Adaptive Model-Based Speech Enhancement

Beth Teresa Logan

Girton College, University of Cambridge,
and
Cambridge University Engineering Department.

Dissertation submitted to the University of Cambridge
for the degree of Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where stated. It has not been submitted in whole or part for a degree at any other university.

The length of this thesis including footnotes and appendices is approximately 37000 words.

Summary

This dissertation details the development and evaluation of techniques to enhance speech corrupted by unknown independent additive noise when only a single microphone is available. It therefore seeks to address a deficiency of many speech enhancement systems which require *a priori* knowledge of the interfering noise statistics. Such a deficiency must be corrected if these systems are to operate in real world situations.

The enhancement systems developed are based on an existing system by Ephraim (Ephraim 1992*a*). This approach models the speech and noise statistics using autoregressive hidden Markov models (AR-HMMs). Two main extensions to this technique are developed in order to make it adaptive. The first estimates the noise statistics from detected pauses. The second forms maximum likelihood estimates of the unknown noise parameters using the whole utterance. Both techniques operate within the AR-HMM framework.

Additional work in this dissertation improves the modelling power of AR-HMM systems by incorporating perceptual frequency. The bilinear transform is used to warp the frequency spectrum of the feature vectors to an approximation of the Bark scale. This modification can be incorporated into both AR-HMM recognition and enhancement systems.

The enhancement techniques are evaluated on the NOISEX-92 and Resource Management (RM) databases, giving indications of performance on simple and more complex tasks respectively. Additional experiments investigating the incorporation of perceptual frequency into AR-HMM systems were conducted on the E-set of the speaker independent ISOLET database.

Both enhancement schemes proposed were able to improve substantially on baseline results. The technique of forming maximum likelihood estimates of the noise parameters was found to be the most effective. Its performance was evaluated over a wide range of noise conditions ranging from -6dB to 18dB and on various types of stationary real-world noises.

The incorporation of perceptual frequency into AR-HMM systems was found to increase recognition performance substantially on both the ISOLET and RM databases. The improvement was less marked for the more complex task, highlighting that AR-HMMs could benefit from the inclusion of more variance information.

Keywords
Speech Enhancement, Autoregressive Hidden Markov Models, Robust Speech Recognition.

Acknowledgements

First I would like to thank my supervisor Tony Robinson. He provided me with the freedom I needed to explore my ideas and yet still was there with helpful suggestions, encouragement and constructive criticism. He also ensured that I was never without the equipment I needed and was helpful on many occasions with obtaining funding for conferences.

Second I would like to thank all those who have contributed to the smooth running of the Fallside Laboratory computer system, no matter what the odds. Particular praise must go to Patrick Gosling for sterling work far above and beyond the requirements of his role as Computer Officer.

I would also like to thank my sponsors without whom I could not have pursued this work. I was funded by a Packer Scholarship administered by the Cambridge Commonwealth Trust and an ORS award. I also received grants from the Cambridge Philosophical Society, Girton College and the Charles Hesterman Merz Fund administered by the Cambridge University Engineering Department which enabled me to attend several conferences which were of great benefit to my work. The latter body also provided me with additional funds in my final year which was much appreciated.

My thanks must also go to Mark Gales for much encouragement, particularly in the early years of my PhD, and helpful suggestions for directions to pursue. I also owe my thanks to Carl Seymour for the use of his code and scripts, and for helpful discussions.

I would also like to acknowledge and thank the following people for proofreading my work and providing much constructive criticism: (in approximate alphabetical order) Gary Cook, Arjen de Vries, Mahesan Niranjan, Gavin Smith, James Christie, Nando de Freitas, Klaus Reinhard, Jaco Vermaak and Ed Whittaker.

Finally, I would like to thank my family and friends for much emotional support, particularly during the final months. I would especially like to thank my father, Patrick Logan and my long-suffering friends Arjen de Vries and Stefan Krauss.

Contents

List of Figures

List of Tables

Notation

Observations

n	—	time index
K	—	frame size
s_n	—	speech sample n
d_n	—	noise sample n
y_n	—	noisy sample n
$\mathbf{s}_t$	—	K-dimensional clean speech observation at time t
$\mathbf{d}_t$	—	K-dimensional noise observation at time t
$\mathbf{y}_t$	—	K-dimensional noisy observation at time t
T	—	length of observation sequence
$\mathbf{S}$	—	entire sequence of clean speech observations $\triangleq \{\mathbf{s}_t, t = 0, \ldots, T\}$
$\mathbf{D}$	—	entire sequence of noise observations $\triangleq \{\mathbf{d}_t, t = 0, \ldots, T\}$
$\mathbf{Y}$	—	entire sequence of noisy observations $\triangleq \{\mathbf{y}_t, t = 0, \ldots, T\}$

Hidden Markov Models

t	—	time index
N	—	number of states
a_{ij}	—	transition probability from state i at time t to j at time $t+1$
x_t	—	clean state at time t
$\tilde{x}_t$	—	noise state at time t
$\bar{x}_t$	—	noisy state at time t
$a_{x_t x_{t+1}}$	—	transition probability from state x_t to x_{t+1}
μ_{jm}	—	mean of the distribution for state j mixture m
$\mathbf{\Sigma}_{jm}$	—	covariance of the distribution for state j mixture m
M	—	number of mixtures
P	—	order of autoregressive process (for AR-HMMs)
A_i	—	ith filter coefficient for an autoregressive model
σ^2	—	autoregressive model variance parameter

Abbreviations

AR-HMM	—	Autoregressive Hidden Markov Model
EM	—	Expectation Maximisation
HMM	—	Hidden Markov Model
MAP	—	Maximum *A Posteriori*
MMSE	—	Minimum Mean Squared Error
PD	—	Probability Distribution
pdf	—	probability density function
PSD	—	Power Spectral Density
SNR	—	Signal to Noise Ratio

Chapter 1

Introduction

As speech systems have evolved from laboratory demonstrations to real-world applications, the need to maintain performance in a wide variety of situations has emerged. For example, a mobile phone may be used at a construction site, an automatic information system may be installed in a crowded shopping centre or a hearing aid may be used at a party. Without enhancement, the performance of these systems may be unacceptable. Thus researchers are investigating new enhancement techniques so that speech processing systems can be used in a wide range of environments.

1.1 Speech Enhancement

Speech enhancement is defined as the attempt to improve the performance of speech communication systems when their input or output signal is corrupted by noise (Ephraim 1992*c*). This performance improvement generally either takes the form of improving the perceptual aspects of the speech or of changing the speech to match the templates used when training speech and speaker recognition systems.

If the first criteria for performance improvement is used, the enhancement system attempts to improve the quality and/or intelligibility of the corrupted speech. These measures are related to listener fatigue and understanding respectively. Typically such enhancement schemes have applications as front ends to speech coders, hearing aids and forensic analysis systems. The particular application will determine whether quality or intelligibility or both are important. Here the ultimate test of the effectiveness of the enhancement technique is human evaluation.

The second criteria for performance improvement involves evaluating the accuracy of speech or speaker recognition systems. It is well known that the performance of such systems degrades in the presence of noise (e.g. see (Deller, Proakis & Hansen 1993)). This is due to the acoustic mismatch between the features used to train and test these systems and the ability of the models to describe the corrupted speech. Typically clean speech is used to train the acoustic models. Therefore enhancement techniques which remove noise leaving an estimate of the clean signal are useful as a front end.

Although it seems logical that optimising either of these criteria would produce the same enhanced speech, this is not the case. People are able to discern perceptual improvements but as yet the optimal way of mathematically describing this has not been developed. Therefore the distortion measures used by speech and speaker recognition systems may not correspond to a guaranteed quality or intelligibility improvement. Thus improving the speech perceptually may not lead to improved recognition performance. Conversely, optimising a mathematical distortion measure may actually increase listener fatigue or decrease intelligibility.

In this dissertation both criteria for performance improvement are considered. This is because the work here has application to many types of speech systems. Improvements in speech recognition scores and distortion measures show quantitative improvement. Informal listening tests give an indication of qualitative improvement.

1.2 Problem Dimensions

The speech enhancement problem is characterised by the type of noise source, the way in which the noise interacts with the clean signal, the number of input channels or microphones available for enhancement and the nature of the final application (Ephraim 1992*c*).

The interfering noise may result from background sources such as computer fans or road noise, room reverberation or the communications channel including the microphone or the loud speaker. Usually noise from background sources is modelled as additive noise whereas echo and channel noise are modelled as convolutional noise. Environmental influences can also cause changes in speech articulation (Rabiner & Juang 1993). For example, a speaker may shout to be heard in extreme noise. This phenomenon is called the Lombard effect and is difficult to model.

The interfering noise may be stationary or non-stationary. For example, background noise from computer fans and the like can be treated as stationary but noise due to slamming doors and competing speakers cannot. Noise is typically modelled independently of the speech unless the Lombard effect is prominent or there is echo.

The number of microphones refers to the number of sources of the corrupted signal or parts of the signal which are available to be used in the enhancement scheme. Having more than one microphone can simplify the process (Deller et al. 1993). For example, if one microphone only records a signal correlated with corrupting noise, then this signal can be used to cancel the noise in the speech signal. However the need for single microphone techniques is unavoidable in many cases because of cost and other implementation issues.

The application of the enhancement system also affects its design. If an enhanced waveform is required then the system must be able to reconstruct the time domain signal. However, if the enhancement system is to be used as a front end to, for example, a clean speech recogniser, only enhanced versions of the recognition parameters are required.

1.3 Techniques for Speech Enhancement

Many enhancement techniques have been proposed. The techniques are categorised according to the type of models, if any, they use to model the speech and noise, and the amount of prior information included. In general, the use of more accurate speech and noise models and the incorporation of relevant prior information leads to better enhancement at the expense of greater computational complexity. A full description is given in Chapter 3.

1.4 Contribution of this Dissertation

This dissertation considers the problem of speech enhancement when only a single microphone is used and when the statistics of the interfering noise are not available *a priori*. Thus it seeks to address a deficiency of many current enhancement techniques and looks toward a system which would have application in the real world.

The interfering noise is assumed to be additive and statistically independent of the speech. It is also assumed to be stationary over the utterance to be enhanced. Although convolutional noise and the Lombard effect are not considered, the technique still has wide application. For example it could be used to implement hands-free dialling of a mobile phone or to improve a speech recognition system in an office environment.

Several systems are developed. They are based on a proven enhancement scheme by Ephraim (Ephraim 1992*a*) which models the speech and noise using autoregressive hidden Markov models (AR-HMMs). The AR-HMM framework is convenient for enhancement of signals corrupted by additive noise since it models features which are additive.

The schemes developed extend this work by estimating the noise statistics directly from the signal to be enhanced rather than using pre-trained noise models. The first technique estimates the noise using pause detection. The second uses maximum likelihood parameter estimation. Both operate within the AR-HMM statistical framework.

Additional work in this dissertation improves the modelling power of AR-HMM systems by the incorporation of perceptual frequency. The improvement can be incorporated into both AR-HMM recognition and enhancement systems. Recognition tests on the ISOLET (Fanty & Cole 1990) and Resource Management (Price, Fisher, Bernstein & Pallett 1988) databases show that this extension improves performance substantially.

The enhancement schemes are evaluated on the NOISEX-92 (Varga, Steeneken, Tomlinson & Jones 1992) and Resource Management (Price et al. 1988) databases giving indications about the performance on both simple and more complex tasks.

The results show that the enhancement scheme based on maximum likelihood noise parameter estimation is superior to that based on noise estimates from pauses. The maximum likelihood scheme is shown to substantially improve on baseline results in terms of recognition performance and distortion measures. Perceptually, the small vocabulary results were the most pleasing with less improvement on the more difficult speaker independent task. An audio Compact Disk containing enhancement examples is supplied with this dissertation.

1.5 Organisation of Thesis

In Chapter 2, the theory of the hidden Markov model and its application to speech modelling is outlined. Chapter 3 then summarises the main methods of speech enhancement presented in the literature. In general, these methods are not able to adapt to changing noise conditions. Therefore, in Chapter 4 existing techniques to adapt to changing environments are studied.

Chapters 5 and 6 contain the main original contribution of this dissertation. In Chapter 5, adaptive enhancement schemes based on autoregressive hidden Markov models are developed. Chapter 6 improves these models by incorporating perceptual frequency.

Chapters 7 and 8 evaluate the techniques developed on small and medium vocabulary tasks respectively. Chapter 9 presents conclusions and suggestions for future work.

Chapter 2

Hidden Markov Models for Speech Modelling

This chapter describes Hidden Markov Models (HMMs) and their application to speech modelling. Particular focus is given to speech recognition systems. In the interests of brevity, only a short description is given here since HMMs are used extensively in the speech processing and other communities and are well understood. Autoregressive HMMs are described in more detail since they feature prominently in this dissertation yet are less well known.

Comprehensive descriptions of HMM theory and its application to speech modelling can be found in (Rabiner 1989), (Rabiner & Juang 1993) and (Deller et al. 1993) and their references. Discussions of the issues involved in building a practical HMM speech recognition system are given in (Young, Woodland & Byrne 1993) and (Young 1996).

2.1 Basic Concepts

A discrete Markov process models a system that can be in one of N distinct states. When in a particular state, an observation is generated according to a probability density function. At regular time intervals, the system changes state also according to a stochastic process. In a first order Markov process, the probability of being in a state depends only upon the previous state.

If the states of the Markov process do not represent physically observable events they are said to be 'hidden'. In this case, the number of states and the probabilities of state transitions must be estimated by observing the output of the process. The resulting model is known as a Hidden Markov Model.

Figure 2.1 shows a typical HMM topology used for speech modelling. Here speech observations $\mathbf{s}$ are emitted from each state x_t according to probability densities $b_{x_t}(\mathbf{s})$. Transitions between states are governed by transition

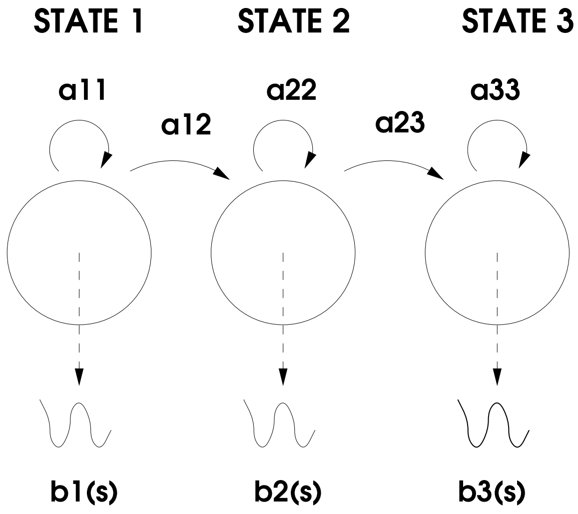

Figure 2.1: Typical HMM for Speech Modelling

probabilities $a_{x_t x_{t+1}}$. Note that not all transitions between the hidden states are legal.

2.2 Probability Distribution

The probability distribution of a HMM is made up of two parts: the distribution of the transition between states and the distribution of the observation given a particular state. The probability density function (pdf) describing a sequence of observations $\mathbf{S}$ given a model with parameters λ is given by

$$p(\mathbf{S}|\lambda) = \sum_X a_{x_0 x_1} \prod_{t=1}^{T} a_{x_t x_{t+1}} b_{x_t}(\mathbf{s}_t). \tag{2.1}$$

Here:

$\mathbf{S}$	$=$ a sequence of K dimensional observations $\{\mathbf{s}_t, t = 1, \ldots, T\}$
X	$=$ a sequence of states $\{x_t, t = 1, \ldots, T\}$
$a_{x_t x_{t+1}}$	$=$ transition probability from state x_t to state x_{t+1}
$b_{x_t}(\mathbf{s}_t)$	$=$ pdf of the observation vector $\mathbf{s}_t$ given the state x_t
N	$=$ the number of states in the model
λ	$=$ the model parameters in general

x_0 is constrained to be the model entry state and x_{T+1} the model exit state.

The pdf $b_{x_t}(\mathbf{s}_t)$ describes the probability of generating $\mathbf{s}_t$ given the model is in state x_t. This function can describe either a discrete or continuous pdf. In this dissertation, $b_{x_t}(\mathbf{s}_t)$ will always describe a continuous function.

In this case, it is usually assumed that $b_{x_t}(\mathbf{s}_t)$ describes a multivariate Gaussian mixture. Therefore

$$b_{x_t}(\mathbf{s}_t) = \sum_{m_t=1}^{M} c_{x_t m_t} b_{x_t m_t}(\mathbf{s}_t). \tag{2.2}$$

Here:

m_t	= denotes the mixture component chosen at time t
M	= the number of mixture components per state
$c_{x_t m_t}$	= the probability of choosing the mixture component m_t given that the process is in state x_t

$b_{x_t m_t}(\mathbf{s}_t)$ is the pdf of the given mixture component m_t in state x_t and is given by

$$b_{x_t m_t}(\mathbf{s}_t) = (2\pi)^{-K/2} |\mathbf{\Sigma}_{x_t m_t}|^{-1/2} \exp\{-\frac{1}{2}(\mathbf{s}_t - \mu_{x_t m_t})' \mathbf{\Sigma}_{x_t m_t}^{-1} (\mathbf{s}_t - \mu_{x_t m_t})\}. \tag{2.3}$$

Here:

K	= dimension of the observations
$\mu_{x_t m_t}$	= mean of the distribution
$\mathbf{\Sigma}_{x_t m_t}$	= covariance of the distribution
$'$	= matrix transpose

2.3 Using HMMs in a Speech Recognition System

In a speech recognition system, T K-dimensional observations $\mathbf{s}_t$ are created from a spoken utterance. These observation vectors are modelled by HMMs. A HMM is built for each recognition template (e.g. phone or word). The problem of speech recognition involves determining a series of recognition templates given the sequence of observation vectors using the given models.

In the isolated word case, each template represents a word. Each sequence of speech observations is assumed to be one word and a template is chosen to match this. The template chosen is the one which maximises the probability of the word given the observation sequence. Thus template w_i is selected from amongst the possible candidate templates in order to maximise $P(w_i|\mathbf{S})$.

$P(w_i|\mathbf{S})$ can be more easily calculated if it is rewritten using Bayes' Rule. Thus

$$P(w_i|\mathbf{S}) = \frac{P(\mathbf{S}|w_i)P(w_i)}{P(\mathbf{S})}. \tag{2.4}$$

The term $P(\mathbf{S}|w_i)$ can be calculated using Equation 2.1 and the model trained for each template. $P(w_i)$ can be used to incorporate prior probabilities of templates.

This concept can be extended to the continuous word case. Here, HMMs are built for sub-word units such as phones. These are then joined together to form a large HMM for each allowable sequence of phones according to the grammar. Depending on the complexity of the grammar, a large number of possible word sequences may be possible for a given utterance. The recognition system must choose the most likely of these sequences.

The choice is still based on maximising $P(w|S)$. However, in this case the maximisation must be performed over all possible sequences of phones or words. Typically, the Viterbi algorithm (e.g. see (Rabiner & Juang 1993)) is used. It efficiently calculates the state sequence which maximises $P(S, X|\lambda)$. Thus the most likely word or phone sequence can be found by constructing a network of all possible state sequences according to the allowable grammar. For large vocabulary systems, pruning of the search space is necessary.

