-1,1,4) & c(0,1,1) & \dots & c(2,1,4) \\ \dots & \dots & \dots & \dots & \dots \\ c(-3,4,2) & c(-3,4,4) & c(-2,4,1) & \dots & c(0,4,4) \end{bmatrix}$$

Similarly, the cross covariance between $\boldsymbol{S}_m$ and $\boldsymbol{S}_o$ is a 7x9 matrix given by

$$\boldsymbol{C}_{mo} = \begin{bmatrix} c(0,1,2) & c(0,1,4) & c(1,1,1) & \dots & c(3,1,4) \\ c(0,3,2) & c(0,3,4) & c(1,3,1) & \dots & c(3,3,4) \\ c(-1,2,2) & c(-1,2,4) & c(-1,2,1) & \dots & c(2,2,4) \\ \dots & \dots & \dots & \dots & \dots \\ c(-3,3,2) & c(-3,3,4) & c(-2,3,1) & \dots & c(0,3,4) \end{bmatrix}$$

The means and covariances of the vector of observed elements $\boldsymbol{S}_o$ and the vector missing elements $\boldsymbol{S}_m$, and the cross covariance between them can now be used to obtain an MAP estimate for $\boldsymbol{S}_m$:

$$\hat{\boldsymbol{S}}_m = \mu_m^S + \boldsymbol{C}_{mo}\boldsymbol{C}_{oo}^{-1}(\boldsymbol{S}_o - \mu_o^S) \tag{5.34}$$

Equation (5.34) would perform global reconstruction of all the missing elements in the damaged spectrogram in a single reconstruction step. However, the direct computation of Equation (5.34) would require inversion and multiplication of extremely large matrices. For example a typical 4 second utterance has 400 frames. If the spectral vectors have 20 frequency components each, there are 8000 components in all in the spectrogram. If 50% of the components are missing, both $\boldsymbol{C}_{oo}$ and $\boldsymbol{C}_{mo}$ are 4000 x 4000 matrices. Direct computation of Equation (5.34) would therefore necessitate the inversion of a 4000x4000 matrix, followed by the multiplication of two 4000x4000 matrices. If the utterance were longer the matrices would become still bigger. Clearly, the matrix operations required are impractical. A much more practical solution would

be to reconstruct the missing elements of the picture incrementally.

5.4.1 Reconstructing missing elements individually

The simplest reconstruction would be to reconstruct each missing element in the spectrogram independently of every other missing element. Let $S(t, k)$ be the missing element being estimated. Note here that $S(t, k)$ is an element of the vector of missing components, $\boldsymbol{S}_m$. The covariances between $S(t, k)$ and the various components of $\boldsymbol{S}_o$ can be used to construct the cross-covariance matrix between the two. We represent this matrix by $\boldsymbol{c}_{om}(t, k)$. Note that the components of $\boldsymbol{c}_{om}(t, k)$ form one row of $\boldsymbol{C}_{mo}$, the cross covariance between $\boldsymbol{S}_m$ and $\boldsymbol{S}_o$.

The MAP estimate of $S(t, k)$ can now be obtained as

$$\hat{S}(t, k) = \mu(k) + \boldsymbol{c}_{mo}(t, k)\boldsymbol{C}_{oo}^{-1}(\boldsymbol{S}_o - \mu_o^S) \tag{5.35}$$

where $\mu(k)$ is the expected value of $S(t, k)$ as given in Equation (5.33). Initially there does not appear to be any advantage to using Equation (5.35) since the dimensionality of $\boldsymbol{C}_{oo}$, the matrix being inverted in Equation (5.35), is no different from that in Equation (5.34) and the estimate of $S(t, k)$ obtained from the two equations is identical. However, the estimation can be considerably simplified by taking advantage of the fact that all components of $\boldsymbol{S}_o$ do not contribute equally to the estimate of $S(t, k)$.

The relative covariance between two components $S(t, k_1)$ and $S(t + \tau, k_2)$ of the spectrogram is defined as

$$r(\tau, k_1, k_2) = \frac{c(t, t + \tau, k_1, k_2)}{\sqrt{c(t, t, k_1, k_1)c(t + \tau, t + \tau, k_2, k_2)}} = \frac{c(\tau, k_1, k_2)}{\sqrt{c(0, k_1, k_1)c(0, k_2, k_2)}} \tag{5.36}$$

If $S(t, k)$ were to be estimated based on only one component of $\boldsymbol{S}_o$, say $S(t + \tau, k_1)$, then it can be shown that the estimate of $S(t, k)$ is given by

$$\hat{S}(t, k) = \mu(k) + r(\tau, k, k_1)\sqrt{\frac{c(0, t, t)}{c(0, t_1, t_1)}}(S(t + \tau, k_1) - \mu(k_1)) \tag{5.37}$$

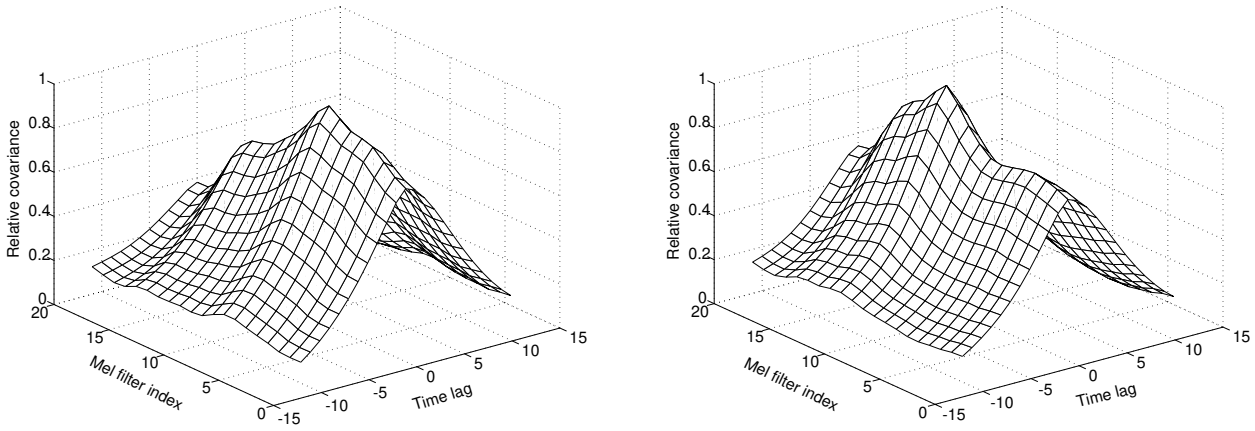


Figure 5.38 The left panel shows the relative covariance between the energy in the 8th frequency component (k=8) of any spectral vector and other elements of the spectrogram. The right panel shows the relative covariance between the energy in the 12th frequency component (k=12) of any spectral vector and other elements in the spectrogram

Clearly, as the relative covariance $r(\tau, k, k_1)$ between $S(t, k)$ and $S(t + \tau, k_1)$ decreases, the contribution of $S(t + \tau, k_1)$ to the estimate of $S(t, k)$ decreases linearly. For very small values of $r(\tau, k, k_1)$ therefore, its contribution to the estimate of $S(t, k)$ becomes negligible. In general, the contribution of any element $S(t_1, k_1)$ of $\boldsymbol{S}_o$ to the estimate of $S(t, k)$ is low if it has low relative covariance with $S(t, k)$, provided $S(t_1, k_1)$ also has low relative covariance with other elements of $\boldsymbol{S}_o$ that have high relative covariance with $S(t, k)$. In this situation $S(t_1, k_1)$ can be removed from the conditioning vector $\boldsymbol{S}_o$ without significant increase in the MSE of estimation.

It is observed that the relative covariance $r(\tau, k_1, k_2)$ between two elements $S(t, k_1)$, and $S(t + \tau, k_2)$ of a spectrogram falls off very quickly as either τ or $k_1 - k_2$ increases. Figure 5.38 shows the variation of $r(\tau, k, k')$ as a function of τ and k' for two different values of k. In both cases we observe that $r(\tau, k, k')$ falls very rapidly from its peak value of 1.0 as both $|\tau|$ and $|k - k'|$ increase, falling below 0.5 for $|\tau| > 5$ or $k - k' > 10$.

As a result, most elements of $\boldsymbol{S}_o$ have very low relative covariance with $S(t, k)$. Additionally, these elements also have low relative covariance with those elements of $\boldsymbol{S}_o$ that have a high relative covariance

with $S(t, k)$. The vector of observed elements that is used to estimate $S(t, k)$ can therefore be constructed from only those observed components of the spectrogram that have a high relative covariance with it, *i.e.* components that have a relative covariance above some threshold R. If we denote the vector constructed of observed components that are used to estimate $S(t, k)$ by $\boldsymbol{S}_o(t, k)$, we would get the following rule for the construction of $\boldsymbol{S}_o(t, k)$

$$S(t_1, k_1) \in \boldsymbol{S}_o(t, k), \quad \text{if } (r(t - t_1, k_1, k) > R) \tag{5.38}$$

Note that the vector of observed components $\boldsymbol{S}_o(t, k)$ is specific to $S(t, k)$. $\boldsymbol{S}_o(t, k)$ typically has much fewer components than $\boldsymbol{S}_o$. We refer to the set of elements in $\boldsymbol{S}_o(t, k)$ as the *neighborhood* of $S(t, k)$. We refer to $\boldsymbol{S}_o(t, k)$ as the *neighborhood vector* of $S(t, k)$. Once $\boldsymbol{S}_o(t, k)$ has been constructed, its mean vector $\mu_o(t, k)$ can be constructed using the expected values of its components, and its autocovariance matrix $\boldsymbol{C}_{oo}(t, k)$, and the cross-covariance matrix between $S(t, k)$ and $\boldsymbol{S}_o(t, k)$, $\boldsymbol{c}_{mo}(t, k)$ can be constructed using the covariance between their components. The estimate for $S(t, k)$, the missing component, is now obtained as

$$\hat{S}(t, k) = \mu(k) + \boldsymbol{c}_{mo}(t, k)\boldsymbol{C}_{oo}^{-1}(t, k)(\boldsymbol{S}_o(t, k) - \mu_o(t, k)) \tag{5.39}$$

All the missing elements in the spectrogram can be estimated in this manner to reconstruct the complete spectrogram.

A simple example of constructing $\boldsymbol{S}_o(t, k)$ for the estimation of a missing element

We illustrate the construction of $\boldsymbol{S}_o(t, k)$, and the corresponding mean and covariance parameters with an example. Figure 5.39 shows a small spectrogram of 16 elements. All elements shaded grey in the picture are missing.

In order to estimate $S(2, 2)$ (shown in a lighter shade of grey), all elements $S(t, k)$ in the spectrogram, such that $r(t - 2, 2, k) \geq 0.5$ are identified. These are represented by the dotted elements in the spectrogram. The vector of observed elements $\boldsymbol{S}_o(t, k)$ is now constructed as

S(1,1)	S(2,1)	S(3,1)	S(4,1)
S(1,2)	S(2,2)	S(3,2)	S(4,2)
S(1,3)	S(2,3)	S(3,3)	S(4,3)
S(1,4)	S(2,4)	S(3,4)	S(4,4)

Figure 5.39 An example spectrogram with 4 spectral vectors, each with 4 elements. The grey elements are missing. The neighborhood vector and the various statistical parameters for the estimation of S(2,2), the element shaded light grey, are to be constructed.

$$\boldsymbol{S}_o(2,2) = [S(1,2), S(2,1), S(2,3), S(3,2), S(3,3)]^T$$

The mean vectors for $\boldsymbol{S}_o(2,2)$ and $S(2,2)$ are constructed as

$$E[\boldsymbol{S}_o(2,2)] = \mu_o(2,2) = [\mu(2), \mu(1), \mu(3), \mu(2), \mu(3)]^T$$

$$E[S(2,2)] = \mu(2)$$

The autocovariance matrix of $\boldsymbol{S}_o(2,2)$ is a 5x5 matrix constructed as

$$\boldsymbol{C}_{oo}(2,2) = \begin{bmatrix} c(0,2,2) & c(1,2,1) & c(1,2,3) & c(2,2,2) & c(2,2,3) \\ c(-1,1,2) & c(0,1,1) & c(0,1,3) & c(1,1,2) & c(1,1,3) \\ c(-1,3,2) & c(0,3,1) & c(0,3,3) & c(1,3,2) & c(1,3,3) \\ c(-2,2,2) & c(-1,2,1) & c(-1,2,3) & c(0,2,2) & c(0,2,3) \\ c(-2,3,2) & c(-1,3,1) & c(-1,3,3) & c(0,3,2) & c(0,3,3) \end{bmatrix}$$

The cross covariance between $S(2,2)$ and $\boldsymbol{S}_o(2,2)$ is a 1x5 matrix given by

$$\boldsymbol{c}_{mo}(2,2) = [c(-1,2,2), c(0,2,1), c(0,2,3), c(1,2,2), c(1,2,3)]^T$$

The estimate of $S(2,2)$ would be given by

$$\hat{S}(2,2) = \mu(2) + \boldsymbol{c}_{mo}(2,2)\boldsymbol{C}_{oo}^{-1}(2,2)(\boldsymbol{S}_o(2,2) - \mu_o(2,2))$$

The optimal relative-covariance threshold R has to be empirically determined. Figure 5.40 shows recognition accuracy obtained using reconstructed spectrograms for incomplete spectrograms with 90% of their elements (randomly) missing, as a function of the relative-covariance threshold. As can be seen from the figure a relative-covariance threshold of around 0.5 seems to be optimal. In fact, this was found to be the optimal relative-covariance threshold at all drop fractions. Including elements with relative covariance below 0.5 in the reconstruction is actually seen to result in poorer reconstruction.

Therefore, using 0.5 as the threshold, $\boldsymbol{S}_o(t, k)$ would, in principle, contain all the elements that are observed in the spectrogram and have a relative covariance greater than 0.5 with $S(t, k)$. In practice it has been observed that when a 20 dimensional mel-spectral representation is used for the spectrograms, it is sufficient to include the 16 observed elements of the spectrogram with the greatest relative covariance with $S(t, k)$ in $\boldsymbol{S}_o(t, k)$ and the inclusion of any more elements does not improve the reconstruction further. The complete procedure for reconstructing a complete spectrogram from an incomplete one therefore consists of constructing $\boldsymbol{S}_o(t, k)$ with upto 16 elements and the corresponding statistical parameter vectors and matrices for every missing element $S(t, k)$ in the spectrogram and computing the estimate for $S(t, k)$ using Equation (5.39).

We refer to this procedure of estimating individual missing elements of the spectrogram as *covariance individual reconstruction*. The nomenclature indicates the fact that covariance-based reconstruction is being performed, and that missing elements are being individually estimated.

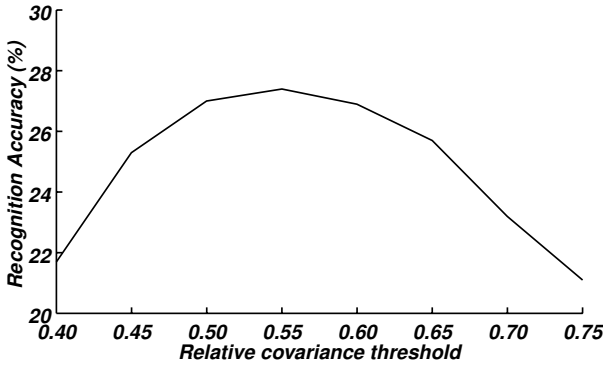


Figure 5.40 Recognition accuracy with spectrograms reconstructed by covariance-based estimation of individual missing elements, as a function of the relative-covariance threshold used to select elements for the neighborhood vector for missing elements. The incomplete spectrograms had 90% of their elements missing.

5.4.2 Jointly reconstructing all missing elements in a vector

Instead of estimating all the individual missing elements in the incomplete spectrogram separately we could reconstruct all the missing elements in a spectral vector simultaneously. We refer to this procedure as *covariance joint reconstruction*. This procedure is a compromise between reconstruction of individual elements and global reconstruction of the entire picture. Let $\boldsymbol{S}(t)$ be the t^{th} spectral vector in the spectrogram. The missing components in $\boldsymbol{S}(t)$ can be separated into a vector of missing elements, the *missing element vector* $\boldsymbol{S}_m(t)$. A vector $\boldsymbol{S}_o(t)$ can now be constructed of all observed elements in the spectrogram that have a relative covariance of at least 0.5 with at least one of the elements in $\boldsymbol{S}_m(t)$. The threshold of 0.5 is applied for the same reason that it was used in the earlier section where missing elements were being individually estimated - this eliminates all components whose contribution to the reconstruction is unreliable, while reducing the dimensions of $\boldsymbol{S}_o(t)$ greatly. We refer to the elements of $\boldsymbol{S}_o(t)$ as the *neighborhood* of $\boldsymbol{S}_m(t)$, and $\boldsymbol{S}_o(t)$ as the *neighborhood vector* of $\boldsymbol{S}_m(t)$. Once again, while $\boldsymbol{S}_o(t)$ in principle contains all observed elements in the spectrogram with a relative covariance greater than 0.5 with any of the elements of $\boldsymbol{S}_m(t)$, in practice limiting $\boldsymbol{S}_o(t)$ to include no more than 16 elements that have a high relative covariance with any one of the elements in $\boldsymbol{S}_m(t)$ does not result in any degradation in performance for the case of 20 mel filter based spectrograms.

The mean vector and covariance matrix of the elements in $\boldsymbol{S}_o(t)$, $\mu_o^S(t)$ and $\boldsymbol{C}_{oo}(t)$ can be constructed as before. Similarly, the mean vector of $\boldsymbol{S}_m(t)$, $\mu_m^S(t)$, and the cross-covariance matrix between $\boldsymbol{S}_m(t)$ and $\boldsymbol{S}_o(t)$, $\boldsymbol{c}_{mo}(t)$ can be constructed. The missing elements in the t^{th} vector of the spectrogram can now be estimated using the MAP Equation:

$$\hat{\boldsymbol{S}}_m(t) = \mu_m^S(t) + \boldsymbol{c}_{mo}(t)\boldsymbol{C}_{oo}^{-1}(t)(\boldsymbol{S}_o(t) - \mu_o^S(t)) \tag{5.40}$$

An example of constructing $\boldsymbol{S}_o(t)$ for joint estimation of all missing elements in a vector

We illustrate the construction of $\boldsymbol{S}_o(t)$, and the corresponding mean and covariance parameters with an

example. 5.41 shows a simple spectrogram with 4 spectral vectors, each with 4 elements. All elements shaded grey are missing. It is desired to estimate all missing elements in the second spectral vector. The elements to be estimated are shaded light grey.

S(1,1)	S(2,1)	S(3,1)	S(4,1)
S(1,2)	S(2,2)	S(3,2)	S(4,2)
S(1,3)	S(2,3)	S(3,3)	S(4,3)
S(1,4)	S(2,4)	S(3,4)	S(4,4)

Figure 5.41 The figure represents a small spectrogram with 4 spectral vectors, each with 4 elements.The grey elements are missing. We wish to estimate all the missing elements in the second spectral vector jointly. These are shown in a lighter shade of grey in the figure.

The missing element vector for the second spectral vector is constructed as

$$\boldsymbol{S}_m(2) = [S(2,2), S(2,4)]^T$$

The neighborhood vector for $\boldsymbol{S}_m(2)$ is constructed of all the elements $S(t,k)$ in the spectrogram, such that either $r(t-2,2,k) \geq 0.5$, or $r(t-2,4,k) \geq 0.5$. These are represented by the dotted elements in the spectrogram. This gives us

$$\boldsymbol{S}_o(2) = [S(1,2), S(1,4), S(2,1), S(2,3), S(3,2), S(3,3), S(3,4)]^T$$

The mean vectors for $\boldsymbol{S}_o(2)$ and $\boldsymbol{S}_m(2)$ are constructed as

$$E[\boldsymbol{S}_o(2)] = \mu_o^S(2) = [\mu(2), \mu(4), \mu(1), \mu(3), \mu(2), \mu(3), \mu(4)]^T$$

$$E[\boldsymbol{S}_m(2)] = \mu_m^S(2) = [\mu(2), \mu(4)]$$

The autocovariance matrix of $\boldsymbol{S}_o(2)$ is a 7x7 matrix constructed as

$$\boldsymbol{C}_{oo}(2) = \begin{bmatrix} c(0,2,2) & c(0,2,4) & c(1,2,1) & c(1,2,3) & c(2,2,2) & c(2,2,3) & c(2,2,4) \\ c(0,4,2) & c(0,4,4) & c(1,4,1) & c(1,4,3) & c(2,4,2) & c(2,4,3) & c(2,4,4) \\ c(-1,1,2) & c(-1,1,4) & c(0,1,1) & c(0,1,3) & c(1,1,2) & c(1,1,3) & c(2,4,4) \\ c(-1,3,2) & c(-1,3,4) & c(0,3,1) & c(0,3,3) & c(1,3,2) & c(1,3,3) & c(1,1,4) \\ c(-2,2,2) & c(-2,2,4) & c(-1,2,1) & c(-1,2,3) & c(0,2,2) & c(0,2,3) & c(1,3,4) \\ c(-2,3,2) & c(-2,3,4) & c(-1,3,1) & c(-1,3,3) & c(0,3,2) & c(0,3,3) & c(0,2,4) \\ c(-2,4,2) & c(-2,4,4) & c(-1,4,1) & c(-1,4,3) & c(0,4,2) & c(0,4,3) & c(0,4,4) \end{bmatrix}$$

The cross covariance between $\boldsymbol{S}_m(2)$ and $\boldsymbol{S}_o(2)$ is a 2x7 matrix constructed as

$$\boldsymbol{C}_{mo}(2) = \begin{bmatrix} c(-1,2,2) & c(-1,2,4) & c(0,2,1) & c(0,2,3) & c(1,2,2) & c(1,2,3) & c(1,2,4) \\ c(-1,4,2) & c(-1,4,4) & c(0,4,1) & c(0,4,3) & c(1,4,2) & c(1,4,3) & c(1,4,4) \end{bmatrix}^T$$

The MAP estimate for the two missing elements in the second vector would now be obtained as

$$\boldsymbol{S}_m(2) = \mu_m^S(2) + \boldsymbol{C}_{mo}(2)\boldsymbol{C}_{oo}(2)^{-1}(\boldsymbol{S}_o(2) - \mu_o^S(2))$$

To reconstruct the complete spectrogram, the vector of missing components $\boldsymbol{S_m}(t)$, the corresponding vector of observed elements with high relative covariance, $\boldsymbol{S}_o(t)$ and the associated mean vectors and covariance matrices would be constructed for each spectral vector in the spectrogram. This missing components in the spectral vector, $\boldsymbol{S_m}(t)$ would then be estimated using Equation (5.40).

5.4.3 Experimental results with covariance based reconstruction

Covariance-based reconstruction was evaluated using the DARPA RM database, using the random-drop paradigm and the experimental setup used in all other experiments described in this chapter. The statistical parameters used for the reconstruction, *i.e.* the means of frequency components, $\mu(k)$ and the various covariance values $c(\tau, k_1, k_2)$ were all learned from the training corpus that was used to train the HMMs.

Figure 5.42 shows the mean squared error of reconstruction as a function of the drop fraction in the incomplete spectrograms for both covariance individual reconstruction and covariance joint reconstruc-

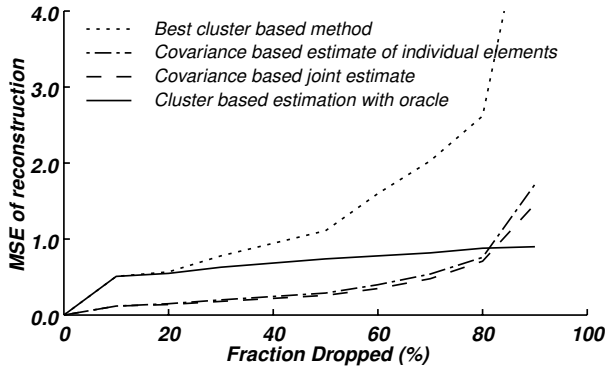


Figure 5.42 MSE of reconstruction for covariance individual reconstruction, covariance joint reconstruction, the best cluster-based reconstruction method (time interpolation based estimation), and the ideal cluster-based method (with oracle knowledge of cluster membership of spectral vectors).

tion. The figure also shows the MSE obtained using the best cluster-based method (cluster time-interpolated reconstruction) as well as the oracle MSE for cluster-based reconstruction with a codebook size of 512. We observe that covariance-based reconstruction results in better MSE than the best cluster-based method, and is in fact comparable with the MSE of cluster oracle reconstruction, except at very high drop rates. Also, the MSE obtained using joint estimation of the missing elements in a vector is marginally better than that obtained when the missing elements are individually estimated.

Figure 5.43 shows an incomplete spectrogram with a drop fraction of 90%, the reconstructed spectrogram obtained with covariance individual reconstruction and covariance joint reconstruction. We observe that even at this high drop rate the reconstruction is quite good. At similar drop rates the best cluster-based reconstruction technique was ineffective (Section 5.3.5).

Figure 5.44 shows the recognition accuracy obtained using reconstructed spectrograms as a function of the fraction of elements that were missing in the spectrogram. Both covariance individual reconstruction and covariance joint reconstruction are evaluated. Recognition accuracies obtained using the best cluster-based reconstruction method, *i.e.* cluster time-interpolated reconstruction, are also shown. Covariance-based reconstruction methods clearly result in the best recognition accuracies. For these test conditions the recognition accuracy obtained with reconstructed spectrograms, when 80% of the elements in the incomplete spectrogram are missing, is not much worse than the recognition accuracy obtained with uncorrupted spectrogram. The superior performance of covariance-based reconstruction methods is attributable to the fact that many more neighboring points are available to reconstruct any point in covariance-based recon-

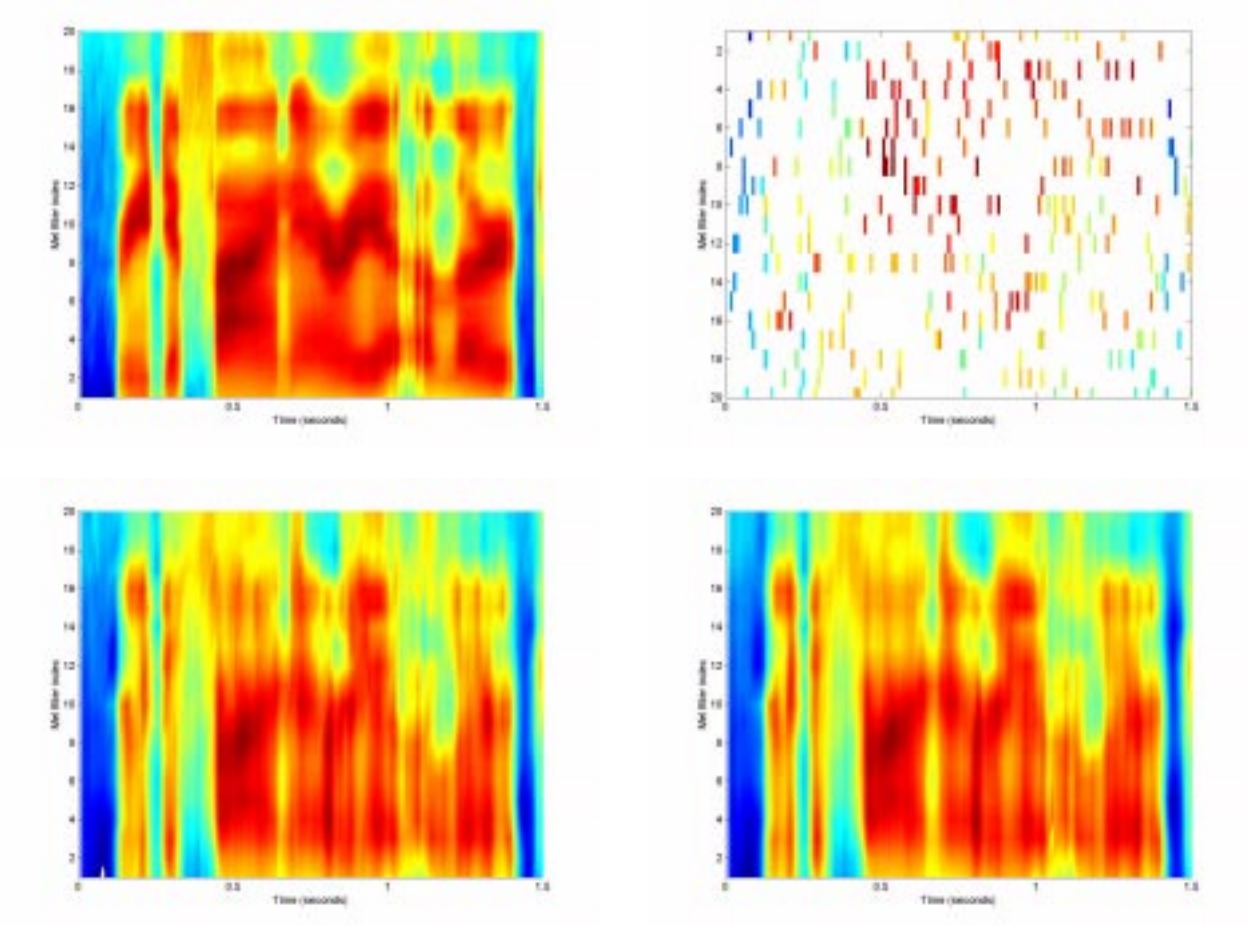


Figure 5.43 Spectrograms reconstructed by covariance-based estimation of missing elements

Panel 1: Original spectrogram

Panel 2: Spectrogram with 90% of its elements randomly deleted

Panel 3: Reconstructed spectrogram obtained by estimating missing elements individually

Panel 4: Reconstructed spectrogram obtained by estimating all missing elements in a vector jointly

structing, than in cluster-based reconstruction.

Joint estimation of missing elements in a vector is seen to result in better reconstruction than reconstruction of individual elements at high drop rates. We hypothesize that joint global reconstruction of all the missing elements in the picture would result in still better recognition accuracies.

5.5 Comparison with classifier-compensation techniques

Figure 5.45 compares the recognition accuracy obtained using the best cluster-based and covariance-based methods with that obtained using class-conditional imputation and marginalization. We observe that spectrogram reconstruction methods result in much better recognition accuracies than those obtained by

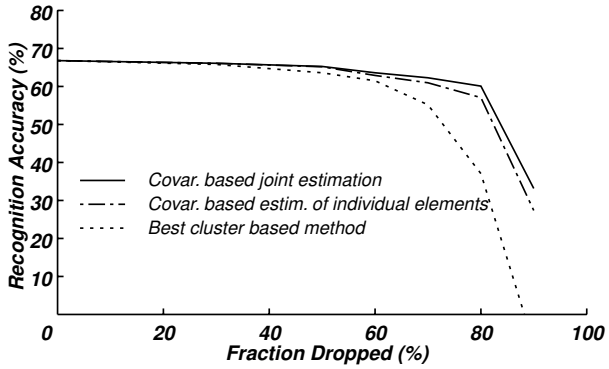


Figure 5.44 Recognition accuracy for covariance-based estimation of individual missing elements, covariance-based joint estimation of missing elements in a vector, and the best cluster-based reconstruction method (cluster time-interpolated reconstruction).

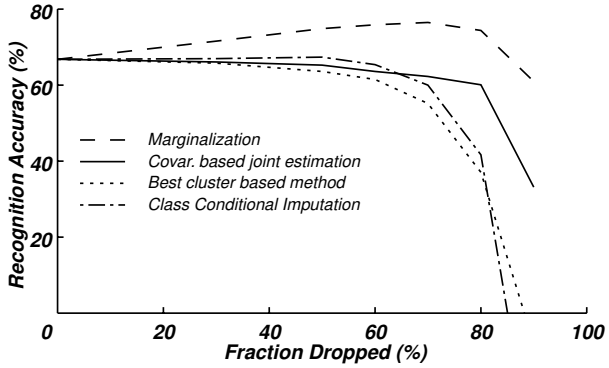


Figure 5.45 Comparison of recognition accuracies obtained with various incomplete-spectrogram methods, as a function of fraction of elements missing in the spectrogram. The methods compared are the best spectrogram reconstruction methods, *i.e.* covariance joint reconstruction and cluster time-interpolated reconstruction, with those obtained with classifier-modification methods, *i.e.* marginalization, and class-conditional imputation.

class-conditional imputation. Marginalization still results in the best recognition accuracies.