Phones can appear in many different contexts within words and this causes considerable acoustic variation of these sounds (Young 1996). Therefore, if models are constructed for only the 45 or so phones needed for English (or the corresponding number for other languages), these will be poor. The usual solution is to construct triphone models. Here, a model is created for a phone in the context of the preceding and following phone (thus three phones define a triphone hence the name).

However, the number of possible triphones is too great for a separate model to be trained for each given a reasonable amount of training data. Therefore, some form of clustering of acoustically similar triphones or triphone states is required (e.g. see (Young 1996) and associated references).

2.3.1 Training the model

The model parameters are mostly determined using training speech data. Notable exceptions are N, the number of states and M, the number of mixture components per state for which no closed form solution exists. These are chosen according to *a priori* assumptions.

The training process adjusts the model parameters to maximise the likelihood $P(\mathbf{S}|w_i)$. Standard techniques allow this to be calculated efficiently. The forward-backward algorithm (Rabiner & Juang 1993) can be used to calculate $P(\mathbf{S}, x_t = x|\lambda)$. The Baum-Welsh algorithm (Baum, Petrie, Soules & Weiss 1970) iteratively uses these probabilities to find maximum likelihood estimates of the HMM parameters.

2.3.2 Observation Vector Parameterisation

An important factor determining the performance of a HMM is the choice of observation vector parameterisation. The observation vectors are formed by dividing the sampled speech into possibly overlapping frames and applying a transformation to each frame. As a minimum, the parameterisation process involves applying a window to each frame to remove boundary effects.

However, much work has concentrated on deriving a suitable transform which encapsulates the most important features of the speech signal (Rabiner & Juang 1993). Typically, features based on the short-time amplitude spectrum of the signal are used since this has been found to be perceptually more important than the short-time phase. One of the most successful parameterisations is Mel-frequency cepstral coefficients (MFCCs).

Mel-Frequency Cepstral Coefficients

The first step in computing the Mel-frequency cepstral coefficients is to generate the Fourier transform on a Mel scale. This scale is related to perceived frequency and has been shown to improve recognition accuracy when used. Typically the spectrum is also smoothed using overlapping windows.

The next step is to take the logarithm of the Fourier components. This makes the statistics of the resulting vector approximately Gaussian (Young 1996). Finally, the discrete cosine transform is taken. The resulting cepstral coefficients are then decorrelated enough to allow them to be modelled by Gaussian distributions with diagonal covariance matrices. This has significant computational benefits.

The lower order cepstral coefficients represent the slowly-varying part of the speech spectrum. This is perceptually more important than the quickly-varying part and as a result the cepstral feature vector typically consists of only 10-20 of the lower order coefficients.

To incorporate information about changes in spectral information over time, delta coefficients are often used. These are needed because HMMs assume that each vector is independent of the next. This is a poor assumption but including the first order differences alleviates it somewhat. Researchers have also found the incorporation of acceleration or 'delta-delta' coefficients to be beneficial. These incorporate information about changes in the delta coefficients.

2.4 Autoregressive Hidden Markov Models

The class of autoregressive HMMs are used extensively in this dissertation. The main difference between this type of HMM and other types is that the choice of observation vector allows the pdf $b_{x_t m_t}(\mathbf{s}_t)$ to be simplified to an autoregressive process with Gaussian error. A full derivation of the theory is given by Juang (Juang 1984) with the multiple mixture case described in (Juang & Rabiner 1985).

The observation vectors for these HMMs are windowed speech frames. It is assumed that each vector is generated by a Pth order zero mean autoregressive process. That is, the elements s_i of the K dimensional observation

$\mathbf{s}_t$ are assumed to satisfy

$$s_i = -\sum_{n=1}^{P} A_n s_{i-n} + u_i \quad i \epsilon \{1, \ldots, K\}. \tag{2.5}$$

Here, A_n is the nth filter coefficient of the autoregressive process. In the literature, A_n is also known as a linear predictive coding (LPC) coefficient. u_i is the ith term of the error sequence. This error sequence is assumed to be Gaussian with zero mean and variance σ^2.

The autoregressive filter coefficients are usually augmented by an extra term A_0 which is defined to be unity and allows Equation 2.5 to be rewritten

$$\sum_{n=0}^{P} A_n s_{i-n} = u_i \quad i \epsilon \{1, \ldots, K\}. \tag{2.6}$$

Standard techniques exist to determine these filter coefficients and the variance given the observation (e.g. see (Deller et al. 1993)).

Using this autoregressive assumption, the expression for the distribution of the observation vector given the state and mixture component, $b_{x_t m_t}(\mathbf{s}_t)$, can be simplified following the method of Juang (Juang 1984). Thus

$$b_{x_t m_t}(\mathbf{s}_t) = (2\pi)^{-K/2} |\mathbf{\Sigma}_{x_t m_t}|^{-1/2} \exp\{-\frac{1}{2}\mathbf{s}'\mathbf{\Sigma}_{x_t m_t}^{-1}\mathbf{s}\} \tag{2.7}$$

becomes

$$b_{x_t m_t}(\mathbf{s}_t) \approx (2\pi)^{-K/2} (\sigma_{x_t m_t}^2)^{-K/2} \exp\{-\frac{1}{2}\alpha(\sigma_{x_t m_t}^{-1}\mathbf{s}_t; \mathbf{A}_{x_t m_t})\}. \tag{2.8}$$

Here $\mathbf{A} = [A_0 \ldots A_P]$ is the vector of autoregressive filter coefficients.

$\alpha(\sigma^{-1}\mathbf{s}_t; \mathbf{A})$ is a function of r_A and $r_{\mathbf{s}}$, the correlations of $\mathbf{A}$ and $\mathbf{s}$ respectively. It is defined as

$$\alpha(\sigma^{-1}\mathbf{s}_t; \mathbf{A}) = \frac{r_A(0) r_{\mathbf{s_t}}(0)}{\sigma^2} + 2\sum_{i=1}^{P} \frac{r_A(i) r_{\mathbf{s_t}}(i)}{\sigma^2} \tag{2.9}$$

where

$$r_A(i) = \sum_{n=0}^{P-i} A_n A_{n+i} \tag{2.10}$$

$$r_{\mathbf{s_t}}(i) = \sum_{n=1}^{K-i} \mathbf{s}_{t,n}\mathbf{s}_{t,n+i}. \tag{2.11}$$

The term $\alpha(\sigma^{-1}\mathbf{s}_t; \mathbf{A})$ is sometimes known in the literature as the residual error since it calculates the power remaining after $\mathbf{s}_t$ is filtered by the autoregressive filter. This can be demonstrated by writing Equation 2.9 in the power spectral density domain:

$$\alpha(\sigma^{-1}\mathbf{s}_t;\mathbf{A}) = \frac{1}{\sigma^2}\int_{-\pi}^{\pi} S(\omega)|A(e^{j\omega})|^2 \frac{d\omega}{2\pi} \tag{2.12}$$

The residual reflects how closely the autoregressive filter models $\mathbf{s}_t$. Therefore, AR-HMMs compare observation vectors to trained models according to how closely the observations fit the autoregressive filter model for each mixture component of each state.

2.5 Summary

This chapter has briefly described the theory of HMMs and their application to speech modelling. The class of autoregressive HMMs has been described in more detail. The following chapters will describe how HMMs can be used in speech enhancement and recognition systems.

Chapter 3

Techniques for Speech Enhancement

This chapter describes approaches to enhance speech signals degraded by statistically independent additive noise when only one microphone is available. The discussion is also limited to wide-band noise. Most of the techniques described produce an enhanced speech waveform. Some approaches however produce enhanced speech parameters suitable for input directly into, for example, a clean speech recogniser. These latter techniques thus fall into the category of robust speech recognition as well as speech enhancement. It should be noted however that robust speech recognition is a complete research topic in itself, and that using speech enhancement as a front-end to a recogniser is only one solution to the problem. Recent reviews of approaches to robust speech recognition can be found in (Gong 1995) and (*ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels* 1997).

The major approaches to speech enhancement are:

- spectral subtraction
- methods utilising the periodicity of speech
- noise masking
- filter-model-based approaches
- enhancement by synthesis
- statistical-model-based approaches.

The first three techniques incorporate assumptions about clean and noisy speech but do not use explicit speech models. Filter-model-based approaches model speech as an autoregressive process and determine the model parameters from the signal to be enhanced. The last two techniques incorporate prior information typically from training examples and use speech and noise models of increasing sophistication. Each of these approaches is described in more detail below.

3.1 Spectral Subtraction

Spectral subtraction (Boll 1979) is one of the most influential speech enhancement algorithms favoured for its simplicity and performance in stationary noise. This approach capitalises on the fact that the amplitude of the short-time speech spectrum is perceptually more important than the phase (e.g. see (Deller et al. 1993)). It estimates the amplitude of the spectrum of the speech signal given a noisy signal and the noise statistics.

Specifically, for speech s_n corrupted by additive noise d_n, the corrupted speech y_n is given by

$$y_n = s_n + d_n. \tag{3.1}$$

Assuming that the speech and noise statistics are stationary and that the noise is uncorrelated with the speech, the energy spectrum of the noisy signal is given by

$$|Y(\omega)|^2 = |S(\omega)|^2 + |D(\omega)|^2. \tag{3.2}$$

The classic spectral subtraction technique (Boll 1979) estimates the speech spectrum as

$$|\hat{S}(\omega)|^2 = |Y(\omega)|^2 - E\left\{|D(\omega)|^2\right\} \tag{3.3}$$

where $E\left\{|D(\omega)|^2\right\}$ denotes the expected value of the noise energy spectrum. This is obtained by averaging the energy spectrums during non-speech activity.

Since the phase of speech is assumed perceptually less important than the amplitude, the noisy phase is used to reconstruct the enhanced signal. Therefore the enhanced speech is estimated by

$$\hat{s}_n = F^{-1}[|\hat{S}(\omega)| \cdot \exp(j\angle Y(\omega))] \tag{3.4}$$

where $\angle Y(\omega)$ is the phase of y_n and F^{-1} denotes the inverse Fourier Transform.

Since speech is only stationary over short time periods, the analysis is performed over short 20-40ms sections of speech called frames. Typically overlapping frames are used. Windowing in the time domain compromises the quality of estimates in the spectral domain. Therefore, a customised smoothing window such as the Hamming window (e.g. see (Rabiner & Schafer 1978)) is used rather than simply 'blocking' the speech into frames.

Many variations of the spectral subtraction technique can be found in the literature (Lim & Oppenheim 1979*a*). The most general form of the subtraction is given by

$$|Y(\omega)|^a = |S(\omega)|^a + kE\left\{|D(\omega)|^a\right\} \tag{3.5}$$

where the parameters a and k are chosen to optimise the enhancement.

The major problem with these techniques is that residual noise remains in the enhanced spectrum because only the average noise spectrum is subtracted. Fluctuations in the actual noise spectrum for each frame mean that spurious spectral components are introduced at the frame rate. This noise, known as 'musical noise', decreases speech quality both for human listeners and recognition systems alike.

Another problem is that the speech spectrum estimate in Equation 3.3 may be negative. In this case some form of ad hoc solution must be applied which invariably involves a non-linear operation such as reflecting the negative spectral component or setting it to zero.

A further problem is that the noise statistics must be known or estimated from the non-speech portions of the signal to be enhanced. Since no information about the speech is known, this may be non-trivial.

A related spectral domain method is Wiener filtering. The Wiener filter is the optimal filter which gives the minimum mean squared error (MMSE) estimate of the speech in the time domain (Anderson & Moore 1979). The frequency response of this filter is

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} \tag{3.6}$$

where P_s is the power spectral density (PSD) of the speech and P_d is the PSD of the noise. Again the analysis is performed on a per frame basis with only the spectral magnitude being estimated and the noisy phase being used when the signal is reconstructed. The speech and noise PSDs must be known *a priori* or estimates made from the noisy speech. Enhancement systems based on Wiener filters will be discussed in the following sections.

In (Ephraim & VanTrees 1995), a technique is described which is shown to be a generalisation of the spectral subtraction approach. The noisy signal is decomposed using the Karhunen-Loève transform into a signal-plus-noise subspace and a noise subspace. Enhancement involves removing the noise subspace and estimating the clean speech from the remaining subspace. Estimates are developed according to perceptually meaningful criteria such as minimising signal distortion while curtailing residual noise energy. Listening tests show improvement over the basic spectral subtraction approach.

3.2 Methods Utilising the Periodicity of Voiced Speech

The techniques in this subsection rely on the fact that voiced speech is almost periodic. They include comb filtering and single channel adaptive noise cancelling.

Comb filtering (Malah & Cox 1982) capitalises on that fact that voiced speech segments have a harmonic spectrum whereas the spectrum for white noise is essentially flat. A filter is constructed which passes the harmonic components and attenuates the noise. To take into account the fact that speech is not precisely periodic, some techniques use warping to improve the periodicity of the speech before applying the filter, and then unwarp it afterwards (Graf & Hubing 1993), (Ramalho & Mammone 1994).

Single channel adaptive noise cancelling (Sambur 1978) is a one microphone implementation of adaptive noise cancelling (ANC). ANC estimates the clean speech given two sources: the noisy speech and a signal correlated with the noise. The single channel approach uses a delayed version of the original signal as a second source. It can be shown that if a delay of T, the pitch period, is used, then an estimate of the clean speech $\hat{s}_n$ from the noisy signal y_n is given by (Deller et al. 1993)

$$\hat{s}_n = 0.5 y_n + 0.5 y_{n-T}. \tag{3.7}$$

One disadvantage of these methods is that they rely heavily on the accuracy of the determination of pitch and degree of voicing in noise, which is far from simple. A further problem is that they generally perform poorly on unvoiced speech. This implies that intelligibility may not be improved since consonants are generally more important than vowels in conveying information. In fact evaluation of a comb filter technique in (Lim & Oppenheim 1979*b*) showed that although the signal to noise ratio (SNR) was improved, intelligibility tended to decrease. Given these shortcomings, these techniques are normally not used to attenuate broadband additive noise (Deller et al. 1993).

3.3 Noise Masking

People cannot detect sounds below a certain threshold if they are masked by another sound such as noise. This phenomenon is known as noise masking (Klatt 1976). It can be used to enhance speech parameters for recognition or coding and hence make them more robust to noise.

For example, in (Van-Compernolle 1989) a noise-robust speech recogniser is developed. Here the speech is pre-processed using spectral subtraction. Noise is then added before parameterising the speech for recognition. The advantage is two-fold. First, the environment-dependent residuals from the spectral subtraction front-end are masked. Second, low-level speech events are masked so that the recogniser does not learn features which may not be distinguishable in noisy conditions.

Other work (Nadas, Nahamoo & Picheny 1989) develops a filterbank speech recognition system assuming that the energy observed in each noisy filterbank is the maximum of the speech and noise energies. This assumption is based on visual observation of spectrums corrupted by additive noise. The speech is modelled by a mixture of Gaussian densities and techniques to estimate the parameters of this model and the likelihood of a particular model given a noisy observation are developed. The approach is shown to substantially decrease the recognition error rate of a speaker independent medium vocabulary task corrupted by noise at relatively high SNRs (around 30dB). A major drawback though is that it is necessary to train new models for different noise conditions.

3.4 Filter-Model-Based Approaches

Filter-model-based techniques assume a linear filter model for the speech and estimate its parameters given a noisy signal. The use of a model for speech is advantageous because it gives the enhancement system more information about the signal to extract.

The speech is modelled by a linear filter representing the vocal tract. This filter is excited by either a turbulent source (unvoiced sounds) or a periodic source (voiced sounds). In the z-domain, the speech $S(z)$ is represented by

$$S(z) = H(z) \cdot U(z) \tag{3.8}$$

where $H(z)$ denotes the filter and $U(z)$ the excitation. Typically $H(z)$ is an all-pole linear filter so that Equation 3.8 becomes

$$S(z) = \frac{1}{A(z)} \cdot U(z). \tag{3.9}$$

In the time domain this is written as

$$s_n = \sum_{k=1}^{P} A_k s_{n-k} + u_n \tag{3.10}$$

where $A_k, k = 1, \ldots, P$ are the filter parameters and P is the filter order. This equation is equivalent to Equation 2.5 introduced earlier. It should be noted thus that u_n can be viewed either as the excitation of the filter or as the modelling error. By letting u_n be normally distributed with mean zero and variance σ^2

$$s_n = \sum_{k=1}^{P} A_k s_{n-k} + \sigma e_n \tag{3.11}$$

where e_n is normally distributed with zero mean and unit variance.

The model is known as an autoregressive or linear predictive model. It is used extensively in many speech processing systems and particularly in speech coders. Standard techniques exist to find the autoregressive parameters $\{A_k, k = 1 \ldots P, \sigma\}$ which minimise the modelling error (Deller et al.

1993). Again, since the characteristics of the vocal tract are not stationary, the speech signal is segmented into frames and the filter parameters are calculated on a per frame basis.

3.4.1 The Basic Technique

A filter-model-based enhancement system was first presented in (Lim & Oppenheim 1978). Here the filter parameters are determined from noisy speech using maximum *a posteriori* (MAP) estimation (Van-Trees 1968). Since this procedure leads to non-linear equations, an iterative scheme is proposed to estimate the parameters. This is shown under certain conditions to be equivalent to iteratively applying Wiener filters to each frame.

As described in Section 3.1, the transfer function for the Wiener filter is given by

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)}. \tag{3.12}$$

where $P_s(\omega)$ is the power spectral density of the speech and $P_d(\omega)$ is the power spectral density of the noise.

$P_s(\omega)$ is estimated using Equations 3.9 to 3.11 as

$$\hat{P}_s(\omega) = |S(\omega)|^2 = \frac{\sigma^2}{|A(\omega)|^2} \tag{3.13}$$

σ^2 can be estimated using Parseval's Theorem (Deller et al. 1993). $P_d(\omega)$ is typically estimated from non-speech portions of the signal.

Once determined, the filter parameters can be used to reconstruct the enhanced speech. Alternatively these parameters (or transforms of them) can be used directly in a speech coding or recognition system. If a speech coding system is being developed, the pitch and degree of voicing must be determined for each frame. Lim and Oppenheim (Lim & Oppenheim 1979*a*) warn that these parameters (particularly the degree of voicing) are not easily estimated in the presence of noise.

Aside from these considerations, this basic enhancement technique has several other problems. It is computationally expensive and a heuristic convergence criterion must be applied. As additional iterations are performed, formants decrease in bandwidth and shift in location. Also, frame to frame pole jitter is observed. Another problem is that the MAP estimator is biased to certain speech classes. The fact that the statistics of the interfering noise are assumed to be stationary and must be determined *a priori* is a further deficiency.

In light of these problems many extensions to and variations on this enhancement method have been proposed. These are discussed in the following section.

3.4.2 Extensions and Variations

Hansen and Clements (Hansen & Clements 1991) solve some of the convergence problems by enforcing inter-frame and intra-frame constraints on the speech generated by the algorithm. This work is further extended in (Arslan & Hansen 1994) using HMMs to partition the noisy speech into broad phone classes which further determine which constraints are to be used. These techniques however require more computation.

Musicus and Lim (Musicus & Lim 1979) consider modelling the speech using a pole-zero model. Again the filter coefficients are estimated using MAP estimation.

In (Ephraim, Wilpon & Rabiner 1987) autoregressive models determined directly from noisy speech are used in a robust recognition system. The scheme iteratively chooses model parameters such that the Itakura-Saito distortion measure (Gray, Buzo, Gray & Matsuyama 1980) is minimised.

The use of Kalman rather than Wiener filters has also been investigated. These filters also determine the MMSE estimator but have the advantage that they intrinsically deal with non-stationary signals. Therefore, slight variations in speech statistics during a frame can be accommodated.

In (Paliwal & Basu 1987), it is shown that if the ideal autoregressive parameters are used then the Kalman filter outperforms the Wiener filter in terms of output SNR values. In (Gibson, Koo & Gray 1991), scalar and vector Kalman filters are applied to the additive coloured noise enhancement problem. Here, the autoregressive parameters are estimated from the noisy speech using a MAP approach similar to (Lim & Oppenheim 1978). When the noise is coloured, it can be assumed to be modelled by a known autoregressive process.

Some researchers have addressed the problem of finding the noise statistics from the utterance to be enhanced (Paliwal 1988), (Lee, Lee & Ann 1997), (Saleh & Niranjan 1998). In (Paliwal 1988), an overdetermined system is considered. An estimate of the autoregressive parameters is formed using the high order Yule-Walker equations and then used to estimate the noise variance from the low order Yule-Walker equations. The technique estimates the noise variance well, even in the presence of non-stationary noise. However, the scheme is only applicable to Gaussian noise.

Saleh (Saleh 1996), (Saleh & Niranjan 1998) introduces a Bayesian framework to the original Lim enhancement technique (Lim & Oppenheim 1978). The advantage of this approach is that the variance of the additive noise and the gain of the all-pole model appear as *hyper-parameters* within this framework and can be estimated using standard Bayesian assumptions. This means that it is not necessary to know the interfering noise statistics *a priori*. The approach also works in non-stationary Gaussian noise. Although the enhancement still suffers from residual noise problems, preliminary experiments show that it outperforms the original Lim and Oppenheim technique in terms of SNR improvement and informal listening tests.

In (Lee et al. 1997), the parameters of coloured interfering noise are estimated using a maximum likelihood approach. Autoregressive models are assumed for both the speech and the noise. This technique will be further discussed in Chapter 4 along with other maximum likelihood approaches.

Although filter-model-based enhancement schemes produce effective enhancement, their performance is always limited by the quality of the autoregressive models which can be obtained from the corrupted speech. Some of the following techniques use models trained from speech samples and are thus able to provide superior performance.

3.5 Enhancement by Synthesis

With the exception of Noise Masking, all the other enhancement techniques described in this chapter involve *estimation* of the clean speech waveform or parameters from a noisy observation. They are thus concerned with finding the optimal *filter* which will produce an enhanced signal. The following two techniques however consider enhancement as a *detection* problem. In this case the aim is to determine the clean speech (or clean speech parameters) that corresponds to the noisy observations. The enhanced speech (or its parameters) is then *synthesised* from a combination of these clean templates.

The major difference between a filtering and synthesis system is that if both are applied to clean speech, the operation of the ideal filter is transparent whereas the synthesised clean speech will always differ from the original. Therefore the performance of an Enhancement by Synthesis system is upper bounded by the quality of the synthesis technique (Ephraim 1992*c*).

3.5.1 Parameter Mapping

This approach tries to learn a mapping from noisy speech parameters to clean parameters. These parameters can then be fed directly into a recognition or coding system. No explicit model of the speech or noise or the way in which they are combined is used. Approaches vary from using linear transformations to artificial neural networks to implement the mapping.

For example, in (Mokbel & Chollet 1992), noisy MFCC vectors are mapped to clean MFCC vectors using a linear transformation. The parameters of this transformation are estimated to minimise the mean squared error between the estimated vector and actual clean vectors. The transformation is shown to improve recognition performance in a car and to be superior to a Kalman filter approach.

A disadvantage of some of these techniques is that their performance may be highly dependent on the noise level at which the mapping is learnt. Also, to learn the mapping an amount of clean speech must be available in stereo with noisy data. This may be unrealistic in some circumstances.

3.5.2 Template-Based Enhancement

Template-Based enhancement techniques are similar to parameter mapping in that they find the best sequence of clean speech templates given the noisy data using some form of correspondence between the two. In this case some form of speech model and information about the manner in which the speech and noise are combined may be included. The parameters to be estimated are restricted to a combination of clean templates.

In (Juang & Rabiner 1987), clean spectral vectors are vector-quantised and noisy versions determined for each cluster. The nearest neighbour of each observed noisy vector in the noisy parameter space is then determined and this is mapped to the average of the corresponding clean speech cluster. The distance measure used to compare vectors is either the likelihood ratio or the cepstral distortion measure. A similar vector-quantised system using formant-based distance measures for the noisy speech vectors is developed in (O'Shaughnessy 1988).

Other template-based approaches include using a linear combination of sine wave templates (Quatieri & McAulay 1990) and phone-based templates (Gong 1993). The work in (Gong 1993) has been extended to use an extra mapping to determine the level of interfering noise (Treurniet & Gong 1994).

3.6 Statistical-Model-Based Approaches

The class of statistical model-based techniques estimates the clean speech signal from the noisy signal using statistical models of the speech and noise. Given a distortion measure and the joint statistics of the speech and noise, enhancement is performed using the estimator which minimises the expected value of the distortion measure between the clean and estimated signals (Ephraim 1992*c*).

For a given distortion measure $d(\mathbf{s}_t, \hat{\mathbf{s}}_t)$ between the clean speech $\mathbf{s}_t$ and enhanced speech $\hat{\mathbf{s}}_t$, and given noisy observations $\mathbf{Y} = \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T$, the approach is to determine $\hat{\mathbf{s}}_t = f(\mathbf{Y})$ such that $E\{d(\mathbf{s}_t, \hat{\mathbf{s}}_t)|\mathbf{Y}\}$ is minimised. If the optimal distortion measure and the true speech and noise statistics were known, optimal enhancement could be performed. Since this is not the case, tractable distortion measures and parametric models trained from existing speech and noise data are used.

3.6.1 Distortion Measures

Two commonly used distortion measures are squared error cost and uniform cost. Although more perceptually meaningful distortion measures exist (Gray et al. 1980), these two measures are often used because they allow mathematically tractable solutions to be found.

The squared error cost function yields Minimum Mean Squared Error (MMSE) estimation which gives the conditional mean of the parameter given the observation. Thus in this case $\hat{\mathbf{s}}_t$ is given by

$$\begin{aligned}\hat{\mathbf{s}}_t &= E\{\mathbf{s}_t|\mathbf{Y}\} && (3.14)\\ &= \int \mathbf{s}_t p_{\mathbf{s}}(\mathbf{s}_t|\mathbf{Y})d\mathbf{s}_t. && (3.15)\end{aligned}$$

Here $p_{\mathbf{s}}(\mathbf{s}_t|\mathbf{Y})$ is the pdf of the clean speech at time t given the complete noisy observation.

The uniform cost distortion measure yields maximum *a posteriori* (MAP) estimation which is the value that maximises the probability density function of the parameter given the observation. Here

$$\hat{\mathbf{s}}_t = \arg\max_{\mathbf{s}_t} p_{\mathbf{s}}(\mathbf{s}_t|\mathbf{Y}). \quad (3.16)$$

If the observations and parameters are jointly Gaussian, these estimates are identical.

3.6.2 Speech and Noise Statistical Models

For reasons of reliability of model estimation and tractability it is best to use parametric models for the speech and noise probability distributions (PDs) (Ephraim 1992*c*). This is advantageous for three reasons. First, a sensibly chosen model can provide a better estimate of a PD than a sample distribution estimate because it will capture intrinsic information about the signal which may not be present in a particular set of training data. Second, since generally only a small number of parameters are required to be estimated they can be reliably estimated from a reasonable amount of training data. Third, closed form expressions of the enhanced signal can be determined as functions of the model parameters thus only these parameters and not the entire training data need to be stored as part of the speech enhancement system.

3.6.3 Spectral Amplitude Models

In (Ephraim & Malah 1984) enhancement is performed using a MMSE estimator of the short-time spectral amplitude. The amplitude of each spectral component is assumed independent with a Gaussian distribution. Strictly, this assumption only holds as the analysis frame length approaches infinity. However the authors justify its use on the grounds that it leads to simple estimators and good enhancement.

The MMSE estimator can be improved by conditioning the estimate on the presence or absence of speech, an idea first proposed in (McAulay & Malpass 1980). This work is closely related to (Ephraim & Malah 1984) but differs in that a maximum likelihood rather than MMSE estimate of the spectral amplitude is formed. The major contribution of (McAulay & Malpass 1980) is the concept of forming different estimators according to whether speech is present or absent, and then forming the enhanced speech as a weighted sum of these estimators. Here the weighting is given by the probability of speech presence or absence. This concept of using a weighted sum of conditioned estimators is an important technique which appears in many statistical model-based and other enhancement approaches.

The technique in (Ephraim & Malah 1984) produces better enhancement than (McAulay & Malpass 1980) although both suffer from residual noise problems. In the case of (Ephraim & Malah 1984) though, this noise is colourless and considered less annoying. Later work (Ephraim & Malah 1985) looks at a MMSE log spectral estimator on the grounds that distortion measures based on the log spectrum are perceptually more meaningful. This results in a superior but computationally more complex enhancement scheme. Other work (Porter & Boll 1984) develops a non-parametric implementation of (Ephraim & Malah 1984) which, although it does not make the assumption that the speech spectral components have a Gaussian distribution, suffers from the disadvantages of non-parametric systems (see Section 3.6.2).

Erell and Weintraub (Erell & Weintraub 1993) implement MMSE estimators to estimate the filterbank log energies used in a speech recognition system. To incorporate information about the correlations between filter channels, the estimator is conditioned on the total frame energy. The algorithm is effective in improving recognition performance.

Another scheme by the same authors (Erell & Weintraub 1994) alternatively tries to solve the problem of the correlations by incorporating information about the pitch period. However, this estimator only improves recognition for voiced speech. Since most errors occur in the unvoiced sections, the recognition accuracy is not significantly improved.

3.6.4 SNR-Based System

In (Gish, Chow & Rohlicek 1990), a probabilistic model conditioned on the instantaneous SNR is used. Clean speech is partitioned using vector quantisation and a mixture of Gaussians is used to model each cluster. Using MMSE estimation, a linear mapping is then determined for each cluster to map noisy vectors to clean speech vectors given the instantaneous SNR γ. The enhancement process for each noisy observation $\mathbf{y}_t$ with SNR γ is as follows. First, cluster $\hat{k}$ is chosen to maximise $p(k|\mathbf{y}_t, \gamma)$. Then the mapping for cluster $\hat{k}$ is used to obtain the enhanced vector.

The technique is shown to improve the performance of a word spotting system. It is an attempt to subvert some of the problems of the systems in Section 3.5.

3.6.5 HMM-Based Systems

A popular choice of parametric models for speech signals are HMMs. These models (described in Chapter 2) have dominated the field of speech processing in recent years. They are able to model the time-varying characteristics of the vocal tract. A contributing factor to their popularity has been the availability of tractable algorithms to train the model parameters and evaluate model probabilities.

The class of Gaussian mixture models also falls into the HMM category since a Gaussian mixture model is equivalent to a one state HMM (Ephraim 1992*c*). Here each mixture component represents a cluster of spectrally similar speech signals. However, no Markovian constraint is placed on the evolution of mixtures as in the case of a multistate HMM. Hence a multistate HMM is better for speech modelling (Ephraim 1992*c*).

The choice of observation vector modelled by the HMM determines which type of HMM is used. Many state-of-the-art speech recognisers use cepstral features (see Section 2.3.2). However, these vectors are formed by non-linear transformations on the raw speech vectors and are inherently less suitable for the additive noise case. Therefore, many enhancement schemes are based on either spectral features or autoregressive HMMs which model raw speech vectors. The various techniques are described below.

Autoregressive HMM-Based Enhancement

Ephraim has developed various enhancement schemes based on autoregressive HMMs. These models have been described in Section 2.4. Each mixture component of each HMM state represents observations that have similar autoregressive or equivalently spectral parameters. Thus a non-stationary autoregressive process is modelled. The autoregressive parameters are used to construct Wiener filters to perform enhancement similar to the Lim and Oppenheim technique (Lim & Oppenheim 1978). The difference here though is that the speech autoregressive parameters are estimated from clean speech training data rather than from the noisy signal as in (Lim & Oppenheim 1978). Two main schemes have been developed. These are discussed below.

Clean HMM Scheme

In (Ephraim, Malah & Juang 1989), the Expectation Maximisation (EM) algorithm (e.g. see (Little & Rubin 1987)) is used to iteratively find the the optimal MAP estimator of the clean speech. It can be shown that this is equivalent to maximising $p(\mathbf{S}|\mathbf{Y})$ where $\mathbf{S}$ and $\mathbf{Y}$ represent clean speech and noisy observation sequences as before. The estimator for the kth iteration is given by

$$\hat{\mathbf{s}}_t^k = \left[\sum_{\beta,\gamma} q_t(\beta, \gamma | \hat{\mathbf{S}}_t^{k-1}) H_{\gamma|\beta}^{-1} \right]^{-1} y_t. \tag{3.17}$$

Here $q_t(\beta, \gamma | \hat{\mathbf{S}}_t^{k-1})$ is the conditional probability of being in state β and choosing mixture component γ given observation $\hat{\mathbf{S}}_t^{k-1}$ at time t. This probability can be calculated using the clean speech HMM. $H_{\gamma|\beta}$ is the Wiener filter for the output Gaussian process given state β and mixture component γ and the noise process statistics:

$$H_{\gamma|\beta} = \frac{S_{\gamma|\beta}}{S_{\gamma|\beta} + D}. \tag{3.18}$$

Here $S_{\gamma|\beta}$ and D are the power spectral densities of the speech mixture component and noise process respectively. The noise statistics are trained from silence portions of the signal. The algorithm is initialised by setting $\hat{\mathbf{s}}_t^0$ to $\mathbf{y}_t$ for all t.

An approximate MAP estimation method is also presented which uses the most likely Wiener filter for each frame rather than the weighted sum.

Experiments show that the algorithm is very dependent on the accuracy of choosing the correct state and mixture for each frame. For low SNRs, this can be difficult since a clean model is used to calculate the probability of each mixture of each state given initially the noisy observation.

This scheme is implemented in (Sheikhzadeh, Sameti, Deng & Brennan 1994) and compared with the spectral subtraction method on noise from 0 to 20dB. The HMM-based system was seen to be superior in terms of SNR improvement and Mean Opinion Score (MOS) tests. The improved performance is due to the use of filters based on prior speech and noise information.

Compensated HMM Scheme

The work in (Ephraim et al. 1989) is extended in (Ephraim 1992*a*). In this later work, the problem of estimating the clean state and mixture component corresponding to each noisy observation frame is addressed and solved using a compensated HMM formed by combining speech and noise statistics. This alleviates the need for an iterative scheme and provides a better state-frame alignment and hence a better selection of filters.

The pdfs describing the clean speech and noise processes are each given by the usual HMM pdf first introduced as Equation 2.1. Thus

$$p(\mathbf{S}) = \textstyle\sum_{\mathbf{X}} a_{x_0 x_1} \prod_{t=1}^{T} a_{x_t x_{t+1}} b_{x_t}(\mathbf{s}_t) \tag{3.19}$$

$$p(\mathbf{D}) = \textstyle\sum_{\tilde{\mathbf{X}}} a_{\tilde{x}_0 \tilde{x}_1} \prod_{t=1}^{T} a_{\tilde{x}_t \tilde{x}_{t+1}} b_{\tilde{x}_t}(\mathbf{d}_t) \tag{3.20}$$

Here $\mathbf{S}$ is a sequence of clean speech observations, $\mathbf{X}$ is a sequence of clean speech states, $a_{x_t x_{t+1}}$ is the transition probability from state x_t to state x_{t+1}

and $b_{x_t}(\mathbf{s}_t)$ is the pdf of the output vector $\mathbf{s}_t$ from the state x_t. Similarly $\mathbf{D}$ is a sequence of noise observations and $\tilde{\mathbf{X}}$ is a sequence of noise states.

The pdfs $b_{x_t}(\mathbf{s}_t)$ and $b_{\tilde{x}_t}(\mathbf{d}_t)$ are assumed Gaussian with zero mean and covariance matrices $\mathbf{\Sigma}_{x_t}$ and $\mathbf{\Sigma}_{\tilde{x}_t}$ respectively. Since the processes are assumed autoregressive, these covariance matrices are dependent on only $P+1$ parameters where P is the order of the autoregressive process (Juang 1984).

For additive noise, these speech and noise models can be combined to produce a model for noisy speech. The pdf for this model is given by Equation 3.21.

$$p(\mathbf{Y}) = \sum_{\bar{\mathbf{X}}} a_{\bar{x}_0\bar{x}_1} \prod_{t=1}^{T} a_{\bar{x}_t\bar{x}_{t+1}} b_{\bar{x}_t}(\mathbf{y}_t) \tag{3.21}$$

Here $\mathbf{Y}$ is a sequence of noisy observations, $\bar{\mathbf{X}}$ is a sequence of composite states with $\bar{x}_t \equiv (x_t, \tilde{x}_t)$, $a_{\bar{x}_t\bar{x}_{t+1}}$ is the transition probability from state $\bar{x}_t$ to state $\bar{x}_{t+1}$ and $b_{\bar{x}_t}(\mathbf{y}_t)$ is the pdf of the output vector $\mathbf{y}_t$ from the state $\bar{x}_t$. For additive noise

$$\begin{aligned} \mathbf{Y} &= \mathbf{S} + \mathbf{D} && (3.22) \\ a_{\bar{x}_t\bar{x}_{t+1}} &= a_{x_t x_{t+1}} a_{\tilde{x}_t\tilde{x}_{t+1}} && (3.23) \\ b_{\bar{x}_t}(\mathbf{y}_t) &= \int b(\mathbf{y}_t - \mathbf{s}_t) b(\mathbf{s}_t) d\mathbf{s}_t. && (3.24) \end{aligned}$$

The pdf $b_{\bar{x}_t}(\mathbf{y}_t)$ is Gaussian with zero mean and covariance matrix $\mathbf{\Sigma}_{\bar{x}}$ given by

$$\mathbf{\Sigma}_{\bar{x}} = g^2 \mathbf{\Sigma}_x + \mathbf{\Sigma}_{\tilde{x}}. \tag{3.25}$$

Here g^2 is a gain to take into account the mismatch between training data (for the clean speech models) and testing data. The calculation of g^2 and a mathematically tractable technique to calculate the determinant and inverse of $\mathbf{\Sigma}_{\bar{x}}$ are described in (Ephraim 1992*b*).