Nevertheless, spectrogram reconstruction methods hold the advantage that it is now possible to use the reconstructed spectrograms to derive other parameters/features such as cepstra which can be used to perform recognition. Recognition accuracies obtained using cepstra are typically much greater than those obtained with log spectra. Marginalization, on the other hand, requires the recognition system to be trained on spectrographic features. As a result cepstra derived from the spectrograms reconstructed by spectrogram reconstruction techniques can be used to obtain greater recognition accuracies than those obtainable with marginalization (using log-spectra-based recognition). Figure 5.46 shows recognition accuracy obtained with cepstra derived from spectrograms reconstructed by covariance joint reconstruction and

compares it with the recognition accuracy obtained with and marginalization and log-spectra-based recognition.

5.6 The short list of useful methods

We have proposed and evaluated several methods of estimating missing elements in incomplete spectrograms in this chapter. While evaluation on the basis of the random-drop paradigm is not comprehensive in any sense, it permits us to short-list the techniques that show promise of being useful.

Among geometrical reconstruction techniques it was found that linear interpolation methods outperform non-linear interpolation methods. Further, interpolation along time was superior to interpolation along frequency. Among geometrical reconstruction techniques therefore linear interpolation along time is the most useful.

Among cluster-based reconstruction techniques single cluster reconstruction, cluster marginal reconstruction and cluster time-interpolated reconstruction were all seen to be useful.

Among covariance-based reconstruction techniques covariance joint reconstruction was superior to covariance individual estimation.

We therefore short-list linear interpolation along time, single cluster reconstruction, cluster marginal reconstruction, cluster time-interpolated reconstruction, and covariance joint reconstruction as possibly useful methods worthy of further investigation. All of these methods differ from each other in a fundamental manner. Other methods have not been considered any further in this thesis, and where presented have

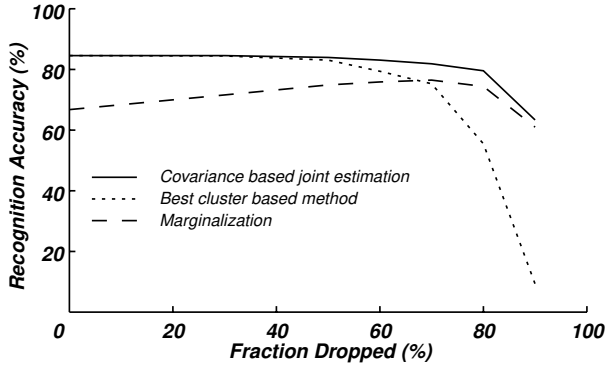


Figure 5.46 Recognition accuracy using cepstra computed from reconstructed spectrograms as a function of drop fraction. The recognition accuracy obtained using marginalization on log-spectra based recognition is also shown.

only been presented for reasons of comparison.

5.7 Summary and conclusions

Spectrogram reconstruction methods are seen to be as effective as classifier-modification methods (class-conditional imputation and marginalization) for handling spectrograms with missing elements. Even simple, purely geometrical, reconstruction methods such as linear and non-linear interpolation are seen to result in fairly effective reconstruction of missing elements when the missing elements are missing completely at random. The best geometrical reconstruction is achieved with linear interpolation. Reconstruction methods that utilize temporal relations between elements, *e.g.* linear interpolation across time, are in general more useful than methods that utilize relationships across frequency bands. This leads us to believe that there is greater continuity between elements across time than there is across frequency.

The use of prior statistical information about the correlations between elements in the spectrogram is very beneficial to the reconstruction. Cluster-based reconstruction techniques utilize the distribution of spectral vectors in the spectrogram. Cluster marginal reconstruction, a cluster-based reconstruction technique that works only with frequency components within a vector, results in significantly superior reconstruction to linear interpolation across frequency, a method that uses purely local information about frequency components. Cluster time-interpolated reconstruction, which combines linear interpolation across time with cluster-based reconstruction, is superior to that obtained with linear interpolation across time alone. In other words, the combination of geometrical reconstruction based on temporal continuity and statistical reconstruction based on statistical relationships across frequency bands results in superior performance to that obtained with local reconstruction based on temporal continuity alone.

The introduction of prior statistical information regarding the relationship between elements both across frequency and across time improves the reconstruction still further. Covariance-based reconstruction methods use statistical correlations between elements, both across time and across frequency explicitly. They are seen to result in the best reconstruction. Covariance-based reconstruction is generally observed to be better when multiple elements are jointly estimated than when they are estimated individually. We speculate that the best reconstruction would be obtained when all the missing elements in the spectrogram are jointly estimated. However, this is computationally infeasible.

In all of the methods described in this chapter the statistical information about the relationship between elements in the spectrogram is represented by very simple models. Cluster-based reconstruction uses a very simple cluster-based representation of the statistics. Within each cluster the distributions are further represented by a very simple Gaussian distribution. Covariance-based reconstruction uses an even simpler statistical representation - the entire spectrogram is represented as the output of a single WSS Gaussian random process. The reconstruction is performed using only the statistical parameters of this process.

Better statistical representations are likely to result in better reconstructions. The simple cluster-based representation used by cluster-based methods treats the sequence of vectors that constitute the spectrogram as independent. Consequently the model permits any vector to follow any other vector and retains no information regarding the sequentiality of the vectors. A superior model would be to model the individual clusters as the states of a Markov chain, *i.e.* modeling the sequence of vectors as the output of a hidden Markov model (HMM) where the individual clusters are the states of the HMM. In an HMM the state (cluster) that generates the current vector is dependent on the state that generated the previous vector. As a result the HMM model captures some of the temporal relationship between vectors, modeling the manner in which vectors can follow one another.

A more constrained model for the sequentiality of vectors would be to model the vectors as the output of a higher-order HMM. While standard HMMs condition the probability of a vector on the state that generated the previous vector, an N^{th} order HMM conditions it on the states that generate the previous N vectors.

The statistical model used with covariance-based reconstruction can also be improved. Covariance-based reconstruction models the sequence of spectral vectors as the output of a single wide-sense stationary random process. A more detailed representation would model the sequence of vectors as the output of a process that switches between a set of random processes. A simpler model would be to treat blocks of vectors as the basic unit in a cluster-based representation. Either model would capture the statistical relationships between elements in different vectors, *i.e.* relationships across time, with greater detail than the single Gaussian representation used by the covariance-based reconstruction methods described in this chapter. However, both representations would have the additional problem of identifying the cluster or random process associated with each of the vectors.

The speech recognition system itself encodes the acoustic, phonetic and linguistic information in speech corpora using various statistical models and can therefore be used to estimate the missing regions of the spectrogram. The statistical representation used by speech recognition systems is extremely detailed, including statistical models for the acoustic parameters derived from the speech, lexical representations of the data [Rabiner 1993], and N-gram statistical models to model the language [Katz 1987]. The lexical and language models provide additional sources of information not available to any of the other statistical models described earlier in this section. Consequently, using the speech recognition system itself to reconstruct damaged regions of spectrograms is likely to give the best reconstructions. One could use marginalization to obtain the best state sequence to represent the vectors. The distribution of the state associated with each vector can then be used to reconstruct the missing components of that vector. This would however necessitate performing recognition on the damaged utterance in order to obtain the best state sequence, which is a computational overhead that we would prefer to avoid.

All the spectrogram reconstruction methods described in this chapter, as well as the classifier-modification techniques described in Chapter 4 have so far been evaluated using the random-drop paradigm. However, the primary goal of developing these techniques was to compensate for the effect of additive noise on speech recognition systems. In the next chapter we evaluate the performance of all these methods as noise compensation techniques.

Chapter 6
Missing feature methods and noisy speech

6.1 Introduction

In the previous chapters we have evaluated several methods to recover data for spectrograms with random regions missing. In this paradigm the probability that any given element in the vicinity of a missing element is observed depends only on the drop fraction, and is independent of the fact that that element is missing. Therefore the probability that any of the observed elements in the spectrogram will have a relative covariance greater than a given threshold with the missing element also depends exclusively on the drop fraction. If there are T elements in the spectrogram that have a relative covariance greater than 0.5 with a missing element the probability that at least one of them will be observed is $1 - \alpha^T$, where α is the drop fraction. For example, if there are only five elements in the spectrogram with a relative covariance greater than 0.5 with the missing element, *i.e.* if $T = 5$, the probability that at least one of them is observed when the drop fraction is 90% ($\alpha = 0.9$) is 0.41. Thus, when elements of the spectrogram are deleted at random there is a relatively high probability that a missing element is well correlated to at least one of the observed elements in the spectrogram even at very high drop fractions. Obviously, for an incomplete spectrogram method to be useful it *has* to work well on the random-drop paradigm. Methods that do not work well even in this situation can, in general, be expected to perform worse in situations where the errors are more systematic and some of the missing components have a very low probability of being well correlated with any of the observed elements. The random-drop paradigm is therefore a very useful paradigm for preliminary evaluation of missing feature methods.

However, except for some special situations such as spectrograms that have been stored on a medium in which random regions have been corrupted, or transmitted spectrograms where elements have been lost in transmission, the random-drop model is unrealistic. When deletions in the spectrogram are due to the effect of corrupting noise the patterns of the missing components are usually much more systematic.

As explained in Sections 3.3 and 3.4, the effect of corrupting noise on speech can be modeled as missing features by deleting all regions of the spectrogram where the local SNR is below a threshold, leaving only the cleaner portions of the spectrogram behind. Figure 6.1 shows two such examples of *spectro-*

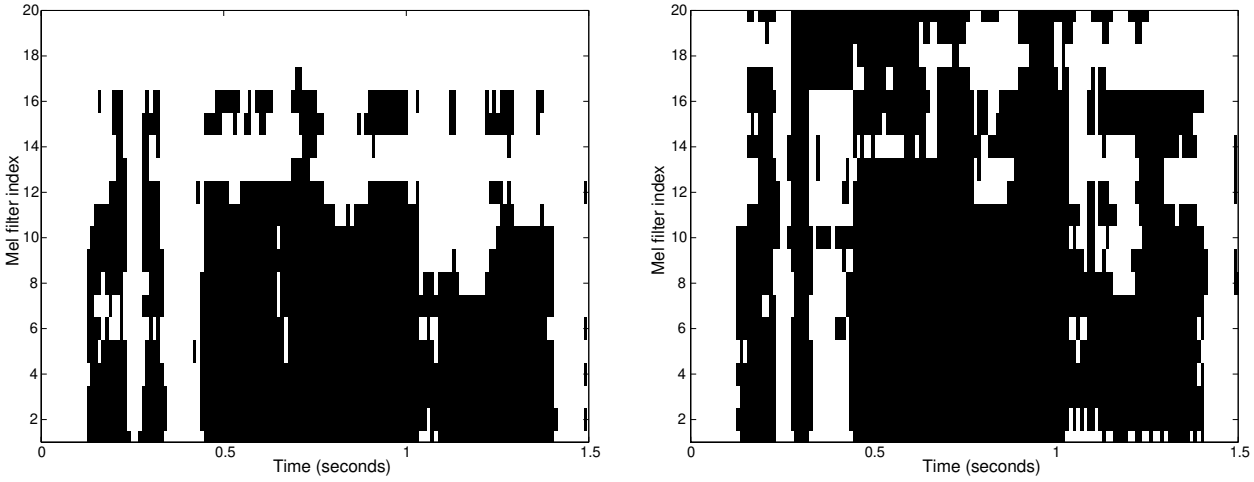


Figure 6.1 Two spectrographic masks. The left panel shows the mask for speech corrupted by white noise to 10 dB where all regions with a local SNR less than 0 dB have been deleted. The white regions in the picture have been deleted. The black regions are the "clean" regions and have been retained. The right panel shows a similar mask for speech that has been corrupted by music to 10 dB. The white regions are the unreliable regions with local SNR less than 0 dB and have been deleted.

graphic masks, or *deletion patterns* in spectrograms, where all elements with local SNR less than 0 dB have been erased. All white regions represent regions that have been deleted and the black regions represent regions that have been retained. In one of the examples the speech has been corrupted by white noise to a global SNR of 10 dB. In the other example the speech has been corrupted by a segment of music, also to a global SNR of 10 dB.

We observe in these spectrograms that the pattern of missing components is not completely random. Missing regions in such spectrograms occur in blocks. If any element in the spectrogram is missing it is highly probable that its neighbors are missing too. Another characteristic of the missing regions is that they are correlated to the underlying speech spectrum. Valleys in the spectrum are more likely to be corrupted to lower SNRs than peaks in the spectrum, and are therefore more likely to be deleted. Thus, the pattern of deleted elements is not only likely to be systematic, it can also favor the deletion of some patterns of spectral features over others.

One consequence of systematic block deletions such as these is that is that the elements of the spectrogram that have a high relative covariance with any of the missing elements are likely to be missing as well. The performance of any incomplete spectrogram method being applied, either for classification or for

reconstruction, is likely to be adversely affected by the block nature of the deletions. Thus, it is not definite that any missing feature method that performs well on spectrograms with random elements missing will also perform well with the kinds of deletion patterns seen in noisy speech as well.

An additional factor affecting the direct application of incomplete spectrogram techniques for noise compensation is that, even after the highly noisy portions of the spectrogram have been deleted, the remaining regions are not completely noise free. They continue to have low levels of noise. As a result the performance of any classification or reconstruction methods that are conditioned on these regions is likely to be worse than their performance on spectrograms of clean speech with identical missing regions.

In this chapter we evaluate and compare the performance of classifier-modification methods (class-conditional imputation and marginalization) and the spectrogram reconstruction methods described in Chapter 5 on speech that has been corrupted by white noise. The goal of the spectrogram reconstruction methods here is not to recreate the corresponding *noisy* spectrograms from the incomplete spectrograms, but to estimate what the value of these regions would have been had the spectrogram been clean. In this situation, the MSE between the reconstructed spectrogram and the complete (noisy) spectrogram is an inappropriate metric to measure the performance of the missing feature methods. We therefore evaluate the performance of the spectrogram reconstruction methods solely on the basis of the recognition accuracy obtained with the reconstructed spectrograms. Another important factor for consideration in noise compensation algorithms is the computational complexity of the algorithms. Procedures that take less time to perform are preferable to those that take more time. We evaluate the computational complexity of incomplete spectrogram reconstruction methods in terms of the total time taken to recognize an average utterance when they are used, and compare the computational complexity of spectrogram reconstruction methods with that of classifier-modification methods.

In many of the recognition experiments reported in the rest of this chapter the recognition accuracy obtained is shown to be *less* than 0%. This is not a paradox. In all experiments recognition accuracy has been measured in terms of the standard NIST metric. According to this metric errors in recognition are categorized into three types: substitutions, deletions, and insertions. A substitution is an error where the recognizer has recognized a word wrongly in an utterance. A deletion is an error where the recognizer has failed to hypothesize a word that has occurred in the utterance. An insertion is an error where the recognizer has hypothesized a word where there was no word at all in the utterance. The total error is the sum of

all three types of errors. Since the recognizer can make many insertion errors, the total number of errors can be much greater than the number of words that were actually uttered. When expressed as a percentage, this would be much greater than 100%. Accuracy is measured as 100 - (Error percentage). When errors are greater than 100%, this would become negative.

6.2 Performance of missing feature methods on speech corrupted by noise

The effectiveness of incomplete-spectrogram methods for noise compensation was evaluated by performing recognition experiments on speech corrupted by white noise. Continuous HMMs with 2000 tied states, each modeled by a single Gaussian density, were trained on the mel spectrograms of 2880 utterances of clean speech. The test set consisted of 1600 utterances from the RM test set. The utterances in the test set were corrupted by additive white Gaussian noise (AWGN) and mel spectrograms were obtained from the noisy speech. All elements of the spectrogram with a local SNR below a threshold were deleted. The optimal SNR threshold was empirically determined.

An important point to note is that in all the experiments reported in this section the local SNR of each element in the spectrogram was assumed to be known. This was possible because noisy speech signals were obtained by corrupting clean speech signals with additive noise. Thus, both the clean speech signal and the noise-corrupted speech signal were available, and therefore the spectrogram of clean speech and the spectrogram of the corresponding noisy speech could be compared to evaluate local SNR values. In a real-life situation the local SNR of noisy speech signals would not be known *a priori* and would have to be estimated. In general, this is very difficult problem in itself, and has not been satisfactorily solved, to date. We address the problem of estimating the local SNR in an unsupervised manner without the use of clean speech spectrograms in Chapter 8, and propose some solutions to the problem.

6.2.1 Obtaining the optimal threshold

The first step in applying incomplete spectrogram methods to noisy speech is that of deleting all elements of the spectrogram that have a local SNR below a particular threshold. We refer to this threshold as the *deletion threshold*. The value of this threshold affects the patterns of the missing regions and thereby the performance of the incomplete spectrogram methods. Figure 6.2 and Figure 6.3 show the variation of the recognition accuracy obtained using class-conditional imputation and marginalization respectively as a

function of the deletion threshold, on speech corrupted to a global SNR of 15 dB and 25 dB by white noise.

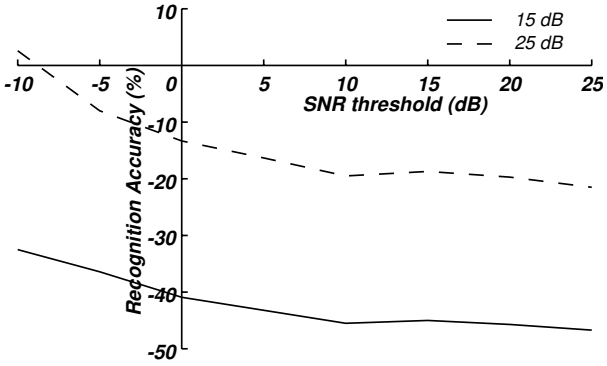


Figure 6.2 Recognition accuracy vs. deletion threshold using class-conditional imputation on speech corrupted to 15 dB and 25 dB by white noise.

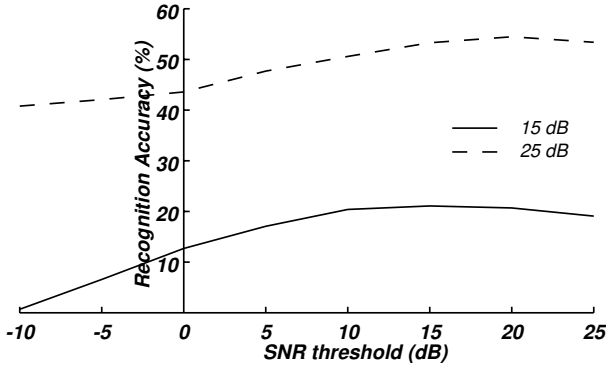


Figure 6.3 Recognition accuracy vs. deletion threshold using marginalization on speech corrupted to 15 dB and 25 dB by white noise.

Class-conditional imputation and marginalization are both seen to be extremely sensitive to the threshold. Unlike in the case of random deletions the performance of class-conditional imputation is seen to be very poor, resulting in negative recognition accuracies. Furthermore, as the deletion threshold increases in terms of SNR, the performance degrades rapidly. No optimal deletion threshold can be identified. Marginalization, on the other hand, results in positive recognition accuracies. The optimal threshold for marginalization is observed to vary with the global SNR of the speech. When the global SNR is 15 dB the optimal deletion threshold is found at 15 dB. When the global SNR is 25 dB, the optimal deletion threshold is 20 dB. However, since the difference in performance between using a 15 dB deletion threshold and a 20 dB deletion threshold is relatively small in both cases the generic deletion threshold for all noise conditions has been chosen to be 15 dB. This is the deletion threshold used in all experiments with marginalization reported later in this section. Since no optimal deletion threshold is identifiable for class-conditional imputation, and also because a minor “bump” is visible in the plots in Figure 6.3 at 15 dB, we use 15 dB as the SNR threshold for class-conditional imputation as well. We note that this is a high value for the deletion threshold since now we delete all elements from the spectrogram where the energy of the corrupting noise is even a thirtieth of that of the underlying speech. Cooke et. al. [Cooke 1999] also report that the optimal deletion threshold for marginalization found in their experiments was very high. Their estimate of the optimal value of the deletion threshold also translates to about 15 dB.

Figure 6.4 and Figure 6.5 show the recognition accuracy obtained with spectrograms reconstructed by

cluster marginal reconstruction and covariance joint reconstruction respectively as a function of the deletion threshold, on speech corrupted to a global SNR of 15 dB and 20 dB by white noise. The performance

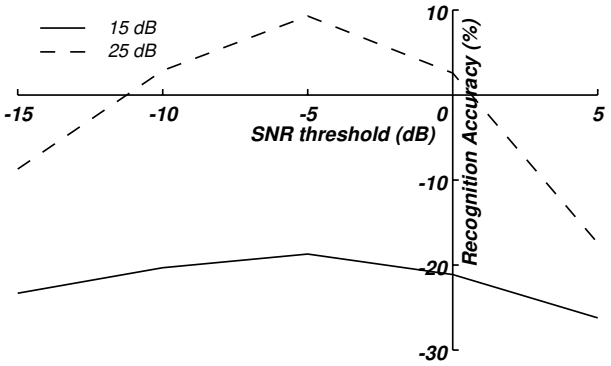


Figure 6.4 Recognition accuracy vs. deletion threshold using cluster marginal reconstruction, for speech corrupted to 15 dB and 25 dB by white noise. A codebook size of 512 was used for the reconstruction

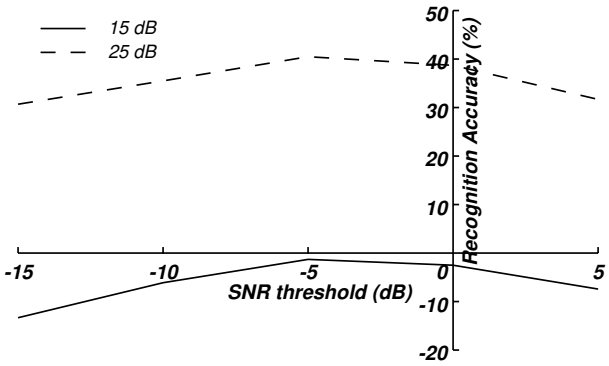


Figure 6.5 Recognition accuracy vs. deletion threshold using covariance joint reconstruction of missing elements in a vector, for speech corrupted to 15 dB and 25 dB by white noise.

obtained with geometrical reconstruction methods was extremely poor at all thresholds and is not shown. The performance of cluster time-interpolated reconstruction, although very high when the missing regions were randomly dropped, is extremely poor on speech corrupted by noise. Presumably this is because the preliminary estimate of missing regions is obtained by linear interpolation across time, a geometrical reconstruction method that also performs very poorly on noisy speech.

The optimal deletion threshold for cluster marginal reconstruction and covariance joint reconstruction is seen to be -5 dB, irrespective of the global SNR of the speech. Therefore, in the rest of this thesis this is the deletion threshold used for all spectrogram reconstruction methods evaluated.

6.2.2 Performance on noisy speech spectrograms

The effectiveness of incomplete-spectrogram methods as noise compensation techniques was measured on speech corrupted by white noise. Utterances from the RM test corpus were corrupted by white noise to a variety of SNRs. The noisy portions of the spectrograms of these utterances were deleted and incomplete spectrogram methods applied to these incomplete spectrograms. The SNR threshold used to delete noisy regions was 15 dB for all classifier-modification techniques, and -5 dB for all spectrogram reconstruction methods.

Figure 6.6 shows the recognition accuracy obtained with marginalization and class-conditional imputa-

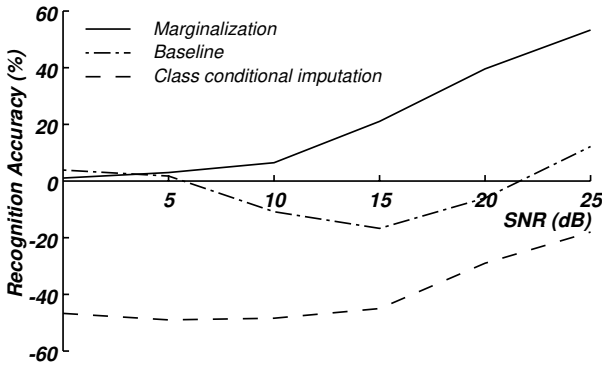


Figure 6.6 Recognition accuracy obtained with marginalization and class-conditional imputation on spectrograms of noisy speech as a function of the global SNR of the noisy speech. The baseline recognition accuracy on noisy spectrograms is also shown.

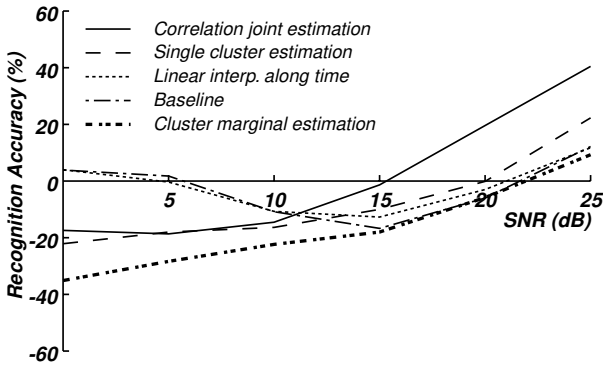


Figure 6.7 Recognition accuracy with noisy spectrograms reconstructed by several spectrogram reconstruction methods as a function of the global SNR of the noisy speech. The baseline recognition accuracy obtained with noisy spectrograms is also shown

tion on spectrograms of speech corrupted by white noise to different levels. The deletion threshold used in all cases was 15 dB. The figure also shows the recognition performance obtained when the complete noisy spectrograms are used for recognition directly (without deleting any elements). This is the performance that would normally have been obtained had no noise compensation been attempted. We refer to this situation as the *baseline.* Marginalization is seen to be a very effective compensation method resulting in large improvements over baseline recognition accuracy at all SNRs. However, class-conditional imputation is seen to be completely ineffective. This result is in variance with the results reported by Cooke et. al. [Cooke 1994], where they reported improvements even with class-conditional imputation, albeit on a different task.

Figure 6.7 shows the recognition accuracy obtained with several spectrogram reconstruction methods. Linear interpolation along time, single cluster reconstruction, cluster marginal reconstruction and covariance joint reconstruction are all represented. The recognition accuracies obtained with cluster time-interpolated reconstruction is not shown since its performance was far inferior to those of the methods represented here, at all SNRs.

Covariance-based reconstruction is seen to result in improvements in recognition accuracy at most SNRs. Single cluster reconstruction results in improvements at some SNRs. Multiple-cluster-based reconstruction methods are, however, seen to be ineffective in general. Reconstruction methods that involve any

geometrical reconstruction, *i.e.* linear interpolation along time, and cluster time-interpolated reconstruction, are observed to perform even more poorly. In general, the improvement obtained using spectrogram reconstruction methods over baseline is not large. Comparison of Figures 6.6 and 6.7 also shows that the performance obtained with spectrogram reconstruction methods is significantly poorer on noise-corrupted speech than that obtained with marginalization.

However, it has been observed that the CMU Sphinx-III, which has been used in these experiments, generates a large number of insertion errors when recognition of noisy speech is performed in the log spectral domain. Simply put, the recognizer tends to hypothesize many more words than actually occur in the utterance. These insertions are usually enumerated as errors. It has also been observed that the problem of the large numbers of insertions is not usually present when recognition is performed with cepstra instead of log spectra. Figure 6.8 shows the recognition accuracies obtained with cepstra derived from spectrograms reconstructed with several spectrogram reconstruction methods. It also shows the baseline performance obtained when recognition is performed with cepstra obtained directly from the spectrograms of noisy speech (without any compensation). We observe that significant improvement in recognition accuracy is

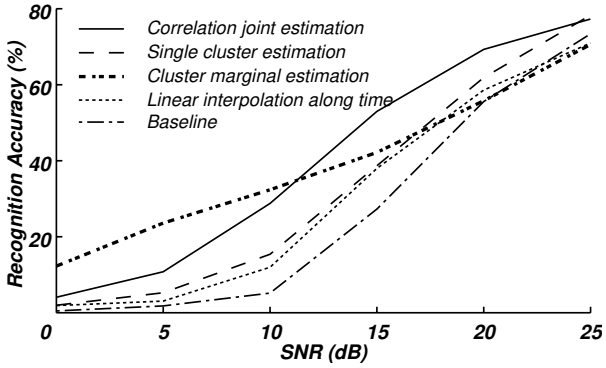


Figure 6.8 Recognition accuracy obtained using cepstra derived from spectrograms reconstructed by four spectrogram reconstruction methods, at several SNRs. Baseline recognition accuracy with cepstra derived directly from the noisy spectrograms is also shown.

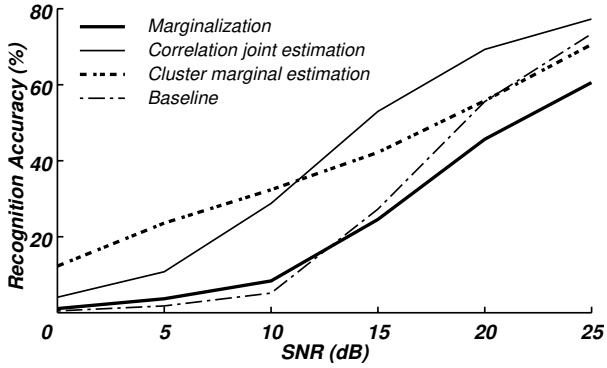


Figure 6.9 Comparison of recognition accuracies in the cepstral domain, obtained with the best cluster-based and covariance-based reconstruction methods, with the recognition accuracy obtained using marginalization in the log-spectral domain.

obtained over baseline with all the spectrogram reconstruction methods. Even simple linear interpolation along time results in improvements in recognition accuracy at all SNRs. Interestingly, the performance of cluster-based reconstruction is superior to that of covariance-based reconstruction at low SNRs, but the situation gets reversed at higher SNRs.

Figure 6.9 compares the recognition accuracy obtained in the cepstral domain with the best cluster-based and covariance-based reconstruction methods - cluster marginal reconstruction, and covariance joint reconstruction - with the performance obtained in the log-spectral domain with the best classifier-modification method, marginalization (since marginalization cannot be performed in the cepstral domain). We observe that the performance of spectrogram reconstruction methods in the cepstral domain is significantly superior to that of marginalization in the log-spectral domain.

6.2.3 Computational complexity of incomplete spectrogram methods

The computational complexity of an algorithm is a measure of the number of mathematical operations required by the computer to perform it. The greater the complexity of the method, the greater the number of operations required, and therefore the greater the amount of time needed to perform it. Ideally we would require any noise compensation algorithm to be minimally complex and take very little computation time.

To be able to compare the computational complexity of the various incomplete-spectrogram methods accurately we would need to know the precise number of additions, multiplications, and other mathematical operations required to perform them. However, this number is not a constant for any of these methods for several reasons:

1) The size of the covariance matrices being inverted in the MAP estimation procedure used by the spectrogram reconstruction methods is not constant and varies from vector to vector. Consequently, the number of multiplications needed to invert these matrices is not a constant number.
2) In cluster marginal reconstruction, the number of mathematical operations required to marginalize out missing components is dependent on the number of elements missing in any vector. This is not a constant.
3) The total number of missing elements in any noisy spectrogram is dependent on the characteristics of the noise corrupting the signal, and can vary from utterance to utterance.
4) The speech recognition system does not evaluate all possible hypotheses during recognition, but restricts itself to a small subset of hypotheses through a procedure called pruning [Ney 1992]. The precise number of hypotheses evaluated varies from utterance to utterance. Thus, the total number of mathematical operations performed by the recognition system is not a constant either.

As a result, the only realistic manner in which the computational complexity of any set of noise compensation algorithms can be compared is on the basis of the total time taken to recognize an utterance,

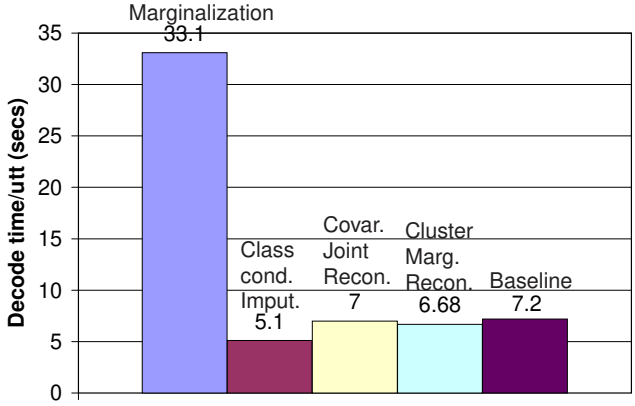


Figure 6.10 Average time taken to recognize an utterance of speech corrupted by white noise to 10 dB when various incomplete spectrogram methods are applied. Recognition was performed using log spectra. The same utterances were used to obtain these numbers in each case. The utterances were 5 seconds long on average.

when these methods are applied.

Figure 6.10 shows the average time taken to recognize an utterance of speech corrupted to 10 dB by white noise, when marginalization, class-conditional imputation, cluster marginal reconstruction, and covariance joint reconstruction are used to compensate for noise. The time taken to recognize noisy speech, without any compensation, is also shown. Recognition was performed using log spectra in all cases.

When considering the numbers in Figure 6.10 it is important to note that the behavior of the recognition system is not invariant across all cases. The recognizer usually takes more time to recognize a noisy utterance than it does to recognize a clean one, because many more hypotheses are considered when the speech is noisy. As a result, the average time taken to recognize an utterance using the log spectra of noisy speech is actually longer than the time taken to recognize an utterance when spectrogram reconstruction methods are used to compensate for the noise, although the latter includes the time taken to actually estimate the missing elements in the spectrogram. Therefore, it would be incorrect to infer, based on the numbers in Figure 6.10 that the relative differences between the time taken by class-conditional imputation, covariance joint reconstruction, and cluster marginal reconstruction would remain the same at all SNRs. However, the variation in the time taken by the recognizer to recognize an utterance is not usually large enough to account for the difference between the time taken by marginalization and that taken by the other methods. It is therefore reasonable to infer that while cluster marginal reconstruction and covariance joint

reconstruction are approximately equivalent in terms of computational complexity, both of them are far less complex than marginalization.

In all cases the true SNR of the elements of the spectrograms was known *a priori*, and was used to compute the spectrographic masks. In a real situation SNRs would not be known *a priori*, and spectrographic masks would have to be estimated. The time taken to estimate these masks would also have to be considered in measuring the computational complexity of any of the incomplete spectrogram methods. However, since mask estimation would have to be performed irrespective of the method being used, the relative complexities of the various methods would not change.

6.3 Summary and conclusion

In this chapter we have evaluated the performance of various incomplete-spectrogram methods on speech corrupted by noise. We have found that the optimal SNR threshold for deleting noisy elements of spectrograms is greater for classifier-modification methods than it is for spectrogram reconstruction methods. For classifier-modification methods the optimal threshold is found to be around 15 dB, whereas for spectrogram reconstruction methods it is -5 dB. When recognition is performed in the log-spectral domain we find that while marginalization is very effective in compensating for noise, among spectrogram reconstruction methods only covariance-based reconstruction is effective. It is found, however, that the recognizer makes a large number of “insertion” errors when recognition of noisy speech is performed in the log-spectral domain, which account for the bad performance of the spectrogram reconstruction methods. When recognition is performed in the cepstral domain using cepstra obtained from the spectrograms reconstructed by the these methods, significant improvements are obtained over baseline. The recognition performance obtained using cepstra derived from the reconstructed spectrograms is also superior to the best performance obtained with classifier-modification methods. This reaffirms our hypothesis that the improvement obtained by transforming the spectrograms to cepstra far outweighs the advantages of the optimal classification performed by classifier-modification methods.

It was also observed that spectrogram reconstruction methods are rather less computationally expensive than the best classifier-modification method, marginalization. This is an additional advantage to using spectrogram reconstruction methods over classifier-modification methods.

In all of the methods described thus far in this thesis, the noisy regions of the spectrogram have been completely erased and treated as totally unknown. However, in most situations where speech has been corrupted by noise, even the noisy regions of the spectrogram retain some information about the true value of the spectrogram at that point. In the case of additive noise they give us an upper bound on the true value. This information can be exploited to improve the performance of missing-feature methods even further.

The next chapter deals with the subject of missing-feature methods that exploit the information in noisy regions of the spectrogram to improve recognition performance.

Chapter 7
Recognition using spectrograms with unreliable data

7.1 Introduction

In Chapter 3 we described how the effect of corrupting noise on speech can be modeled by the deletion of elements in the spectrogram of the corrupted speech signal. In this chapter we extend this approach to tag noise corrupted regions of the spectrogram as “unreliable” instead of deleting them from the spectrogram entirely. The implication of tagging an element as being “unreliable” is that the observed value of the element is not necessarily the same as its true value although it may be related to the true value in some manner. The spectrograms resulting from such tagging are still *incomplete* in the sense that they have several elements whose true value is unknown. However, the relation between the observed value of these elements and their true values provides some information regarding the true value. In the event that the relationship between the values of the unreliable elements and their actual uncorrupted values is completely unknown, or that the values of the unreliable elements are completely independent of the uncorrupted value of the elements, the observed values of the elements convey no information and the elements can be treated as missing. The advantage with denoting components as “unreliable”, instead of deleting them altogether, is that the relation between the observed value of these components and their true value can be used in recognition, or in the estimation of these components.

In order to distinguish these spectrograms from spectrograms where nothing is known about the missing regions we refer so them as *unreliable spectrograms* (rather than incomplete spectrograms). We refer to all methods dealing with the problem of recognition based on such spectrograms as *unreliable-spectrogram methods*.

We would like to establish some definitions relating to data sets with unreliable elements before we proceed. We distinguish between the *observed* value of a data element, and the *true* value of the data element. The observed value of a data element is its measured value. On the other hand the true value of a data element is the value it would have had, had it not been corrupted in any manner. We further distinguish between reliable and unreliable data elements. A data element is *reliable* if its observed value is known with certainty to be identical to true value of the element and unreliable otherwise. We call the mechanism that renders the observed value of the data different from its true value the *unreliability mecha-*

nism. The unreliability mechanism may be any non-invertible transformation that does not permit us to infer the precise true value of the data from its observed value.

In this chapter we are specifically interested in unreliability mechanisms that ensure that the observed value of an unreliable data point is guaranteed to be greater than, or equal to its true value. We call such a mechanism as a *bounding unreliability mechanism.* Let us represent the observed value of a data set by $\boldsymbol{Y}$ and the *true* value of these data by $\boldsymbol{X}$. In reliable regions of the data set $\boldsymbol{X}$ is known to be the same as $\boldsymbol{Y}$ with certainty, and we refer to the corresponding set of data elements in these regions as $\boldsymbol{X}_r$ and $\boldsymbol{Y}_r$. In the unreliable regions $\boldsymbol{X}$ may not be the same as $\boldsymbol{Y}$ and we denote these regions as $\boldsymbol{X}_u$ and $\boldsymbol{Y}_u$

The effect of a bounding unreliability mechanism can now be written as

$$\begin{aligned} \boldsymbol{X}_r &= \boldsymbol{Y}_r \\ \boldsymbol{X}_u &\leq \boldsymbol{Y}_u \end{aligned} \tag{7.1}$$

We refer to the problem of estimating the value of $\boldsymbol{X}_u$ based on the values of $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_u$ as the *inference* of unreliable data. The MAP procedure for estimation of missing elements (Section 2.5.4) can easily be modified to estimate the true value of unreliable elements $\boldsymbol{X}_u$ when the unreliability mechanism is of the kind described in Equation (7.1). *Classification* with unreliable data, on the other hand, is the problem of identifying which of a set of classes the data $\boldsymbol{X}$ belong to, based only on $\boldsymbol{Y}$.

In speech recognition systems the effect of additive noise can be modeled as the rendering of some regions of the spectrogram unreliable. Classifier-modification methods such as marginalization and class-conditional imputation can be modified for recognition with spectrograms with unreliable regions. The spectrogram reconstruction methods described in Chapter 5 can also be modified to re-estimate the unreliable regions of corrupted spectrograms.

The next section describes the bounded MAP estimation procedure to estimate the true value of unreliable elements in a data set corrupted by a bounding unreliability mechanism. In the following section we describe how the effect of additive noise on speech can be modeled as the rendering of some regions of its spectrogram unreliable. In subsequent sections we describe how conventional classifier-modification methods and the spectrogram reconstruction methods presented in this thesis can all be modified to work

with spectrograms with unreliable regions, and how the bounded MAP estimation procedure can be applied to these cases.

7.2 Bounded MAP estimation

Consider a data set $\boldsymbol{X}$, consisting of two subsets $\boldsymbol{X}_a$ and $\boldsymbol{X}_b$ such that $\boldsymbol{X} = (\boldsymbol{X}_a, \boldsymbol{X}_b)$. Assume that the distribution of $\boldsymbol{X}$ is known and is given by $P(\boldsymbol{X})$. $\boldsymbol{X}_a$ is known and $\boldsymbol{X}_b$ is unknown. It is, however, known that $\boldsymbol{X}_b \le \boldsymbol{Y}_b$, where $\boldsymbol{Y}_b$ is the observed value of the data elements where $\boldsymbol{X}_b$ is unknown. The expression $\boldsymbol{X}_b \le \boldsymbol{Y}_b$ means that each element of $\boldsymbol{X}_b$ is less than or equal to the corresponding element in $\boldsymbol{Y}_b$. $\boldsymbol{Y}_b$ is therefore the upper bound on $\boldsymbol{X}_b$. The *a posteriori* distribution of $\boldsymbol{X}_b$ is now given by $P(\boldsymbol{X}_b | \boldsymbol{X}_a, \boldsymbol{X}_b \le \boldsymbol{Y}_b)$. The MAP estimate of $\boldsymbol{X}_b$ is therefore given by

$$\hat{\boldsymbol{X}}_b = \operatorname{argmax}_{\boldsymbol{X}_b}\{P(\boldsymbol{X}_b | \boldsymbol{X}_a, \boldsymbol{X}_b \le \boldsymbol{Y}_b)\} \tag{7.2}$$

We refer to the estimation described by Equation (7.2) as *bounded MAP estimation*. We can expand $P(\boldsymbol{X}_b | \boldsymbol{X}_a, \boldsymbol{X}_b \le \boldsymbol{U}_b)$ using Bayes' theorem to obtain

$$P(\boldsymbol{X}_b | \boldsymbol{X}_a, \boldsymbol{X}_b \le \boldsymbol{Y}_b) = \frac{P(\boldsymbol{X}_b, \boldsymbol{X}_b \le \boldsymbol{Y}_b | \boldsymbol{X}_a)}{P(\boldsymbol{X}_b \le \boldsymbol{Y}_b | \boldsymbol{X}_a)} = \begin{cases} \dfrac{P(\boldsymbol{X}_b | \boldsymbol{X}_a)}{P(\boldsymbol{X}_b \le \boldsymbol{Y}_b | \boldsymbol{X}_a)}, & \text{if } \boldsymbol{X}_b \le \boldsymbol{Y}_b \\ 0, & \text{else} \end{cases} \tag{7.3}$$

This is a constrained variant of the standard (unbounded) MAP estimate, which is given by

$$\hat{\boldsymbol{X}}_b = \operatorname{argmax}_{\boldsymbol{X}_b}\{P(\boldsymbol{X}_b | \boldsymbol{X}_a)\} \tag{7.4}$$

Comparing Equations (7.2) and (7.3) with Equation (7.4), it is easy to see that when the peak value of $P(\boldsymbol{X}_b | \boldsymbol{X}_a)$ occurs for some $\boldsymbol{X}_b \le \boldsymbol{Y}_b$, the bounded and the unbounded MAP estimates are identical. They only differ when the unbounded MAP estimate lies outside the region bounded by $\boldsymbol{Y}_b$. Figure 7.1 shows two examples of bounded MAP estimation.

When the distribution of $\boldsymbol{X}$ is Gaussian and $\boldsymbol{X}_b$ has only one component, it is easy to see that the

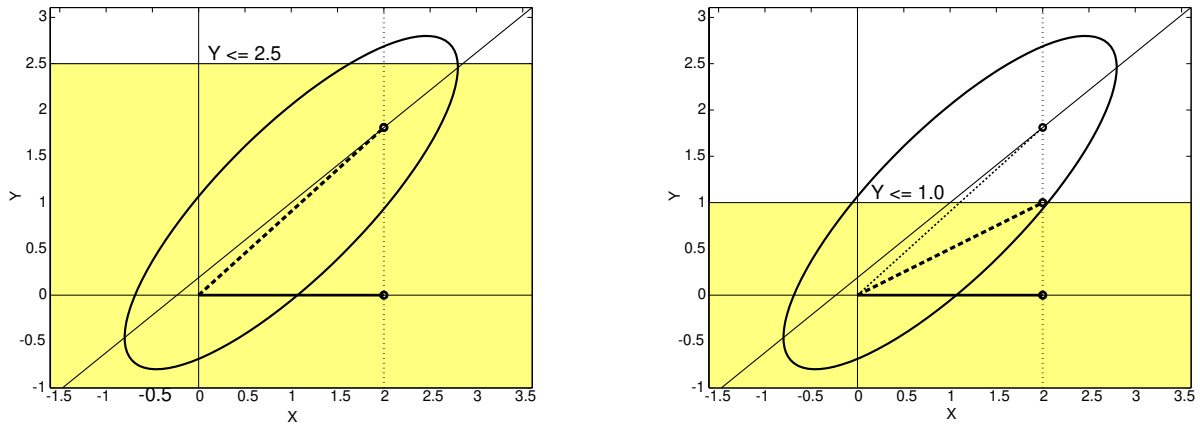


Figure 7.1 Two examples of bounded MAP estimation. In both figures the ellipse represents the cross section of the Gaussian distribution of the data. The X component of a vector has been observed and is represented by the solid line along the X axis. The Y component has not been observed and has to be estimated. The regression line representing the regular (unbounded) MAP estimates for various values of X is given shown by the diagonal line.
Left panel: The upper bound on Y is 2.5. The bounded MAP estimate of Y (and the complete vector) therefore has to lie within the shaded region and is given by the point where the distribution of all vectors with X=2 is highest, within the shaded region. In this case the regular MAP estimate of Y (given by the point where the regression line intersects the dotted vertical line at the observed value of X) lies within the shaded region. Therefore, the bounded MAP estimate for the complete vector is identical to the regular MAP estimate. This is shown by the thick dashed line.
Right panel: The upper bound on Y is 1.0. The regular MAP estimate of the complete vector (shown by the thin dotted line) lies outside the permitted region. The point where the distribution of vectors with X=2 peaks lies on the actual bound in this situation. The MAP estimate for the complete vector is shown by the thick dashed line.

bounded MAP estimate of $\boldsymbol{X}_b$ is given by

$$\hat{\boldsymbol{X}}_b = BMAP(\boldsymbol{X}_b \mid \boldsymbol{X}_a, \boldsymbol{X}_b \le \boldsymbol{Y}_b) = \begin{cases} MAP(\boldsymbol{X}_b) & \text{if } MAP(\boldsymbol{X}_b) \le \boldsymbol{Y}_b \\ \boldsymbol{Y}_b, & \text{else} \end{cases} \tag{7.5}$$

where $MAP(\boldsymbol{X}_b)$ stands for the unbounded MAP estimate of $\boldsymbol{X}_b$, and $BMAP(\boldsymbol{X}_b \mid \boldsymbol{X}_a, \boldsymbol{X}_b \le \boldsymbol{Y}_b)$ stands for the bounded MAP estimate of $\boldsymbol{X}_b$, conditioned on $\boldsymbol{X}_a$ and subject to the bound $\boldsymbol{X}_b \le \boldsymbol{Y}_b$. In other words, the bounded MAP estimate of $\boldsymbol{X}_b$ lies either on the unbounded MAP estimate or on the bound $\boldsymbol{Y}_b$. However, when $\boldsymbol{X}_b$ has more than one component the situation is more complicated. In this case the bounded MAP lies on the bounds of some of the components of $\boldsymbol{X}_b$ and these components would condition the unbounded MAP estimate of the rest of the components. Figure 7.2 explains this with an example where $\boldsymbol{X}_b$ has two components.

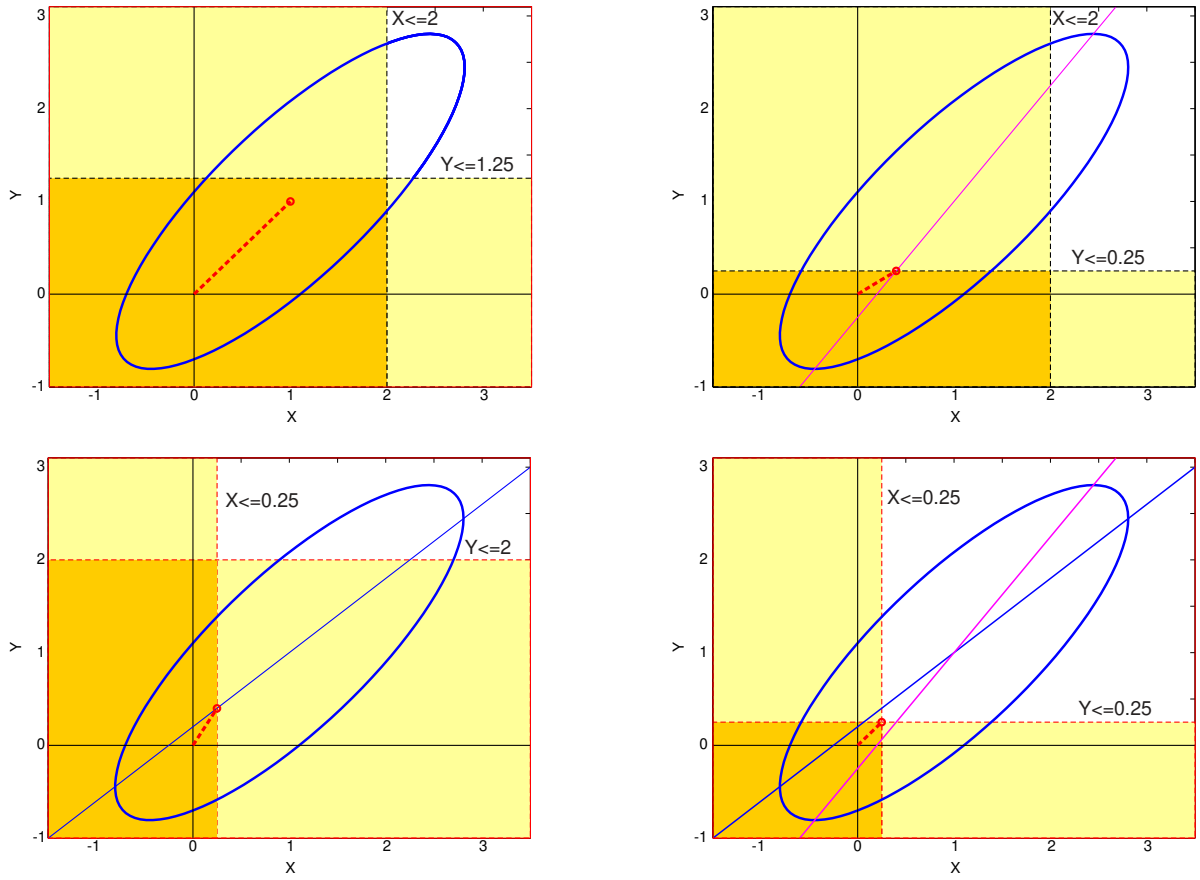


Figure 7.2 Examples of bounded MAP estimation when more than one element is to be estimated. In all cases the ellipse represents a cross section of the Gaussian distribution of the random vector. In all cases both components of the vector are unknown and are to be estimated. The regions shaded lightly represent the regions permitted by the individual bounds on X and Y. The darkly shaded region is the intersection of both bounds. All valid MAP estimates must lie in this region.

Panel 1: The bounded MAP estimate lies within the bounded region, but on neither bound

Panel 2: The diagonal line represents the regression line relating the regular MAP estimate of X to the corresponding Y. The bounded MAP estimate lies where this line intersects the Y bound

Panel 3: The diagonal line represents the regression line relating the regular MAP estimate of Y to the corresponding X. The bounded MAP estimate lies where this line intersects the X bound.

Panel 4: Here the bounded MAP estimate lies on neither regression line. Instead it lies at the point where the X and Y bounds intersect.

However, when $P(\boldsymbol{X}_b \middle| \boldsymbol{X}_a)$ is Gaussian, it can be shown [APPENDIX B] that the bounded MAP estimate of all the components of $\boldsymbol{X}_b$ can be found iteratively. Let us represent the k^{th} element of $\boldsymbol{X}_b$ as $\boldsymbol{X}_{b,\,k}$, the corresponding element of $\boldsymbol{Y}_b$ as $\boldsymbol{Y}_{b,\,k}$, and the *current* estimate of $\boldsymbol{X}_{b,\,k}$ as $\overline{\boldsymbol{X}}_{b,\,k}$. Then, initializing $\overline{\boldsymbol{X}}_{b,\,k} = \boldsymbol{Y}_{b,\,k}$, $1 \le k \le K_b$, where K_b is the total number of elements in $\boldsymbol{X}_b$, the bounded MAP esti-

mate of all the components of $\boldsymbol{X}_b$ can be found by iterating the following equation until it converges.

$$\bar{\boldsymbol{X}}_{b,k} = \text{argmax}_x\{P(x|\boldsymbol{X}_{b,k} \le \boldsymbol{Y}_{b,k}, \boldsymbol{X}_a, \bar{\boldsymbol{X}}_{b,j} \ \forall j, j \ne k)\}, 1 \le k \le K_b \tag{7.6}$$

Equation (7.6) states that if bounded MAP estimates are obtained for each of the elements of $\boldsymbol{X}_b$ iteratively, conditioned on both $\boldsymbol{X}_a$ and the current estimate of the rest of the elements of $\boldsymbol{X}_b$, the estimated value of $\boldsymbol{X}_b$ will converge to its unbounded MAP estimate. The bounded MAP estimates for individual components can be found using Equation (7.5).

The bounded MAP estimation of unreliable components of data is used to estimate unreliable regions of spectrograms in several of the methods described in this chapter.

7.3 The effect of additive noise on spectrograms

In spectrograms of noisy speech the values of noisy regions of the spectrogram, while being too noisy to be used directly for classification or recognition, are nevertheless related to the "true" value of the spectrogram in those regions (*i.e.* the value the spectrogram would have had, had the speech not been noisy). The precise relation of these values is dependent on the particular noise corruption mechanism. In this chapter we assume that the noise corrupting the speech is additive (irrespective of whether it is stationary or non-stationary) and that it is uncorrelated with the speech signal. *i.e.*

$$y[l] = s[l] + n[l] \tag{7.7}$$

where $s[l]$ represents the clean speech signal, $n[l]$ represents the corrupting noise signal, and $y[l]$ represents the observed noisy speech signal. Let us denote the spectrogram of $y[l]$ as $\boldsymbol{Y}$, the spectrogram of $s[l]$ as $\boldsymbol{S}$, and the spectrogram of $n[l]$ as $\boldsymbol{N}$. Then $Y(t, k)$, the value of $\boldsymbol{Y}$ at any point in the time-frequency plane, is related to $S(t, k)$, the value of $\boldsymbol{S}$, and $N(t, k)$ to the value of $\boldsymbol{N}$ at the same point in the time-frequency plane as [Papoulis 1991]

$$Y(t, k) = S(t, k) + N(t, k) \tag{7.8}$$

Since the spectrogram of a signal is guaranteed to be positive at all points on the time-frequency plane, it follows that

$$Y(t, k) \geq S(t, k) \tag{7.9}$$

In other words, the value of the spectrogram of the noisy speech signal gives us an upper bound on the spectrogram of the underlying clean speech signal. Using the reasoning described in Section 3.3 we can assume that all elements in the noisy spectrogram that have a relatively high local SNR are good approximations to the corresponding values in the clean spectrogram. The elements with low SNR, on the other hand, merely give us an upper bound on the value of the clean spectrogram. Considering all elements whose local SNR is greater than a threshold T as having a high SNR, our assumption gives us

$$\begin{aligned} S(t, k) &\cong Y(t, k), \quad SNR_y(t, k) > T \\ S(t, k) &\leq Y(t, k), \quad otherwise \end{aligned} \tag{7.10}$$

where $SNR_y(t, k)$ is the local SNR of $Y(t, k)$. The noisy spectrogram $\boldsymbol{Y}$ can therefore be separated into two components, $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_u$, where $\boldsymbol{Y}_r$ consists of all the regions of the noisy spectrogram $\boldsymbol{Y}$ whose SNR lies above the threshold, and $\boldsymbol{Y}_u$ consists of all the regions of $\boldsymbol{Y}$ whose SNR lies at, or below the threshold. The components of $\boldsymbol{Y}_r$ are assumed to be the *reliable regions* of the spectrogram, since they are assumed to be good approximations of the corresponding regions in the true spectrogram $\boldsymbol{S}$, which we denote as $\boldsymbol{S}_r$. The components of $\boldsymbol{Y}_u$ are the *unreliable regions* of $\boldsymbol{Y}$, since their values cannot be used to approximate the corresponding regions of $\boldsymbol{S}$, which we denote by $\boldsymbol{S}_u$. We refer to $\boldsymbol{S}_r$ and $\boldsymbol{S}_u$ as the *reliable components* of $\boldsymbol{S}$ and the *unreliable components* of $\boldsymbol{S}$ respectively. Together they represent $\boldsymbol{S}$ completely. The relation between $\boldsymbol{Y}_r$ and $\boldsymbol{Y}_u$, and the corresponding regions of $\boldsymbol{S}$, $\boldsymbol{S}_r$ and $\boldsymbol{S}_u$, is given by

$$\begin{aligned} \boldsymbol{S}_r &\cong \boldsymbol{Y}_r \\ \boldsymbol{S}_u &\leq \boldsymbol{Y}_u \end{aligned} \tag{7.11}$$

Note that the only difference between this situation and that described in Section 3.3 is that instead of erasing the regions whose local SNR lies below a threshold, we are marking them as “unreliable”. We use the same terminology as that used for the case of missing components and refer to the patterns of regions marked unreliable in the spectrogram as *deletion patterns*, or *spectrographic masks*.

As given by Equation (4.3), restated below for clarity, optimal speech recognition of an utterance is per-

formed by evaluating

$$\hat{W} = \arg\max_W \{P(\boldsymbol{S}|W)P(W)\} = \arg\max_W \{P(\boldsymbol{S}_r, \boldsymbol{S}_u|W)P(W)\} \quad (7.12)$$

where $\hat{W}$ is the estimated sequence of words in the utterance, and W is any arbitrary sequence of words. Ideally, we would want to perform recognition with the true spectrogram of the speech $\boldsymbol{S}$, *i.e.* with the true values of $\boldsymbol{S}_r$ and $\boldsymbol{S}_u$. However, we have access only to $\boldsymbol{Y}$, the spectrogram of the noisy observations (speech). In Equation (7.11) the value of $\boldsymbol{S}_r$ can be assumed to be known and equal to $\boldsymbol{Y}_r$. The value of $\boldsymbol{S}_u$ however is uncertain, and only its upper bound $\boldsymbol{Y}_u$ is known.

The similarity between Equation (7.11) and Equation (7.1) is apparent. If we were to attempt to recognize speech based directly on the relation in Equation (7.11), the problem would be that of classification with unreliable data. This would be analogous to classification with incomplete data, except that we now have the additional constraint imposed by the upper bound on the unreliable components. Alternatively, we could attempt to estimate the value of $\boldsymbol{S}_u$, based on the value of the reliable components $\boldsymbol{S}_r$, constrained to the upper bound $\boldsymbol{Y}_u$, and use this estimated value for recognition. This would be analogous to the spectrogram reconstruction methods described in Chapter 5, except that we would be inferring the *true value of unreliable elements*, rather than inferring the value of *missing* elements. We refer to these methods also as unreliable spectrogram reconstruction methods, or simply as spectrogram reconstruction methods.

7.4 Classifier modification methods: Recognizing speech directly with unreliable spectrograms

Recognition with unreliable spectrograms is similar to recognition with incomplete spectrograms. The only difference is that the upper bound on the value of the unknown elements is known.

Conventional classifier-modification methods of classification with incomplete spectrograms, *i.e.* class-conditional imputation, and marginalization, can be modified to perform classification with unreliable spectrograms. Cooke et. al. [Cooke 1999], and Josifovski et. al. [Josifovski 1999] report in detail on class-conditional imputation and marginalization with unreliable spectrograms. We describe some of these details in the following sections. More can be found in [Cooke 1999] and [Josifovski 1999].

7.4.1 Class-conditional imputation of unreliable regions in spectrograms

Class-conditional imputation of unreliable regions of spectrograms estimates the value of $\boldsymbol{S}_u$ conditioned on the upper bound $\boldsymbol{Y}_u$ and uses this estimate for recognition. The bounded MAP estimation procedure is used for the estimation. As in the case of class-conditional imputation of missing regions (Section 4.2), a separate estimate of $\boldsymbol{S}_u$ is specific to the word hypothesis being considered. Recognition is performed as

$$\hat{W} = \arg\max_W \{P(\boldsymbol{S}_r, \hat{\boldsymbol{S}}_{u,W} | W) P(W)\} \tag{7.13}$$

where $\hat{\boldsymbol{S}}_{u,W}$ is the bounded MAP estimate of the unreliable components $\boldsymbol{S}_u$,:

$$\hat{\boldsymbol{S}}_{u,W} = \arg\max_S \{P(S | \boldsymbol{S}_r, \boldsymbol{S}_u \le \boldsymbol{Y}_u, W)\} \tag{7.14}$$

where $\boldsymbol{Y}_u$ are the values of the unreliable regions of the noisy spectrogram $\boldsymbol{Y}$.

For HMM-based speech recognition systems where the best state sequence associated with the word sequence is estimated along with the word sequence, Equation (7.13) gets modified to

$$\hat{W} = \arg\max_{W,s} \{P(\boldsymbol{S}_r, \hat{\boldsymbol{S}}_{u,W,s} | \boldsymbol{s}, W) P(\boldsymbol{s} | W) P(W)\} \tag{7.15}$$

where $\boldsymbol{s} = [s_1, s_2, s_3, \ldots, s_N]$ represents any valid state sequence that can be generated by the HMM for W. $\hat{\boldsymbol{S}}_{u,W,\boldsymbol{s}}$, the estimate for $\boldsymbol{S}_u$, is given by

$$\hat{\boldsymbol{S}}_{u,W,\boldsymbol{s}} = \arg\max_S \{P(S | \boldsymbol{S}_r, \boldsymbol{S}_u \le \boldsymbol{Y}_u, s_1, s_2, s_3, \ldots, s_N)\} \tag{7.16}$$

We refer to the individual spectral vectors of the true spectrogram $\boldsymbol{S}$ as $\boldsymbol{S}(t)$, and separate the reliable and unreliable components of $\boldsymbol{S}(t)$ into $\boldsymbol{S}_r(t)$ and $\boldsymbol{S}_u(t)$, respectively. Similarly, we refer to the individual spectral vectors of the noisy spectrogram $\boldsymbol{Y}$ as $\boldsymbol{Y}(t)$ and separate the reliable and unreliable components of $\boldsymbol{Y}(t)$ into $\boldsymbol{Y}_r(t)$ and $\boldsymbol{Y}_u(t)$ respectively. The estimate of $\boldsymbol{S}_u$ can now be expressed in terms of the estimates of the individual terms $\boldsymbol{S}_u(t)$ as

$$\hat{\boldsymbol{S}}_{u,W,s} = [\hat{\boldsymbol{S}}_{u,W,s}(1), \hat{\boldsymbol{S}}_{u,W,s}(2), \hat{\boldsymbol{S}}_{u,W,s}(3), \ldots, \hat{\boldsymbol{S}}_{u,W,s}(N)] \tag{7.17}$$

where $\hat{\boldsymbol{S}}_{u,W,s}(t)$ refers to the estimates of $\boldsymbol{S}_u(t)$ when the word hypothesis being considered is W and the state sequence being considered is $\boldsymbol{s}$. Since the HMM assumes that the individual vectors of the spectrogram are independent, Equation (7.16) leads to

$$\hat{\boldsymbol{S}}_{u,W,s}(t) = \arg\max_S\{P(S|\boldsymbol{S}_r(t), \boldsymbol{S}_u(t) \le \boldsymbol{Y}_u(t), s_t)\} \tag{7.18}$$

The right hand side of Equation (7.18) is dependent only on s_t, and is independent of both the word sequence W and the complete state sequence $\boldsymbol{s}$. The implication of this is that bounded MAP estimates of the unreliable components of a spectral vector are estimated separately for each state considered during recognition, using the distribution of that state. To compute the likelihood of any state for any vector, the estimates for the unreliable components of that vector obtained using the distribution of that state are used.