The conditional pdf of $\mathbf{s}_t$ given $\mathbf{Y}$ can now be written

$$p(\mathbf{s}_t|\mathbf{Y}) = \sum_{\bar{x}_t} p(\bar{x}_t|\mathbf{Y}) b_{\bar{x}_t}(\mathbf{s}_t|\mathbf{y}_t). \tag{3.26}$$

The MMSE estimate of $\mathbf{s}_t$ given $\mathbf{Y}$ is therefore given by

$$\begin{aligned} \hat{\mathbf{s}}_t &= E\{\mathbf{s}_t|\mathbf{Y}\} \\ &= \sum_{\bar{x}_t} p(\bar{x}_t|\mathbf{Y}) E\{\mathbf{s}_t|\mathbf{y}_t, \bar{x}_t\} \\ &= \sum_{\bar{x}_t} p(\bar{x}_t|\mathbf{Y}) H_{\bar{x}_t}\mathbf{y}_t && (3.27) \end{aligned}$$

where $H_{\bar{x}_t}$ is the MMSE estimator of the signal $\mathbf{s}_t$ given $\mathbf{y}_t$. Here $H_{\bar{x}_t}$ is the Wiener filter formed using composite state $\bar{x}_t$. This corresponds to forming the Wiener filter from speech state x_t and noise state $\tilde{x}_t$. The filter is given by

$$H_{(x_t,\tilde{x}_t)} = \frac{\Sigma_x}{\Sigma_x + \Sigma_{\tilde{x}}}. \tag{3.28}$$

This is similar to the Wiener filter formed in Equation 3.18. Thus the MMSE estimator is given by the weighted sum of Wiener filters for each combination of speech and noise states, weighted by the probability of that combination.

The model parameters for the clean signal and noise process can be determined from training data as in a standard HMM training problem. The $p(\bar{x}_t|\mathbf{Y})$ term in the conditional process can be calculated using the forward-backward formulas for HMMs.

Other estimators, such as the MMSE spectral amplitude estimator or MMSE log spectral amplitude estimator can also be used instead of Wiener filters. These estimators are similar to those described in Section 3.6.3 except that each estimator is conditioned on the speech and noise state.

The MAP estimator of $\mathbf{s}_t$ given $\mathbf{Y}$ is obtained by maximising $p(\mathbf{s}_t|\mathbf{Y})$ over $\mathbf{s}_t$. This is done numerically using the EM algorithm to iteratively maximise $p(\mathbf{s}_t|\mathbf{Y})$. This results in a reestimation formula similar to that developed for the clean HMM scheme above.

$$\hat{\mathbf{s}}_t^k = \left[\sum_{\bar{x}_t} p(\bar{x}_t|\hat{\mathbf{s}}_t^{k-1}, \mathbf{Y}) H_{\bar{x}_t}^{-1}\right]^{-1} \mathbf{y}_t. \tag{3.29}$$

Additionally, an approximate MAP estimator can be formed as

$$\hat{\mathbf{s}}_t = \max_{\bar{x}_t} p(\bar{x}_t|\mathbf{Y}) H_{\bar{x}_t}\mathbf{y}_t. \tag{3.30}$$

The approach is evaluated using clean speech models trained on 5 minutes of conversational speech from 6 speakers. White noise at SNRs ranging from 5dB to 10dB is added to test sentences. The enhanced speech is evaluated in terms of SNR improvement and informal listening tests. Not a great deal of variation is noted between the different algorithms in terms of SNR improvement although all were successfully able to improve upon the SNR. Listening tests on a MMSE waveform estimator demonstrate that although some residual noise is still present after enhancement at 10dB, it is less annoying than the 'musical' noise typical of spectral subtraction enhancement systems. The MAP estimator has more residual noise than the MMSE estimator and is thus judged inferior. This is fortuitous since the MMSE estimator is less computationally expensive than the MAP estimator, the latter being a iterative scheme.

Enhancement Based on Autoregressive HMM Variations

Lee et al. have developed enhancement schemes based on variations to the basic AR-HMM. In (Lee, Lee, Song & Yoo 1996) and (Lee & Shirai 1996), schemes based on the Hidden Filter HMM (HF-HMM) are studied. This model (Sheikhzadeh & Deng 1994) is a special case of the autoregressive HMM with a frame size of one sample. It is assumed that such HMMs are better able to model quickly-varying parts of the speech signal since information is not lost at the frame boundaries.

In (Lee & Shirai 1996), the general coloured noise case is considered. HF-HMM parameters and noise parameters are trained from *a priori* information. The noise is modelled as an autoregressive process. Since the HF-HMM lends itself naturally to time domain analysis, the system is represented by a state-space model as follows.

Using the notation previously introduced, and letting $\mathbf{s}_t = [s_t, s_{(t-1)}, \ldots, s_{(t-P)}]'$ be the speech observation and $\mathbf{d}_t = [d_t, d_{(t-1)}, \ldots, d_{(t-P)}]'$ be the noise observation, the augmented state is given by $\mathbf{z}_t = [\mathbf{s}_t : \mathbf{d}_t]'$ where $'$ denotes matrix transpose and P is the order of the state-space analysis. This augmented state space formulation is also used in (Gibson et al. 1991).

In (Gibson et al. 1991), a Kalman filter is then used to form a MMSE estimator of the speech given the state-space model. In (Lee & Shirai 1996) however, an estimator $\bar{\mathbf{z}}_{t,j}$ for each Markov state j of the HF-HMM is formed. These estimators are then combined in the same way as in AR-HMM systems above. That is, if the HF-HMM has N states, the MMSE of $\bar{\mathbf{z}}_t$ is given by

$$\hat{\bar{\mathbf{z}}}_t = \sum_{j=1}^{N} \bar{\mathbf{z}}_{t,j} p(j|\mathbf{Y}) \tag{3.31}$$

where $p(j|\mathbf{Y})$ is the probability of state j given the noisy observation at time t and the HF-HMM.

The fact that this algorithm does not divide the speech into frames makes it intuitively appealing. However, this strength has computational consequences. In order to train the parameters of a clean HF-HMM system, $P+1$ autocorrelation coefficients must be calculated using $P+1$ speech samples for each of the T speech samples in the training data (Sheikhzadeh & Deng 1994). Conversely, frame-based AR-HMMs require $P+1$ autocorrelation coefficients to be calculated for each frame. K speech samples are used for the calculation where K is the frame size, but the coefficients need only be calculated for the $\frac{T}{K}$ frames (Juang 1984). Thus the former training procedure is of the order of a factor of $P+1$ slower. This would restrict the application of HF-HMM-based enhancement algorithms to systems which only required limited training data.

Another potential problem with this enhancement algorithm is that the pdf $p(j|\mathbf{Y})$ is not rigorously calculated as in (Ephraim 1992*a*) but approximated by a normal distribution. Also, there is little scope for forming anything other than a MMSE time domain estimator which may be perceptually inferior to MMSE spectral estimators.

Other enhancement techniques based on different AR-HMM variations are presented in (Lee, Lee, Song & Yoo 1996), (Lee, Rheem & Shirai 1996) and (Lee & Rheem 1997). These follow the same pattern of MMSE estimation using a Kalman filter and Equation 3.31 or appropriate variations. These HMMs are even more computationally intensive.

Experimental support for this work is limited and consists mostly of SNR value improvements for two enhanced sentences. In addition, it is stated that the enhanced speech has good intelligibility.

Cepstral HMM-Based Enhancement

Because many state of the art HMM recognition systems use cepstral features (e.g. see (Woodland, Gales, Pye & Young 1997)), researchers have investigated enhancement schemes which adapt or use cepstral-based models.

In (Beattie & Young 1992), the goal is to improve recognition performance of a cepstral-based HMM system in additive noise. Here it is noted that a Wiener filter applied in the spectral domain corresponds to an additive correction in the cepstral domain. Thus using trained spectral domain speech HMMs and given noise conditions, a Wiener filter can be designed for each clean speech state. This can be used to correct the mean of the corresponding state in the cepstral domain. The technique is shown to improve performance on small vocabulary tasks.

Seymour has developed an enhancement system along the lines of Ephraim which uses cepstral-based HMMs to determine $p(\bar{x}_t|\mathbf{Y})$ (Seymour 1996). Enhancement can then be performed using statistics trained in the linear spectral domain for each cepstral HMM state. Thus the enhanced speech is determined using

$$\hat{\mathbf{s}}_t = \sum_{\bar{x}_t} p(\bar{x}_t|\mathbf{Y}_{\text{cepstral}}) H_{\bar{x}_t,\text{linear}} \mathbf{y}_{t,\text{linear}} \tag{3.32}$$

where the subscripts 'cepstral' and 'linear' denote features in the cepstral and linear domains respectively and H is a Wiener filter or a MMSE amplitude estimator. This system has been evaluated using clean speech recognition scores of enhanced utterances and shown to perform well on both small and medium vocabulary tasks.

3.7 Summary

This chapter has described the main techniques of speech enhancement. Generally, it was seen that techniques which incorporated more prior information about the speech were superior. A deficiency of many techniques however was the need for a method to determine the statistics of the interfering noise. The next chapter reviews techniques of automatically or adaptively estimating these statistics.

Chapter 4

Techniques for Adaptive Environmental Compensation

Few of the enhancement schemes presented in the previous chapter are able to operate when the statistics of the degrading noise are unknown. This has proved problematic when employing these systems in real world situations. In this chapter, approaches to adaptive environmental compensation are reviewed.

There are two main classes of approaches. The first involves detecting speech-free portions of the signal and using these to estimate the noise. The second class builds a statistical model for the system and reestimates the noise statistics within this model using maximum likelihood estimation.

4.1 Noise Estimated from Non-speech Regions

The classic way to detect non-speech sections or 'pauses' is to combine the information provided by the the zero crossing count and energy measure for each frame (Deller et al. 1993). Typically the energy is higher during speech regions than non-speech regions, while the zero crossing count is high during fricatives (which may occur at the beginning of words) and low during voiced regions. The combination of this information can lead to a good estimate of the speech/non-speech divide.

Obviously however, this method relies heavily on thresholds which may need to change if the background noise is non-stationary. Also, the assumptions about energy and zero crossings may be invalid for large amounts or certain types of background noise.

A more recent technique (Abdallah, Montrēsor & Bauding 1997) uses the 'degree of organisation' of the signal. Although some setting of thresholds is still necessary, this technique can operate in situations where energy measures are not sufficient such as when the SNR is very low.

A related area in which there has been a great deal of research deals with the problem of detecting word boundaries for isolated word speech recognition systems (see review in (Junqua, Mak & Reaves 1994)). The problem here is slightly different to pause detection because the assumption is made that there is a single beginning and single endpoint of the speech. However, much of the work in this area is applicable and methods based on energy levels, zero crossing counts, duration information, pitch information, linear prediction residual energy, and Viterbi search have been developed. Some of these have been applied to enhancement systems.

For example, in (Sheikhzeheh, Brennan & Sameti 1995), a pause detector is developed for use in a real-time speech enhancement system. Non-speech activity is detected using the combination of information from a voiced/unvoiced detector based on the periodicity of the speech and energy levels before and after enhancement. Again however, thresholds must be determined empirically for the various anticipated noise conditions. An energy-based pause detector was used in an earlier system (Sheikhzadeh et al. 1994).

Recently, researchers have investigated the use of HMMs to detect non-speech activity in additive noise (McKinley & Whipple 1997). The advantage here is two-fold. First, a HMM incorporates far more information about speech than could be incorporated by energy levels, pitch information and zero crossing information alone and most importantly, the temporal evolution of this information is modelled by the Markov process. Of course this additional information adds complexity to the system and must be trained on speech prototypes. These may not be major considerations if dedicated hardware is used and suitable speech databases are available.

The second advantage is that a probabilistic framework can be used to decide whether a particular frame is speech or noise. In (McKinley & Whipple 1997), the framework of (Ephraim 1992*a*) is used. This models the speech and noise as autoregressive HMMs and develops a model for the noisy speech. The noisy speech model can then be used to determine the probability of a particular frame being speech-free. Results presented in (McKinley & Whipple 1997) show that this approach outperforms other techniques. This was also the case in earlier work on endpoint detection using HMMs (Wilpon & Rabiner 1987). Both these techniques train the noise model on the first few frames of the speech.

Parallel work developed in this dissertation and previously published as (Logan & Robinson 1996) also detects pauses using the AR-HMM framework. This work does not assume that the first frames of the utterance are speech. It will be discussed in the following chapter.

4.2 Maximum Likelihood Parameter Estimation

The approach here is to build a statistical model for the speech and noise and then to estimate the unknown parameters such that the likelihood of the model given the observations is maximised. The various methods are characterised by the type of models used and the amount of prior speech information which is incorporated into these models. Many of the techniques have been developed solely for robust speech recognition and these are described first. This is followed by descriptions of adaptive techniques for speech enhancement.

4.2.1 Robust Speech Recognition

A great deal of work in the robust speech recognition community has focused on schemes to adapt existing speech models to different environments, or alternatively to adapt speech parameters from new environments to existing models (Lee 1997). Here the term 'environments' is used very loosely to mean any difference between the conditions used to train and test speech recognition systems. This difference could be caused by a change in speaking style, accent, speaking rate, communications channel or background noise. While distortion caused by additive noise is the prime concern in this dissertation, it is still instructive to review approaches that compensate for other degradations. The focus will mostly be on 'blind' or 'automatic' model-based adaptation schemes.

A landmark paper is (Rose, Hofstetter & Reynolds 1994). Here, a general procedure is derived for estimating speech model parameters from noise corrupted observations. In this work, the noise statistics were known *a priori* but the technique illustrates a way of estimating model parameters in noise. A brief description is given here.

The speech is modelled as a mixture of Gaussians. Thus the probability density function is given by

$$p(\mathbf{S}|\lambda_s) = \prod_{t=1}^{T} \sum_{i=1}^{M} p_i b_i(\mathbf{s}_t). \tag{4.1}$$

Here $\mathbf{S}$ is the entire speech observation for time $t = 1, ..., T$. It is modelled by model λ_s with M Gaussian mixture components. p_i is the weight of the ith mixture and $b_i(\mathbf{s}_t)$ is a Gaussian pdf. It is assumed that the components of each observation vector are independent so that this Gaussian has a diagonal covariance matrix. Thus the model parameters are the means $\mu_{i,k}$ and variances $\sigma^2_{i,k}$ of each of the K components of the observation vector $\mathbf{s}_t$ and the mixture weights p_i.

The noise observation is also modelled by a mixture of Gaussians. Hence

$$p(\mathbf{D}|\lambda_d) = \prod_{t=1}^{T}\sum_{j=1}^{N} q_j a_j(\mathbf{d}_t). \tag{4.2}$$

Here $\mathbf{D}$ is the entire noise observation for time $t = 1, ..., T$ which is modelled using N Gaussian mixture components. q_j is the weight of the jth mixture and $a_j(\mathbf{d}_t)$ is a Gaussian pdf. Again it is assumed that this Gaussian has a diagonal covariance matrix.

Using these speech and noise models, a general expression for a noise corrupted model is then derived. The noise parameters are assumed known and maximum likelihood estimators are determined for the unknown speech means, variances and mixture weights. Since no closed-form solution exists, the EM algorithm is employed to iteratively reestimate the unknown parameters. Because the covariance matrices of the Gaussian pdfs are diagonal, a separate estimator can be derived for each component of the mean and covariance vectors.

The reestimation formulas are as follows. The kth component of the mean of the ith speech mixture component, $\mu_{i,k}$, is reestimated using

$$\mu_{i,k}{}' = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i_t\!=\!i, j_t\!=\!j|y_{t,k},\lambda)E\{s_{t,k}|y_{t,k}, i_t\!=\!i, j_t\!=\!j,\lambda\}}{\sum_{t=1}^{T} p(i_t\!=\!i, j_t\!=\!j|y_{t,k},\lambda)}. \tag{4.3}$$

Thus the speech mean for mixture component i is reestimated as the sum over all observations and all noise mixture components of the expected value of the speech given speech mixture component i and noise mixture component j, weighted by the likelihood of speech mixture component i and noise mixture component j given the noisy observation at time t. The sum is then normalised by the sum of the likelihood of each possible combination of speech and noise mixture components.

Similarly, the kth component of the variance of the ith speech mixture component, $\sigma_{i,k}$, is reestimated using

$$\sigma_{i,k}^{2}{}' = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N} p(i_t\!=\!i, j_t\!=\!j|y_{t,k},\lambda)E\{s_{t,k}^2|y_{t,k}, i_t\!=\!i, j_t\!=\!j,\lambda\}}{\sum_{t=1}^{T} p(i_t\!=\!i, j_t\!=\!j|y_{t,k},\lambda)} - (\mu'_{i,k})^2. \tag{4.4}$$

Rose et al. then derive different expressions for the expectations in Equations 4.3 and 4.4 according to the type of interfering noise. A closed-form solution is only easily obtained for the case when the speech and noise vectors are assumed additive and independent and, as mentioned, when the components of each observation vector are assumed statistically independent.

Sankar and Lee (Sankar & Lee 1995), (Sankar & Lee 1996) apply this technique to speech recognition in unknown noise when the speech model parameters are assumed known. They investigate the convolutional noise case and use cepstral feature vectors. For convolutional noise, the speech and noise vectors are additive in the cepstral domain and can be modelled using mixtures of Gaussians with diagonal covariance matrices. Therefore, the equations developed in (Rose et al. 1994) are directly applicable (albeit for noise model reestimation rather then speech model reestimation). They call the approach Stochastic Matching.

It is significantly harder to adapt to additive noise though if cepstral features are used. This is because if the speech and noise cepstral features are assumed normally distributed, then the feature vectors of the corrupted process are log-normally distributed (Gales 1995). Thus Equations 4.3 and 4.4 cannot be used. In fact no closed form expressions exist to reestimate the unknown noise parameters (Afify, Gong & Haton 1997). Thus the unknown parameters must be reestimated using numerical integration or from speech pauses as described in (Afify et al. 1997).

Despite this, some researchers have investigated schemes which assume that the corrupted cepstral feature vectors *are* normally distributed. In this work, a simplified function models the effect of additive and convolutional noise. In the spectral domain, the corrupted speech $Y(\omega)$ is formed by applying an unknown filter $H(\omega)$ to clean speech $S(\omega)$ and adding noise $D(\omega)$ such that

$$Y(\omega) = S(\omega)|H(\omega)|^2 + D(\omega). \tag{4.5}$$

This can be rewritten in the log-spectral or the cepstral domain as

$$\begin{aligned} \mathbf{y}_t &= \mathbf{s}_t + \mathbf{q}_t + \log(1 + e^{\mathbf{d}_t - \mathbf{s}_t - \mathbf{q}_t}) \qquad (4.6) \\ &= \mathbf{s}_t + f(\mathbf{s}_t, \mathbf{d}_t, \mathbf{q}_t) \qquad (4.7) \end{aligned}$$

where $\mathbf{y}_t$, $\mathbf{s}_t$, $\mathbf{q}_t$ and $\mathbf{d}_t$ are in either the log-spectral or cepstral domain respectively and $\mathbf{q}_t = 2\mathbf{h}_t$.

In Codeword-Dependent Cepstral Normalisation (CDCN) (Acero & Stern 1990), $p(\mathbf{y}_t|\mathbf{s}_t)$ is modelled as a multivariate Gaussian. A MMSE estimator of $\mathbf{y}_t$ is formed as the weighted sum of estimates where the weighting depends on classes of speech called 'codewords'. Maximum likelihood is used to estimate $\mathbf{q}_t$ and $\mathbf{d}_t$ within this framework. The enhanced speech vector is then tested using a clean speech recogniser.