We refer to the procedure of class-conditional marginalization of unreliable elements as *bounded class-conditional imputation.*

7.4.2 Marginalization of unreliable regions in spectrograms

In marginalization, recognition with unreliable spectrograms is performed directly by redefining the recognizer to use both the reliable components of the spectrogram, and the bounds on the unreliable elements. Recognition with unreliable spectrograms is performed as

$$\hat{W} = \arg\max_W\{P(\boldsymbol{S}_r, \boldsymbol{S}_u \le \boldsymbol{Y}_u|W)P(W)\} \tag{7.19}$$

$P(\boldsymbol{S}_r, \boldsymbol{S}_u \le \boldsymbol{Y}_u|W)$ is derived from $P(\boldsymbol{S}|W)$ as

$$P(\boldsymbol{S}_r, \boldsymbol{S}_u \le \boldsymbol{Y}_u|W) = \int_{-\infty}^{\boldsymbol{Y}_u} P(\boldsymbol{S}_r, \boldsymbol{S}_u|W)d\boldsymbol{S}_u = \int_{-\infty}^{\boldsymbol{Y}_u} P(\boldsymbol{S}|W)d\boldsymbol{S}_u \tag{7.20}$$

The optimal recognition would now be defined over the marginal distributions so obtained as

$$\hat{W} = \arg\max_W \left\{ P(W) \int_{-\infty}^{\boldsymbol{Y}_u} P(\boldsymbol{S}|W) d\boldsymbol{S}_u \right\} \tag{7.21}$$

For HMM based systems where the best state sequence is also estimated Equation (7.21) becomes

$$\begin{aligned} \hat{W} &= \arg\max_W \arg\max_{\boldsymbol{s}} \left\{ P(W) \int_{-\infty}^{\boldsymbol{Y}_u} P(\boldsymbol{S}|\boldsymbol{s}, W) P(\boldsymbol{s}|W) d\boldsymbol{S}_u \right\} \\ \hat{W} &= \arg\max_W \arg\max_{\boldsymbol{s}} \left\{ P(\boldsymbol{s}|W) P(W) \int_{-\infty}^{\boldsymbol{Y}_u} P(\boldsymbol{S}|\boldsymbol{s}) d\boldsymbol{S}_u \right\} \end{aligned} \tag{7.22}$$

The HMM assumption that individual vectors of the spectrogram are independent leads to

$$\begin{aligned} P(\boldsymbol{S}|\boldsymbol{s}) &= P(\boldsymbol{S}_r, \boldsymbol{S}_u|\boldsymbol{s}) = P(\boldsymbol{S}_r(1), \boldsymbol{S}_u(1), \boldsymbol{S}_r(2), \boldsymbol{S}_u(2), \ldots, \boldsymbol{S}_r(N), \boldsymbol{S}_u(N)|s_1, s_2, \ldots, s_N) \\ P(\boldsymbol{S}_r, \boldsymbol{S}_u|\boldsymbol{s}) &= \prod_{n=1}^{N} P(\boldsymbol{S}_r(n), \boldsymbol{S}_u(n)|s_n) \end{aligned} \tag{7.23}$$

Combining Equation (7.22) and Equation (7.23) we get

$$\hat{W} = \arg\max_W \arg\max_{\boldsymbol{s}} \left\{ P(\boldsymbol{s}|W) P(W) \prod_{n=1}^{N} \int_{-\infty}^{\boldsymbol{Y}_r(n)} P(\boldsymbol{S}_r(n), \boldsymbol{S}_u(n)|s_n) d\boldsymbol{S}_u(n) \right\} \tag{7.24}$$

Since the terms being multiplied in the right hand side of Equation (7.24) are dependent only on the particular state s_n, the implication of this equation is that in computing the likelihood of any state for any spectral vector during recognition, $\int_{-\infty}^{\boldsymbol{Y}_r(n)} P(\boldsymbol{S}_r(n), \boldsymbol{S}_u(n)|s_n) d\boldsymbol{S}_u(n)$ would be computed instead of $P(\boldsymbol{S}_r(n), \boldsymbol{S}_u(n)|s_n)$. Recognition would be performed using these modified likelihoods.

We refer to this procedure as *bounded marginalization* since the missing elements are marginalized only with the bound.

7.5 Compensating the data: spectrogram reconstruction methods

In this thesis we approach the problem of recognition with unreliable spectrograms as a data compensa-

tion problem. We estimate the true value of the unreliable regions of spectrograms, conditioned both on the reliable regions and the bounds on the unreliable regions. Recognition is performed with the reconstructed spectrogram, or with features derived from it.

The spectrogram reconstruction methods described in Chapter 5 can all be modified to reconstruct unreliable regions of spectrograms. In the following sub sections we describe geometrical and statistical methods of estimating the true value of the unreliable regions of spectrograms using bounding information.

We refer to spectrogram reconstruction methods applied to unreliable spectrograms as *bounded spectrogram reconstruction methods*, or simply as spectrogram reconstruction methods for brevity.

7.5.1 Geometric estimation of unreliable spectrographic components

Geometric reconstruction methods estimate missing components of incomplete spectrograms by linear or non-linear interpolation between, or extrapolation of the values of, the observed components of the spectrogram (Section 5.2). When applied to the estimation of unreliable regions these methods would have to be modified to take the upper bound on the unreliable element into account. Let $S(t, k)$, the k^{th} component of the t^{th} spectral vector be an unreliable component that has to be estimated. Let $Y(t, k)$ be the upper bound on the value that $S(t, k)$ can have. The simplest manner in which geometric reconstruction methods can be used to estimate $S(t, k)$ would be as

$$\hat{S}(t, k) = \begin{cases} geom(S(t, k)), & geom(S(t, k)) < Y(t, k) \\ Y(t, k), & else \end{cases} \tag{7.25}$$

where $geom(S(t, k))$ is the geometric estimate we would have had for $S(t, k)$, had it been missing. $\hat{S}(t, k)$ is the estimated value of $S(t, k)$. Here $geom(S(t, k))$ could be any of the geometrical reconstruction methods described in Section 5.2 such as linear interpolation across frequency or linear interpolation across time.

However, when the deletion pattern (*i.e.* the pattern of unreliable regions in the spectrogram) has been induced by noise it tends to be related to the energy in the signal. For example, when speech is corrupted by white noise all low energy regions would be marked as unreliable, while all high energy regions sur-

rounding these low energy regions would be marked “reliable”. In this situation, when the values of the low energy regions are estimated by interpolation between the high energy regions, the interpolation-based estimate almost always lies above the bound (or the observed value of the unreliable regions) and gets replaced by the unreliable value itself. As a result, estimates of the unreliable regions frequently become the observed values of the unreliable regions themselves. Figure 7.3 explains this with an example. As a result, we do not, in general, expect geometrical reconstruction methods to be effective on speech corrupted by noise.

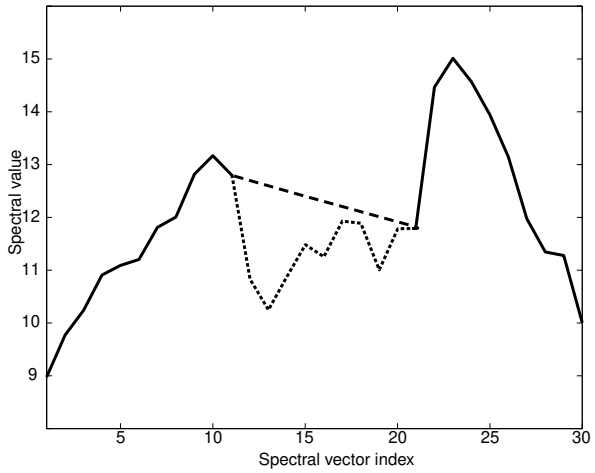


Figure 7.3 Plot showing an example of bounded linear interpolation along time. The dotted region represents the unreliable region that has to be estimated. The dashed line represents the standard estimate obtained by linear interpolation along time. All the observed unreliable values lie below the linear interpolation based estimate. As a result, the bounded estimates are simply the original values themselves when the estimate in Equation (7.25) is used.

7.5.2 Cluster-based reconstruction of unreliable regions

Cluster-based reconstruction of missing regions of spectrograms was explained in detail in Section 5.3. We recapitulate the important points in brief here. In cluster-based reconstruction of missing regions of spectrograms (Section 5.3) we use the distributions of the spectral vectors to estimate missing regions of spectrograms. Each spectral vector is assumed to be independent of every other vector. Vectors are assumed to be segregated into a number of clusters. The distribution of the vectors belonging to each of the clusters (the distribution of the cluster) is further assumed to be Gaussian. The overall distribution of vectors is given by

$$P(S) = \sum_{k=1}^{K} c_k (2\pi |\Theta_k|)^{-\frac{d}{2}} e^{-\frac{1}{2}(S-\mu_k)^T \Theta_k^{-1} (S-\mu_k)} \tag{7.26}$$

where S represents any spectral vector, d is the dimensionality of the vector, K is the total number of clusters in the distribution (codebook size), c_k is the *a priori* probability that S belongs to the k^{th} cluster, and μ_k and Θ_k are the mean vector and the covariance matrix respectively of the distribution of vectors belonging to the k^{th} cluster. In order to estimate the missing regions of incomplete spectral vectors, the cluster that the vector belongs to is identified, and the distribution of that cluster is used to obtain MAP estimates of missing regions.

Cluster-based reconstruction techniques can be easily modified to estimate unreliable regions of spectrograms (rather than missing regions). The modifications needed are somewhat different when the distribution of spectral vectors is represented by a single cluster, from when it is represented by multiple clusters.

Single cluster based estimation of unreliable regions: In single cluster based reconstruction all spectral vectors are assumed to belong to one cluster which is represented by a single distribution. The parameters of this distribution are simply the global mean and covariance of all spectral vectors. Thus the distribution of spectral vectors is given by

$$P(S) = (2\pi |\Theta|)^{-\frac{d}{2}} e^{-\frac{1}{2}(S-\mu)^T \Theta^{-1} (S-\mu)} \tag{7.27}$$

where μ and Θ are the global mean and covariance of spectral vectors. $\boldsymbol{S}_u(t)$, the unreliable components of the t^{th} spectral vector $\boldsymbol{S}(t)$, can now be estimated simply as the bounded MAP estimate

$$\hat{\boldsymbol{S}}_u(t) = BMAP(\boldsymbol{S}_u(t) | \boldsymbol{S}_r(t), \boldsymbol{S}_u(t) \le \boldsymbol{Y}_u(t)) \tag{7.28}$$

where $\hat{\boldsymbol{S}}_u(t)$ is the estimate of $\boldsymbol{S}_u(t)$, $\boldsymbol{S}_r(t)$ is the vector of reliable components of $\boldsymbol{S}(t)$, and $\boldsymbol{Y}_u(t)$ is the upper bound on the value of $\boldsymbol{S}_u(t)$. The bounded MAP estimate for the unreliable regions can be obtained using the iterative procedure described in Section 7.2.

We refer to this procedure as *bounded single cluster reconstruction*.

Multiple cluster based reconstruction of unreliable regions: In multiple-cluster-based reconstruction, the distribution of spectral vectors is modeled by multiple clusters. *i.e.* the codebook size (K in Equation (7.26)) is greater than 1. Here, reconstruction proceeds in two steps. In the first step the cluster that the vector belongs to, *i.e.* the *cluster membership* of the vector, is identified. In the second step the unreliable regions of the vector are estimated using the distribution of the cluster. Each of these steps has to take the upper bound on the value of the unreliable region into consideration. There are two ways of incorporating this bound into the estimation:

1) Bounded marginalization based estimation

2) Preliminary estimate based estimation

The following subsections describe each of these methods in greater detail.

7.5.2.1 Bounded marginalization based estimation

In bounded marginalization based estimation we estimate the cluster membership of vectors with unreliable components as

$$\hat{k}_{S(t)} = \arg\max_k \{P(\boldsymbol{S}_r(t), \boldsymbol{S}_u(t) \le \boldsymbol{Y}_u(t) \big| k) P(k)\} \tag{7.29}$$

where $\hat{k}_{S(t)}$ is the *estimated* cluster membership of $\boldsymbol{S}(t)$, $\boldsymbol{S}_r(t)$ and $\boldsymbol{Y}_u(t)$ is the upper bound on the value of $\boldsymbol{S}_u(t)$. $P(\boldsymbol{S}_r(t), \boldsymbol{S}_u(t) \le \boldsymbol{Y}_u(t) \big| k)$ has to be obtained by integrating the distribution of cluster as

$$P(\boldsymbol{S}_r(t), \boldsymbol{S}_u(t) \le \boldsymbol{Y}_u(t) \big| k) = \int_{-\infty}^{\boldsymbol{Y}_u(t)} P(\boldsymbol{S}_r(t), \boldsymbol{S}_u(t) \big| k) d\boldsymbol{S}_m(t) = \int_{-\infty}^{\boldsymbol{Y}_u(t)} P(\boldsymbol{S}(t) \big| k) d\boldsymbol{S}_u(t) \tag{7.30}$$

Equation (7.30) is similar to obtaining the marginal distribution of $\boldsymbol{S}_r(t)$, except that instead of integrating $\boldsymbol{S}_u(t)$ from minus infinity to infinity, we only integrate it up to the bound $\boldsymbol{Y}_u(t)$. Hence we refer to this procedure as *bounded cluster marginal reconstruction*.

When the cluster distributions are Gaussian, Equation (7.30) cannot generally be solved easily, especially when $\boldsymbol{Y}_u(t)$ has more than one component (*i.e.* when more than one component of $\boldsymbol{S}(t)$ is unreli-

able). However, when the covariance matrix of the Gaussian distribution of the cluster is assumed to be diagonal (*i.e.*when the various elements of the spectral vectors *within any cluster* are assumed to be independent of each other) the problem becomes simpler. In this case, if we represent the individual components of $\boldsymbol{S}_u(t)$ as $S_u(t, l)$, and the components of $\boldsymbol{Y}_u(t)$ as $Y_u(t, l)$,

$$\begin{aligned} \boldsymbol{S}_u(t) &= [S_u(t, 1), S_u(t, 2), \ldots, S_u(t, U)] \\ \boldsymbol{Y}_u(t) &= [Y_u(t, 1), Y_u(t, 2), \ldots, Y_u(t, U)] \end{aligned} \tag{7.31}$$

where U is the number of components in $\boldsymbol{S}_u(t)$ (*i.e.* the total number of unreliable components in $\boldsymbol{S}(t)$). We now have

$$P(\boldsymbol{S}(t)\big|k) = P(\boldsymbol{S}_r(t)\big|k)P(S_u(t, 1)\big|k)P(S_u(t, 2)\big|k)\ldots P(S_u(t, U)\big|k) \tag{7.32}$$

where the distribution of each of the components is also a Gaussian given by

$$P(S_u(t, l)\big|k) = \frac{1}{\sqrt{2\pi\sigma_k^2(l)}}\exp\left(-0.5\frac{(S_u(t, l) - \mu_k(l))^2}{\sigma_k^2(l)}\right) \tag{7.33}$$

where $\mu_k(l)$ and $\sigma_k^2(l)$ are mean and variance of $S(t, l)$, given that it belongs to the k^{th} cluster. Equation (7.30) now simply becomes

$$P(\boldsymbol{S}_r(t), \boldsymbol{S}_u(t) \le \boldsymbol{Y}_u(t)\big|k) = P(\boldsymbol{S}_r(t)\big|k)\prod_{l=1}^{U}\int_{-\infty}^{Y_u(t, k)}\frac{1}{\sqrt{2\pi\sigma_k^2(l)}}\exp\left(-0.5\frac{(s - \mu_k(l))^2}{\sigma_k^2(l)}\right)ds \tag{7.34}$$

Each of the integral terms in the right hand side of Equation (7.34) is a form of the error function (erfc) and can be looked up from standard tables.

In order to take advantage of the simplicity of Equation (7.34), in multiple-cluster-based representations it is convenient to model the distributions of the individual clusters as having a diagonal covariance matrix.

Equation (7.34) and Equation (7.29) can now be used to estimate the cluster membership of the spectral vectors. Once the cluster membership of the vector has been estimated, the distribution of that cluster can be used to obtain a bounded MAP estimate of $\boldsymbol{S}_u(t)$, the unreliable components of the vector.

7.5.2.2 Preliminary estimate based estimation

In preliminary estimate based estimation, a preliminary estimate $\bar{\boldsymbol{S}}_u(t)$ for the unreliable components of the spectral vector $\boldsymbol{S}(t)$ is obtained by bounded linear interpolation along time as described in Section 7.5.1. The preliminary estimates of the unreliable regions are used along with the reliable components to identify the cluster membership of $\boldsymbol{S}(t)$.

$$\hat{k}_{S(t)} = \arg\max_k \{P(\boldsymbol{S}_r(t), \bar{\boldsymbol{S}}_u(t) \big| k) P(k)\} \tag{7.35}$$

Once the cluster membership of a vector is identified, the distribution of the cluster is used to obtain bounded MAP estimates of $\boldsymbol{S}_u(t)$, the unreliable components of the vector.

7.5.3 Covariance-based reconstruction of unreliable regions

Covariance-based reconstruction methods assume that the sequence of spectral vectors that constitute a spectrogram are the output of a Gaussian wide-sense stationary (WSS) random process. We recapitulate the salient points for this model for convenience.

In a WSS process the expected value (the mean) of the k^{th} element of a t^{th} spectral vector $\mu(t, k)$ is independent of where the vector occurs in the spectrogram. The covariance between the $k_1{}^{\text{th}}$ element of the $t_1{}^{\text{th}}$ spectral vector $S(t_1, k_1)$ and the $k_2{}^{\text{th}}$ element of the $t_2{}^{\text{th}}$ spectral vector $S(t_2, k_2)$, $c(t_1, t_2, k_1, k_2)$ is only dependent on the distance between the two vectors, and not on their actual positions in the spectrogram.

$$\begin{gathered} \mu(t, k) = \mu(t + \tau, k) = \mu(k) \\ c(t_1, t_2, k_1, k_2) = c(t_1 + \tau, t_2 + \tau, k_1, k_2) = c(\tau, k_1, k_2) \end{gathered} \tag{7.36}$$

Since the random process is assumed to be Gaussian, the joint distribution of any subset of components in a sequence of spectral vectors is Gaussian. This permits us to estimate the values of the unreliable components of a spectrogram using the bounded MAP estimation procedure for Gaussian distributions described in Section 7.2. The unreliable elements in the spectrogram can either be individually estimated, or jointly estimated. In the following subsections we describe the individual and joint estimation of unreliable elements in a spectral vector.

7.5.3.1 Estimation of individual unreliable elements in a spectrogram

Let $S(t, k)$ be an unreliable component of the spectrogram with upper bound $Y(t, k)$. Let $\boldsymbol{S}_r(t, k)$ be a vector constructed with all those reliably observed components of the spectrogram that have a relative covariance greater than or equal to a preset threshold T with $S(t, k)$. *i.e.* $\boldsymbol{S}_r(t, k)$ is constructed of elements $S(t_i, k_i)$ as

$$\boldsymbol{S}_r(t, k) = [S(t_1, k_1)S(t_2, k_2)S(t_3, k_3)\ldots] \tag{7.37}$$

such that

$$r(t_i - t, k_i, k) \geq T \tag{7.38}$$

for all $S(t_i, k_i)$ included in $\boldsymbol{S}_r(t, k)$, where $r(t_i - t, k_i, k)$ is defined as

$$r(t_i - t, k_i, k) = \frac{c(t_i - t, k_i, k)}{\sqrt{c(t_i, k_i, k_i)c(t, k, k)}} \tag{7.39}$$

$S(t, k)$ and $\boldsymbol{S}_r(t, k)$ have a jointly Gaussian distribution. Thus, $S(t, k)$ can be estimated as the bounded MAP estimate

$$\hat{S}(t, k) = BMAP(S(t, k)|\boldsymbol{S}_r(t, k), S(t, k) \leq Y(t, k)) \tag{7.40}$$

We refer to this procedure as *bounded covariance individual reconstruction*.

7.5.3.2 Joint estimation of all unreliable elements in a spectral vector

In joint estimation of unreliable elements of vectors, we find a bounded MAP estimate for all the entire vector $\boldsymbol{S}_u(t)$ with the upper bound $\boldsymbol{Y}_u(t)$ jointly. We construct a vector $\boldsymbol{S}_r(t)$ of all elements in the spectrogram that have a relative covariance greater than a preset threshold T with at least one of the elements in $\boldsymbol{S}_u(t)$. *i.e.*, $\boldsymbol{S}_r(t)$ is constructed of elements $S(t_i, k_i)$ as

$$\boldsymbol{S}_r(t) = [S(t_1, k_1)S(t_2, k_2)S(t_3, k_3)\ldots] \tag{7.41}$$

such that

$$r(t_i - \tau, k_i, \kappa) \geq T \tag{7.42}$$

for some τ, and κ, such that

$$S(\tau, \kappa) \in \boldsymbol{S}_u(t) \tag{7.43}$$

$\boldsymbol{S}_u(t)$ and $\boldsymbol{S}_r(t)$ have a jointly Gaussian distribution. Therefore a bounded MAP estimate for $\boldsymbol{S}_u(t)$ can be obtained as

$$\hat{\boldsymbol{S}}_u(t) = BMAP(\boldsymbol{S}_u(t) | \boldsymbol{S}_r(t), \boldsymbol{S}_u(t) \leq \boldsymbol{Y}_u(t)) \tag{7.44}$$

The unreliable elements in each of the spectral vectors in the spectrogram would be jointly estimated in this manner to reconstruct the entire spectrogram. We refer to this procedure as *bounded covariance joint reconstruction*.

7.6 Experimental results

It is to be expected that recognition performance obtained with the *unreliable-spectrogram methods* described in this chapter should be superior to those obtained using the *incomplete-spectrogram methods* described in Chapters 4 and 5 since the bounding information present in the noisy observations is used in the former. In this section we report some experiments that demonstrate the validity of this assumption.

The recognition performance of all the methods described in this chapter were evaluated on speech corrupted by white noise. Continuous HMMs with 2000 tied states, each modeled by a Gaussian density, were trained on the mel spectrograms of 2880 utterances of clean speech. The test set consisted of 1600 utterances from the RM test set. The utterances in the test set were corrupted by additive white Gaussian noise (AWGN) and mel spectrograms using 20 mel filters were obtained from the noisy speech. All elements of the spectrogram with a local SNR below a threshold were marked as unreliable. The observed noisy value of these regions therefore provided the upper bound on the value of the elements in these regions.

The SNR threshold used for tagging elements are “reliable” or “unreliable” were the optimal thresholds determined in Section 6.2.1. A threshold of 15 dB was used with marginalization and class-conditional imputation. For all spectrogram reconstruction methods a threshold of -5 dB was used.

In all experiments it was assumed that the local SNR of every element in the spectrogram was known a *priori*. This was possible because the noisy speech was obtained by corrupting clean speech with white noise. Thus, the spectrograms of the clean speech and the noisy speech were both available, facilitating

computation of the local SNR in each element of the spectrogram. In a real situation (where only the noisy utterance is available) the local SNR would not be known. However, here we are only interested in evaluating the performance of the methods described in this chapter in the ideal situation where the local SNRs are known.

7.6.1 Recognition using log spectra

In this section, we compare the recognition performances of the various methods described in this chapter using a speech recognition system trained with spectrographic features. The recognition system was trained using the log spectra of clean speech. Only spectral features were used; no difference or double difference features were used. Recognition was performed either directly with the spectrograms of noisy speech with some regions tagged as unreliable using classifier-modification methods (marginalization and class-conditional imputation), or with the reestimated spectrograms (for the spectrogram reconstruction methods).

Figure 7.4 and Figure 7.5 show the recognition accuracy obtained with classifier-modification methods, bounded class-conditional imputation and bounded marginalization, as a function of the global SNR of the noisy speech being recognized and compares them with the performance obtained with regular (unbounded) class-conditional imputation and marginalization of *missing* regions in spectrograms (*i.e.* when the noisy regions are deleted, rather than being marked unreliable). In all experiments, the local SNR threshold for marking spectrographic elements as unreliable (or missing) was 15 dB. As can be seen, the tagging of regions as “unreliable” and using the bounding information present in the noisy observations of these regions results in large improvements over simply deleting these regions from the spectrogram. In all cases, recognition accuracies obtained using unreliable spectrogram methods are much greater than those obtained when recognition is performed with the noisy spectrograms directly (baseline). Similar results have been reported for these techniques by Cooke et. al. [Cooke 1999]. In fact our experiments show that bounded class-conditional imputation is a very effective algorithm whereas unbounded class-conditional imputation is not effective at all. Bounded marginalization, by virtue of being an optimal classification method, is still more effective than bounded class-conditional imputation.

Figure 7.6 shows the recognition accuracy obtained with spectrograms reconstructed by several bounded spectrogram reconstruction methods as a function of the global SNR of the noisy speech. Figure

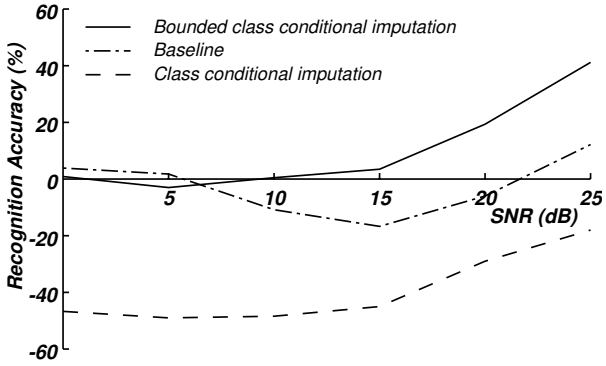


Figure 7.4 Comparison of the performance of bounded class-conditional imputation and (unbounded) class-conditional imputation on speech corrupted by white noise.

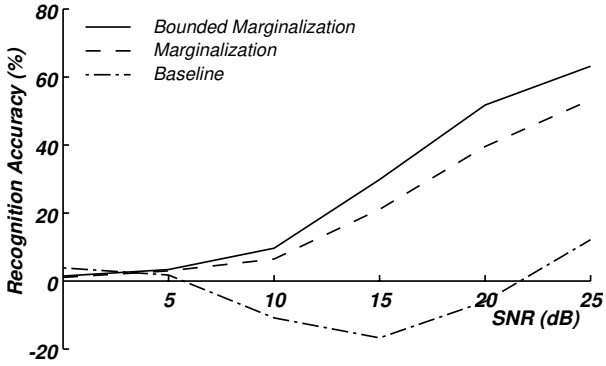


Figure 7.5 Comparison of the performance of bounded marginalization and (unbounded) marginalization on

7.7 shows the recognition accuracies obtained with the corresponding unbounded spectrogram reconstruction methods. In all experiments the local SNR for tagging elements of the spectrograms as unreliable, or missing was -5 dB. For the multiple-cluster-based methods a codebook size with 512 clusters was used.

Geometrical reconstruction methods are not effective. Cluster-based reconstruction methods with preliminary estimate based cluster membership estimation perform poorly and are not shown. However, we observe that the recognition performance obtained with statistical bounded spectrogram reconstruction methods is much better than that obtained with unbounded spectrogram reconstruction methods. It is interesting to note that the best recognition is obtained with bounded cluster marginal reconstruction, whereas unbounded cluster marginal reconstruction is not effective as a noise compensation technique. On the other hand, the difference between bounded an unbounded reconstruction is not so large either for single cluster

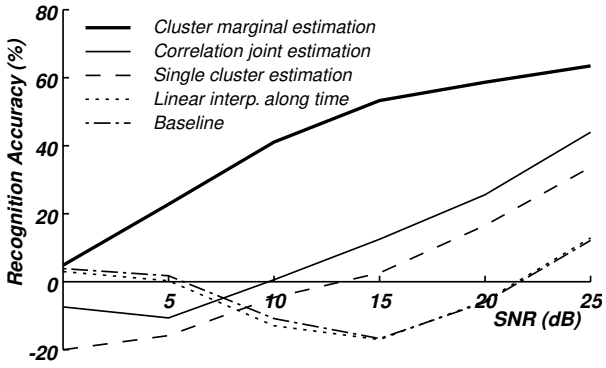


Figure 7.6 Recognition performance with spectrograms reconstructed using several unreliable spectrogram methods (bounded estimation) on speech corrupted by white noise.

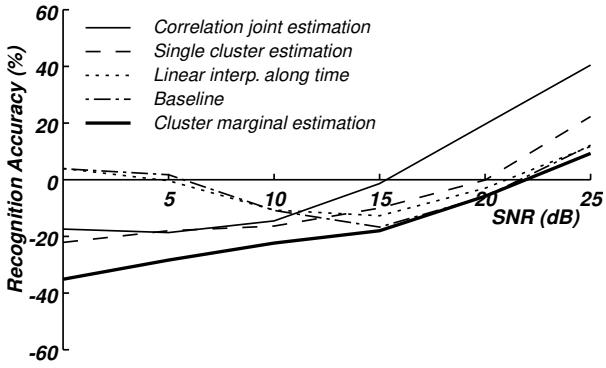


Figure 7.7 Recognition performance with spectrograms reconstructed using several incomplete spectrogram methods (using unbounded estimation) on speech corrupted by white noise.

reconstruction or for covariance-based reconstruction.

Figure 7.8 compares the best classifier-modification method, bounded marginalization, with the best reconstruction methods, bounded covariance joint reconstruction and bounded cluster marginal reconstruction. It is interesting to note that the performance achieved by bounded cluster marginal reconstruction, which is a spectrogram reconstruction method, is superior to the performance obtained with bounded marginalization, which is an optimal classification procedure. However, it may not be possible to make any inferences regarding the comparative performance of the two procedures in general since the two procedures vary in many aspects including the SNR thresholds, the fact that marginalization of unreliable elements is performed directly by the recognizer making the performance subject to the idiosyncrasies of the particular search algorithm used by the recognizer, etc.

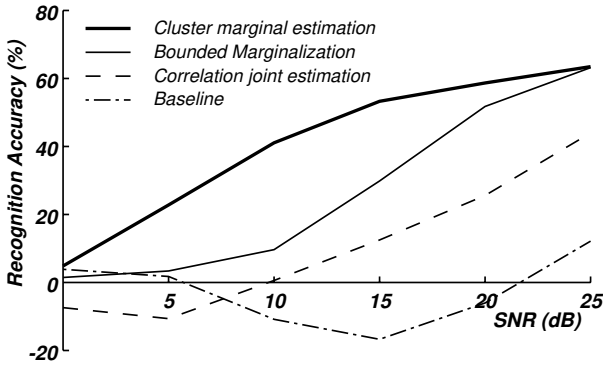


Figure 7.8 Comparison of the recognition performance of the best classifier-modification methods with performance obtained with the best spectrogram reconstruction methods on speech corrupted by white noise. Baseline recognition accuracy obtained with noisy speech spectrograms is also shown.

7.6.2 Recognition using cepstra

Recognition experiments with log spectra only give us the relative performance of classifier-modification methods and spectrogram reconstruction methods in a perfectly fair setting. However, the true test of the spectrogram reconstruction methods is the performance of recognition using cepstra derived from the reconstructed spectrograms, where much better recognition accuracies can be expected.

The experiments reported in this section were performed on a speech recognition system trained with cepstra. 13 dimensional cepstra obtained from the 20 dimensional mel-spectral vectors of clean speech were used to train the recognizer. The reconstructed spectrograms used for recognition in the experiments

reported in Section 7.6.1 were transformed to 13 dimensional cepstra for recognition. The setup used was identical to that used for the log-spectrum based experiments. Continuous HMMs with 2000 tied states, each modeled by a Gaussian density, were trained. No delta or double-delta features were used.

Figure 7.9 shows the recognition accuracy obtained with cepstra computed from the spectrograms reconstructed by spectrogram reconstruction methods. For comparison, the recognition accuracy obtained using bounded marginalization with a log-spectra-based recognizer is also shown (since marginalization cannot be performed on a cepstra-based recognizer).

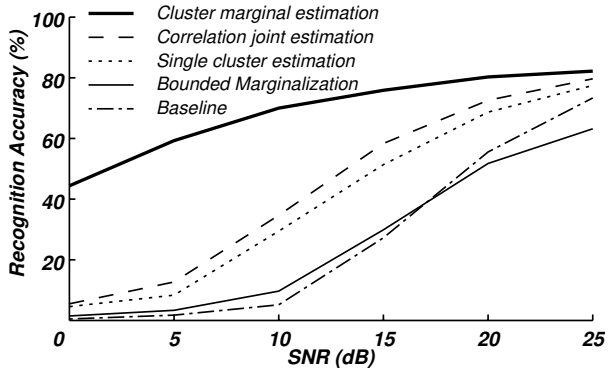


Figure 7.9 Recognition accuracy obtained with cepstra derived from spectrograms of speech corrupted by white noise reconstructed by several bounded spectrogram reconstruction methods. The performance obtained with bounded marginalization, and baseline recognition accuracy obtained with cepstra derived directly from noisy spectrograms are also shown.

We note that in the case of unreliable spectrogram methods the trends in the recognition accuracy in log-spectra-based recognition are repeated in cepstra-based recognition. Methods that result in improvement in the log-spectral domain result in improvement in the cepstral domain as well. We further note that even the simplest statistical reconstruction technique, *i.e.* single cluster reconstruction, results in better recognition accuracy overall with cepstra-based recognition than the best classifier-modification method with log-spectra based recognition. In general, the superior performance due to performing recognition in the cepstral domain outweighs the advantages of the optimal classification being performed by marginalization since the latter has to be performed in the log-spectral domain.

Among the spectrogram reconstruction methods, geometrical reconstruction is completely ineffective and is not shown. All statistical reconstruction techniques are effective. However, the relative differences between some of them are seen to be reduced. The difference between bounded single cluster reconstruction and bounded covariance joint reconstruction is much lesser when recognition is performed in the cep-

stral domain than when it is performed in the log-spectral domain. Our experiments show that bounded cluster marginal reconstruction remains by far the best method in the cepstral domain as well.

7.6.3 Computational complexity of bounded methods

The application of the bounds increases the computational complexity of all incomplete-spectrogram methods. Bounded marginalization and bounded cluster marginal reconstruction require the computation of error functions in order to obtain bounded marginal distributions of observed elements in spectral vectors. This is not required when bounds are not considered. Bounded class-conditional imputation, bounded covariance joint reconstruction, and bounded cluster marginal reconstruction require the computation of bounded MAP estimates, which can be an iterative process in the worst case and can involve several comparisons against bounds in the best case. As a result, the time taken to perform all of these methods increases.

Figure 7.10 shows the average time taken to recognize an utterance of speech corrupted to 10 dB by white noise when marginalization, class-conditional imputation, cluster marginal reconstruction, and covariance joint reconstruction are used along with bounds. Recognition was performed using log spectra in every case.

Comparison with Figure 6.10 affirms that the usage of bounds does indeed increase the computational complexity of all the methods shown in the figure. As in the case of unbounded methods, we observe that

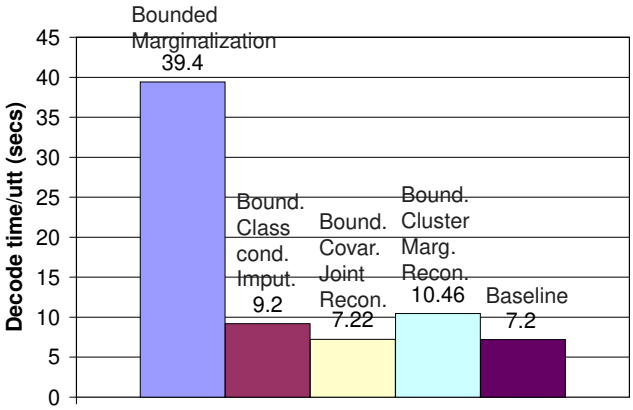


Figure 7.10 Average time taken to recognize and utterance of speech corrupted by white noise to 10 dB when various unreliable-spectrogram methods are applied. Recognition was performed using log spectra.

bounded spectrogram reconstruction methods are significantly less expensive than bounded marginalization (the best classifier-modification method). Among spectrogram reconstruction methods we note that bounded cluster marginal reconstruction is significantly more expensive than bounded covariance joint reconstruction. It was observed that bounded cluster marginal reconstruction took four times longer than covariance joint reconstruction to reconstruct spectrograms. The time taken to perform recognition with the reconstructed spectrograms was approximately the same in both cases.

7.7 Improving the reliability of the *reliable* regions of spectrograms

So far we have tagged regions of the spectrogram as being reliable if the local SNR exceeds a threshold T, and unreliable if it does not. Representing individual elements of the observed noisy spectrogram $\boldsymbol{Y}$ as $Y(t, k)$, and the individual elements of the clean spectrogram $\boldsymbol{S}$ as $S(t, k)$, we have

$$\begin{aligned} S(t,k) &\cong Y(t,k), \quad SNR_y(t,k) > T \quad (reliable) \\ S(t,k) &\le Y(t,k), \quad otherwise \qquad (unreliable) \end{aligned} \tag{7.45}$$

where $SNR_y(t, k)$ is the local SNR of $Y(t, k)$. Therefore

$$\begin{aligned} \boldsymbol{S}_r &\cong \boldsymbol{Y}_r \\ \boldsymbol{S}_u &\le \boldsymbol{Y}_u \end{aligned} \tag{7.46}$$

All the methods described so far in this chapter have attempted to deal with the uncertainty in the value of $\boldsymbol{S}_u$, assuming the value of $\boldsymbol{S}_r$ was known. The value of $\boldsymbol{S}_r$ is *approximated* by $\boldsymbol{Y}_r$. However, the elements of $\boldsymbol{Y}_r$ are not free of noise. In fact, the SNR of its elements can be as low as the SNR threshold T, which is -5 dB for the spectrogram reconstruction methods. If the value of $\boldsymbol{S}_r$ could be better approximated, the speech recognition performance of unreliable-spectrogram methods can be expected to improve. This would imply estimating $\boldsymbol{S}_r$ from the value of $\boldsymbol{Y}_r$, instead of simply approximating $\boldsymbol{S}_r$ by $\boldsymbol{Y}_r$.

Several methods have been proposed in the literature that attempt to estimate the spectrum of the underlying clean speech from the spectrum of noisy speech [Boll 1979] [Moreno 1996]. While any one of these

can be used to estimate $\boldsymbol{S}_r$ from $\boldsymbol{Y}_r$, we use spectral subtraction [Boll 1979] to do so.

Spectral subtraction is a method of canceling additive uncorrelated noise from a noisy speech signal. A running estimate of the spectrum of the corrupting noise signal is maintained, and subtracted from the power spectrum of the noisy speech. Spectral subtraction takes advantage of the fact that the transition from the non-speech regions to speech regions in any utterance is usually abrupt, indicated by a sudden increase in the energy in the signal. Thus, any quick increase in the energy in the speech signal is assumed to indicate the onset of speech. All regions deemed to be non-speech can be used to estimate the noise spectrum.

The initial portion of any utterance is assumed to contain only noise, and the spectrum of this region, *i.e.* the average of the first few spectral vectors in a spectrogram, is used to initialize the estimate of the noise spectrum. Thereafter, the estimate of the k^{th} frequency band of the noise spectrum in the t^{th} analysis window is given by

$$\hat{N}(t, k) = \begin{cases} (1-\lambda)\hat{N}(t-1, k) + \lambda Y(t, k), & \text{if}(Y(t, k) < \beta N(t, k)) \\ \hat{N}(t-1, k), & \text{otherwise} \end{cases} \tag{7.47}$$

where $Y(t, k)$ is the k^{th} frequency band of the t^{th} spectral vector of the noisy speech. λ is the *noise update factor.* β is the *threshold factor* used to identify the onset of speech. Once the estimate of the noise spectrum is known the estimate of the clean speech spectrum is obtained from the noisy spectrum as

$$\hat{S}(t, k) = Y(t, k) - \alpha\hat{N}(t, k) \tag{7.48}$$

α is an *oversubtraction factor*, and is incorporated in Equation (7.48) to account for the possibility that the noise spectrum may be underestimated. We use the simpler notation

$$\begin{aligned} \hat{S}(t, k) &= Specsub(Y(t, k)) \\ \hat{\boldsymbol{S}} &= Specsub(\boldsymbol{Y}) \end{aligned} \tag{7.49}$$

to indicate that $S(t, k)$ has been estimated from $Y(t, k)$ using spectral subtraction, as given in Equation (7.48), and that the spectrogram $\hat{\boldsymbol{S}}$ has been obtained by performing spectral subtraction on every compo-

nent of $\boldsymbol{Y}$. The relation between the true spectrogram and the noisy spectrogram can now be stated as

$$\begin{aligned} \hat{\boldsymbol{S}}_r &= Specsub(\boldsymbol{Y}_r) \\ \boldsymbol{S}_u &\leq \boldsymbol{Y}_u \end{aligned} \tag{7.50}$$

The estimate of $\boldsymbol{S}_r$ and the bound on $\boldsymbol{S}_u$ given in Equation (7.50) can be used in the unreliable spectrogram methods, instead of the relations in Equation (7.46). The only modification would be that the value associated with the "reliable" regions of the spectrogram would be $Specsub(\boldsymbol{Y}_r)$, instead of $\boldsymbol{Y}_r$. We refer to unreliable spectrogram methods that use spectral subtraction to estimate the reliable regions of spectrograms as *unreliable-spectrogram methods with spectral subtraction.* In particular, we refer to spectrogram reconstruction based methods that use spectral subtraction to estimate reliable regions of spectrograms as *spectrogram reconstruction methods with spectral subtraction*.

Figure 7.11 shows the recognition accuracies obtained with the best classifier-modification method, bounded marginalization, and with the best spectrogram reconstruction methods, bounded covariance joint reconstruction and bounded cluster marginal reconstruction, when spectral subtraction was used to improve the estimate of $\boldsymbol{S}_r$. The recognition accuracy obtained when recognition is performed directly with spectrally-subtracted speech (and no unreliable spectrogram methods are applied) is also shown. Figure 7.12 shows the absolute improvement in recognition accuracy in each of these methods due to using

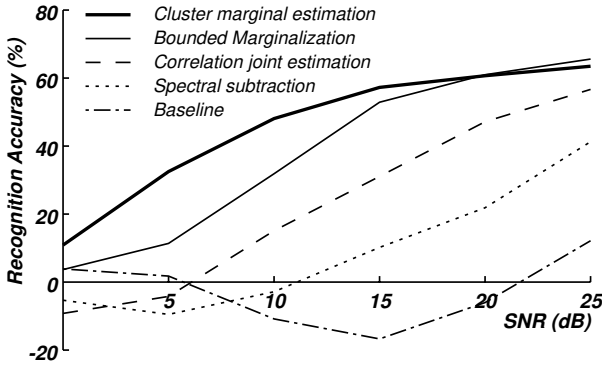


Figure 7.11 Recognition accuracy obtained with several unreliable spectrogram methods on speech corrupted by white noise, when the reliable portions of the spectrogram are estimated using spectral subtraction. The recognition accuracy obtained using spectrally-subtracted logspectra, and the baseline are also shown.

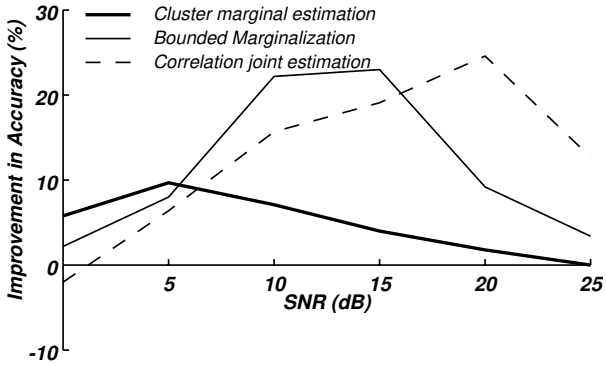


Figure 7.12 Absolute improvement in recognition accuracy due to estimating reliable portions of spectrograms using spectral subtraction. This is the difference between the recognition accuracy shown in Figure 7.11 and the recognition accuracy shown in Figure 7.8

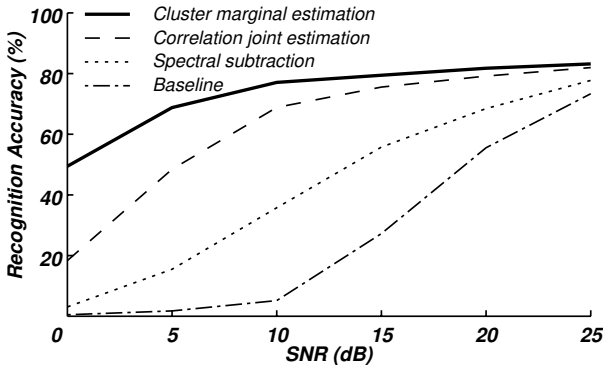


Figure 7.13 Recognition accuracy obtained with cepstra derived from spectrograms reconstructed with the combination of bounded spectrogram reconstruction methods and spectral subtraction. Recognition performance with cepstra derived directly from spectrally-subtracted speech and baseline recognition accuracy with cepstra derived from noisy speech are also shown.

the spectrally-subtracted estimate of $\boldsymbol{S}_r$ instead of approximating it with $\boldsymbol{Y}_r$.

Improving the estimate of $\boldsymbol{S}_r$ by spectral subtraction is seen to result in significant improvement in the recognition accuracy obtained with all the methods. In all cases, recognition accuracy obtained with unreliable spectrogram methods was far greater than the baseline recognition accuracy (obtained by performing recognition directly on noisy spectrograms), as well as that obtained with spectrally-subtracted speech.

Figure 7.13 shows the recognition accuracy obtained when spectrograms reconstructed by spectrogram reconstruction methods with spectral subtraction, i.e bounded cluster marginal reconstruction and bounded covariance joint reconstruction, were transformed into cepstra and recognition was performed using a cepstra-based recognizer. The baseline recognition accuracy obtained when recognition is performed directly with cepstra of noisy speech and the recognition accuracy obtained with cepstra of spectral subtracted speech are also shown for comparison. Comparison with Figure 7.9 shows that large improvements in performance are obtained by preliminary spectral subtraction of reliable regions of spectrograms, even when recognition is performed in the cepstral domain.

Overall, we see from Figure 7.13 that very large improvements in recognition accuracy are achievable when bounded spectrogram reconstruction methods with spectral subtraction are used to compensate for corrupting noise, when the local SNR of elements of the spectrogram are known *a priori*.

7.8 Recognition of speech corrupted with non-stationary noises

We had mentioned at the outset in Chapter 1 that one of the important goals of attempting to compensate for noise with missing-feature methods was to be able to compensate for non-stationary noises. However, in all the experiments reported with incomplete-spectrogram and unreliable-spectrogram methods so far in this thesis we have used stationary white noise as the corrupting signal. It is therefore important to determine the extent to which unreliable spectrogram methods are effective on speech corrupted by non-stationary noise.

Figure 7.14 shows the recognition accuracy obtained when bounded spectrogram reconstruction methods (with spectral subtraction) were applied to speech corrupted with music. The local SNR of each element of the spectrogram was assumed to be known. Recognition was performed in the cepstral domain with a recognizer trained on cepstra. The baseline recognition accuracy obtained when recognition was performed directly with the music-corrupted speech, as well as the recognition accuracy obtained with spectrally-subtracted speech are shown.

We note that unreliable spectrogram methods are highly effective on speech corrupted by music as well, when the local SNR of the elements in the spectrogram are known. For example, the recognition accuracy obtained with spectrograms reconstructed by cluster-based reconstruction when the global SNR of the noisy speech is 5 dB is very close to that obtained with clean, uncorrupted speech. Note that spectral subtraction is not effective here, due to the non-stationary nature of music. Spectral subtraction and other conventional techniques are only effective when the corrupting signal is stationary or slowly varying.

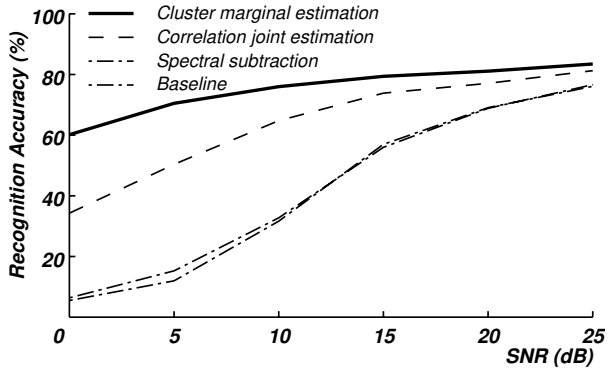


Figure 7.14 Recognition accuracy obtained when bounded spectrogram reconstruction methods are applied to speech corrupted by music to several SNRS. Baseline recognition accuracy, and recognition accuracy obtained with spectral subtraction alone are also shown. Recognition was performed in the cepstral domain in all cases.

7.9 Summary and conclusions

We have seen in this chapter that by tagging noisy regions of spectrograms as “unreliable” instead of deleting them altogether, and by appropriately modifying incomplete-spectrogram methods to consider the upper bound on the value of the spectrogram provided by the noisy observations large improvements can be made in recognition accuracy. The performance of both classifier-modification methods, and the spectrogram reconstruction methods proposed in this thesis is observed to improve significantly with this approach.

The best recognition performance obtained with the spectrogram reconstruction methods proposed in this thesis are seen to be comparable with, or better than, the performance obtained with the best current classifier-modification method, bounded marginalization, *even* when recognition is performed using log spectra. When recognition is performed using cepstra derived from reconstructed spectrograms, the performance obtained with the simplest spectrogram reconstruction method, bounded single cluster reconstruction, is superior to that obtained with bounded marginalization on log spectra. Large improvements are obtained with spectrogram reconstruction methods on speech corrupted by music as well, when the local SNR of the elements of the spectrogram are known. The performance of these methods appears to be independent of the kind of corrupting noise, once the local SNRs are known.

It is interesting to observe that the improvement in the performance of marginalization and cluster marginal reconstruction due to the usage of the upper bound on unreliable elements is far greater than the improvement in either single cluster reconstruction or covariance-based reconstruction methods. In fact, bounded marginalization based reconstruction is the most effective of all spectrogram reconstruction methods, whereas all cluster-based methods were ineffective when bounds were not used. We observe that both, marginalization and cluster marginal reconstruction involve classification of some kind. In marginalization, the optimal state sequence representing the utterance is identified. In cluster marginal reconstruction the cluster that the clean spectral vector belongs to is identified. It may therefore be inferred that the incorporation of the upper bound on the value of the unreliable elements improves classification performance far more than it improves the estimation of their values. Thus, all methods which involved classification were seen to improve much more than methods that did not involve classification of any kind.

As in the case of incomplete spectrograms, bounded geometrical reconstruction methods were com-

pletely ineffective in compensating for noise. It is clear that noisy regions of spectrograms cannot simply be estimated by simple interpolation or extrapolation of the reliable portions of the spectrogram. The prior statistical information used by the statistical reconstruction methods is essential for effective reconstruction. Among statistical reconstruction methods, bounded cluster marginal reconstruction was seen to be significantly superior to bounded covariance-based reconstruction methods. However, the latter are computationally less expensive than the former.

Recognition accuracy can be further improved by *estimating* the true value of reliable regions, instead of simply approximating them with the less noisy portions of the noisy spectrogram. Overall, large improvements in recognition performance are achievable on noise corrupted speech by using the unreliable spectrogram methods (in combination with spectral subtraction). For example, the recognition accuracy of speech corrupted by music to 0 dB goes up from less than 10% when the noisy speech is used directly for recognition to over 60% when cluster-based reconstruction of unreliable regions is performed.

However, the results reported in this chapter are all subject to the condition that the spectrographic masks that distinguish the reliable regions of the spectrogram from the unreliable ones are known perfectly. These masks were obtained using perfect knowledge of the local SNR of each of the elements in the spectrogram. As such, they only establish an upper bound on what is achieveable using missing feature methods. It is therefore more correct to say that large improvements in recognition accuracy are *potentially* achievable on noise corrupted speech by using unreliable-spectrogram methods.

In a real situation, the local SNR of the spectrographic elements would not be known and the spectrographic masks would have to be estimated. Needless to say, any procedure that estimates spectrographic masks is likely to make errors, and therefore the performance of unreliable spectrogram methods can be expected to be worse when estimated masks are used, than when masks are obtained with perfect knowledge of the local SNR of the elements of the spectrogram. Needless also to say, unreliable-spectrogram methods that do not function with perfect knowledge of the local SNR and perfect knowledge of deletion patterns cannot be expected to perform with estimated deletion patterns. Thus, geometrical reconstruction methods cannot, in general, be expected to perform well on noise corrupted speech. We do not consider them any further in this thesis.

Estimation of deletion patterns can be a very complicated task. At the greatest detail, this would entail

estimating the local SNR of each element of every spectral vector in the spectrogram. At the coarsest level, we only need to distinguish between the components of the spectrogram that are heavily corrupted and those that are relatively less corrupted. *i.e.* we only need to be able to decide whether the local SNR in the elements lies above the threshold T or below it. Even this latter estimation can be very difficult.

In the next chapter we discuss the estimation of deletion patterns, and the performance of unreliable spectrogram methods with estimated deletion patterns.

Chapter 8
Estimating the locations of corrupt regions in spectrograms

8.1 Introduction

In the preceding chapters we have described several techniques that improve the recognition performance on noisy utterances of speech by reconstructing the noisy regions of their spectrograms. We have demonstrated that considerable improvements in recognition accuracy can be obtained with these methods, even when the corrupting noise is non-stationary. However, in all the experiments reported thus far, we have assumed that the spectrographic masks that distinguish the reliable regions of the spectrogram from the unreliable regions were known *a priori*. In any real situation the true spectrographic masks would not be available. For any solution based on missing-feature methods to be complete it is therefore also necessary to estimate the spectrographic masks themselves.

We refer to the true spectrographic masks as *oracle masks*, and estimated spectrographic masks as *estimated masks*, for brevity.

The estimation of spectrographic masks only involves the estimation of very simple, binary information about every element in the spectrogram - we only need to determine whether any element is noisy or not. However, even this simple binary assessment can be a very difficult task. This is especially so when the noise corrupting the speech is non stationary. Other researchers working on missing-feature-based approaches to noise compensation have all attempted to estimate these masks based on running estimates of the spectrum of the noise [Cooke 1997][Cooke 1999], and have reported varying degrees of success with these methods, depending on the kind of noise being considered. Another popular 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 [Hirsch 1995]. The histogram of spectral elements in any frequency band over a given time window would therefore exhibit two peaks, one each representing the noisy elements and the clean elements respectively. Spectrographic masks are derived based on estimates of the noise spectra obtained as the difference in the positions of the two peaks [Cooke 1999]. No other method has been employed to identify masks to the best of our knowledge.

In this chapter we address the problem of automatically estimating the spectrographic masks for noisy

speech. We first analyze the effect of errors in the spectrographic masks. Thereafter we discuss three methods of estimating spectral masks. In the first method we use a running estimate of the spectrum of the corrupting noise to identify low SNR regions on the spectrograms. This is essentially the method described in [Cooke 1997] and [Cooke 1999]. Since the running noise estimate is obtained using the noise estimation method in spectral subtraction, we refer to this method as *spectral-subtraction-based mask estimation*. In the second method we use the noise spectrum estimate obtained by the most successful noise compensation technique in our repertoire, the vector Taylor series algorithm or VTS, to estimate spectral masks. We refer to this method as *VTS-based mask estimation*. In the third method we train a simple two class classifier to identify noisy regions of the spectrograms, and thereby the spectrographic masks. We refer to this method as *classifier-based mask estimation*. Finally we describe experimental results with spectrographic masks so obtained.

In the rest of this chapter we restrict our discussion to only two of the spectrogram reconstruction methods described so far:

1) Bounded covariance joint reconstruction
2) Bounded cluster marginal reconstruction

All analysis and experimentation has been done with these methods only. However, the results obtained are generalizable to other methods as well. Results with classifier-modification methods are not shown since, in general, baseline recognition accuracy with the cepstra of noisy speech is not significantly worse than the recognition accuracy obtained with bounded marginalization in the log-spectral domain, even with oracle masks.

In all the experiments reported in this chapter the RM database, and the CMU Sphinx-III recognition system was used as in other chapters. All recognition experiments were performed using cepstra derived from reconstructed log spectra.

8.2 The effect of errors in mask estimation

The ability of missing feature methods to compensate for noise depends critically on the accuracy of the spectrographic masks used. Errors in the spectrographic mask can cause the recognition performance of missing feature methods to degrade significantly.

Errors in spectrographic masks can be one of two kinds: reliable elements of the spectrogram may be declared unreliable, or unreliable elements may be tagged as being reliable. We refer to the first type of error as a *false alarm*. We refer to the second type of error as a *miss*. The effect of false alarms is that clean elements of the spectrogram are tagged as being noisy and are therefore reconstructed. The effect of misses is that noisy elements of the spectrogram are treated as being reliable, and are used directly for recognition. The effect of both type of errors is not the same. In the case of misses the worst case would be when all noisy elements are tagged as being clean. The recognition performance in this case (assuming the reliable regions of the spectrogram have all been tagged correctly) is simply the baseline recognition performance that is obtained with noisy speech, since the spectrograms simply remain unprocessed. The worst case scenario for false alarms, however, is much worse. Here, all reliable regions of the spectrogram would get tagged as being unreliable. Therefore, assuming that all unreliable regions of the spectrogram are correctly tagged, the spectrogram would be assumed to have no reliable elements at all. As a result neither recognition with, nor reconstruction of, the spectrograms would be possible.

The effect of false alarms and misses on the performance of missing feature methods is illustrated in Figure 8.1 and Figure 8.2. For the plot in Figure 8.1 random false alarms were introduced into the spectro-

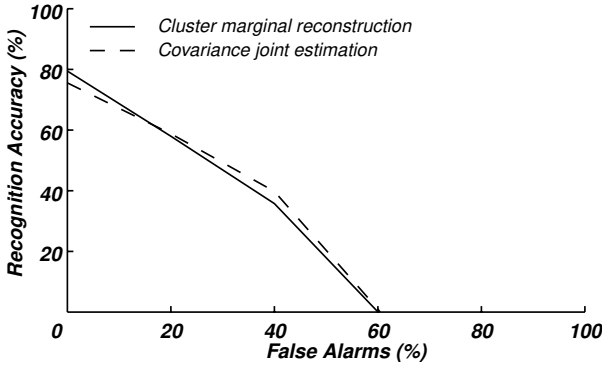


Figure 8.1 Recognition accuracy with cepstra derived from reconstructed spectrograms, as a function of the fraction of reliable elements in the spectrogram that were erroneously tagged as being unreliable

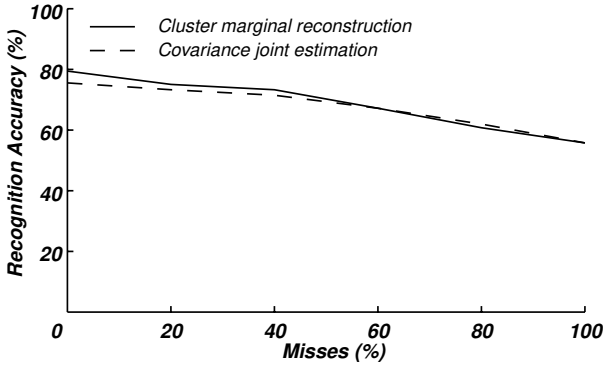


Figure 8.2 Recognition accuracy with cepstra derived from reconstructed spectrograms, as a function of the fraction of unreliable elements in the spectrogram that were erroneously tagged as being reliable

graphic mask of speech corrupted to 15 dB by white noise. No misses were introduced. The figure shows how the recognition performance of the unreliable-spectrogram methods degrades as the fraction of clean elements wrongly identified as being unreliable increases. Figure 8.2 similarly shows how their performance degrades when random misses were introduced into spectrographic masks. We observe that recog-

nition performance degrades very quickly with increasing fraction of false alarms. However, the sensitivity of all missing-feature methods to misses is not so much, and the performance degrades much more slowly as the fraction of noisy elements identified as being reliable increases.

We can infer from Figures 8.1 and 8.2 that it is critical for any algorithm that estimates the spectrographic masks of noisy speech to make minimal false alarm errors. Misses, on the other hand, are not so critical.

8.3 Estimating spectrographic masks using spectral subtraction

As mentioned in Section 7.7, spectral subtraction is a procedure that attempts to cancel additive uncorrelated noise from a noisy speech signal. To do this, a running estimate of the spectrum of the corrupting noise signal is maintained as follows: the initial portion of any utterance is assumed to contain only noise, and the spectrum of this region, *i.e.* the first few spectral vectors in a spectrogram, are used to initialize the estimate of the noise spectrum. Thereafter any sudden increase in the energy in the noisy speech signal is assumed to indicate the onset of speech and regions in the speech whose energy falls lies below a given threshold are assumed to consist only of noise. The estimate of the k^{th} frequency band of the noise spectrum in the t^{th} analysis window is given by

$$\hat{N}(t,k) = \begin{cases} (1-\lambda)\hat{N}(t-1,k) + \lambda Y(t,k), & \text{if}(Y(t,k) < \beta N(t,k)) \\ \hat{N}(t-1,k), & \text{otherwise} \end{cases} \tag{8.1}$$

The noise estimate so obtained can be used to estimate the SNR of spectrographic elements. If $Y(t,k)$ is the observed value of the k^{th} frequency band of the t^{th} spectral vector in the noisy spectrogram, the estimate of the SNR of $Y(t,k)$ would be given by

$$\overline{SNR}(t,k) = \frac{Y(t,k) - \hat{N}(t,k)}{\hat{N}(t,k)} \tag{8.2}$$

Spectrographic masks are estimated simply by tagging all elements of the spectrogram whose estimated SNR is lower than a threshold T. Variants of this method of estimating spectrographic masks have been reported in [Cooke 1997][Cooke 1999].

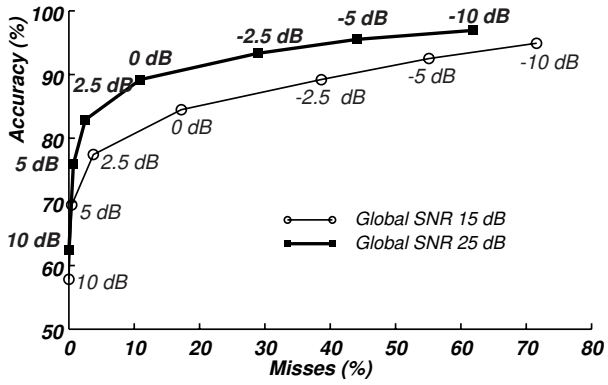


Figure 8.3 Percentage of reliable elements in the spectrogram correctly identified by the spectral-subtraction-based mask estimate as being reliable (accuracy) vs. percentage of unreliable elements falsely identified as being reliable (false alarms). The percentage of misses in the mask would be (100 - accuracy). The number beside each point indicates the deletion threshold used.