One of the main problems with this technique is that it does not model the effects of the environment on the noise variance. This affects the accuracy of the estimation at low SNRs.

A later approach (Moreno, Raj & Stern 1995) approximates the function $f(\mathbf{s}_t, \mathbf{d}_t, \mathbf{q}_t)$ in Equation 4.7 with a vector Taylor series. Again maximum likelihood is used to reestimate the unknown parameters, including the noise variance. The algorithm is shown to perform better than CDCN on a large vocabulary test set over a wide range of SNRs. Further extensions use a vector polynomial function (Raj, Gouvêa, Moreno & Stern 1996) and sequential estimation (Kim 1998).

The problem of adapting to additive noise is easier in the linear spectral domain. Here the corrupted feature vectors can be assumed to be normally distributed and hence the Stochastic Matching technique applied (Lee 1997). Strictly speaking, full covariance matrices should be used to model the correlations between spectral components. This results in a considerable increase in computational complexity (Afify et al. 1997). A further point to note is that the distance measure used in this domain is the spectral difference which is inferior to the log spectral distance used in the cepstral domain (Rabiner & Juang 1993).

A theoretical framework which applies Stochastic Matching in both the cepstral and linear domains to adapt to both convolutional and additive noise is proposed in (Siohan & Lee 1997). The linear spectral domain is modelled by a Gaussian mixture similar to the cepstral domain and Monte Carlo simulations are proposed to map between the cepstral and linear domain. The computational complexity of this system limits its usefulness. No results are published in this paper.

4.2.2 Adaptive Enhancement

Lee et al. have developed several adaptive speech enhancement schemes based on making maximum likelihood estimates of unknown parameters. They operate on time domain observations and are therefore able to use Kalman filtering to perform the enhancement. They are also sequential techniques in that the updating of parameters is performed at each time step (sometimes after a delay for increased accuracy) rather than iterating over the whole signal.

In (Lee, Lee, Song & Yoo 1996), the speech is modelled using a Hidden Filter HMM (HF-HMM) as described in Section 3.6.5. The noise is assumed Gaussian. This assumption allows $p(\mathbf{z}_t|\mathbf{s}_t)$ to be modelled as a Gaussian also and a simpler than otherwise expression to be obtained for the likelihood of the model given the noisy speech. This can be maximised with respect to the unknown speech and noise model parameters and new estimates of these obtained. Experiments show SNR improvement on a small vocabulary task. Computational considerations hinder the extension of this system to a large vocabulary task.

In (Lee et al. 1997), the speech and noise are modelled as autoregressive processes. Maximum likelihood estimates are made of both the speech and noise model parameters. An advantage of this technique is that the unknown coloured noise case can be accommodated. However, the lack of prior speech information is a disadvantage.

4.3 Summary and Conclusions

This chapter has reviewed strategies to automatically obtain the statistics of the interfering noise. The first set of techniques estimates the noise from pauses and so relies on pause detection. Here it was noted that a model-based approach was superior.

The second set of strategies proposes models for the speech and noise and forms maximum likelihood estimates of unknown parameters. It was seen that for the additive noise case, three approaches are possible.

If cepstral feature vectors are used then the corrupted vectors are log-normally distributed. Thus there are no exact solutions to the maximum likelihood reestimation problem and the noise parameters must be reestimated using numerical integration or from speech pauses.

Alternatively, researchers have assumed that the corrupted cepstral features *are* normally distributed. Although this is in conflict with the models proposed in (Varga & Moore 1990) and (Gales 1995), good results have been obtained.

The third approach is to use linear spectral features which results in a mathematically simpler system since in this case, the corrupted feature vectors are normally distributed. The main disadvantage here is that a linear spectral distance measure must be used rather than the superior log-spectral distance measure used to compare cepstral features.

Enhancement schemes have traditionally studied the additive noise case. Therefore it comes as no surprise that the adaptive enhancement schemes studied in this chapter operate on time domain observations which pose no mathematical difficulties for additive noise. These schemes use autoregressive models of the speech and noise. As will be seen in Section 6.1, the use of these models implies that the Itakura-Saito distortion measure is used. This is related to the log spectral distance measure which is known to be superior to a linear spectral measure (e.g. see (Gray et al. 1980)).

However, the adaptive enhancement schemes studied here either train the speech models using the HF-HMM or estimate them directly from the noisy speech. As was seen in the previous chapter, it is highly desirable to use speech models trained using prior information. However, the HF-HMM is too computationally expensive to contemplate using it to build a vocabulary independent enhancement system.

Referring back to Section 3.6.5, it can be seen that two of the most successful enhancement schemes in terms of perceptual improvement are the model-based schemes of Ephraim (Ephraim 1992*a*) and Seymour (Seymour 1996). These techniques use the Itakura-Saito and cepstral distortion measures respectively and incorporate prior speech and noise information at reasonable computational cost. Neither of these schemes however is adaptive.

In (Seymour 1996), cepstral features are used to calculate the probability of each state given the noisy observation. Thus an adaptive form of this scheme would be mathematically difficult for the reasons described above.

The AR-HMM system proposed by Ephraim however is more suited to adaptation. Its strengths are

- linearly combinable feature vectors
- Itakura-Saito distortion measure
- reasonable computational requirements.

The remainder of this dissertation will thus be concerned with extending Ephraim's system such that it can adapt to unseen noise.

Chapter 5

Adaptive Speech Enhancement Schemes Based on Autoregressive Hidden Markov Models

Many of the enhancement systems described in Chapter 3 require estimates of the statistics of the corrupting noise. Although it may be possible in some circumstances to estimate these statistics from training data, in many real-world situations this is not feasible. Therefore in Chapter 4, techniques to estimate the noise from a corrupted signal were reviewed. It was seen that these techniques fall into two broad categories: estimating the noise from detected pauses and making a maximum likelihood estimate of the noise parameters given a statistical model.

In both Chapter 3 and Chapter 4, a recurrent theme is that superior performance can be achieved if a model-based approach is taken. In many cases, the preferred model was the HMM since this is a well established model for speech. The benefits of using a model come at the cost of increased computational complexity.

At the end of Chapter 4, it was concluded that it would be advantageous to develop an adaptive enhancement scheme based on the MMSE compensated autoregressive enhancement technique described by Ephraim (Ephraim 1992*a*) and in Section 3.6.5. This is a model-based enhancement scheme. As discussed in Chapter 4, it is a suitable starting point for an adaptive enhancement scheme because it uses linearly combinable feature vectors, a proven distortion measure (Itakura-Saito) and has reasonable computational requirements.

Interestingly, the technique uses autoregressive HMMs. These have been neglected in recent years in favour of other HMM configurations. A further motivation then for this approach was to investigate the performance of these models using computing resources which were unavailable when these models were first proposed.

Two main extensions to the work of Ephraim are developed in this chapter in order to make it adaptive. These correspond to the two main categories of adaptation discussed in Chapter 4. The first reestimates the noise statistics using portions of the corrupted signal which have been identified as non-speech or 'silence'. The autoregressive probability framework is used for the silence detection. The second scheme makes a maximum likelihood estimate of the noise parameters within this framework. In this chapter, the notation used in Section 3.6.5 will be followed.

5.1 Adaptive Speech Enhancement Using Recognised Silences

As described in Section 3.6.5, the technique presented in (Ephraim 1992*a*) models the clean speech $\mathbf{S}$ and noise $\mathbf{D}$ using autoregressive HMMs. These models are combined to give a model for noisy speech $\mathbf{Y}$. A MMSE and MAP estimator for the clean speech are then formed using this model.

Both the clean speech model and noise model are trained using *a priori* information. In (Ephraim 1992*a*), a general clean speech model is trained. This model has a multiple mixture state representing speech and a single mixture state representing silence. Transitions between the two states are freely allowed. The clean speech state is trained using a vector quantisation approach in which speech frames are clustered according to the similarity of their autoregressive parameters. The silence state is trained using speech frames with low energy.

An alternative approach to modelling speech is to train phone- or word-based HMMs as in a conventional speech recogniser. Here each mixture component of each HMM state still represents clusters of speech with similar autoregressive parameters, but the features used to train each HMM are (loosely) constrained to be those corresponding to each phone or word. A separate silence HMM is also trained.

Regardless of whether a silence state or a silence HMM is trained, this information can be used to make a decision about whether a given frame is speech or noise. Consider for example a word-based HMM system. By combining clean speech and noise HMMs as described in Section 3.6.5, a compensated model can be created to represent each word in the presence of noise. These models can then be used to recognise the noisy utterance using conventional HMM speech recognition techniques.

If perfect speech recognition in noise could be achieved, the frames labelled as silence would correspond to estimates of the noise. Even if recognition errors were made these frames could perhaps be used to give a better estimate of the noise. This could then be combined with the clean speech models to form a new compensated model, and the process repeated. These ideas can be used to develop a simple adaptive speech enhancement system. The work described here has been previously published (Logan & Robinson 1996).

Figure 5.1 shows the algorithm for the system. Frames labelled as silence by the recognition pass are used to reestimate the noise statistics. These statistics are then used to form a new compensated model, and the process repeated. When the likelihood of the utterance given the compensated model has converged, the speech is enhanced using Wiener filters or a spectral estimator as described in (Ephraim 1992*a*).

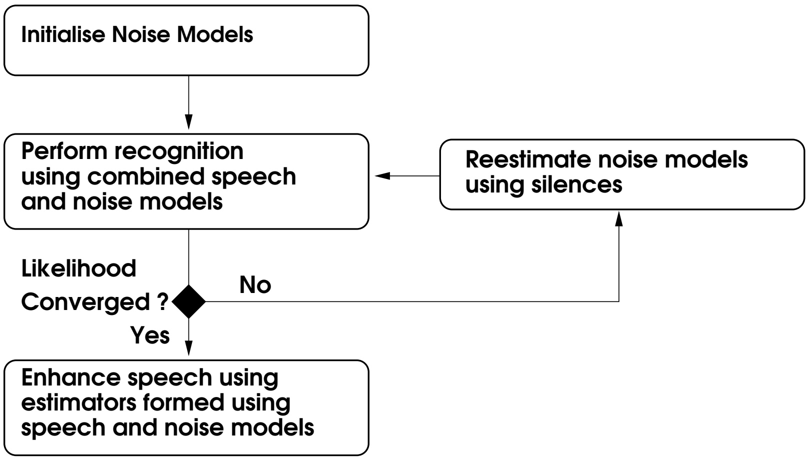

Figure 5.1: Basic Adaptive Enhancement Algorithm

The work in this dissertation will consider two of the estimators proposed in (Ephraim 1992*a*): the Wiener filter and a MMSE PSD estimator. As described previously, the Wiener filter for state $(x_t, \tilde{x}_t)$ is given by

$$H_{x_t,\tilde{x_t}} = \frac{f_{x_t}}{f_{x_t} + f_{\tilde{x}_t}}. \tag{5.1}$$

Here f_{x_t} and $f_{\tilde{x}_t}$, the power spectral densities of the speech and noise are estimated using the autoregressive parameters of state x_t and $\tilde{x}_t$ respectively similar to Equation 3.13.

The expected PSD of the clean speech at time t given the noisy observations is given by (Ephraim 1992*a*)

$$E\{|S_t|^2|\mathbf{y}_t, x_t, \tilde{x}_t, \lambda\} = H_{x_t,\tilde{x_t}} f_{x_t} + |H_{x_t,\tilde{x_t}} Y_t|^2. \tag{5.2}$$

Here Y_t is the Fourier transform of the noisy observation at time t.

The required noise statistic is thus the noise autocorrelation function since this is used to generate autoregressive parameters for the noise model.

Only stationary noise modelled by single state HMMs is considered so this statistic is obtained by simply averaging the frames labelled as silence.

As described in Section 3.6.5, the estimator used for each frame is the weighted sum of estimators for each possible combination of speech and noise states, weighted by the probability of that compensated state given the observation. If word or phone rather than general speech HMMs are used, then Viterbi alignment is used to determine the most probable compensated state for each frame. Thus in this case, the most likely estimator is used for enhancement.

5.1.1 Weighted Noise Reestimation

A simple extension to the above ideas yields a weighted noise reestimation scheme described here.

The pdf for the noisy speech is given in Equation 3.21 and reproduced below.

$$p(\mathbf{Y}|\lambda) = \sum_{\bar{\mathbf{X}}} a_{\bar{x}_0} a_{\bar{x}_1} \prod_{t=1}^{T} a_{\bar{x}_t \bar{x}_{t+1}} b_{\bar{x}_t}(\mathbf{y}_t) \tag{5.3}$$

As discussed in Chapter 2, standard techniques exist to calculate $P(\bar{x}_t|\mathbf{Y},\lambda)$, the probability of a particular state at time t given this model and the noisy observation. A probability of particular interest is the probability that observation $\mathbf{y}_t$ is speech-free or noise. This is given by the probability that the composite state $\bar{x} \equiv (x_t, \tilde{x})$ is $(x_t \equiv \mathsf{silence}, \tilde{x})$ i.e. $P((x_t \equiv \mathsf{silence}, \tilde{x}_t)|\mathbf{Y},\lambda)$. Once determined, either a likelihood ratio test or a Neyman-Pearson test can be used to decide whether a given frame is silence (McKinley & Whipple 1997).

The likelihood ratio test labels the frame at time t as silence if

$$P((x_t \equiv \mathsf{silence}, \tilde{x}_t)|\mathbf{Y},\lambda) \geq P((x_t, \tilde{x}_t)|\mathbf{Y},\lambda) \ \ \forall x_t. \tag{5.4}$$

The Neyman-Pearson labels a frame as silence if

$$P((x_t \equiv \mathsf{silence}, \tilde{x}_t)|\mathbf{Y},\lambda) \geq \eta \tag{5.5}$$

where η is empirically chosen to minimise the probability of false alarms.

$P(\bar{x}_t|\mathbf{Y},\lambda)$ can also be used to weight each frame thus giving a simple weighted noise reestimation scheme. Here the reestimated noise autocorrelation function $\mathbf{r}'_{\tilde{x}_t}$ at time t is given by

$$\mathbf{r}'_{\tilde{x}_t} = \frac{\sum_{t=1}^{T} P((x_t \equiv \mathsf{silence}, \tilde{x}_t)|\mathbf{Y},\lambda)\mathbf{r}_{\mathbf{Y_t}}}{\sum_{t=1}^{T} P((x_t \equiv \mathsf{silence}), \tilde{x}_t|\mathbf{Y},\lambda)} \tag{5.6}$$

where $\mathbf{r}_{\mathbf{Y_t}}$ is the autocorrelation coefficients of the noisy observation for frame t.

It should be noted that if phone or word HMMs are used, the calculation of $P(\bar{x}_t|\mathbf{Y}, \lambda)$ for all possible state sequences may be intractable. This is because for continuous speech recognition, the noisy state sequence $\bar{\mathbf{X}}$ represents all possible paths through all possible combinations of HMMs. For large vocabulary systems, this search space can be quite large and pruning of improbable sequences will be necessary.

5.2 Maximum Likelihood Noise Estimation

The algorithms described in the previous section do not guarantee convergence of the likelihood of the speech utterance given the models. They are also sensitive to recognition errors, particularly the recognition of silence. To this end a more formal noise estimation scheme has been developed. It has been previously published as (Logan & Robinson 1997*a*).

The scheme is related to several adaptation techniques described in Chapter 4 and in particular to the work of Rose et al. (Rose et al. 1994) and Sankar and Lee (Sankar & Lee 1996). The basic principle is to form a maximum likelihood estimate of the unknown parameters given a model. In the original paper by Rose et al., the speech and noise are modelled using mixtures of Gaussians with diagonal covariance matrices. The noise models are assumed known and a procedure is developed to estimate the speech model parameters. A closed-form solution is only easily obtained for the case when the speech and noise vectors are assumed additive and independent.

In later work (Sankar & Lee 1996), the technique is applied to the estimation of noise parameters rather than speech parameters. Here, the effect of convolutional noise on cepstral parameters is compensated for. In this domain, the speech and noise observations are additive so the procedure developed in (Rose et al. 1994) is directly applicable.

However, this dissertation is concerned with the problem of additive rather than convolutional noise. As discussed in the conclusion of Chapter 4, working with cepstral features is less appropriate for additive noise. This is because the non-linearity introduced by the logarithm when forming cepstral features makes reestimation of unknown parameters using maximum likelihood mathematically unattractive (Afify et al. 1997).

Ephraim has demonstrated that good enhancement of speech corrupted by additive noise can be achieved using AR-HMMs (Ephraim 1992*a*). These HMMs model feature vectors which are additive. However, the approach is unable to adapt to unknown noise.

In this section, it will be shown that the technique of Rose et al. can be combined with the work of Ephraim to develop an enhancement scheme which can adapt to unknown noise. This is possible because the required likelihood function is a linear function of the autocorrelation coefficients.

The procedure closely follows the work in (Rose et al. 1994) but applies it for the first time to AR-HMMs. A full description of the technique is given in Appendix A. A summary is given here. To simplify notation, a single mixture component per HMM state is assumed. The extension to multiple mixture systems is straightforward.

Following the method of (Rose et al. 1994), a model for a sequence of noisy observations $\mathbf{Y}$ is derived as

$$P(\mathbf{Y}|\lambda) = \sum_{\mathbf{X}} \sum_{\tilde{\mathbf{X}}} \int \int_{\mathbf{C}} P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}}|\lambda) d\mathbf{S} d\mathbf{D}. \tag{5.7}$$

Here $\mathbf{S}$ is a sequence of clean speech observations, $\mathbf{X}$ a sequence of clean speech states, $\mathbf{D}$ a sequence of noise observations and $\tilde{\mathbf{X}}$ a sequence of noise states. λ refers generally to the model parameters. The contour of integration $\mathbf{C}$ is taken over all possible combinations of speech and noise which can form the noisy observation. In this case, additive combinations are considered.

Given this model, the noise parameters are chosen to maximise the likelihood of the observed data. This is the method of maximum likelihood parameter estimation. The model parameters λ are: $\{a_{x_t x_{t+1}} \forall x_t x_{t+1}\}$, $\{a_{\tilde{x}_t \tilde{x}_{t+1}} \forall \tilde{x}_t \tilde{x}_{t+1}\}$, $\{\Sigma_x \forall x\}$ and $\{\Sigma_{\tilde{x}} \forall \tilde{x}\}$. (Here the notation of Section 3.6.5 has been used.) Thus it is required to find a new estimate of λ, λ', which maximises $P(\mathbf{Y}|\lambda)$.

Since no closed form solution exists for this maximisation problem, the method of Baum et al. (Baum et al. 1970) is used to find a solution. The technique iteratively maximises an auxiliary function $Q(\lambda, \lambda')$ with respect to λ'. For the case considered here, $Q(\cdot)$ is given by

$$Q(\lambda, \lambda') = E\{\log P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}}|\lambda')\}. \tag{5.8}$$

To reestimate the noise parameters, it is only necessary to maximise $Q(\cdot)$ with respect to $\{a_{\tilde{x}_\tau \tilde{x}_{\tau+1}} \forall \tilde{x}_\tau \tilde{x}_{\tau+1}\}$ and $\{\Sigma_{\tilde{x}} \forall \tilde{x}\}$.

Consider first the maximisation of $Q(\cdot)$ with respect to $\{a_{\tilde{x}_\tau \tilde{x}_{\tau+1}} \forall \tilde{x}_\tau \tilde{x}_{\tau+1}\}$. As described in Appendix A, the new estimate of of $a_{\tilde{x}_\tau \tilde{x}_{\tau+1}}$ is obtained using

$$a'_{\tilde{x}_\tau \tilde{x}_{\tau+1}} = \frac{\sum_{t=0}^{T} p(\tilde{x}_t = \tilde{x}_\tau, \tilde{x}_{t+1} = \tilde{x}_{\tau+1}, \mathbf{Y}|\lambda)}{\sum_{t=0}^{T} p(\tilde{x}_t = \tilde{x}_\tau, \mathbf{Y}|\lambda)}. \tag{5.9}$$

Thus the transition probability $a_{\tilde{x}_\tau \tilde{x}_{\tau+1}}$ is reestimated as the sum over all observations of the joint likelihood of state x_τ at time t and state $x_{\tau+1}$ at time $t+1$ and the observation sequence $\mathbf{Y}$, scaled by the sum over all observations of the joint likelihood of state x_τ at time t and the observation sequence $\mathbf{Y}$.

Now consider the reestimation of $\{\Sigma_{\tilde{x}} \forall \tilde{x}\}$. Because the noise is assumed to come from an autoregressive process, each $\Sigma_{\tilde{x}}$ can be calculated from its corresponding autocorrelation function. This is therefore the required statistic and is denoted by $\mathbf{r}_{\tilde{x}}$. This can be reestimated as described in Appendix A using the following equation for each noise state $\tilde{x}$

$$\mathbf{r}'_{\tilde{x}} = \frac{\sum_{t=1}^{T}\sum_{\forall x} p(x_t = x, \tilde{x}_t = \tilde{x}, \mathbf{Y}|\lambda) E\{\mathbf{r}_{\tilde{x}}|\mathbf{y}_t, x_t = x, \tilde{x}_t = \tilde{x}, \lambda\}}{\sum_{t=1}^{T}\sum_{\forall x} p(x_t = x, \tilde{x}_t = \tilde{x}, \mathbf{Y}|\lambda)}. \tag{5.10}$$

Thus the reestimated autocorrelation function is given by the sum over all observations and all clean speech states of the expected value of the autocorrelation function given the particular speech state and noise state weighted by the likelihood of being in that speech and noise state at time t and given the observation $\mathbf{Y}$.

Note that the forms of Equations 5.9 and 5.10 are similar to the usual parameter reestimation formulae for AR-HMMs (Juang 1984) and to the reestimation equations for Gaussian mixture models presented in (Rose et al. 1994).

Once each $\mathbf{r}_{\tilde{x}}$ has been reestimated, it is used to form a model of the noise spectrum which is required to form estimators of the enhanced speech and for the construction of the noisy speech model.

For stationary noise, only maximisation with respect to $\Sigma_{\tilde{x}}$ is required. In this case, $p(x_t = x, \tilde{x}_t = \tilde{x}, \mathbf{Y}|\lambda)$ can be calculated using the usual forward-backward equations (e.g. see (Rabiner & Juang 1993)).

$p(x_t = x, \tilde{x}_t = \tilde{x}, \mathbf{Y}|\lambda)$ can also be approximated by noting that one state sequence dominates $P(\mathbf{Y}, \mathbf{X}, \tilde{\mathbf{X}}|\lambda)$ (Merhav & Ephraim 1991). Thus $p(x_t = x, \tilde{x}_t = \tilde{x}, \mathbf{Y}|\lambda)$ can be replaced by by either one or zero depending on whether or not x_t is part of the dominant state sequence. Therefore Equation 5.10 can be rewritten

$$\mathbf{r}'_{\tilde{x}} = \frac{\sum_{t=1}^{T} E\{\mathbf{r}_{\tilde{x}}|\mathbf{y}_t, x_t = x_t^*, \tilde{x}_t = \tilde{x}, \lambda\}}{T}. \tag{5.11}$$

Here, $x^* = \{x_t^*, t = 1, \ldots, T\}$ is the most likely clean speech state sequence. This can be found by performing Viterbi alignment using the compensated model on the noisy observations.

The expected value of the autocorrelation function given the composite state $(x_t, \tilde{x}_t)$ and $\mathbf{y}_t$ is most easily obtained from the expected value of the noise PSD function $|D|^2$. $|D|^2$ forms a Fourier transform pair with the autocorrelation function which is convenient since it is simpler to work in the frequency domain. Specifically,

$$\begin{aligned} &E\{r_{\tilde{x}_\tau}(i)|\mathbf{y}_\tau, x_\tau = x_t, \tilde{x}_\tau = \tilde{x}_t, \lambda\} \\ &= E\left\{ \frac{1}{K}\sum_{k=0}^{K-1} |D_k|^2 \exp\left(\frac{j2\pi ik}{K}\right) \middle| \mathbf{y}_\tau, x_\tau = x_t, \tilde{x}_\tau = \tilde{x}_t, \lambda \right\} \end{aligned} \tag{5.12}$$

$$= \frac{1}{K}\sum_{k=0}^{K-1} E\left\{ |D_k|^2 \middle| \mathbf{y}_\tau, x_\tau = x_t, \tilde{x}_\tau = \tilde{x}_t, \lambda \right\} \exp\left(\frac{j2\pi ik}{K}\right). \quad (5.13)$$

Thus the term $E\{\mathbf{r}_{\tilde{x}}|z_t, x_t = x_t^*, \tilde{x}_t = \tilde{x}, \lambda\}$ in Equation 5.11 can be evaluated as the inverse Fourier transform of $E\left\{ |D_k|^2 \middle| \mathbf{y}_\tau, x_\tau = x_t^*, \tilde{x}_\tau = \tilde{x}_t, \lambda \right\}$.

This can be estimated as shown in (Ephraim 1992*a*). Here, each component k of $|D|^2$ is evaluated using Wiener filters designed to return the noise PSD given the composite state information. Thus

$$E\left\{ |D_k|^2 \middle| \mathbf{y}_\tau, x_\tau = x_t^*, \tilde{x}_\tau = \tilde{x}_t, \lambda \right\} = H_{x_t^*,\tilde{x}_t,k} f_{x_t^*,k} + |H_{x_t^*,\tilde{x}_t,k} Y_{t,k}|^2. \quad (5.14)$$

Here $H_{x_t,\tilde{x}_t,k}$ is the kth component of the Wiener filter for the composite state $(x_t, \tilde{x}_t)$, $f_{x_t,k}$ is the kth component of the Fourier transform of the autoregressive coefficients for clean speech state x_t and $Y_{t,k}$ is the kth component of the Fourier transform of the noisy observation at time t. Since the Wiener filter is designed to return the MMSE estimator of the noise, its transfer function is given by

$$H_{x_t^*,\tilde{x}_t,k} = \frac{f_{\tilde{x}_t,k}}{f_{x_t^*,k} + f_{\tilde{x}_t,k}}. \quad (5.15)$$

Equation 5.14 is derived assuming that the covariance matrices $\mathbf{\Sigma}_x$ and $\mathbf{\Sigma}_{\tilde{x}}$ of the speech and noise processes (see Equation 3.25) are circulant (Ephraim 1992*b*). This implies that the covariance matrices can be replaced by a matrix of the form

$$\Sigma \approx C_K(f) = K^{-1} U D(f) U^T. \quad (5.16)$$

Here K is the frame size, U is a KxK matrix whose (k, n)th element is the complex exponential $\exp(-j2\pi kn/K)$ and $D(f)$ is a KxK diagonal matrix whose Kth diagonal element is given by the PSD of the autoregressive process associated with that covariance matrix. $C_K(f)$ assymtotically approaches Σ as K approaches infinity. Thus the assumption holds assuming sufficiently large K.

The new adaptive enhancement algorithm thus operates as shown in Figure 5.2. Comparison of Figure 5.1 with Figure 5.2 shows that the only difference between the systems is the technique of noise reestimation.

5.3 Summary

In this chapter, techniques to automatically estimate the noise from the speech to be enhanced have been described.

The first technique estimates the noise statistics from detected pauses. The AR-HMM framework is used for the pause detection. A variation of this technique which weights the noise estimation according to how likely it is that a particular frame is noise was also developed. The second approach is to use maximum likelihood reestimation to estimate the noise parameters.

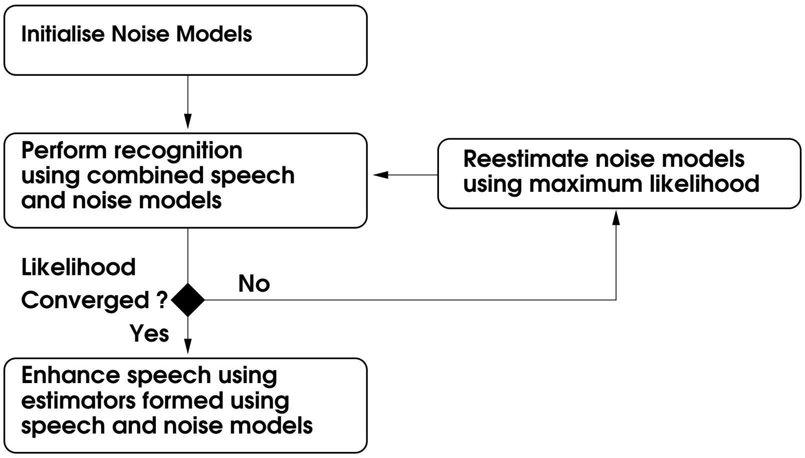

Figure 5.2: Improved Adaptive Enhancement Algorithm

Chapter 6

Autoregressive Hidden Markov Models Using a Perceptual Frequency Scale

The adaptive enhancement algorithms developed in the proceeding chapter use autoregressive hidden Markov models. These are commonly used in the speech enhancement community because they are mathematically well suited to additive noise. This is in contrast to MFCC-based models which have inherent non-linearities which make adaptation to additive noise difficult (Afify et al. 1997).

However, MFCC-based models form the basis of many large vocabulary speech recognition systems (e.g. (Woodland et al. 1997)). Clearly then the modelling power of cepstral based features is greater than an autoregressive system and it would be highly desirable to incorporate any advantages of the first system into the second. This provides the motivation for this section of work.

In this chapter, the distortion measures used in the two systems are examined. A major difference between them, the use of a perceptual frequency scale in the MFCC system, is incorporated into the AR-HMM system. This is found to improve recognition performance substantially.

6.1 Autoregressive HMM and MFCC HMM Distortion Measures

As noted in Chapter 2, one difference between AR-HMMs and 'standard' HMMs is the form of the state dependent pdf ($b_{x_t}(\mathbf{s}_t)$ in Equation 2.1). In both cases this pdf is a (possibly multiple mixture) Gaussian. However, in the AR-HMM case, the covariance matrix of each Gaussian takes a special form. The Gaussian pdf is then is a function of the $P+1$ autocorrelation coefficients of the waveform where P is the order of the autoregressive process (Juang 1984).

The equation for this pdf is reproduced here. A single mixture component per state has been assumed for simplicity.

$$b_{x_t}(\mathbf{s}_t) \approx (2\pi)^{-K/2}(\sigma^2_{x_t})^{-K/2} exp\{-\frac{1}{2}\alpha(\sigma^{-1}\mathbf{s}_t; \mathbf{A}_{x_t})\} \tag{6.1}$$

Here

$$\alpha(\sigma^{-1}_{x_t}\mathbf{s}_t; \mathbf{A}_{x_t}) = r_{A_{x_t}}(0) r_{\mathbf{s_t}}(0) + 2\sum_{n=1}^{P} r_{A_{x_t}}(i) r_{\mathbf{s_t}}(i). \tag{6.2}$$

Refer to Section 2.4 or (Juang 1984) for more detail.

The difference between the log likelihood for a given observation $\mathbf{s}_t$ using its optimal autoregressive parameters and an arbitrary set is given by (Juang 1984)

$$\begin{aligned} L &= \log b_{x_t}(\mathbf{s})_{\max} - \log b_{x_t}(\mathbf{s}) \\ &= \frac{K}{2}\frac{1}{K}\left[\alpha(\sigma^{-1}\mathbf{s}, \mathbf{A}) - \log\sigma_o^2 + \log\sigma^2 - 1\right]. \end{aligned} \tag{6.3}$$

Here σ_o^2 is the residual energy obtained when the optimal autoregressive parameters corresponding to $\mathbf{s}_t$ are used. As noted in (Juang 1984), the bracketed term in Equation 6.3 is equivalent to the Itakura-Saito (I-S) distortion measure. Thus this measure is used to compare candidate models to each observation.

The I-S distortion measure between speech spectrum $S(\omega)$ and autoregressive spectrum $\frac{\sigma^2}{|A(e^{j\omega})|^2}$ can also be written (Rabiner & Juang 1993)

$$\begin{aligned} d_{I-S}\left(S(\omega), \frac{\sigma^2}{|A(e^{j\omega})|^2}\right) &= \frac{1}{\sigma^2}\int_{-\pi}^{\pi} S(\omega)|A(e^{j\omega})|^2 \frac{d\omega}{2\pi} - \log\sigma_o^2 + \log\sigma^2 - 1 \\ &= \int_{-\pi}^{\pi} \frac{S(\omega)}{\frac{\sigma^2}{|A(e^{j\omega})|^2}} \frac{d\omega}{2\pi} - \log\sigma_o^2 + \log\sigma^2 - 1. \end{aligned} \tag{6.4}$$

Thus it can be seen that a major component of this distance measure is the average of the ratio of the power spectral densities.

The cepstral distortion measure used in cepstral-based HMMs is also related to a ratio of spectrums. Specifically, using the definition of the cepstrum (Rabiner & Juang 1993)

$$\log S(\omega) = \sum_{n=-\infty}^{\infty} c_n e^{-jn\omega} \tag{6.5}$$

where c_n is the nth cepstral coefficient, the cepstral difference can be shown to be

$$\sum_{n=-\infty}^{\infty} (c_n - c_n')^2 = \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 \frac{d\omega}{2\pi} \tag{6.6}$$

$$= \int_{-\pi}^{\pi} \left| \log \frac{S(\omega)}{S'(\omega)} \right|^2 \frac{d\omega}{2\pi}. \tag{6.7}$$

Thus the distortion measures used in AR-HMMs and cepstral-based HMMs are related. However the distortion measure used in a typical cepstral-based system differs from Equation 6.7 in two important ways. First, if the cepstral features are modelled by Gaussians, a weighted distortion measure is used. This aspect will be discussed in more detail in Section 8.2.2.

Second, the cepstral coefficients are generally derived using a warped frequency spectrum yet AR-HMMs as described in (Juang 1984) use a linear frequency spectrum. It is well known (Deller et al. 1993) that it is more appropriate to use a warped frequency scale such as the Mel or Bark scale since these correspond more closely to the frequency resolution of the human ear. Therefore, in this chapter AR-HMMs are extended such that they use a non-linear frequency scale. This work has been previously published as (Logan & Robinson 1997*b*).

6.2 Incorporation of Perceptual Frequency

In this section, frequency warping is incorporated into an AR-HMM recognition system. The approach taken here is to use the bilinear transform (Oppenheim & Johnson 1972) to perform the warping. The inspiration came from its use to improve the performance of linear prediction coding systems (Strube 1980) and LPC-cepstral recognition systems (Shikano 1985).

The bilinear transform converts a time sequence to a new sequence with a warped spectrum. By adjusting the so-called warping factor, the degree of warping can be made to be a very good approximation to the Bark scale. This is known to be related to perceptual frequency in a similar manner to the Mel-scale (Deller et al. 1993). Further description of the bilinear transform is given in Appendix B.

There are two main ways in which the bilinear transform can be implemented. Using the recursion presented in (Oppenheim & Johnson 1972), samples in the time domain, autocorrelation domain or autoregressive parameter domain can be transformed. In these cases, a new infinite sequence results. The Fourier transform of this sequence yields a spectrum or PSD on a warped frequency scale.

The second technique is described in (Strube 1980). Here the warped autocorrelation coefficients are given directly by

$$\bar{R}_n = \sum_{k=0}^{K/2} \cos(n\bar{\Omega}_k) \left(\frac{d\bar{\Omega}}{d\Omega} \right)_k P_k. \tag{6.8}$$

Here $\bar{R}_n$ is the nth warped autocorrelation coefficient, $\bar{\Omega}_k$ is the warped frequency corresponding to Ω_k and P_k is the non-warped PSD.

There is a significant difference in computational complexities between the two schemes. The first approach requires $O(N^2)$ calculations per frame where N is the number of samples in the sequence to be transformed. N may be the frame size, the number of autocorrelation coefficients or the order of the autoregressive process according to which sequence is transformed. The second technique requires calculations $O(PK)$ per frame where P is the number of autoregressive coefficients required and K is the frame size.

Figure 6.1 shows the way in which the bilinear transform is integrated into a clean AR-HMM recognition system. It is seen that this can be easily achieved by warping the autocorrelation coefficients of the testing and training data.

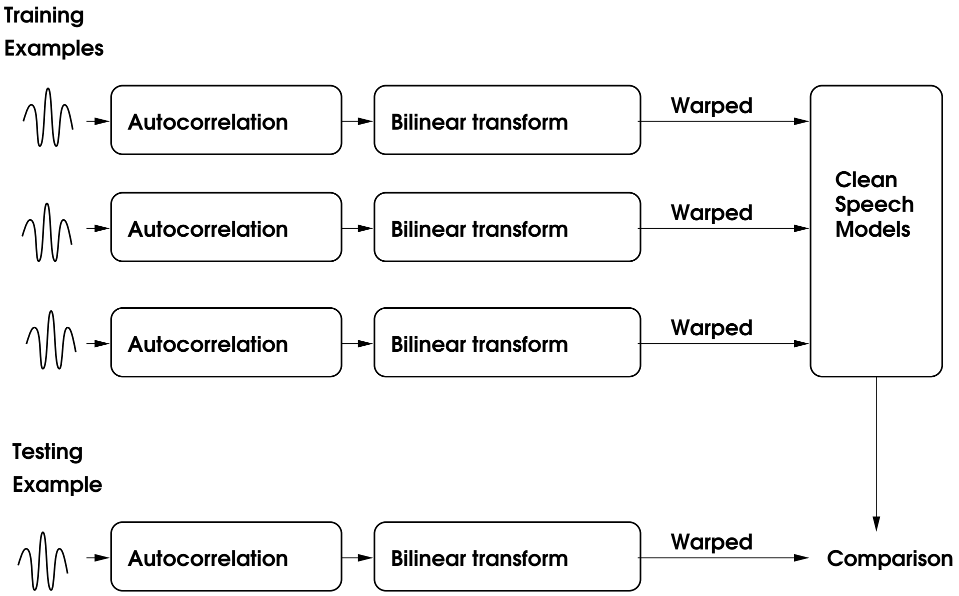

Figure 6.1: A perceptual frequency AR-HMM recognition system. Here the autocorrelation functions of both testing and training examples are warped using the bilinear transform such that all comparisons of testing and training data are performed on a warped frequency scale.

6.2.1 Determination of the Warping Factor

The bilinear transform has one parameter: the warping factor. This determines how closely the transform approximates the Bark scale and must be chosen according to the sampling rate. Figure 6.2 shows the approximation for various warping factors at a 16kHz sampling rate. It can be seen that the approximation is reasonable for warping factors in the range 0.5-0.6.