The estimate SNR given by Equation (8.2), by nature of being an estimate, is not identical to the true value of the SNR. Any spectrographic mask derived on the basis of these estimates is likely to be erroneous as well. The degree of error, as measured by the fraction of misses and false alarms in the estimated spectrogram, depends on the value of T used. It would therefore have to be carefully chosen.

Figure 8.3 plots the relation between the percentage of reliable elements correctly identified and the false alarm percentage for various values of T[1] for speech corrupted with white noise to 15 dB and 25 dB. The knee of the curve is seen to be at T = 2.5 dB for both cases. At higher thresholds the fraction of false alarms increases greatly. At lower thresholds the misses increase. T was therefore chosen to be 2.5 dB: any element $Y(t, k)$ whose local estimated SNR, $\overline{SNR}(t, k)$, was below 2.5 dB was assumed to be unreliable. Note that this threshold is different from the optimal deletion threshold for obtaining spectrographic masks when the true SNR of the spectrographic elements was known (Section 6.2.1).

8.3.1 Experimental results with spectral-subtraction-based mask estimation

In order to evaluate spectral-subtraction-based mask estimation experiments were run on speech corrupted by white noise and music to several different SNRs. Spectral-subtraction-based masks were estimated and bounded spectrogram reconstruction methods applied to these masks.

1. These were obtained by comparing the estimated spectrographic mask with the true spectrographic mask for the noisy speech.

Figure 8.4 shows an example of the estimated spectrographic mask for an utterance corrupted to 10 dB by white noise. Figure 8.5 shows the oracle (true) spectrographic mask for the same utterance. Visual comparison of the two figures shows that the estimated mask resembles the oracle mask, at least at a gross level. Figure 8.6 shows the recognition accuracy obtained with unreliable spectrogram methods on speech corrupted by white noise using estimated masks. Figure 8.7 shows the recognition accuracies obtained on the same utterances when oracle masks were used with these missing feature methods. We note that the recognition accuracy obtained with the estimated masks is much greater than the baseline recognition accuracy obtained with the cepstra of noisy speech. This is indicative that spectral-subtraction-based mask

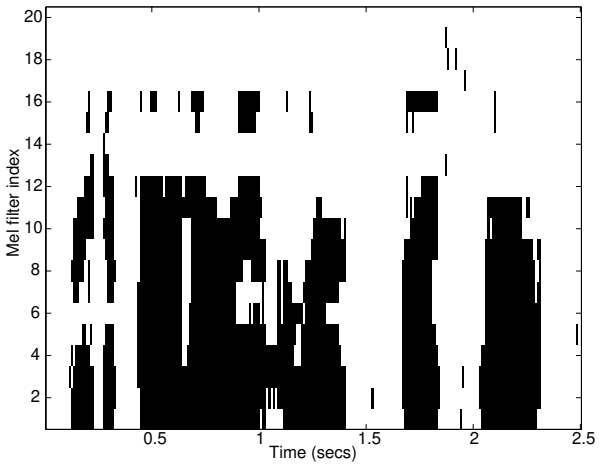


Figure 8.4 Spectrographic mask estimated using spectral-subtraction-based estimation for an utterance of speech corrupted to 10 dB by white noise.

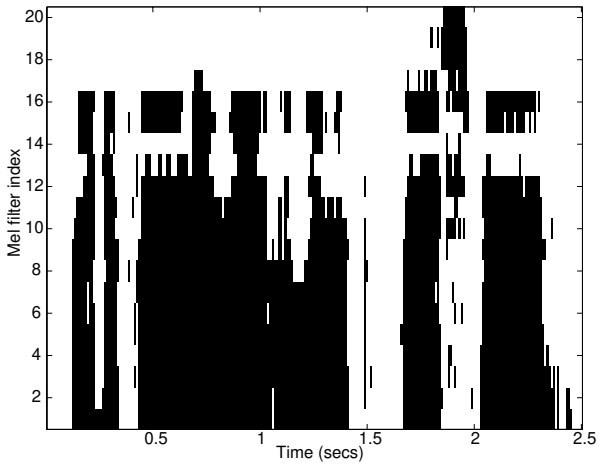


Figure 8.5 Oracle spectrographic mask for the same utterance.

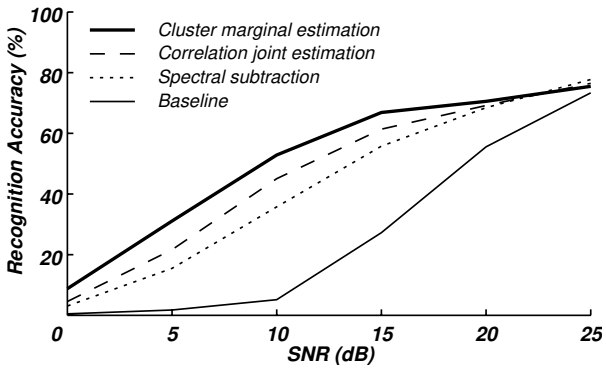


Figure 8.6 Recognition accuracy obtained by applying incomplete spectrogram methods with spectrographic masks estimated by spectral-subtraction-based estimation, for speech corrupted by white noise. Baseline recognition accuracy for the noisy speech, and the performance obtained when only spectral subtraction is used to compensate for the noise are also shown.

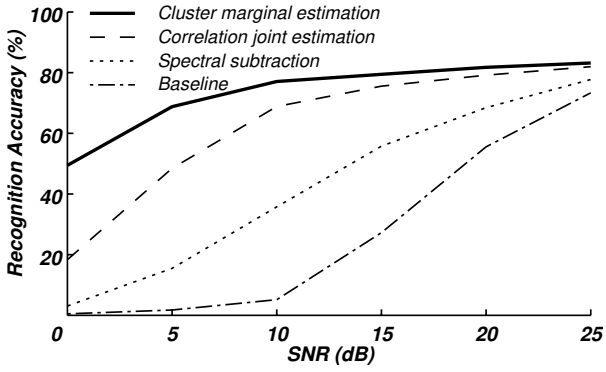


Figure 8.7 Recognition accuracy obtained when incomplete spectrogram methods are used with oracle masks to compensate for additive white noise.

estimation can be effective on speech corrupted by white noise. In general, it can be expected that spectral-subtraction-based estimation of spectrographic masks will be effective in situations where spectral subtraction itself is effective. Spectral subtraction is known to be effective when the noise corrupting the speech is stationary or slowly varying. It can therefore be expected that for such noises spectrographic masks can be estimated and missing feature methods can effectively be used to compensate for the effect of the noise on speech recognition systems.

However, Figures 8.6 and 8.7 also show that the recognition accuracy obtained with the estimated masks is much poorer than that obtained with oracle masks, especially at low SNRs. There is, therefore, considerable scope for improvement in the masks even when the corrupting noise is white.

Figures 8.8 and 8.9 show the estimated mask and the oracle mask for an utterance of speech that has been corrupted to 10 dB by music. It is clear from these figures that spectral subtraction is completely unable to estimate the mask when the corrupting noise is music. Figure 8.10 shows the recognition accuracy obtained with spectrogram reconstruction methods on speech corrupted by music, when estimated masks are used. Figure 8.11 shows the recognition performance obtained on the same utterances when oracle masks are used. Spectrogram reconstruction methods are completely ineffective at compensating for music when the estimated spectrographic masks are used. Once again, it is clear from these figures that spectral subtraction is completely ineffective as a mask estimation method when the corrupting signal is music.

8.4 Estimating spectrographic masks with VTS

Vector Taylor Series (VTS) is a noise compensation algorithm that attempts to reduce the effect of linear filtering and additive noise on the log-spectral vectors of noisy speech [Moreno 1996]. If $\boldsymbol{Y}(t)$ represents the t^{th} log spectral vector for the utterance that has been corrupted by linear filtering and additive noise, and $\boldsymbol{X}(t)$ is the value that would have been observed had the speech not been corrupted in any manner, then it can be shown that the relation between the two is given by [Acero 1991]:

$$\boldsymbol{Y}(t) = \boldsymbol{X}(t) + \boldsymbol{H} + \log(\boldsymbol{N} - \boldsymbol{H} - \boldsymbol{X}(t)) \tag{8.3}$$

where $\boldsymbol{H}$ is the logarithm of the squared magnitude of the spectrum of the impulse response of the linear

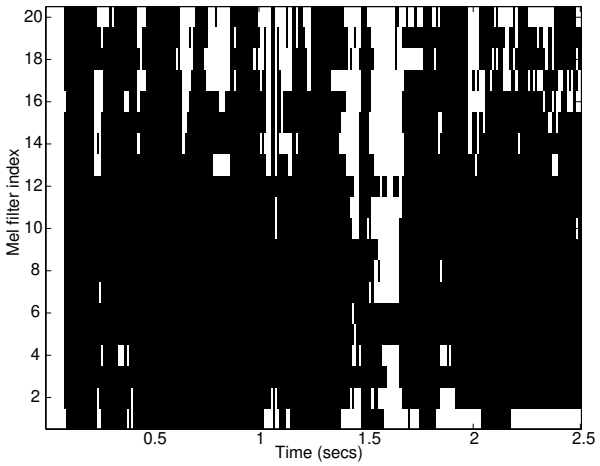


Figure 8.8 Spectrographic mask estimated using spectral-subtraction-based estimation for an utterance of speech corrupted to 10 dB by music.

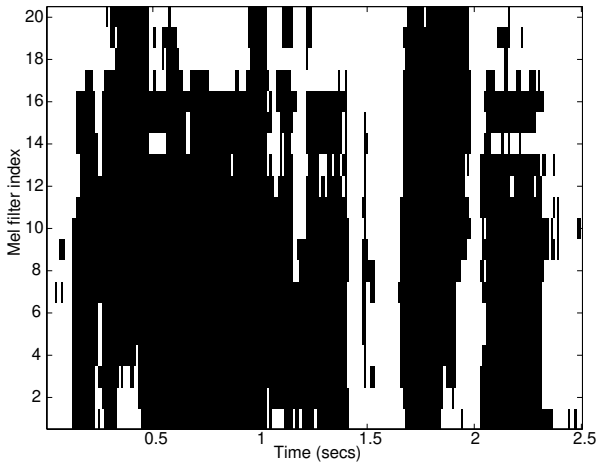


Figure 8.9 Oracle spectrographic mask for the same utterance.

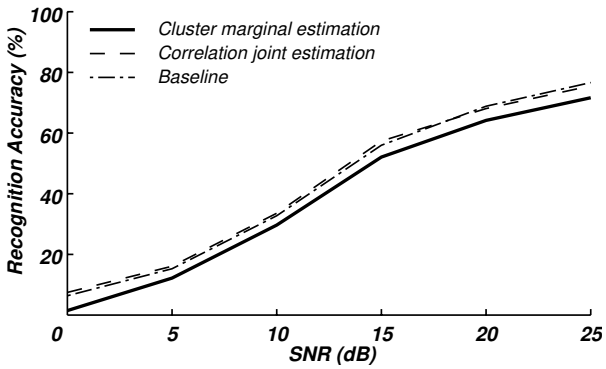


Figure 8.10 Recognition accuracy obtained with spectrographic masks estimated by spectral-subtraction-based estimation, for speech corrupted by music. Baseline recognition accuracy for the corrupted speech is also shown.

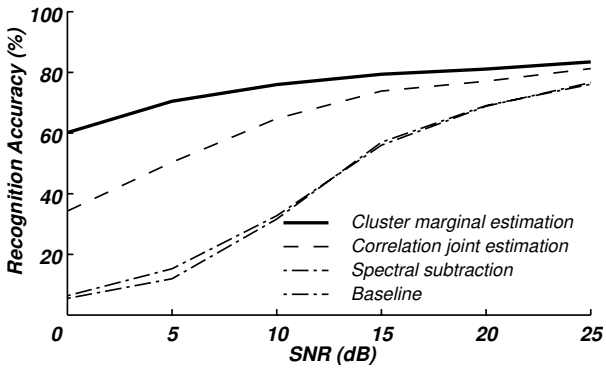


Figure 8.11 Recognition accuracy obtained when incomplete spectrogram methods are used with oracle masks to compensate for music.

filter, and $\boldsymbol{N}$ is the log spectrum of the noise. It is assumed that the noise is stationary, and that differences in the spectrum of the noise corrupting individual spectral vectors (each representing one analysis window of speech) are attributable only to differences in realization of the same random process (*i.e.* estimation error). It is further assumed that the distribution of the log spectrum of the noise in the various analysis windows is Gaussian, with a mean μ_N, which also represents the estimate of the true log spectrum of the noise, and variance Σ_N.

The distribution of the log spectra of clean speech is assumed to be a Gaussian mixture. The set of parameters of the Gaussian mixture, Φ, are learned from a training corpus of clean speech. The problem

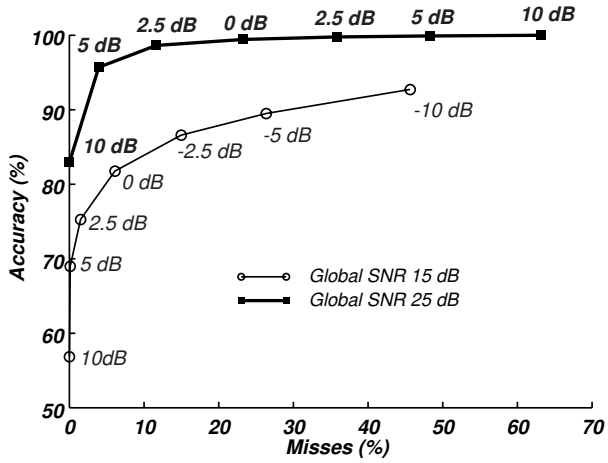


Figure 8.12 Percentage of reliable elements in the spectrogram correctly identified by the VTS-based mask estimate as being reliable (Accuracy) vs. percentage of unreliable elements falsely identified as being reliable (false alarms). The number beside each points indicates the deletion threshold used.

addressed in VTS, within the framework of this formulation, is the maximum likelihood estimation of the channel parameter $\boldsymbol{H}$, and the mean and the variance of the noise, μ_N and Σ_N. Representing the set of log spectral vectors of the noisy utterance as $\boldsymbol{Y}$, the estimate is given by

$$\boldsymbol{H}, \mu_N, \Sigma_N = \arg\max_{H, \mu, \Sigma}\{P(\boldsymbol{Y}|\boldsymbol{H}, \mu, \Sigma, \Phi)\} \tag{8.4}$$

Once $\boldsymbol{H}$, μ_N, and Σ_N have been estimated $\boldsymbol{X}(t)$ is estimated from $\boldsymbol{Y}(t)$ using an MMSE estimator.

The mean value μ_N of the noise log spectrum is also the estimate of the true log spectrum of the noise. It can be used to estimate the local SNR of the elements of the spectrogram of the noisy speech. If $Y(t, k)$ is the value of the k^{th} frequency band of the t^{th} spectral vector in the noisy spectrogram, and we represent the k^{th} frequency component of μ_N by $\mu_N(k)$, the estimate of the SNR of $Y(t, k)$ would be given by

$$\overline{SNR}(t, k) = \frac{Y(t, k) - \mu_N(k)}{\mu_N(k)} \tag{8.5}$$

Spectrographic masks are computed based on the estimated SNR values by tagging all elements in the spectrogram for which $\overline{SNR}(t, k)$ lies below a threshold T as unreliable. Figure 8.12 plots the percentage of reliable elements correctly identified against the false-alarm percentage for various values of T for speech corrupted with white noise to 15 dB and 25 dB. The knee of the curves is seen to be between 5 dB

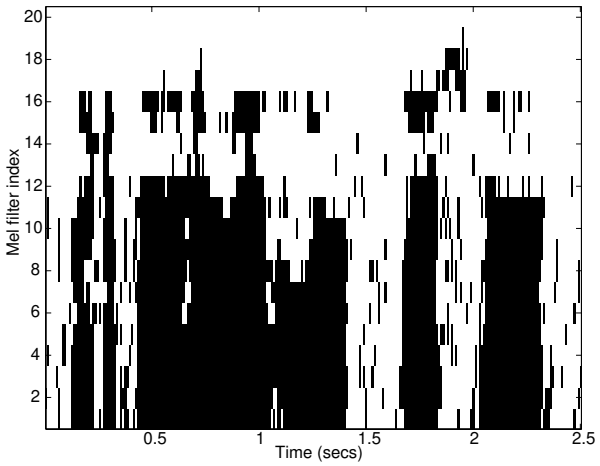


Figure 8.13 Spectrographic mask estimated using VTS-based estimation for an utterance of speech corrupted to 10 dB by white noise.

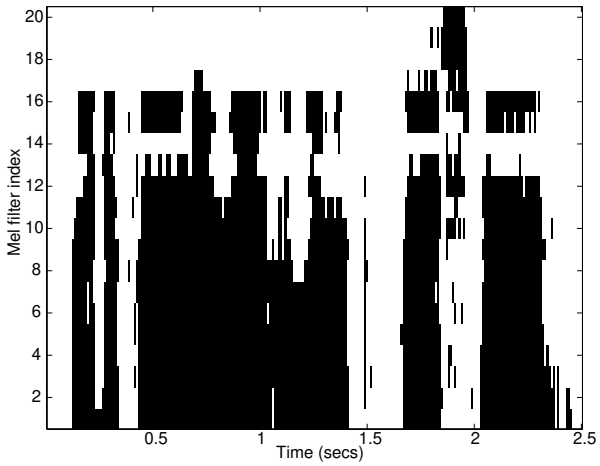


Figure 8.14 Oracle spectrographic mask for the same utterance.

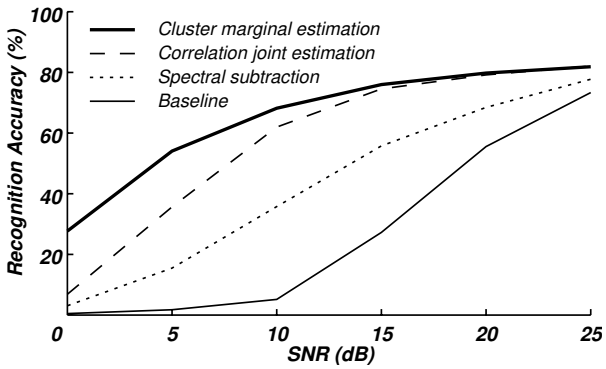


Figure 8.15 Recognition accuracy obtained with spectrographic masks estimated by VTS-based estimation, for speech corrupted by white noise. Baseline recognition accuracy for the corrupted speech is also shown.

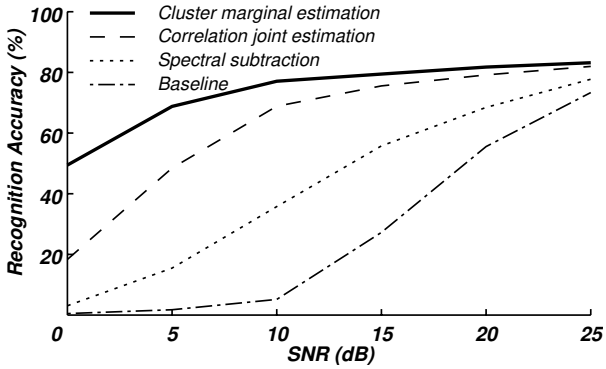


Figure 8.16 Recognition accuracy obtained when incomplete spectrogram methods are used with oracle masks to compensate for white noise.

and 0 dB. The threshold T was therefore set to be at 2.5 dB

8.4.1 Experimental results with VTS-based mask estimation

Figure 8.13 shows an example of the spectrographic mask estimated by VTS-based mask estimation for an utterance corrupted to 10 dB by white noise. We observe that the spectrographic mask obtained using VTS-based estimation is a very good approximation to the oracle mask shown in Figure 8.14. Figure 8.15 shows the recognition accuracy obtained with unreliable-spectrogram methods using masks estimated by VTS-based estimation. We observe that VTS-based mask estimation is also very effective in terms of the recognition accuracy obtained when these masks are used with unreliable spectrogram methods. Large

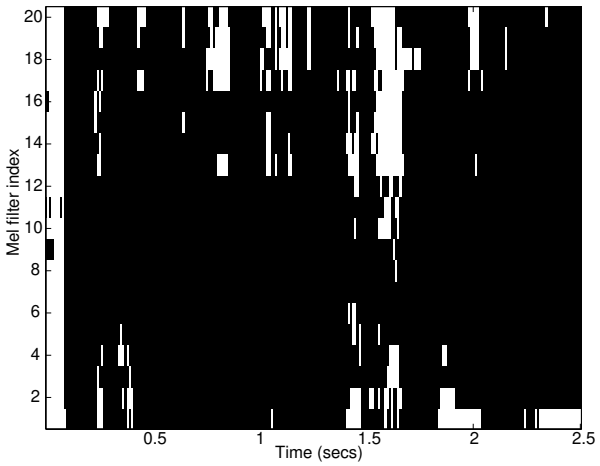


Figure 8.17 Spectrographic mask estimated using VTS-based estimation for an utterance of speech corrupted to 10 dB by music.

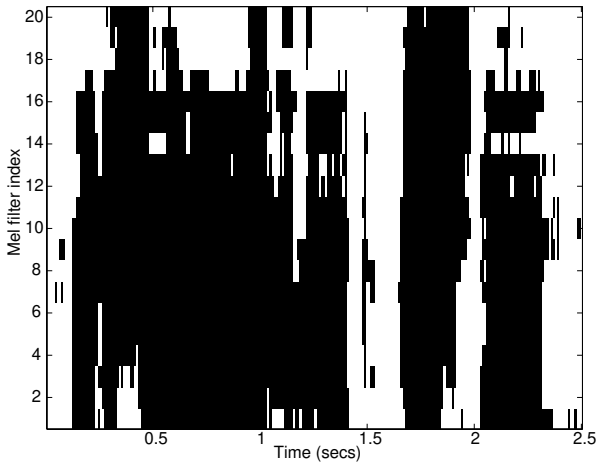


Figure 8.18 Oracle spectrographic mask for the same utterance.

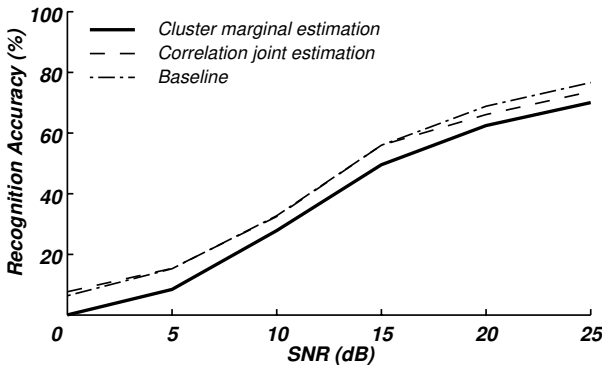


Figure 8.19 Recognition accuracy obtained with spectrographic masks estimated by VTS-based estimation, for speech corrupted by music. Baseline recognition accuracy for the corrupted speech is also shown

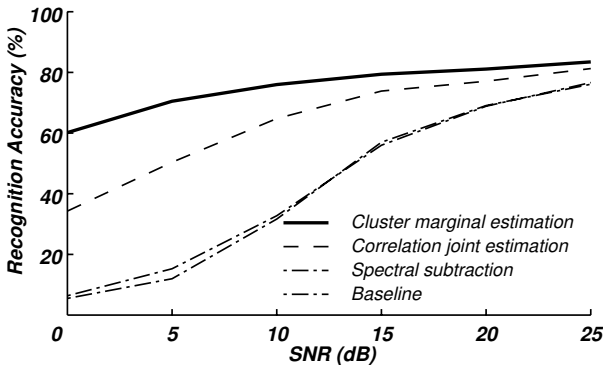


Figure 8.20 Recognition accuracy obtained when incomplete spectrogram methods are used with oracle masks to compensate for music.

improvements in recognition accuracy over baseline are achieved at all SNRs. Comparison with Figure 8.6 also shows that the recognition accuracy obtained using VTS-based spectrographic mask estimates is significantly greater than that obtained with spectral-subtraction-based mask estimates. The difference between the performance obtained with oracle masks and the performance with estimated masks is much smaller when the masks are estimated using VTS-based estimation.

Figure 8.17 and Figure 8.18 show the mask obtained with VTS-based estimation and the oracle mask respectively for an utterance of speech that has been corrupted to 10 dB by music. As in the case of spectral-subtraction-based mask estimation, the mask obtained by VTS-based estimation is a very poor approximation to the oracle mask. Figure 8.19 shows the recognition accuracy obtained on speech corrupted with

music, when masks estimated using VTS are used in conjunction with spectrogram reconstruction methods. We observe that the performance obtained with unreliable spectrogram methods is very poor, frequently resulting in recognition accuracies *lower* than the baseline. VTS-based estimation is ineffective when the corrupting noise is music.

8.5 Estimating spectrographic masks using a classifier

Spectrographic masks essentially separate the elements of the spectrogram out into two classes - the class of unreliable elements, and the class of reliable elements. Each element of the spectrogram belongs to one of these two classes. In classifier-based estimation of spectrographic masks we therefore treat the problem of estimating spectrographic masks as one of classification.

Each element of the spectrogram is represented by a vector of features for the purpose of this classification. We refer to this vector as the *classification* vector. In our experiments the classification vector $\vec{Y}(t, k)$ representing each element $Y(t, k)$ of the spectrogram was constructed as

$$\vec{Y}(t, k) = \begin{bmatrix} Y(t, k) \\ Y(t+1, k) - Y(t-1, k) \\ Y(t, k+1) - Y(t, k-1) \\ Y(t+1, k+1) - Y(t-1, k-1) \\ Y(t-1, k+1) - Y(t+1, k-1) \end{bmatrix} \tag{8.6}$$

While there are other ways in which the classification vector representing any element of the spectrogram can be constructed, it is expected that such a vector would capture information about the variation of the elements in the spectrogram that would be useful for classification.

We use a simple bayesian classifier to classify each element of the spectrogram as belonging either to the reliable or the unreliable class. Separate classifiers are used for each frequency component in the spectral vector. Individual elements of the spectrogram are assumed to be uncorrelated to each other for the purpose of classification and classification of the individual elements of the spectrogram is done independently of other elements in the spectrogram. If we represent the parameters of the distribution of reliable elements in the k^{th} frequency band of the spectral vectors in the spectrograms as $\Phi_{r,k}$ and the parameters of the distribution of unreliable elements as $\Phi_{u,k}$, the value $M(t, k)$ of the spectrographic mask in the k^{th} frequency band for the t^{th} spectral vector is given by

$$M(t,k) = \begin{cases} reliable & \text{if } P(r)P(Y(t,k)|\Phi_{r,k}) \geq P(u)P(Y(t,k)|\Phi_{u,k}) \\ unreliable & \text{if } P(u)P(Y(t,k)|\Phi_{u,k}) > P(r)P(Y(t,k)|\Phi_{r,k}) \end{cases} \quad (8.7)$$

where $P(r)$ and $P(u)$ are the *a priori* probabilities of the reliable and unreliable class, respectively. Ideally, the *a priori* probabilities of the classes would be specific to the global SNR of the speech - at low SNRs the fraction of elements that are noise corrupted (and thereby unreliable) can be expected to be higher than at high SNRs. However, since the global SNR of the utterance being recognized is not known beforehand in most real situations, the same values of $P(r)$ and $P(u)$ have to be used at all SNRs.

Figure 8.21 plots percentage of reliable elements correctly identified against the false alarm percentage for various values of $P(r)$ for speech corrupted with white noise to 5 dB, 15 dB and 25 dB. We observe that the best value of $P(r)$, given by the knee in the curve is between 0.7 and 0.8 in all cases with some variation. $P(r)$ was therefore chosen to be 0.8.

In general, since misses are less expensive (in terms of their effect on recognition accuracy) than false alarms, it is better to choose a high *a priori* probability for the class of reliable elements, $P(r)$.

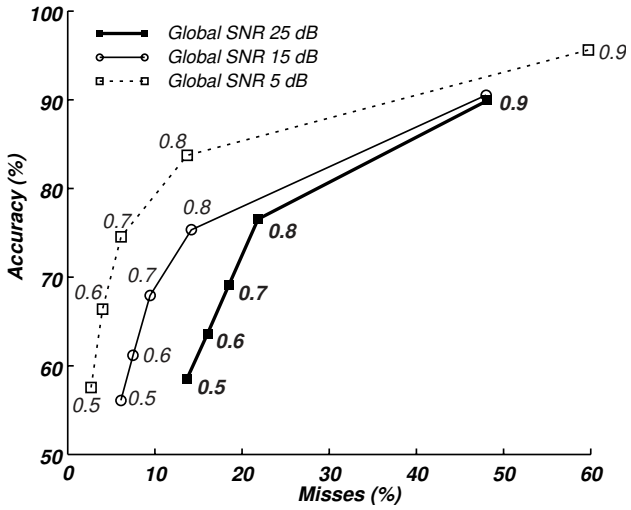


Figure 8.21 Percentage of reliable elements in the spectrogram correctly identified by the mask as being reliable (Accuracy) vs. percentage of unreliable elements falsely identified as being reliable (false alarms). The number beside each points shows the value used for the *a priori* probability of reliable regions.

8.5.1 Experimental results with classifier-based mask estimation

Classifier-based mask estimation was evaluated on both speech corrupted by white noise, and speech corrupted by music. Ideally, the mask estimation procedure would be independent of the type of noise corrupting the speech and the same distributions would be used to represent the reliable and unreliable classes irrespective of the kind of noise corrupting the speech signal. In our experiments, however, it was assumed that the type of noise corrupting the speech was known *a priori*. Therefore, for experiments with white noise the classifier was trained with speech corrupted with white noise. For experiments with music the classifier was trained with speech corrupted by music.

8.5.1.1 Experiments with white noise

To estimate spectrographic masks for speech corrupted with white noise a single reliable/unreliable classifier was trained for each frequency band using speech corrupted by white noise to several SNRs between 0 dB and 30 dB. The spectrographic masks for all utterances being recognized were estimated using this classifier. We refer to such a classifier as a *fair* classifier since the global SNR of the speech being recognized is not assumed to be known beforehand. Figure 8.22 shows an example of a spectrographic mask estimated by classification for an utterance of speech corrupted to 10 dB by white noise. Figure 8.23 shows the corresponding oracle mask for the utterance.

Recognition experiments show that spectrograms reconstructed with masks estimated using such a classifier result in recognition accuracies that are comparable with those obtained with spectral-subtraction-based mask estimation. Figure 8.24 shows the recognition accuracies obtained using masks estimated by classifier-based estimation. Comparison with Figure 8.6 (recognition performance with spectral-subtraction-based masks) shows that the two are very similar.

If the classifier used to estimate spectrographic masks for a noisy utterance is trained using speech corrupted to the same SNR as the speech being recognized, the performance of the mask estimation can be improved even further. We call such a classifier a *cheating* classifier since it is assumed that the global SNR of the speech being recognized is known *a priori*. Figure 8.25 shows the recognition performance obtained when spectrographic masks are estimated using such a cheating classifier. Masks obtained with cheating classifiers are seen to result in much greater accuracies than masks obtained with a fair classifier.