Smith and Abel (Smith & Abel 1995) have developed a formula to determine the optimal warping factor for a given sampling frequency. Their technique is based on minimising the error between the Bark scale and the bilinear transform approximation. This formula is reproduced here.

$$\alpha \approx 1.0211 \left[\frac{2}{\pi} \tan^{-1}(0.076 f_s) \right]^{\frac{1}{2}} - 0.19877 \qquad (6.9)$$

Here α is the warping factor and f_s is the sampling frequency measured in kHz. For the sampling frequency of 16kHz, Equation 6.9 gives a warping factor of 0.57.

A final way to determine the warping factor is to build a warped AR-HMM recognition system and investigate its performance for different warping factors. This is particularly interesting because it shows how sensitive the system is to different warping factors and how critical it is that the Bark scale is used. The experiments are described below.

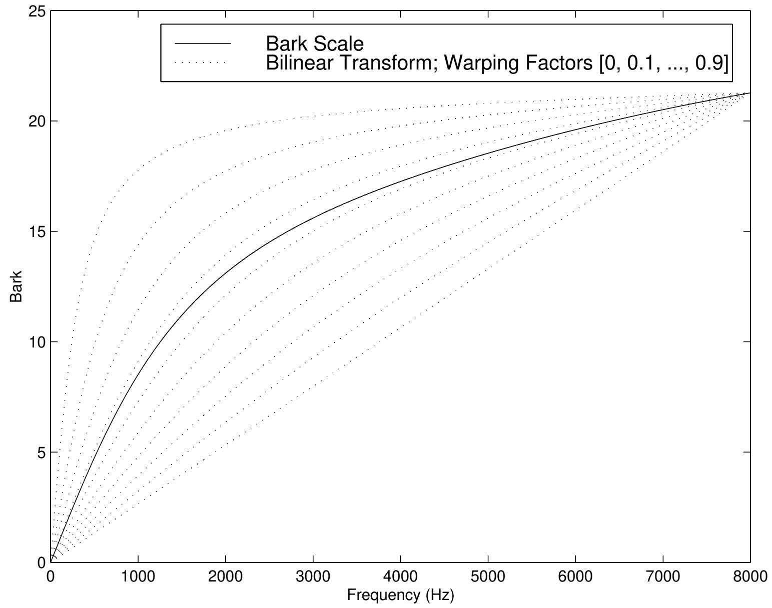

Figure 6.2: Bilinear transform approximation to the Bark scale for various warping factors at 16kHz sampling rate. A warping factor of zero implies no warping and is represented by the rightmost dotted line.

Initial Recognition Experiments

Experiments were conducted using the ISOLET database collected for use in (Fanty & Cole 1990). This database contains two isolated tokens of each letter of the alphabet for 150 American English speakers, 75 male and 75 female. 120 speakers were used for training and 30 for testing. The speech is sampled at 16kHz.

Experiments were performed using the English E-set letters {"B", "C", "D", "E", "G", "P", "T" and "V"} only. This is considered to be a difficult task due to the high level of confusibility between these letters. The small number of classes however, means that recognition experiments can be performed in a reasonable time.

The implementation of the clean AR-HMM system was as follows. The speech was divided into frames of 32ms with a 16ms overlap. One 13 emitting state HMM was trained for each letter using AR models of order 20. Warping was performed in the frequency domain according to Equation 6.8. The clean autocorrelation coefficients were normalised by their autoregressive variance so that energy information was removed.

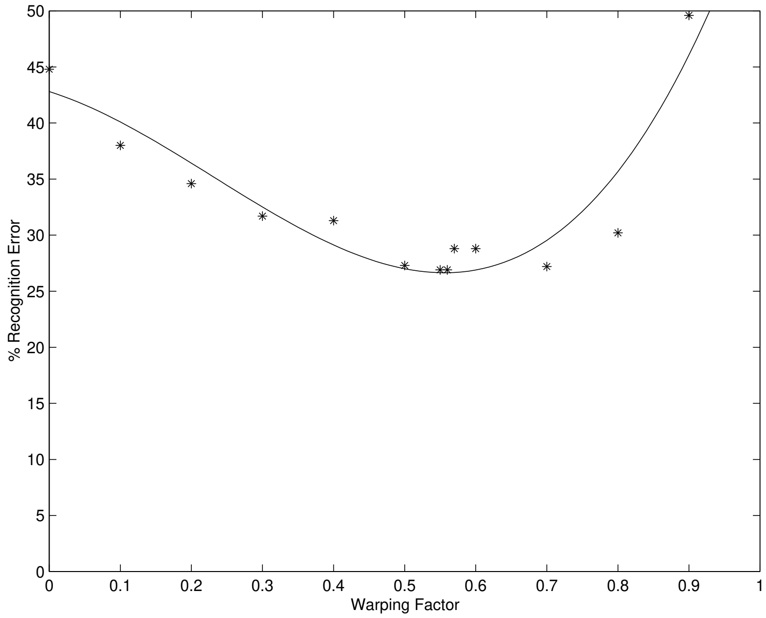

Figure 6.3: Clean speech recognition error rates for various warping factors. Experiments performed on ISOLET data using 3 mixture AR-HMM systems. A 4th-order polynomial is fitted to the data.

Recognition results for various warping factors using a 3 mixture system are plotted in Figure 6.3 with a 4th-order polynomial fitted. The results indicate that the choice of warping factor is not critical and indeed any value in the range 0.5-0.6 will suffice. This is in agreement with the plot of warping factors versus the Bark scale. In light of these observations and given that the warping factor returned by Equation 6.9 is within this range, 0.57 is the chosen warping factor and is used for all subsequent experiments unless stated. Figure 6.4 shows the approximation to the Bark scale for this warping factor.

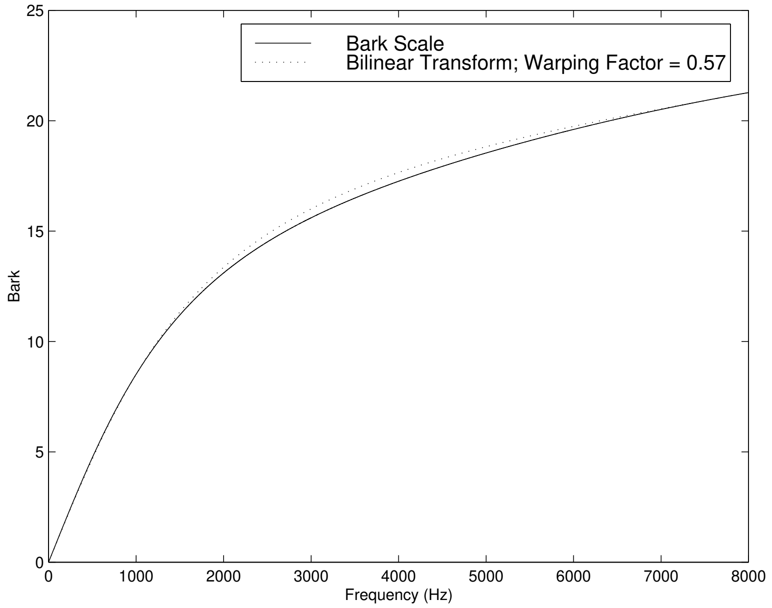

Figure 6.4: Bilinear transform approximation to the Bark scale for the chosen warping factor of 0.57.

6.2.2 Comparison with MFCC Clean Recognition

The recognition results for various numbers of mixture components using the chosen warping factor of 0.57 are shown in Table 6.1. Also in this table are results for an AR-HMM system without warping and a MFCC-based system.

The MFCC-based system had a comparable number of parameters to the AR-HMM system. Specifically, one 13 emitting state HMM was trained for each letter with 12 MFCC coefficients and one energy coefficient per state. Recognition was performed with and without the 13 delta coefficients and using various numbers of mixture components. The Gaussian mixture components had diagonal covariance matrices.

The '% Error' figure in Table 6.1 was calculated using

$$\%\ \mathrm{Error} = \frac{D+S+I}{N} \cdot 100\%. \tag{6.10}$$

Here D, I and S represent the number of deletions, insertions and substitutions respectively and N is the number of letters in the test set.

Model	Number Mixture Components	% Error (D,S,I)
AR no warping	1	40.4 (0,192,2)
	2	38.5 (0,192,0)
	3	41.0 (0,137,0)
AR with warping	1	26.7 (0,127,1)
	2	28.8 (0,138,1)
	3	28.8 (0,138,1)
MFCC no deltas	1	24.4 (0,117,0)
	2	25.4 (0,122,0)
	3	25.4 (0,122,0)
MFCC with deltas	1	17.1 (0,82,0)
	2	12.9 (0,62,1)
	3	11.0 (0,53,0)

Table 6.1: Recognition results for clean speech tests using the E-Set from the ISOLET database. AR-HMMs with and without frequency warping are compared to MFCC systems with and without delta parameters.

It can be seen that warping the frequency scale decreases the recognition error of the AR-HMM system quite considerably. In fact the results now approach those of the MFCC-based system without delta parameters.

One obvious difference between the MFCC-no-delta and AR-HMM system is the lack of energy information in the AR-HMM system. To investigate the effect of this, energy information was incorporated into the AR-HMM framework similar to (Juang & Rabiner 1985). Here, the likelihood expression for an observation given the current state was modified by the addition of a term describing the probability of the state having a given energy. This extra term was scaled by an empirically chosen factor.

Table 6.2 shows the error rates obtained as a function of this scaling factor. It is seeing that by an appropriate choice of scaling factor, the error rate decreases by about 1-2% absolute. It would appear then that in this domain, the information in the delta parameters is now the main information missing from this modified AR-HMM system.

6.2.3 A Perceptual Frequency AR-HMM Enhancement System

The work in the previous section has demonstrated that the incorporation of perceptual frequency can improve the modelling power of AR-HMM clean recognition systems. These benefits can also be incorporated into AR-HMM

Model	Number Mixture Components	Scale Factor	% Error (D,S,I)
AR with warping	1	0	26.7 (0,127,1)
	1	10	26.5 (0,126,1)
	1	20	24.8 (0,118,1)
	1	30	25.0 (0,119,1)

Table 6.2: Recognition results for clean speech tests using the E-Set from the ISOLET database. AR-HMMs with frequency warping are augmented by the addition of energy information scaled by an empirically determined factor.

enhancement systems as follows.

The enhancement systems described in Chapter 5 calculate which filter to use to enhance each frame of corrupted speech based on the probability of each mixture component given the noisy observation. This probability is calculated using a compensated AR-HMM. Although this AR-HMM traditionally uses a linear frequency scale, it is fairly straightforward to build a compensated AR-HMM using speech and noise models trained on observations with a warped frequency scale. Hence a perceptual frequency compensated AR-HMM can be created. This compensated AR-HMM can then used in conjunction with warped observation vectors to determine the weights required for the filtering stage of the enhancement system.

This system is shown in Figure 6.5. Since it is computationally expensive to warp and unwarp time domain observations, the filters are applied to non-warped observations. Therefore, it is necessary to obtain unwarped versions of each warped filter and hence of each speech and noise model for use with the non-warped (i.e. unprocessed) observations.

However, because the warping process transforms a finite sequence to an infinite sequence (Oppenheim & Johnson 1972), it is never possible to exactly unwarp a given speech or noise model. This problem can be addressed using single pass retraining to train non-warped models.

Given a set of models, single pass retraining (e.g. see (Young, Jansen, Ollason & Woodland 1996)) generates a parallel set of models using different training data. This is achieved by computing the state probabilities using the original models and the original training data, but then switching to a new set of training data to compute parameter estimates for the new model. Thus given parallel warped and non-warped observations and warped models, non-warped models which correspond exactly to the warped models can be trained.

Therefore, in order to build a non-adaptive perceptual frequency enhancement system, the first step is to train warped speech and noise models

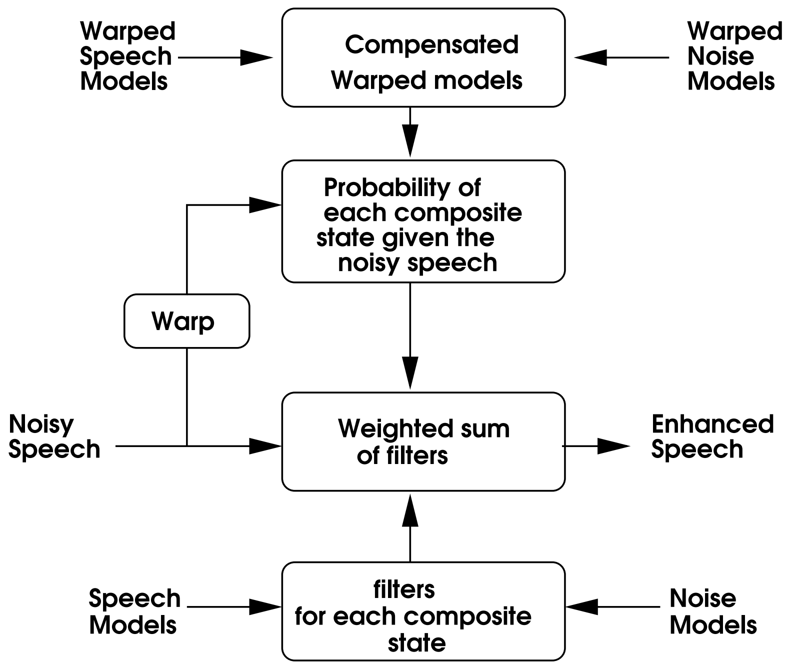

Figure 6.5: A perceptual frequency enhancement system. Here the weights for each filter are determined using perceptual frequency AR-HMMs. The filters themselves are constructed using non-warped AR-HMMs.

using warped observations. These are then used in conjunction with parallel warped and non-warped observations to train non-warped models which correspond exactly to the warped models.

The enhancement process requires warped and non-warped observations. The warped observations are used in conjunction with warped models to calculate the probability of each composite state. The statistics of each state in the corresponding non-warped model are then used to construct filters which operate on non-warped observations.

In order to implement the adaptive enhancement schemes described in the previous chapter, the noise models must now be reestimated in parallel in both the warped and non-warped domains. This can also be achieved using the compensated warped models in conjunction with warped observations for all probability calculations, and warped and non-warped data in parallel with this to reestimate the noise statistics in the warped and non-warped domain respectively.

A final point to note is that it is possible to obtain a version of the unwarped spectrum by sampling the warped spectrum. Consider the warped spectrum corresponding to a set of autocorrelation coefficients warped using Equation 6.8. This can be written

$$S(\bar{\omega}) = \frac{\bar{\sigma}^2}{|B(\bar{\omega})|^2}. \tag{6.11}$$

Here $\bar{\omega}$ is the warped frequency corresponding to ω. As described in Appendix B, these are related by

$$\bar{\omega} = \arctan\left[\frac{(1-a^2)\sin\omega}{(1+a^2)\cos\omega - 2a}\right]. \tag{6.12}$$

$B(\bar{\omega}) = \sum_{n=0}^{P} b_n e^{-jn\bar{\omega}}$ is the transfer function formed from the autoregressive filter corresponding to the warped autocorrelation coefficients. $\bar{\sigma}^2$ is the variance parameter for this filter. For sufficiently large filter order P, $\sigma^2 = \bar{\sigma}^2$ (Strube 1980). Therefore, given that the relationship between ω and $\bar{\omega}$ is known, $S(\bar{\omega})$ can be sampled to give $S(\omega)$.

This unwarped spectrum will not correspond exactly to the original non-warped spectrum. This original spectrum could only be obtained by sampling if $B(\bar{\omega})$ and $\bar{\sigma}^2$ were obtained by transforming the original filter parameters. This would imply that $B(\bar{\omega})$ had infinite order (Strube 1980).

However, a spectrum unwarped in this fashion will model the lower formants with more precision at the expense of features at the higher frequencies. Since low frequency features are perceptually more important, there is reason to expect that using this technique to unwarp spectrums to create filters for enhancement would lead to improved enhancement.

Preliminary experimentation with an enhancement system in which warped models are unwarped by resampling the warped spectrum indicated that there was no perceptual improvement over a similar system using non-warped models trained using single pass retraining. These experiments were performed at the given model order of 20. This agrees with the results in (Strube 1980) where the incorporation of perceptual frequency into a coding system did not improve intelligibility scores if the filter order was sufficient for modelling.

The results in (Strube 1980) do indicate however that if a lower filter order is used, then there could be perceptual advantage in using the re-sampled warped spectrum rather than the more exact single pass retrained spectrum. This could be exploited to construct a system with less parameters. Investigation of this research path is left to future work.

6.3 Summary

In this chapter, perceptual frequency was incorporated into AR-HMM systems using the bilinear transform. Clean speech recognition experiments showed that this improved recognition performance considerably. The construction of an enhancement system incorporating perceptual frequency was also described.

Chapter 7

Evaluation on the NOISEX-92 Database

In order to evaluate the effectiveness of the developed enhancement algorithms, initial experiments were conducted using the NOISEX-92 database (Varga et al. 1992). This database is suitable for small vocabulary experiments and was therefore used to verify the working of the algorithms and to investigate variations.

7.1 The NOISEX-92 Database

The NOISEX-92 is a publicly available database often used for benchmark experiments. It contains isolated digits and digit triplets spoken by both male and female speakers. The speech is corrupted by various noise sources at SNRs ranging from -6 to 18dB. The clean speech signal is also available. The database was generated by adding the noise to the clean speech hence speech artifacts due to speaker stress in noise are not present. Speech corrupted by convolutional noise is also available.

The experiments in this chapter used only the isolated digit task with predominantly the male speaker. The female speaker was used for some tests. Only additive noise was considered.

7.2 Noise Sources

The following four stationary noise sources are used: Lynx helicopter noise, speech noise, car noise and F16 aircraft noise. These noises can be modelled using single state HMMs. Appendix C shows a typical spectrogram and spectrum for each of these noises.

7.3 Evaluation Methods

Several methods were used to evaluate the effectiveness of the enhancement algorithms. An informal qualitative assessment was made of enhanced waveforms and a Compact Disc is supplied as Appendix D in order that the main characteristics of each algorithm may be heard. Quantitative assessment was performed by recognition tests and distortion measures. These are described more fully below.

7.3.1 Distortion Measures

The Itakura distortion measure (Gray et al. 1980) was used to provide a quantitative measure of the quality of the enhanced speech. This measure is related to human auditory perception since it is heavily influenced by mismatch in formant locations. The distortion calculated was for both the speech portions of the signal and the entire waveform.

7.3.2 Recognition Tests

Because the optimal distortion measure is unknown, the enhancement algorithms were also evaluated by investigating the clean speech recognition performance. Specifically, the enhanced speech was converted to MFCC parameters and recognised using a clean speech MFCC system. Here the features were derived directly from the enhanced spectra without resynthesising the speech in the time domain. A MFCC recognition system was used since it gave the best performance overall. These recognition experiments gave an indication of the value of each enhancement scheme as a front-end to a recognition system. They also provided quantitative evidence of performance.