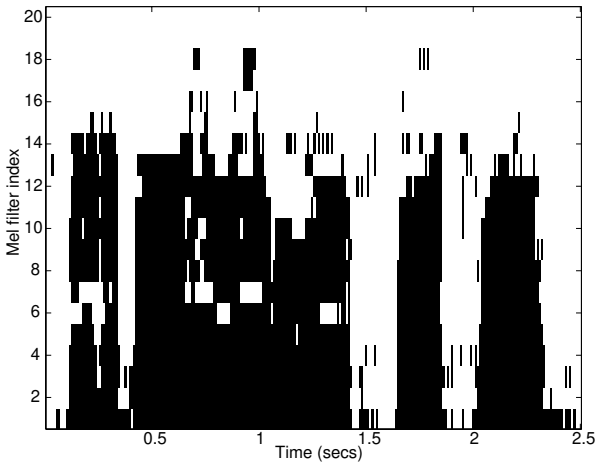


Figure 8.22 Spectrographic mask estimated using a fair classifier for an utterance of speech corrupted to 10 dB by white noise

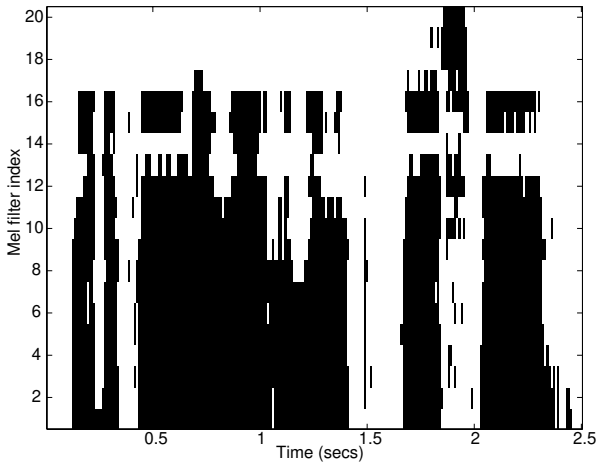


Figure 8.23 Oracle mask for the same utterance

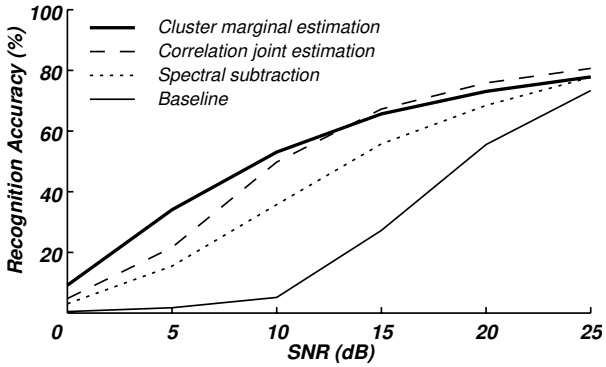


Figure 8.24 Recognition accuracy on speech corrupted by white noise, with unreliable spectrogram methods using masks obtained by a fair classifier

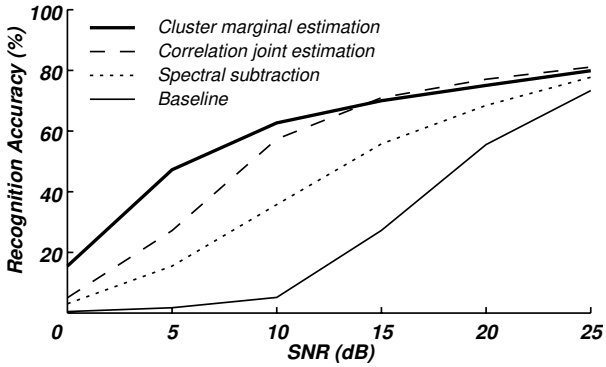


Figure 8.25 Recognition accuracy on speech corrupted by white noise, with unreliable spectrogram methods using masks obtained by a cheating classifier

8.5.1.2 Experiments with music

For experiments with music a single reliable/unreliable classifier was trained for each frequency band using speech corrupted by music to several SNRs between 0 dB and 30 dB. Spectrographic masks for all utterances corrupted by music were estimated using these classifiers. Figure 8.26 shows the estimated spectrographic mask for an utterance of speech corrupted to 10 dB by music. Figure 8.27 shows the recognition accuracy obtained with masks estimated using this classifier. We observe that a small improvement in recognition accuracy is obtained at all SNRs over baseline using covariance-based estimation. This is an improvement over the performance using either spectral-subtraction-based estimation or VTS-based esti-

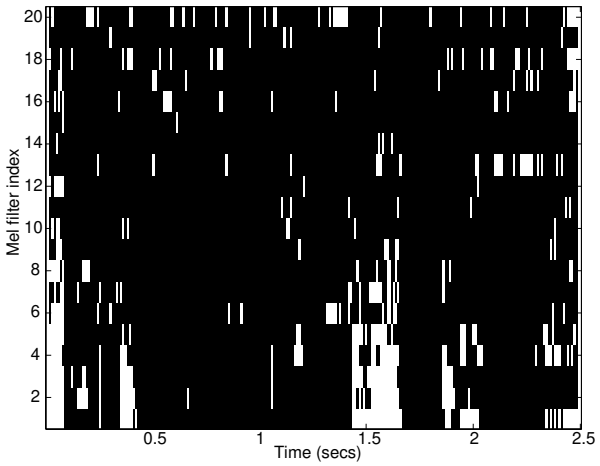


Figure 8.26 Spectrographic mask estimated for an utterance corrupted by music to 10 dB using a "fair" reliable/unreliable classifier.

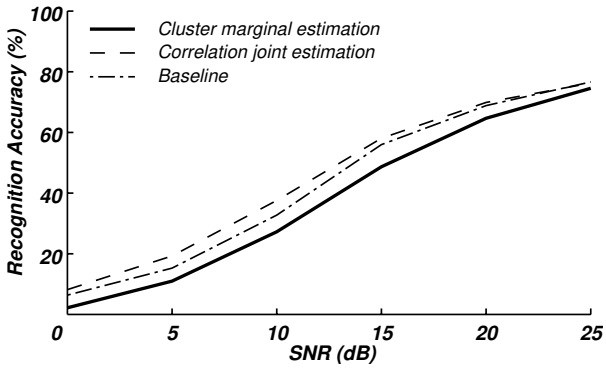


Figure 8.27 Recognition accuracy on speech corrupted with music using masks estimated by a fair classifier.

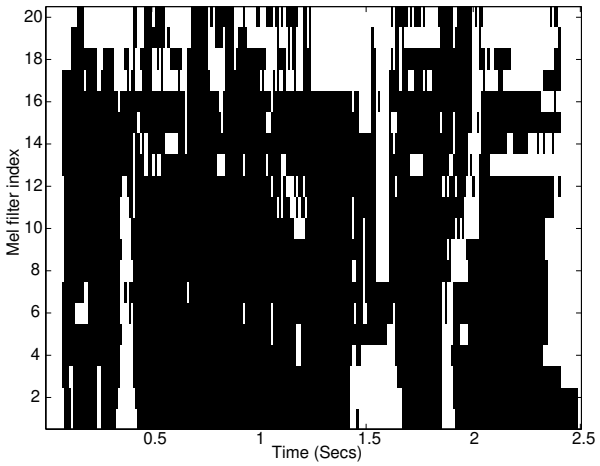


Figure 8.28 Spectrographic mask estimated for the same utterance as the one above using a cheating classifier.

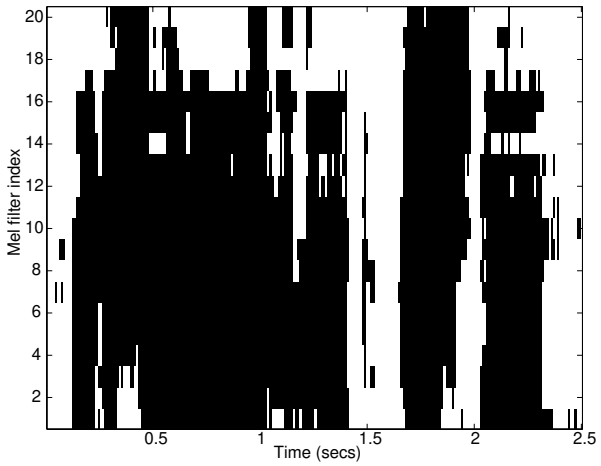


Figure 8.29 Oracle mask for the same utterance

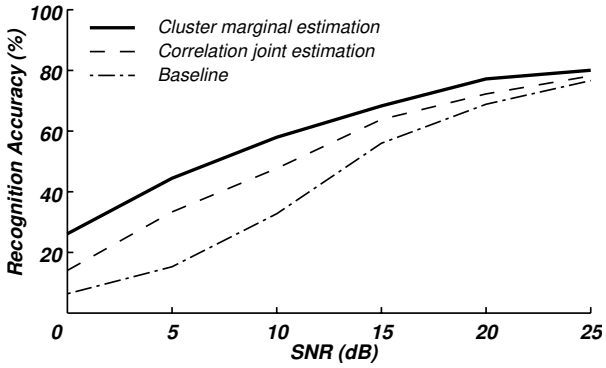


Figure 8.30 Recognition accuracy on speech corrupted with music using masks estimated with a cheating classifier.

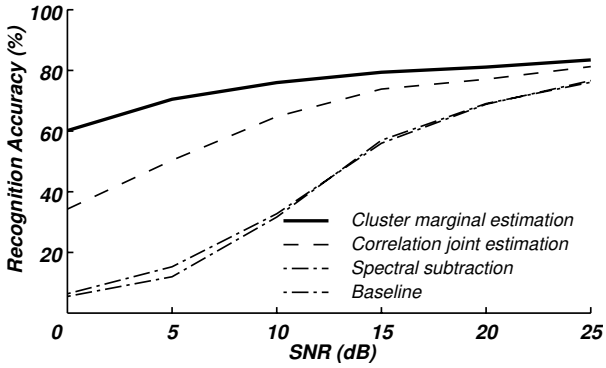


Figure 8.31 Recognition accuracy on speech corrupted with music using oracle masks

mation, where no improvement was obtained at all. However, even for classifier-based estimation, the recognition performance obtained with bounded cluster marginal estimation based on the estimated masks is poorer than the baseline performance. Also, it is doubtful whether the improvement in recognition accuracy seen with covariance-based reconstruction is significant and would carry over to other experiments.

The performance of classifier-based mask estimation for speech corrupted with music can be improved significantly using a *cheating* classifier, where the classifier is trained using speech corrupted to the same global SNR as the speech being recognized. Figure 8.28 shows the mask estimated by a cheating classifier for the same utterance represented in Figure 8.26 The oracle mask for the utterance is shown in Figure 8.29 It can be seen that the “cheating” mask is a much better approximation for the oracle mask than the one obtained using a fair classifier, or any of the other methods described earlier. Figure 8.30 shows the recognition performance obtained by applying incomplete spectrogram methods with the cheating masks on speech corrupted by music. We note that a significant improvement over baseline is obtained using the cheating masks with both cluster-based reconstruction and covariance-based reconstruction. In fact the performance obtained with cluster-based reconstruction using the estimated masks is comparable to the performance obtained with covariance-based reconstruction using oracle masks, shown in Figure 8.31, at most SNRs.

8.6 Discussion and Conclusions

All of the spectrographic mask estimation methods described in this chapter have been reasonably successful at estimating masks for speech corrupted by white noise. The recognition accuracy obtained using spectrogram reconstruction methods with the estimated spectrographic masks are significantly higher than the baseline recognition accuracy obtained with cepstra derived directly from the noisy speech. In fact, the recognition accuracy obtained with cluster marginal reconstruction in conjunction with spectrographic masks using VTS-based estimation is significantly higher than the performance obtained with VTS, our best algorithm to compensate for white noise prior to the work reported in this thesis. Figure 8.32 compares the recognition accuracy obtained using VTS compensation, and cluster marginal reconstruction and covariance joint reconstruction with VTS-based estimation of spectrographic masks, on speech corrupted with white noise to several SNRs. We observe that the performance obtained with covariance joint reconstruction is comparable with that obtained with VTS, and that obtained with cluster marginal reconstruction is

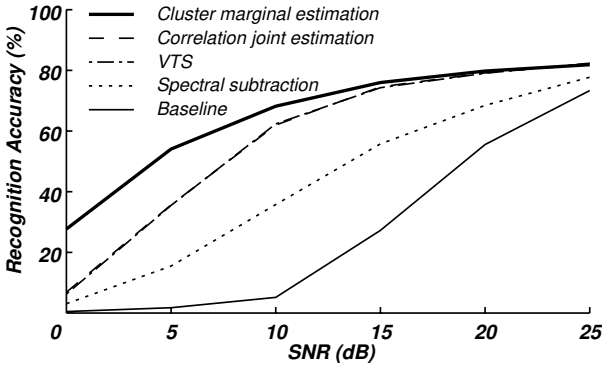


Figure 8.32 Comparison of recognition accuracies obtained on speech corrupted with white noise with VTS compensation, and with incomplete spectrogram methods using spectrographic VTS-based spectrographic masks. The curves for VTS and covariance joint reconstruction are almost coincident and therefore indistinguishable.

in fact significantly higher than that obtained with VTS compensation, especially at low SNRs.

All of these methods of mask estimation can also be expect to perform equally well on other stationary or quasi-stationary noises. However, their performance on speech corrupted by music is very poor. The reason for this poor performance is easy to understand in each of the methods.

The spectral subtraction noise estimate given by Equation (8.1) is based on the assumption that the underlying speech signal varies much faster than the noise [Hirsch 1995]. Music violates this assumption. As a result, the noise estimator described by Equation (8.1) is unable to estimate the noise spectrum, and spectrographic masks based on SNR values computed using these estimates of the noise spectrum are also erroneous.

VTS makes the explicit assumption that the corrupting noise is stationary. In fact, we only obtain a single estimate of the noise spectrum over the entire utterance, and masks are obtained based on this estimate of the noise spectrum. The procedure can be modified to work with short, sliding windows of speech, to compute a time varying estimate for the noise spectrum. However, such a procedure would still be constrained to tracking only slowly varying noises. It would not be able to track noises whose spectrum varies as fast as that of music.

Of the three methods, classifier-based methods of estimating spectrographic masks hold the most promise. They have been seen to perform quite well on speech corrupted by white noise. The performance obtained on white noise with the “cheating” classifier is, in fact, comparable with that obtained with VTS compensation. While the performance of unreliable spectrogram methods on speech corrupted by music

using spectrographic masks obtained with “fair” classifiers is not significantly better than the baseline, a significant improvement in recognition accuracy is obtainable when cheating classifiers are used to obtain the masks. This is, in fact, the first time that any consistent improvement has been obtained on speech corrupted with music.

We note here that although we refer to the case where the classifier has *a priori* knowledge of the global SNR of the speech being recognized as a “cheating” classifier, this may be a misnomer. It is relatively easy to estimate the global SNR of speech corrupted by white noise to within a few dB of the true SNR [Hirsch 1995]. Thus it is quite possible to perform the classification in two steps, the first identifying the global SNR of the speech, and the second using the appropriate classifier for the mask estimation.

A more serious problem is the assumption that the kind of noise corrupting the speech signal is known *a priori*. Implicit in this assumption is the assumption that the kind of noise corrupting the speech that is used to train the classifier is identical to the kind found in the test data. While this is possible for many commonplace noises, such as car noise, or even factory floor noise, the sheer variety of sounds in music makes it highly unlikely that the precise type of musical sounds used to train the classifier will also be found in the test utterance.

However, many possible solutions suggest themselves to this problem such as adapting the classifier to the kind of sounds found in the test utterances. In the following, concluding, chapter of this thesis, we discuss them among several other issues.

Chapter 9
Summary and Conclusions

9.1 Summary of major results and contributions

In this thesis we have tried to improve recognition accuracy of noisy speech by developing data compensation techniques based on the missing-feature paradigm. In the missing-feature paradigm noisy regions of spectrograms are identified and deleted to minimize the effect of corrupting noise, resulting in incomplete spectrograms with missing regions. Recognition is performed based on the information in the incomplete spectrograms.

Conventional missing-feature methods modify the recognizer to perform recognition with the incomplete spectrograms. The missing regions are not reconstructed. Instead, the manner in which the *a posteriori* likelihoods of sound classes is computed is modified. While this is theoretically optimal, it introduces the constraint that recognition has to be performed using the spectrogram itself. It is well known that speech recognition accuracy is much higher when performed with features such as cepstra which are computed from spectrograms by various transformations. As a result, while conventional missing-feature methods result in recognition performance that is fairly robust to corruption by noise, the best recognition performance obtained when using these methods (which is the recognition performance obtained with spectrograms of clean speech) is frequently inferior to the recognition accuracy obtained with the cepstra of noisy speech.

What is unique about the work in this thesis is that we *reconstruct* the missing portions of the spectrogram to get complete spectrograms, so that cepstral (or related) features can be derived from them. The recognition performance obtained using this approach is superior to that obtained using conventional methods. To the best of our knowledge, this approach has not been tried prior to this thesis.

There are several other advantages to this approach. The reconstruction methods we propose are based on very simple statistical models of the distribution of spectrograms and are computationally much simpler than the best current techniques. Also, the recognizer need not be modified in any manner since the entire noise compensation procedure including the identification of noisy regions of spectrograms, reconstruction of the regions, and derivation of features is done independently of the recognizer.

We propose several spectrogram reconstruction methods and focus on the two most effective ones, *cluster marginal reconstruction* and *covariance joint reconstruction*. The proposed spectrogram reconstruction methods in this thesis were found to be extremely effective at compensating for additive white noise. The recognition performance obtained was significantly superior to the performance obtained with our previous best algorithm, VTS. On non-stationary noise it was found that the techniques developed could be very successfully applied, provided the spectrographic masks identifying the noisy regions of the spectrograms could be accurately estimated. Thus, the problem of compensating for non-stationary noises has been reduced to one of reliably estimating spectrographic masks. While the problem of estimating spectrographic masks has not been completely solved for the case of non-stationary noises, it has been shown that classifier-based estimation of spectrographic masks is a viable approach to solving this problem.

The missing-feature-based noise compensation methods developed in this thesis are the best data-compensation solutions to compensating for white noise developed to date. They are also a partial solution to the problem of compensating for non-stationary noises, reducing the problem to one of reliably identifying spectrographic masks. The problem of estimating masks is one of estimating very crude, binary information regarding the degree of corruption in the various elements of the spectrogram, and may be much more tractable than the problem of actually tracking the spectrum of the noise. We therefore consider the methods developed in this thesis to be a first serious step towards compensating for non-stationary noises as well.

The complete noise compensation procedure consists of two steps:

1) Identification and deletion of the noisy regions of the spectrograms

2) Reconstruction of the deleted regions

The following sections describe our findings on these issues in reverse order.

9.2 Reconstruction of missing regions

A spectrogram can be visualized as a surface on a two dimensional support, where the two dimensions are time and frequency. Incomplete spectrograms are surfaces where some regions of the surface are missing. When the missing elements of the spectrogram are randomly distributed it was found that they could be effectively reconstructed by simple geometrical methods such as linear and non-linear interpolation. In

this situation it was found that linear interpolation was generally more effective than non-linear interpolation. Also interpolation along the time axis was much more effective than interpolation along the frequency axis.

Much better reconstruction was obtained when the missing regions were reconstructed on the basis of the statistical properties of the elements of spectrograms of clean speech. The *cluster-based reconstruction methods* proposed in this thesis assume that spectral vectors are segregated into a number of clusters, each of which has a Gaussian distribution. The resulting mixture Gaussian distribution is used to reconstruct the missing regions of spectral vectors. These methods only use the statistical correlations among different elements of a spectral vector (*i.e.* correlations across frequency) to reconstruct the missing components of the vector. *Covariance-based reconstruction methods*, on the other hand, model the sequence of spectral vectors in the spectrogram as the output of a WSS random process and use the statistical parameters of this process to reconstruct missing regions of the spectrograms. These methods use pairwise statistical correlations among all elements of the spectrogram (*i.e.* correlations both across frequency and across time) to reconstruct missing regions. It was found that covariance-based methods resulted in superior reconstruction compared to cluster-based methods when random elements of the spectrogram were missing.

When the missing regions of the spectrogram are induced by corrupting noise they do not occur at random locations. Instead, they occur in blocks and are related both to the spectrum of the corrupting noise causing the deletions and to the spectrum of the underlying speech itself. In this situation it was found that geometrical reconstruction techniques, or any methods that involved reconstruction based only on the geometry of the spectrogram, were completely ineffective. Recognition accuracies obtained with cepstra derived from spectrograms where noisy regions were deleted and reconstructed by geometrical methods were comparable to those obtained with cepstra derived from the noisy spectrogram itself. However reconstruction based on the statistical properties of the spectrogram was more effective. In particular, recognition accuracies obtained with cepstra derived from spectrograms reconstructed by covariance-based reconstruction methods were seen to be significantly superior to the baseline accuracy obtained using the cepstra of noisy speech. For cluster-based reconstruction, it was found that modeling the distribution of spectral vectors by a single cluster resulted in comparable or better recognition accuracies than modeling the distribution by a number of clusters.

When speech is corrupted by additive noise the observed value of any element of the spectrogram is the

upper bound on the true value of that element since the spectrogram now represents the sum of the energies in the speech and the noise. Therefore, while it is still appropriate to delete the noisy regions of spectrograms, the observed value of these regions are an upper bound on their true value and can be used to condition the estimates of the missing regions. It was found that when the estimates of missing regions were conditioned by these upper bounds, they were far superior to those obtained when bounds were not used. In particular, it was found that recognition accuracies obtained when reconstruction was performed with the best cluster-based reconstruction method, *cluster marginal reconstruction*, recognition accuracies on speech corrupted to 10 dB by noise were comparable to the accuracy obtained on clean speech, provided the spectrographic masks identifying the noisy regions of the spectrogram to be deleted were accurately known.

Another factor affecting reconstruction is the fact that even the regions of the spectrogram that have not been deleted are affected by noise. It was found that reducing the noise level in these elements by spectral subtraction prior to reconstruction improved reconstruction still further.

9.2.1 Discussion

Analysis of the covariance between the different elements of the spectrogram shows that covariance across frequency is greater than covariance across time. However, due to the finite length of the spectral vectors (only 20 elements in our experiments), the number of neighboring elements available to reconstruct any point is much more restricted when reconstruction is based only on elements within the same vector, than when it is based on elements of different vectors. As a result, linear interpolation along time results in better reconstruction than interpolation along frequency.

Geometrical reconstruction methods base the reconstruction of missing regions only on the regions that are present in the spectrogram. Since these regions have also been corrupted by noise, even in the best case, the reconstructed regions would be at least as noisy as the remaining regions. Additionally, when blocks of elements are missing, simple interpolation-based reconstruction completely ignores the expected nature of speech spectrograms and the correlations between their elements.

Among statistical reconstruction methods, covariance-based reconstruction methods use the covariances both across time and across frequency to perform reconstruction. Cluster-based methods base the

reconstruction only on covariances between different elements of the same vector. As a result, covariance based techniques are able to identify many more observed elements in the spectrogram to base the reconstruction on and their performance is consequently better than that of cluster-based reconstruction when the bounds on the values of the missing regions are not considered. Among cluster-based reconstruction methods, it was found that increasing the number of clusters does not improve reconstruction in any manner. This seems to indicate that the global distribution of spectral vectors is as well modeled by a single Gaussian as it is by a mixture of Gaussians, for the purpose of reconstruction.

When the observed value of noisy regions is used as an upper bound in the estimation the performance of multiple cluster based reconstruction improves dramatically. The bounding information improves the accuracy of identification of the cluster that any vector belongs to, thereby localizing the region in which the reconstructed vectors can lie very effectively. The identification of clusters is treated as a classification problem. The bounding information is seen to improve the accuracy of classification much more greatly than it does the accuracy of the reconstruction, given the distribution of the complete vector. As a result, the improvement in multiple-cluster-based methods is much greater than that of single-cluster-based reconstruction or covariance-based reconstruction.

9.2.2 Relative merits of the reconstruction techniques

Cluster-based reconstruction techniques are seen to be superior to covariance-based reconstruction when the upper bounds on the values of the missing regions are known. However covariance-based reconstruction methods still hold some advantages. First, they are seen to be the superior reconstruction method when no information about the missing regions is available (*i.e.* no bounding information is available). Second, they are far less computationally expensive than cluster-based methods. Thus they would be the methods of preference where computational expense is an issue.

9.3 Identification and deletion of the noisy regions of the spectrograms

Accurate identification of noisy regions of spectrograms, or the *spectrographic masks*, is crucial for missing-feature based noise compensation methods to be effective. We have shown that if spectrographic masks can be accurately identified spectrogram reconstruction methods can be used to compensate very well for fairly high levels of noise. However, errors in the estimation of the masks can cause the perfor-

mance of these methods to degrade quickly.

Conventional methods use spectral-subtraction-based running estimates of the noise spectrum to identify noisy regions of the spectrogram. We have evaluated three methods of estimating masks: spectral-subtraction-based mask estimation, VTS-based mask estimation, and classifier-based mask estimation. Of the three, spectral-subtraction-based mask estimation is similar to the procedure used in conventional methods. VTS-based estimation and classifier-based estimation are new techniques that have been introduced in this thesis.

It was found that all three methods were effective in estimating masks for speech corrupted by white noise. The best performance was obtained using VTS-based mask estimation. The combination of VTS-based mask estimation and the best cluster-based reconstruction method resulted in the best recognition accuracies obtained with any data compensation method to date.

None of the mask estimation methods were effective on speech corrupted by music. However, it was found that if the type of music corrupting the speech and the global SNR of the corrupted speech were known *a priori*, good estimates of the masks were obtained with classifier-based estimation, and significant improvements could be seen in recognition accuracy. Similar results have been reported by Seltzer [Seltzer 2000].

Discussion

Estimation of the spectrum of a random process is a difficult task. It is necessary to have a sufficiently long sample of the process to obtain reliable estimates. It is important that the spectrum of the noise does not vary much within this segment. Spectral subtraction, VTS, and other methods of estimating the spectra therefore work best when the noise spectrum is stationary or slowly varying. They are very effective when the corrupting noise is white. It can be expected that these methods will be equally effective on other slowly varying or stationary noise. However, when the spectrum being tracked is that of a non-stationary signals such as music, the estimates of the spectrum are very poor, or completely wrong. As a result spectrographic masks estimated using such estimates are very poor.

Classifier-based mask estimation, on the other hand, does not attempt to estimate the noise spectrum. However, the features being used for the classification are sensitive to the global SNR of the speech. As a result, classifier-based estimation is effective only when the global SNR is known *a priori*.

9.4 Topics for further investigation

The methods presented in this thesis have been very successful at compensating for noise when spectrographic masks can be reliably found. When the corrupting noise is white, the masks can be very well estimated by VTS-based estimation. VTS-based mask estimation is dependent on the estimate of the noise spectrum obtained by VTS. It is known that VTS is quite successful at estimating noise spectra when the noise is slowly varying or stationary [Kim 1998]. VTS-based estimation of spectrographic masks will generally perform reliably when the corrupting noise is slowly varying or stationary. We expect, therefore, that the methods presented in this thesis will, in general, result in significant improvements in recognition accuracy on speech corrupted by stationary or slowly varying noises.

However, VTS-based estimation of masks, as presented in this thesis, uses a single estimate for the noise spectrum for the entire utterance. This estimate can be significantly improved by permitting the estimate to vary from frame to frame. There are two possible ways of doing this.

1) Estimate the noise spectrum in a sliding window of the speech
2) Use a Kalman filter formulation of VTS to recursively estimate the noise in each incoming frame for speech

In the first approach the estimate of the noise spectrum within any frame of speech would be obtained based on a small segment of speech, say 1 second long, centered at that frame. It is expected that such an estimator would be able to track the spectrum of slowly varying noises better than the direct formulation of VTS used in this thesis.

In the second approach an *a priori* distribution of the noise spectrum would be assumed and recursively updated based on every incoming spectral vector of noisy speech. It has been shown that this method of estimating the noise spectrum is significantly superior to the standard VTS formulation at tracking time-varying noises [Kim 1998].

Classifier-based estimation of spectral masks has been seen to be quite effective for white noise. It has also been observed to be effective when the global SNR and type of corrupting noise are assumed to be known *a priori*. Both these requirements, however, may be unrealistic. Several possibilities present themselves to improve the performance of classifier based systems.

1) Use features that are based specifically on the characteristics of speech, rather than the nature of the

noise, for classification

2) Adapt the classifier in an unsupervised manner

3) Correlate the classification decisions regarding the elements of the spectrogram

In the first approach we would use computable features of the speech waveform that are known to be corrupted by noise. For example, for voiced speech the ratio of the energy in the harmonics of the pitch frequency to that at other frequencies, within any band of frequencies, is high for clean speech, and lower at other frequencies [Morgan 1997]. However, when speech is corrupted by noise, this ratio would change. Other features that suggest themselves are the average spectral tilt within any frequency band, the phase characteristics of speech spectra, etc. These features are likely to be more invariant to the kind of noise corrupting the speech than the simple power spectral values and their derivatives used in this thesis. Promising results using this approach have been reported by Seltzer [Seltzer 2000].

In the second approach the distributions of the classes would be adapted to the noisy data in an unsupervised manner. Adaptation methods such as MAP [Duda 1973] or MLLR [Leggetter 1994] could be used to adapt the distributions. Classification would be done with the adapted distributions. This method would be expected to result in better masks than classification without adaptation would, provided the baseline classifier is reasonably correct. Also, adaptation could be used even when speech-specific features are used for classification.

In the third approach we would take advantage of the fact that when speech is corrupted by noise, the noisy regions of the spectrogram occur in blocks. Thus, the fact that any particular element is noisy immediately raises the probability that the elements surrounding it are noisy (and to be deleted) too. This correlation could be captured by statistical models such as Markov fields. Use of these models can be combined with adaptation and speech specific features.

Another approach that could be used to estimate spectrographic masks would be to treat noisy regions of spectrograms as outliers in an otherwise normal distribution and use outlier identification techniques, such as those described in [Tukey 1977], to identify them. This method would be useful for speech corrupted by sharp or transient noises such as door slams and phone rings,

Although the reconstruction obtained using the methods developed in this thesis is extremely good, it can be improved further. Cluster-based reconstruction techniques model the sequence of spectral vectors in

the spectrogram as the output of an IID process. The distribution of the vectors is modeled simply as a Gaussian mixture distribution. No information regarding the sequentiality of the vectors is retained. A superior cluster-based representation would be to model the sequence of spectral vectors as the output of an HMM. This is just a simple extension of the cluster-based model, where the *a priori* probability of the various clusters is made dependent on which cluster the previous vector belonged to. However, the introduction of this simple probability enforces temporal constraints on the model and would be expected to improve cluster identification and reconstruction significantly.

An even better model would be to model the sequence of vectors as the output of a higher order HMM. In a higher order HMM of order N the *a priori* probability of any cluster is made conditional to the cluster that the previous N vectors belonged to. As a result, a much greater constraint is placed on the sequentiality of the vector. One serious disadvantage with higher order HMMs is the exponential increase in the number of parameters need for the model with increasing N. The estimates obtained for the parameters, with any amount of finite data, would be very poor. Also, the reconstruction would become extremely expensive computationally. A simple, and intuitively appealing solution to this problem is to use what we term a *tree-structured higher order HMM*. In a standard higher order HMM the clusters that any of the past vectors can belong to are assumed to be identical to the clusters that could be associated with the current vector. In a tree-structured HMM the number of clusters modeling past vectors would be fewer than the clusters modeling the current vector. Figure 9.1 represents such a model schematically.

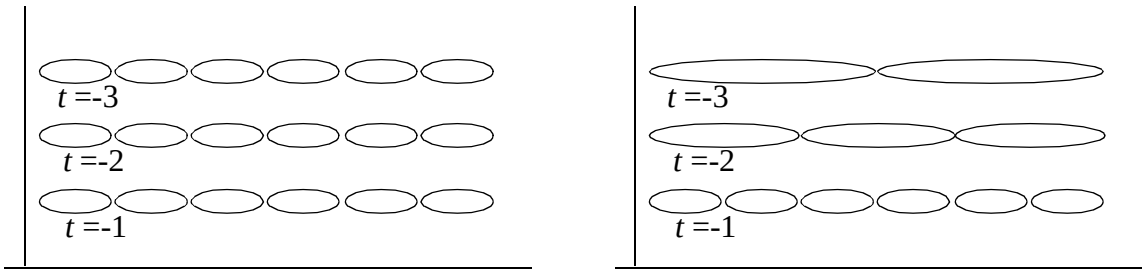


Figure 9.1 The panel to the left represents the manner in which data is modeled in a standard 3rd order HMM. The same 6 clusters covers the space at every time instant. The right panel shows data modeling in a tree-structured HMM. A smaller numbers of clusters are used to represent the distribution of data that occurs further back in time.

This model has the intuitive appeal that while the distribution of data at any instant is dependent on the distribution of data occurring in the past, it is less and less dependent on the precise location of the past data points as they get further away from the current data point. Additionally, the total number of parame-

ters needed in such a model would be much fewer than a standard higher-order HMM and would therefore be much better estimated. Reconstruction would simply proceed by identifying the cluster that the current vector belongs to, and reconstructing the missing portions of the vector based on the distribution of that cluster.

Covariance-based reconstruction was seen to be superior to cluster-based reconstruction both when random elements of the spectrogram were deleted, and when the bounds implied by the observed values of noisy elements were not considered. However, once the bounds were considered it was found to be much better to use multiple-cluster-based reconstruction. The application of the bounds improved the accuracy of cluster identification greatly, resulting in this improved performance. Similar improvement could be expected from covariance-based reconstruction if the spectral vectors in the spectrogram were assumed to be generated by one of a number of WSS random processes. Reconstruction would then consist of identifying which processes generated what vector and using the parameters of that process, as well as the cross-covariance between that process and the processes that generated adjacent vectors, to reconstruct the complete vector. This model would be fairly complex and, possibly, computationally expensive. A simpler model might be to model the distributions of short sequences of vectors (say 5-10 vectors) using a cluster-based representation. Reconstruction would proceed as in the case of cluster-based reconstruction - the cluster that any sequence of vectors belongs to would be identified, and the distribution of that cluster would be used to reconstruct the missing components of the central vector in the sequence.

9.5 Some remaining questions

While we have shown that the methods in this thesis are very effective on speech corrupted by white noise, and expect that they will perform equally well with other slowly varying noises, the only situation where they have been tested is when the recognition system itself has been trained with *clean* speech. This is not such a serious problem as long as the noise remains additive. Experiments show that the recognition accuracies obtained when the best cluster-based reconstruction technique is used with the true (oracle) spectrographic masks for the utterances, the recognition accuracy obtained on speech corrupted by noise to 5 dB SNR is comparable to the accuracy obtained when the recognizer is trained with speech at 5 dB SNR. In other words, the recognition accuracy obtained with 5 dB speech on the clean speech recognizer after missing-feature based compensation is applied is comparable with the recognition accuracy obtained with

a *matched recognizer* where system has been trained to recognize 5 dB speech. Previous experience with other data compensation algorithms indicates that if the missing-feature-based compensation were to be applied to both the data used to train the recognizer and the data being recognized, even better performance may be achieved.

However, in all of this, it has been assumed that the noise is additive, and that a clean uncorrupted corpus of speech exists such that the noisy speech could be modeled as speech from this corpus to which noise has been added. The statistical properties of this clean corpus have been used for the compensation. The question that arises is: what happens when such a clean corpus is not available. In such a situation, both VTS-based spectrographic mask estimation, and cluster or covariance-based reconstruction cannot be performed as described in this thesis. We have not worked out a satisfactory solution to this problem yet.

Another question that remains unanswered in all the experiments reported in this thesis is the effect of linear filtering on the reconstruction. It has been assumed everywhere that the speech has been corrupted solely by additive noise. However, when speech is recorded using arbitrary microphones the filter response of the microphone and the recording environment affect the speech as well. In such a situation the procedure that estimates spectrographic masks would have to estimate the log-spectrum of the impulse response of the filter as well. The filter response would then have to be subtracted out of the log-spectral values before reconstruction is performed. Since VTS has been shown in other work to be extremely effective at estimating these filter characteristics, we hypothesize that the performance of the reconstruction techniques would not be affected greatly by linear filtering. However, this hypothesis remains to be tested.

Finally the effect of non-linear phenomena such as non-linear filtering or clipping cannot be modeled as additive noise. In such a case, while the entire concept of reconstructing the badly damaged regions of the spectrogram remains valid, the precise manner in which bounding or other information is extracted from the observed values of the spectrogram would depend on the non-linear phenomenon affecting the speech. We have not investigated the effect of any non-linear phenomena on our methods.

9.6 Future directions

This thesis has presented a set of data-compensation methods based on the missing-feature paradigm that are seen to be very effective on speech corrupted by slowly varying noises. However, for any noise

compensation solution based on the methods described in this thesis to be complete some of the questions mentioned in this chapter would have to be answered. The primary question is the effect of not having a corpus of clean speech to begin with. Since the best current speech recognition systems rely on "multi-style" training, where the system is trained with speech recorded under various conditions, this is frequently the case. It may be possible to obtain the distributions of the spectrograms of clean speech from the clean regions of the spectrograms of the multi-style training data. However, for this to be possible, it is important to be able to identify these regions of these spectrograms first. Thus, the primary focus of any future work would have to be on improving the estimation of spectrographic masks under these conditions.

Even if the spectrographic masks were perfectly identified, the statistical properties of the spectrograms of clean speech would have to be estimated from these incomplete spectrogram. There has been significant work in the fields of statistical analysis on estimating the statistical properties of incomplete data [Ghahramani 1994][Little 1987]. However, these methods would have to be adapted to work on spectrographic data, to develop the kind of statistical models used with the reconstruction techniques. This would have to be a part of any future work.

Finally, there may be situations where it may be required to perform recognition using log spectra. In such a situation, better recognition accuracies may be obtained using missing feature methods if the recognizer itself were trained using incomplete spectrograms of noisy speech. The mathematics for this are readily available [Ghahramani 1994]. However, the actual implementation of such a solution still remains to be done.

Appendix A
Derivation of selected statistical relationships

A.1 Mean Squared Error (MSE) of an MAP estimate

In this section we derive the formula for the mean squared error of the MAP estimate of a Gaussian random vector.

Let $\boldsymbol{X}_m$ and $\boldsymbol{X}_o$ be jointly Gaussian vectors. Let μ_m and Θ_{mm} be the mean vector and covariance matrix respectively of $\boldsymbol{X}_m$. Let μ_o and Θ_{oo} be the mean vector and covariance matrix of $\boldsymbol{X}_o$. Let Θ_{mo} be the cross-covariance between $\boldsymbol{X}_m$ and $\boldsymbol{X}_o$. The conditional distribution of $\boldsymbol{X}_m$ is seen to be a Gaussian of the form (Section 2.5.4)

$$P(\boldsymbol{X}_m|\boldsymbol{X}_o) = C\exp(-0.5(\boldsymbol{X}_m-\mu_m-\Theta_{mo}\Theta^{-1}{}_{oo}(\boldsymbol{X}_o-\mu_o))^T(\Theta_{mm}-\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om})^{-1} (\boldsymbol{X}_m-\mu_m-\Theta_{mo}\Theta^{-1}{}_{oo}(\boldsymbol{X}_o-\mu_o))) \quad \text{(A1.1)}$$

The MAP estimate of X_m conditioned on X_o is given by

$$\hat{\boldsymbol{X}}_m = \arg\max_{X_m}\{P(X_m|\boldsymbol{X}_o)\} \quad \text{(A1.2)}$$

which gives us

$$\hat{\boldsymbol{X}}_m = \mu_m + \Theta_{mo}\Theta^{-1}{}_{oo}(\boldsymbol{X}_o-\mu_o) \quad \text{(A1.3)}$$

i.e. the MAP estimate of $\boldsymbol{X}_m$ is simply the expected value of $\boldsymbol{X}_m$. *i.e.* $\hat{\boldsymbol{X}}_m = E[\boldsymbol{X}_m]$.

The MSE of the MAP estimate is defined as

$$\begin{aligned} MSE(\hat{\boldsymbol{X}}_m) &= E[trace((\boldsymbol{X}_m-\hat{\boldsymbol{X}}_m)(\boldsymbol{X}_m-\hat{\boldsymbol{X}}_m)^T)|\boldsymbol{X}_o] \\ MSE(\hat{\boldsymbol{X}}_m) &= trace(E[(\boldsymbol{X}_m-\hat{\boldsymbol{X}}_m)(\boldsymbol{X}_m-\hat{\boldsymbol{X}}_m)^T|\boldsymbol{X}_o]) \end{aligned} \quad \text{(A1.4)}$$

However, $E[(\boldsymbol{X}_m-\hat{\boldsymbol{X}}_m)(\boldsymbol{X}_m-\hat{\boldsymbol{X}}_m)^T|\boldsymbol{X}_o]$ is simply the variance of $P(\boldsymbol{X}_m|\boldsymbol{X}_o)$ and is seen from Equation (A1.1) to be $\Theta_{mm}-\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om}$. We therefore get

$$MSE(\hat{\boldsymbol{X}}_m) = trace(\Theta_{mm} - \Theta_{mo}\Theta_{oo}^{-1}\Theta_{om}) \tag{A1.5}$$

$$MSE(\hat{\boldsymbol{X}}_m) = trace(\Theta_{mm}) - trace(\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om}) \tag{A1.6}$$

A.2 MSE increases as length(S_m) increases

In this section we show that the mean squared error of the MAP estimate of the missing components of a Gaussian random vector increases as the number of missing components increases.

Consider two incomplete observations $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ of a Gaussian random vector $\boldsymbol{X}$. $\boldsymbol{X}_2$ is identical to $\boldsymbol{X}_1$, except that it has one more component missing than $\boldsymbol{X}_1$. Let the vector of observed components of $\boldsymbol{X}_1$ be $\boldsymbol{X}_{1,o}$, and the vector of missing components in $\boldsymbol{X}_1$ be $\boldsymbol{X}_{1,m}$. Similarly, let the observed and missing components of $\boldsymbol{X}_2$ be $\boldsymbol{X}_{2,o}$ and $\boldsymbol{X}_{2,m}$ respectively. Since $\boldsymbol{X}_2$ has one more component missing than $\boldsymbol{X}_1$, we would have

$$\begin{aligned} \boldsymbol{X}_{2,m} &= [\boldsymbol{X}_{1,m}, X_a] \\ \boldsymbol{X}_{1,o} &= [\boldsymbol{X}_{2,o}, X_a] \end{aligned} \tag{A2.1}$$

where X_a is the component that is additionally missing in $\boldsymbol{X}_2$.

The *a posteriori* distributions of $\boldsymbol{X}_{1,m}$ and $\boldsymbol{X}_{2,m}$ would be given by

$$P(\boldsymbol{X}_{2,m}|\boldsymbol{X}_{2,o}) = P(\boldsymbol{X}_{1,m}, X_a|\boldsymbol{X}_{2,o}) \tag{A2.2}$$

$$P(\boldsymbol{X}_{1,m}|\boldsymbol{X}_{1,o}) = P(\boldsymbol{X}_{1,m}|X_a, \boldsymbol{X}_{2,o}) \tag{A2.3}$$

and would both be Gaussian. $P(\boldsymbol{X}_{1,m}|\boldsymbol{X}_{2,o})$ and $P(X_a|\boldsymbol{X}_{2,o})$ would also be Gaussian [Papoulis 1991].

Let Θ_{mm2} be the variance of $P(\boldsymbol{X}_{1,m}, X_a|\boldsymbol{X}_{2,o})$. The MSE of the MAP estimate of $\boldsymbol{X}_{2,m}$ is then $trace(\Theta_{mm2})$.

Let Θ_{mm1} be the variance of $P(\boldsymbol{X}_{1,m}|X_a, \boldsymbol{X}_{2,o})$. Let $\overline{\Theta}_{mm1}$ be the variance of $P(\boldsymbol{X}_{1,m}|\boldsymbol{X}_{2,o})$. Let θ_{aa} be the variance of $P(X_a|\boldsymbol{X}_{2,o})$. Let Θ_{ma1} be the cross covariance between $\boldsymbol{X}_{1,m}$ and X_a, condi-

tioned on $\boldsymbol{X}_{2,\,o}$. Then it can be shown (Section A.1) that

$$\Theta_{mm1} \;=\; \overline{\Theta}_{mm1} - \Theta_{ma1}\theta_{aa}^{-1}\Theta_{ma1}^{T} \;=\; \overline{\Theta}_{mm1} - \theta_{aa}^{-1}\Theta_{ma1}\Theta_{ma1}^{T} \tag{A2.4}$$

since θ_{aa} is the variance of a single component and is therefore simply a positive number.

The MSE of the MAP estimation of $\boldsymbol{X}_{1,\,m}$ is the trace of Θ_{mm1}. We get from Equation (A2.4) that

$$trace(\Theta_{mm1}) \;=\; trace(\overline{\Theta}_{mm1}) - \theta_{aa}^{-1}trace(\Theta_{ma1}\Theta_{ma1}^{T}) \tag{A2.5}$$

It is easy to see that $trace(\Theta_{ma1}\Theta_{ma1}^{T})$ has to be a positive number. Therefore

$$\theta_{aa}^{-1}trace(\Theta_{ma1}\Theta_{ma1}^{T}) \geq 0 \tag{A2.6}$$

It is also easy to see (Section A.1) that $trace(\Theta_{mm2}) \;=\; trace(\overline{\Theta}_{mm1}) + \theta_{aa}$. *i.e.*

$$trace(\Theta_{mm2}) \geq trace(\overline{\Theta}_{mm1}) \tag{A2.7}$$

Combining Equations (A2.5) and (A2.7), we get

$$trace(\Theta_{mm1}) \leq trace(\Theta_{mm2}) - \theta_{aa}^{-1}trace(\Theta_{ma1}\Theta_{ma1}^{T}) \tag{A2.8}$$

Combining Equations (A2.7) and (A2.8) we get

$$trace(\Theta_{mm2}) \geq trace(\Theta_{mm1}) \tag{A2.9}$$

In other words, the MSE of estimation of $\boldsymbol{X}_{1,\,m}$ is less than the MSE of estimation of $\boldsymbol{X}_{2,\,m}$. It is easy to extend the above logic to show that in general the MSE of estimation is greater for the vector with the greater number of components missing.

A.3 Average distance to closest element in an incomplete spectrogram with random elements missing, as a function of the drop fraction

In this section we derive the formula for the average distance between any point in a sequence, where random elements have been deleted, and the closest observed point as a function of the drop fraction.

Consider an infinite two-sided sequence where elements are missing with a drop fraction be α. In order

for the nearest neighbor to any element in this sequence to be n points away, it is necessary that the $n-1$ intervening points on either side of the present point are all missing, and that at least one of the two points n locations away from the current point is present. The probability that the $n-1$ points immediately on either side of the current point are all missing is $\alpha^{2(n-1)}$. The probability that at least one of the two points n locations away from the current point is present is $1-\alpha^2$. Thus, the probability that the nearest point to the current point is n locations away is given by

$$P(n) = (1-\alpha^2)\alpha^{2(n-1)} \tag{A3.1}$$

The expected distance of the nearest point to the current point is then given by

$$E[n] = \sum_{n=1}^{\infty} n(1-\alpha^2)\alpha^{2(n-1)} \tag{A3.2}$$

It is easy to show that since $\alpha < 1$

$$\sum_{n=1}^{\infty} n\alpha^{2(n-1)} = \frac{1}{(1-\alpha^2)^2} \tag{A3.3}$$

Combining Equations (A3.2) and (A3.3) we get

$$E[n] = \frac{1}{1-\alpha^2} \tag{A3.4}$$

For finitely long sequences the expected distance of the closest point would be somewhat larger than that given in Equation (A3.4) and would depend on the distance of the boundaries of the sequence from the point in consideration.

A.4 MSE of MAP estimates increases with decreasing covariance between the estimated and conditioning variables

In this section we show that the mean squared error of the MAP estimate of a Gaussian random vector increases as the cross covariance between the elements of the vector and the conditioning variables decreases.

Consider two jointly Gaussian vectors $\boldsymbol{X}_m$ and $\boldsymbol{X}_o$. The MSE of the MAP estimate of $\boldsymbol{X}_m$ is given by

$$MSE(\boldsymbol{X}_m) = trace(\Theta_{mm}) - trace(\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om}) \tag{A4.1}$$

where Θ_{mm} is the covariance matrix of $\boldsymbol{X}_m$, Θ_{oo} is the covariance matrix of $\boldsymbol{X}_o$, and Θ_{mo} is the cross covariance between $\boldsymbol{X}_m$ and $\boldsymbol{X}_o$.

Θ_{oo}^{-1} has the same properties as Θ_{oo}, *i.e.* it is symmetric and positive definite. We can therefore construct a random vector $\boldsymbol{Y}$, such that $E[\boldsymbol{Y}\boldsymbol{Y}^T] = \Theta_{oo}^{-1}$. We can now write

$$trace(\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om}) = trace(\Theta_{mo}E[\boldsymbol{Y}\boldsymbol{Y}^T]\Theta_{om}) = trace(E[\Theta_{mo}\boldsymbol{Y}\boldsymbol{Y}^T\Theta_{om}]) \tag{A4.2}$$

Defining

$$\boldsymbol{Z} = \Theta_{mo}\boldsymbol{Y} \tag{A4.3}$$

we get

$$trace(\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om}) = trace(E[\boldsymbol{Z}\boldsymbol{Z}^T]) \tag{A4.4}$$

If we represent the i^{th} element in the j^{th} row of Θ_{mo} as $\theta_{i,j}$, then it is easy to see from Equation (A4.3) that as $|\theta_{i,j}|$ decreases, $|\boldsymbol{Z}|$ decreases, and consequently, $trace(E[\boldsymbol{Z}\boldsymbol{Z}^T])$ decreases. Thus, as the values of $|\theta_{i,j}|$, the magnitudes of the covariances between the components of $\boldsymbol{X}_m$ and the components of $\boldsymbol{X}_o$ decrease, $trace(\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om})$ decreases, and $trace(\Theta_{mm}) - trace(\Theta_{mo}\Theta_{oo}^{-1}\Theta_{om})$ increases.

Therefore, from Equation , as the covariance between the components of $\boldsymbol{X}_m$ and $\boldsymbol{X}_o$ decreases, the MSE of the MAP estimate of $\boldsymbol{X}_m$ increases.

Appendix B
Iterative procedure for joint bounded MAP estimation

The problem of joint bounded MAP estimation is that of finding a set of values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k$ such that

$$\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_k = \arg\max_{y_1, y_2, \ldots, y_k} \{P(y_1, y_2, \ldots, y_k | y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k)\} \quad \text{(B0.1)}$$

We derive an iterative solution for this estimate in this appendix.

Let $y_1^n, y_2^n, \ldots, y_k^n$ be the estimate obtained after the n^{th} iteration of this procedure. If the $n+1^{\text{th}}$ estimate of y_1 is obtained as

$$y_1^{n+1} = \arg\max_{y_1} \{P(y_1, y_2^n, \ldots, y_k^n | y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k)\} \quad \text{(B0.2)}$$

then it is easy to see that

$$P(y_1^{n+1}, y_2^n, \ldots, y_k^n | y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k) \ge P(y_1^n, y_2^n, \ldots, y_k^n | y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k)$$

Using Bayes' rule and eliminating all irrelevant terms, it can be shown that Equation (B0.2) can be restated as

$$y_1^{n+1} = \arg\max_{y_1} \{P(y_1 | y_1 \le Y_1, y_2^n, \ldots, y_k^n)\} \quad \text{(B0.3)}$$

which is simply the bounded MAP estimate of y_1, conditioned on $y_2^n, \ldots, y_k^n$. Using similar logic, it can be shown that if the $n+1^{\text{th}}$ estimate of y_j is obtained as

$$y_j^{n+1} = \arg\max_{y_1} \{P(y_j | y_1^{n+1}, y_2^{n+1}, \ldots, y_{j-1}^{n+1}, y_j \le Y_j, y_{j+1}^n, \ldots, y_k^n)\} \quad \text{(B0.4)}$$

then

$$\begin{aligned} &P(y_1^{n+1}, y_2^{n+1}, \ldots, y_{j-1}^{n+1}, y_j^{n+1}, y_{j+1}^n, \ldots, y_k^n | y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k) \\ &\ge P(y_1^{n+1}, y_2^{n+1}, \ldots, y_{j-1}^{n+1}, y_j^n, y_{j+1}^n, \ldots, y_k^n | y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k) \end{aligned} \quad \text{(B0.5)}$$

In other words, if we were to begin with some set of initial estimates $y_1^1, y_2^1, \ldots, y_k^1$, and find the $n+1^{\text{th}}$ estimate of each y_j as the bounded MAP estimate of the that component as given by Equation (B0.4), each

step in the iteration would result in an increase in $P(y_1, y_2, \ldots, y_k \mid y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k)$.

When $P(y_1, y_2, \ldots, y_k)$ is Gaussian, $P(y_1, y_2, \ldots, y_k \mid y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k)$ has only one peak. Thus, the iterative solution given by Equation (B0.4) is guaranteed to find this peak, which is the unique solution to Equation (B0.1).

Therefore, the iterative solution to the joint bounded MAP estimation of a set of jointly Gaussian variables $y_1, y_2, \ldots, y_k$ conditioned on the bound $y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k$ is given by the following procedure:

1) Initialize all the y_i values as $y_i^1 = Y_i$

2) Obtain the $n + 1^{\text{th}}$ estimate of y_j as

$$y_j^{n+1} = \text{argmax}_{y_1}\{P(y_j \mid y_1^{n+1}, y_2^{n+1}, \ldots, y_{j-1}^{n+1}, y_j \le Y_j, y_{j+1}^{n}, \ldots, y_k^{n})\}$$

3) Iterate until $P(y_1, y_2, \ldots, y_k \mid y_1 \le Y_1, y_2 \le Y_2, \ldots, y_k \le Y_k)$ converges.

References

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

[2] Ahmed, S., Tresp, V. (1993), “Some solutions to the missing feature problem in vision”, Advances in neural information processing systems 5. Morgan Kaufmann Publishers, San Mateo, CA

[3] Boll, S.F. (1979), “Suppression of Acoustic Noise in Speech Using Spectral Subtraction”, IEEE Transactions on Acoustics, Speech and Signal Processing 27(2) pp:113-120, April, 1979

[4] Bourlard, H., Dupont, S. (1996), “A new ASR approach based on independent processing and recombination of partial frequency bands”, Proc. Intl. Conf. on Speech and Language Processing, 1996.

[5] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J. (1984), *Classification and Regression Trees*, Wadsworth & Brooks, Monterey, CA

[6] Cooke, M.P., Green, P.G., Crawford, M.D. (1994), “Handling missing data in speech recognition”, Proc. Intl. Conf. on Speech and Langugage Processing, 1994

[7] Cooke, M.P., Morris, A., Green, P.D., (1997), “Missing data techniques for robust speech recognition”, Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing, 1997

[8] Cooke, M., Green, P., Josifovski, L., Vizinho, A. (1999), “ Robust ASR with Unreliable Data and Minimal Assumptions”, Proceedings, Robust’99, Tampere, Finland

[9] Cooke, M., Green, P., Josifovski, L., Vizinho, A. (2000), “Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data”, To be published in Speech Communication

[10] Cooke, M., Green, P. (2000), “Auditory Organization and Speech Perception: Pointers for Robust ASR”, to appear in *Listening to Speech*, editors, Greenberg and Ainsworth, Oxford University Press

[11] David, M.H., Little, R.J.A., Samuhel, M.E., Triest, R.K. (1983), “Imputation methods based on the propensity to respond”, American Statistical Association 1983, Proceedings of the Business and Economics Section

[12] Davis, S., Mermelstein, P. (1980), “Comparison of paramteric representation for monosyllable word recognition in continuously spoken sentences”, IEEE Trans. on Acoustics, Speech, and Signal Processing, Vold: 28(4), PP: 357-366

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

[14] Duda, R.O., Hart, P.E. (1973), *Pattern Classification and Scene Analysis*, John Wiley and Sons, Inc. New York

[15] El-Maliki, M., Drygajlo, A. (1999), “Missing Features Detection and Handling for Robust Speaker Verification”, Proc. Eurospeech 1999

[16] Ephraim, Y. (1990), “A Minimum Mean Square Error Approach for Speech Enhancement”, Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, April 1990, pp: 829-832

[17] Ephraim, Y., Malah, D. (1984), “Speech Enhancement using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator”, IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-32(6) pp:1109-1121, Dec., 1984

[18] Fletcher, H., (1953), *Speech and Hearing in Communication*, Van Nostrand, New York 3.

[19] Fururi, S. (1981), “Cepstral Analysis Technique for Automatic Speaker Verification”, IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-20(2), pp:254-272, April 1981

[20] Gales, M.J.F., Young, S.J. (1993), “HMM recognition in noise using parallel model combination”, Proc. EUROSPEECH 1993, pp 837-840

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

[22] Gelb, A. (1974), *Applied Optimal Estimation*, MIT Press, Cambridge, MA

[23] Ghahramani, Z., Jordan, M.I. (1994), “Supervised learning from incomplete data via an EM approach”, Advances in Neural Information Processing Systems 6, (J.D. Cowan, G.Tesaurao & J.Alspector, eds.), Morgan Kaufmann Publishers, San Matero, CA, pp 120-129

[24] Hermansky, H., Tibrewala, S., Pavel, M. (1996), “Towards ASR on partially corrupted speech”, Proc. Intl. Conf. on Speech and Language Processing, 1996

[25] Hermansky, H. (1990), “Perceptual linear predictive (PLP) analysis of speech”, Journal of the Acoustic Society of America, 87, pp:1738-1752

[26] Hirsch, H.G., Ehrlicher, C. (1995), “Noise estimation techniques for Robust Speech Recognition”, Proc. IEEE Conf. on Acoustics, Speech and Signal Processing 1995, pp:153-156

[27] Jefferys, W.H., Berger, J.O (1992), “Ockham’s Razor and Bayesian Analysis”, American Scientist 80, pp: 64-72

[28] Josifovski, L., Cooke, M., Green, P., Vizihno, A. (1999), “State Based Imputation of Missing Data for Robust Speech Recognition and Speech Enhancement”, Proc. EUROSPEECH, 1999

[29] Juang, B.H., Levinson, S.E., Sondhi, M.M. (1986), “Maximum likelihood estimation for multivariate mixture observations of Markov chains”, IEE Transactions on Information Theory, IT-32(2), pp: 307-309, March 1986

[30] Kay, S.M. (1988), *Modern Spectral Estimation: Theory And Application*, Prentice Hall Inc., New Jersey

[31] Katz, S.M. (1987), “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol 35, no. 3, pp: 400-401, March 1987

[32] Kim, N.S. (1998), “IMM-Based Estimation for Slowly Evolving Environments”, IEEE Signal Processing Letters, Vol. 5, No. 6, pp: 146-149, June 1998

[33] 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.

[34] Lim, J.S. (1983), *Speech Enhancement*, Prentice Hall, Englewood Cliffs, New Jersey

[35] Linde, Y., Buzo, A., Gray, R.M. (1980), "An Algorithm for Vector Quantizer Design," IEEE Transactions on Communications, Vol. COM-28, No. 1

[36] Liporace, L.R. (1982), “Maximum Likelihood Estimation for Multivariate Observations of Markov Sources”, IEEE Transactions on Information Theory, IT-28, September 1982, pp: 729-734

[37] Lippmann, R.P. (1997), “Speech recognition by machines and humans”, Speech Communication, 22(1), 1-16

[38] Lippmann, R.P., Carlson, B.A. (1997), “Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering and noise”, Proc. EUROSPEECH 1997

[39] Little, R.J.A and Rubin, D.B (1987), *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., New York

[40] Little, R.J.A., Schluchter, M.D (1985), “Maximum likelihood estimation for mixed continuous and categorical data with missing values”, Biometrika, 72: 497-512

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

[42] Madow, W.G., Nisselson, H., Olkin, I. (1983), *Incomplete Data in Sample Surveys, vol. 1: Report and case studies*, Academic Press, New York

[43] McLachlan, G., Basford, K. (1988), *Mixture models: Inference and applications to clustering*, Marcel-Dekker

[44] McQueen, J. (1967), “Some methods for classification and analysis of multivariate observations”, 5-th Berkeley Symposium on mathematics, Statistics and Probability,1, S. 281-298.

[45] Mendenhall, W., Sinchich, T. (1996), *A second course in statistics: Regression Analysis*, 5th edition, Prentice Hall Inc.

[46] Miller, G.A., Licklider, J.C.R. (1950), “The intelligibility of interrupted speech”, Journal of the Acoustic Society of America, 22: 167-173

[47] Moore, B.C.J. (1997), *An introduction to the Psychology of Hearing*, 4th edition, Academic Press

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

[49] Morgan, D.P., George, E.B, Lee, L.T., Kay, S.M. (1997), “Cochannel speaker separation by harmonic enhancement and suppression”, IEEE Transactions on Speech and Audio Processing, Vol. 5, pp: 407-424, 1997

[50] Neumeyer, L., Weintraub, M. (1994), “Probabilistic Optimal Filtering for Robust Speech Recognition”, Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, 1994

[51] Ney, H., Mergel, D., Noll, A., Paesler, A. (1992), “Data driven search organization for continuous speech recognition”, IEEE Transactions on Signal Processing, 40(2), pp: 272-281, February 1992

[52] Nocerino, N., Soong, F.K., Rabiner, L.R., Klatt, D.H. (1985), “Comparative study of several distortion measures for speech recognition”, Speech Communication, No. 4, pp: 317:331

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

[54] O’Shaughnessy, D. (1987), *Speech Communication - Human and Machine*, Addison-Wesley Publishing Company

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

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

[57] Porter, J.E., Boll, S.F. (1984), “Optimal estimators for spectral estimators of noisy speech”, Proc.

IEEE Conf. on Acoustics, Speech and Signal Processing, 1984, PP: 18A.2.1-18A.2.4

[58] Press, W.H., Teukolsky, S.A., Vetterling, W.T, Flannery, B.P. (1992), *Numerical Recipes in C*, Cambridge University Press

[59] Price, P., Fisher, W.M., Bernstein, J., Pallet, D.S. (1988), "The DARPA 1000 word Resource Management database for continuous speech recognition", Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, pp: 651-654, 1988

[60] Quinlan, J.R. (1986), "Induction of decision trees", Machine Learning, 1:81-106

[61] Quinlan, J.R. (1989), "Unknown attribute values in induction", Proc. of the Sixth International Conference on Machine Learning.

[62] Rabiner, L.R., Juang, B-H. (1993), *Fundamentals of Speech Recognition*, PTR Prentice-Hall, Inc., New Jersey

[63] Rabiner, L.R., Schafer, R.W. (1978), *Digital Processing of Speech Signals*, Prentice Hall Inc., New Jersey

[64] Raj, B., Parikh, V., Stern, R.M. (1997), "The Effects of Background Music on Speech Recognition Accuracy", Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, 1997

[65] Renevey, P., Drygajlo, A. (1999), "Missing Feature Theory and Probabilistic Estimation of Clean Speech Components for Robust Speech Recognition", Proc. EUROSPEECH 1999

[66] Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York

[67] Schafer, J.L, (1997), *Analysis of Incomplete Multivariate Data*, Chapman and Hall, New York

[68] Seltzer, M. (2000), *Automatic Detection of Corrupted Speech Features for Robust Speech Recognition*, Masters Dissertation, Carnegie Mellon University, May 2000

[69] Shanmugam, S., Breipohl, A.M. (1988), *Random Signals: Detection, Estimation, and Data Analysis*, John Wiley & Sons, New York

[70] Sharf, L.L. (1991), *Statistical Signal Processing, Detection, Estimation, and Time Series Analysis*, Addison-Wesley

[71] Stark, H., Woods, J.W. (1994), *Probability Theory, Random Processes, and Estimation Theory for Engineers*, Prentice Hall, New Jersey

[72] Therrien, C.W. (1992), *Discrete Random Signals and Statistical Signal Processing*, Prentice Hall Inc., New Jersey

[73] Tukey, J.W. (1977), *Exploratory Data Analysis*, Addison-Wesley

[74] Varga, A.P., Moore, R.K. (1990), "Hidden markov model decomposition of speech and noise", Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, 1990, pp. 845-848

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

[76] Vizinho, A., Green, P., Cooke, M., Josifovski, L. (1999), "Missing Data Theory, Spectral Subtraction and Signal-to-Noise Estimation for Robust ASR: An Integrated Study", Proc. EUROSPEECH 1999

[77] Warren, R.M., Riener, K.R., Bashford, J.A., Brubaker, B.S. (1995), "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits", Perception and Psychophysics 57(2), pp: 175-182