7.4 Recognition Systems

A clean speech MFCC HMM system is required for analysis. In addition, warped and non-warped clean speech AR-HMM recognition systems are needed for the word-based enhancement system and for baseline experiments. These models were constructed using the HMM Toolkit V1.5 (Young et al. 1993). It was necessary to make substantial additions to the toolkit in order to accommodate AR-HMMs since these are not implemented. The parameter reestimation formulae for these HMMs are provided in standard references (e.g. see (Rabiner & Juang 1993)).

Both the MFCC and AR systems worked with frames of 32ms with overlap of 16ms. The frame size and overlap were chosen to be convenient for enhancement. For the given sampling rate of 16kHz, 32ms corresponds to a frame size of 512 which is a power of 2. Therefore a fast Fourier transform algorithm can be used without zero-padding. The overlap of half the frame size is useful for reconstructing the enhanced speech in the time domain. Since Hamming windows are used to create each frame, the enhanced frames can simply be overlapped and added together.

The MFCC clean recognition system consisted of an 8-emitting state left-to-right HMM model for each digit and a 1-emitting state model for silence. The MFCC feature vectors contained 15 cepstral coefficients. Diagonal covariance matrices were used.

The standard recipe for training and testing models built using the HMM Toolkit was used. Thus training was performed using Baum-Welsh reestimation of parameters from an initial model. Connected word Viterbi decoding was used for recognition (i.e. not isolated word recognition). The syntax for the recognition network was constrained to be a string of digits each followed by silence.

The AR-HMM recognition system also trained 8-emitting state left-to-right digit models and a 1-emitting state silence model. Each state had 2 mixture components. The autoregressive order was 20. Again continuous speech recognition was performed during decoding. Systems were trained using warped and non-warped parameterisations where the frequency warping was implemented as described in Chapter 6.

Similar to (Seymour 1996), it was found that improved recognition performance could be obtained by increasing the log observation probability of the silence model by a fixed value. Particularly in the case of low SNRs, this improves the chance of low energy frames at word boundaries being recognised correctly as silence. It was found though that the optimal value of this parameter is a function of both the SNR and the noise type. Therefore, since the noise type is assumed unknown, the silence probability increment was only varied for the baseline experiments to investigate the empirical best performance.

The models were trained on 100 clean digits and tested on 100 unseen clean digits and 100 digits for each noise type at each SNR from -6dB to 18dB. For processing, the digits were grouped into 5 files of 20 digits each similar to (Gales 1995) and (Seymour 1996).

7.5 Result Presentation

In order to enhance legibility, summary results are shown in this chapter. Full results for each test condition are given in Appendix E. The summary results are the distortion measures and recognition error rates averaged over all noise types for each SNR.

SNR	Distortion		% Error (D,S,I)			
dB	Overall	Speech	MFCC Models		AR Models	
-6	1.18	1.18	95.00	(380,0,0)	95.00	(380,0,0)
0	1.10	1.00	95.00	(380,0,0)	94.50	(378,0,0)
6	0.99	0.78	92.00	(300,0,0)	93.75	(375,0,0)
12	0.84	0.55	77.00	(88,160,60)	80.50	(127,156,26)
18	0.66	0.34	54.50	(57,97,64)	65.75	(27,214,22)
∞	0.00	0.00	0.00	(0,0,0)	0.00	(0,0,0)

Table 7.1: Distortions and word error rates for speech corrupted by the four noises and recognised using clean MFCC and AR models; derived from Table E.1.

7.6 Statistical Analysis

In order to compare whether the differences between the various algorithms are significant, statistical analyses of the recognition errors were performed. The technique used was the Matched-Pairs test described in (Gillick & Cox 1989). The statistic used by this test is the difference in the errors given by two different algorithms on the same set of utterances. The MFCC recognition results were compared since these are available for all test systems. A confidence level of 95% was used.

7.7 Baseline Performance

Table 7.1 shows the summary distortion and recognition error rates for speech corrupted by each of the four noise sources tested with clean models. Results for both clean MFCC and clean non-warped AR models are shown. The insertion penalties and silence probability increments were chosen to maximise the accuracy at each SNR and for each noise type. It can be seen that the performance degrades rapidly with decreasing SNR for both systems.

Table 7.2 summarises the word error rates achievable when the training and testing conditions are matched. This gives an indication of the upper bound performance achievable by any enhancement system. The matched system was obtained using single pass retraining.

Results are shown for AR systems with and without frequency warping. It can be seen that only the warped system approaches the performance of the MFCC-based system.

Table 7.3 summarises the word error rates for the compensated non-warped and warped AR systems. Here the compensated systems were formed using noise models trained on portions of the noisy signal labelled as

SNR	% Error (D,S,I)					
dB	MFCC Models		AR Models		Warped AR Models	
-6	32.50	(37,81,12)	42.00	(33,109,26)	33.50	(14,108,10)
0	2.50	(0,10,0)	24.00	(6,80,10)	6.50	(3,23,0)
6	0.25	(0,1,0)	9.75	(0,39,0)	0.75	(0,3,0)
12	0.00	(0,0,0)	3.75	(0,15,0)	0.00	(0,0,0)
18	0.00	(0,0,0)	1.75	(0,5,0)	0.00	(0,0,0)

Table 7.2: Word error rates for corrupted speech recognised using matched MFCC and AR models; derived from Table E.2.

SNR	% Error (D,S,I)			
dB	AR Models		Warped AR Models	
-6	38.75	(30,100,25)	20.75	(14,57,12)
0	22.5	(13,67,10)	4.75	(2,17,0)
6	8.00	(3,29,0)	0.75	(0,3,0)
12	3.25	(0,13,0)	0.00	(0,0,0)
18	1.00	(0,4,0)	0.00	(0,0,0)

Table 7.3: Word error rates for corrupted speech recognised using compensated AR models; derived from Table E.3.

noise. Therefore they represent the best compensated models available to calculate the probabilities required for enhancement. Again the performance of the warped system is superior.

In light of these results, the remainder of the work in this chapter will focus on enhancement systems based on warped AR-HMMs.

7.8 Enhancement Experiments

The configuration of the enhancement systems studied was as described in Section 6.2.3. That is, warped AR-HMMs were used to calculate the probabilities required for noise reestimation and choice of filters, whereas non-warped AR-HMMs were used to construct the filters required. The frame size, frame overlap and parameterisation of the speech used for the enhancement schemes were identical to that used previously. The noise was modelled using a single state AR-HMM with autoregressive order 20. This model was initialised by assuming the whole utterance was noise.

The variations investigated can be divided into three main categories:

- enhancement using Wiener filters versus enhancement using MMSE PSD estimation

- noise estimation from silences versus maximum likelihood noise parameter estimation
- word-based HMMs versus general speech HMMs.

Two types of recognition results are given in the following tables. The 'MFCC Recoded' column lists recognition results obtained by recognising the reparameterised enhanced speech using the clean MFCC models. The 'AR Compensated' column lists the results obtained when recognising the noisy speech using the compensated AR models. Hence the 'Recoded' column indicates how close the enhanced speech is to clean speech and the 'Compensated' column indicates the quality of the filters used to enhance the speech.

It was found for the word-based HMMs that the optimal insertion penalty for each test varied according to the noise type and SNR. However, the SNR was the dominating factor with deletions more prominent at low SNRs and insertions more prominent at high SNRs. In order to automatically choose the insertion factor, the NIST tool *wavmd* was used to approximate the SNR for each test file. This was then mapped to an insertion penalty. The mapping from SNR to insertion penalty was determined separately for each system tested.

7.9 Word-Based Models

The first set of experiments used systems constructed from word-based HMMs. The warped frequency clean speech word-based HMMs were identical to the warped clean AR-HMM models used for the baseline experiments. The non-warped clean speech models were trained using single pass retraining on the warped models in order that the data used to train each mixture component corresponded in both systems.

For these experiments, Viterbi alignment was used to obtain the most likely speech and noise state for each frame given the noisy observation. The most likely mixture component given this state was determined. The speech and noise statistics for this mixture component were then used to construct Wiener filters or MMSE PSD estimators for enhancement. These statistics were also used for the maximum likelihood noise reestimation scheme.

7.9.1 Noise Estimation Using Recognised Silences

Table 7.4 summarises the recognition performance and distortion measures for the system enhanced by the first (unweighted) scheme described in Section 5.1 in which the noise is reestimated from the recognised silences. These results are for a system using Wiener filters at the enhancement stage. Table 7.5 summarises the results for the same system with a MMSE PSD estimator at the enhancement stage.

	Distortion		% Error (D,S,I)			
SNR dB	All	Speech	MFCC Recoded		AR Compensated	
-6	1.50	1.23	76.75	(122,137,48)	74.25	(103,109,85)
0	1.34	1.06	63.25	(150,89,14)	44.50	(130,73,2)
6	0.94	0.68	39.50	(44,84,30)	17.50	(2,60,8)
12	0.59	0.44	21.75	(23,44,20)	5.75	(0,20,3)
18	0.52	0.38	39.00	(102,37,17)	1.75	(0,5,2)

Table 7.4: Distortions and word error rates for corrupted speech enhanced adaptively using recognised silences to estimate the noise; Wiener filters and word-based HMMs; derived from Table E.4.

	Distortion		% Error (D,S,I)			
SNR dB	All	Speech	MFCC Recoded		AR Compensated	
-6	1.20	1.00	73.50	(100,109,85)	74.25	(103,109,85)
0	0.97	0.77	51.25	(132,71,2)	44.50	(130,73,2)
6	0.61	0.40	17.75	(2,60,9)	17.50	(2,60,8)
12	0.34	0.23	6.00	(0,21,3)	5.75	(0,20,3)
18	0.22	0.17	1.75	(0,5,2)	1.75	(0,5,2)

Table 7.5: Distortions and word error rates for corrupted speech enhanced adaptively using recognised silences to estimate the noise; MMSE PSD estimation and word-based HMMs; derived from Table E.5.

Naturally the 'AR Compensated' columns in both these tables are identical since the two systems differ only in the function applied to perform the enhancement. It can be seen however, that the recognition performance and distortion measures are superior for the MMSE PSD estimator.

The inferior performance of the Wiener filter estimator arises because it is a MMSE time domain estimator yet the Itakura distortion measure and MFCC recognition results are influenced by errors in the short-time spectrum. Appendix F derives the expected short-time spectral amplitude given by a Wiener filter. This is seen to be highly dependent on the SNR. In particular, it is seen that when the signal and noise have approximately the same magnitude, the expected value of the enhanced spectral amplitude is a scaled version of the original signal rather than the original signal itself.

Both systems however have improved considerably on the original baseline in Table 7.1. It should be noted though that even the performance of the MMSE PSD system does not approach the theoretical best performance given in Table 7.2.

Perceptual Characteristics

Listening tests showed that the quality of the enhanced waveform was quite good unless gross recognition errors were made. These errors affect the filters used for enhancement. A particular problem was substitution errors which sometimes caused a digit to sound like the mis-recognised one. In addition, insertion errors and badly aligned digits led to portions of inadequately suppressed noise and deletion errors caused digits to be omitted. These general characteristics were observed for all the enhancement systems studied whenever recognition errors were made.

In general, the Wiener filter system sounded better than the MMSE PSD estimator system. This was because in the latter system, the silence portions of the wave were not completely suppressed, even if the silence filter was correctly chosen. This is an example of a situation in which perceptual improvement does not guarantee improved recognition accuracy.

Figures 7.1 and 7.2 show the spectrograms for a portion of clean speech and the same portion with Lynx noise at 12dB added. Figures 7.3 and 7.4 show the enhanced waveforms for the Wiener and MMSE PSD estimator systems respectively. It can be seen that the Wiener filter system suppresses more of the noise in the silence portions.

The trend of the Wiener filter enhancement system sounding superior yet having inferior recognition performance was observed across all systems studied. Therefore for the remainder of this dissertation, recognition results and distortion measures will only be given for MMSE PSD estimators. Conversely, spectrograms will only be shown for systems using Wiener filters.

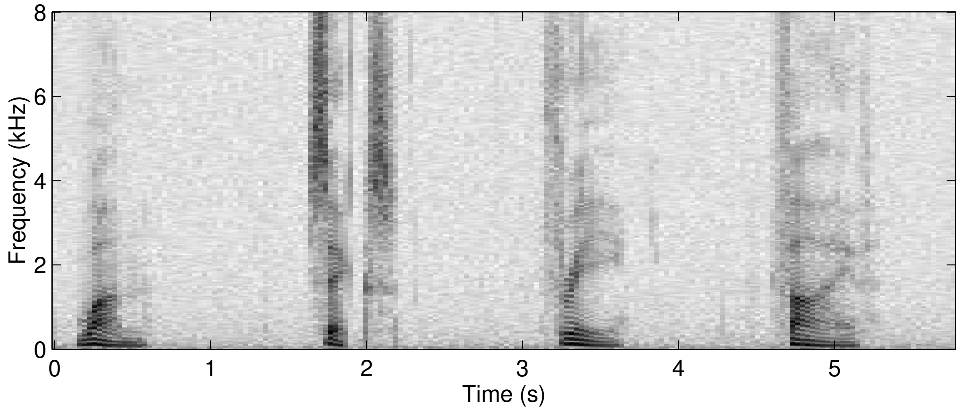

Figure 7.1: Clean speech spectrogram for the male speaker; first 4 test digits 'ONE', 'SIX', 'THREE', 'FIVE'.

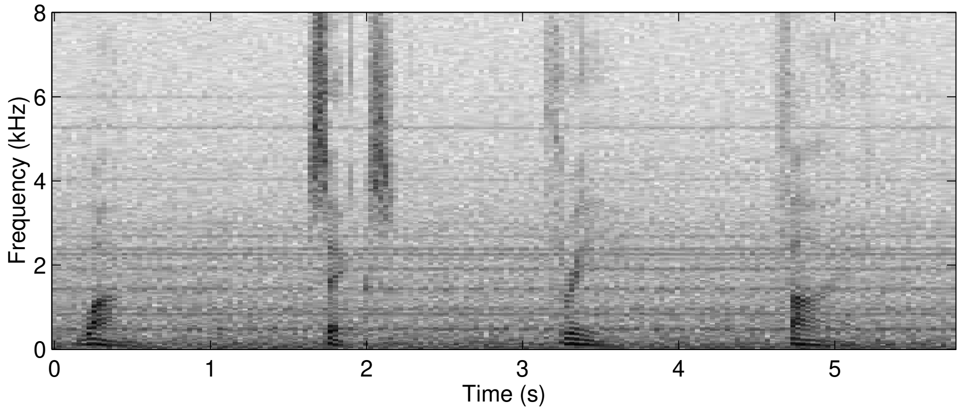

Figure 7.2: Spectrogram for speech with Lynx Noise at 12dB.

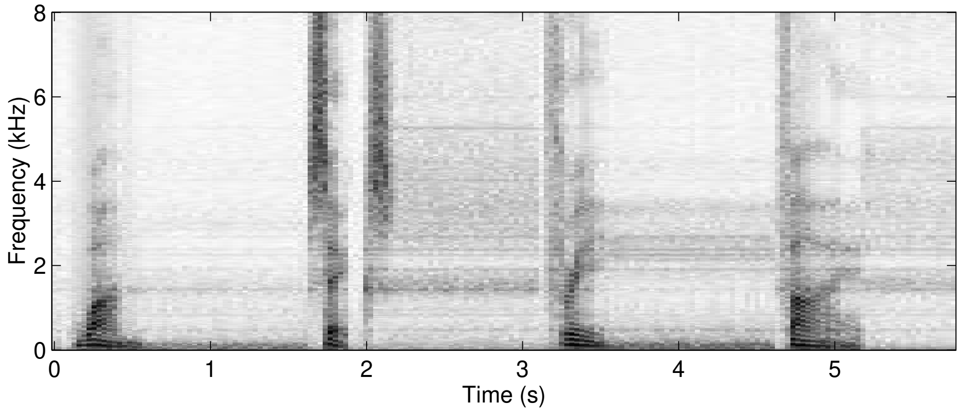

Figure 7.3: Spectrogram for speech with Lynx noise at 12dB; Enhanced using noise from recognised silences and Wiener filters.

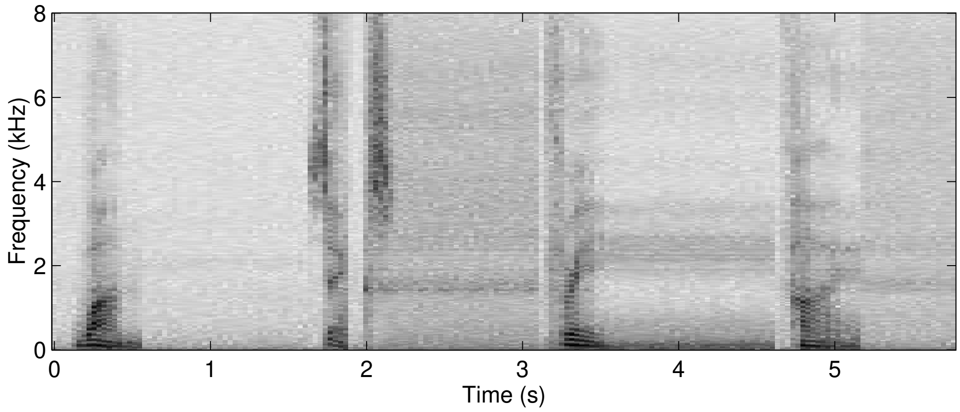

Figure 7.4: Spectrogram for speech with Lynx noise at 12dB; Enhanced using noise from recognised silences and MMSE PSD estimators.

SNR dB	Distortion All	Distortion Speech	% Error (D,S,I) MFCC Recoded		AR Compensated	
-6	0.63	0.69	26.50	(30,53,23)	26.75	(27,57,23)
0	0.39	0.42	6.25	(1,14,10)	6.25	(1,14,10)
6	0.26	0.27	0.50	(0,2,0)	0.75	(0,2,1)
12	0.18	0.18	0.00	(0,0,0)	0.00	(0,0,0)
18	0.13	0.12	0.00	(0,0,0)	0.00	(0,0,0)

Table 7.6: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimation; MMSE PSD estimation and word-based HMMs; derived from Table E.6.

7.9.2 Maximum Likelihood Noise Parameter Estimation

Table 7.6 summarises the results obtained using the maximum likelihood parameter estimation scheme described in Section 5.2. Here it is seen that the results are much improved over the technique of estimating the noise from the recognised silences. Analysis shows that they are significantly better. The results are also comparable to the matched model baseline results in Tables 7.2 and 7.3.

Figure 7.5 shows the convergence of the noise spectrum for the Lynx noise at 12dB. It is seen that convergence is achieved within several iterations.

Listening tests confirmed the superior performance of this algorithm over the previous one in which the noise is estimated from the recognised silence regions. A major benefit was fewer recognition errors giving superior filters for enhancement. Figure 7.6 shows the enhanced spectrogram for the same portion of speech studied earlier.

Variation of Autoregressive Order

As mentioned, an autoregressive order of 20 was used for the experiments. However, recognition systems tend to model the spectrum using less parameters. Therefore, the maximum likelihood MMSE PSD estimator system was implemented using an autoregressive order of 15.

The results from this experiment are summarised in Table 7.7. Although these seem comparable to the results for order 20, analysis shows that they are significantly worse. Therefore, for the remainder of this dissertation, an autoregressive order of 20 was maintained.

Comparison with a Linear Spectral System

In order to determine whether the AR-HMM system has any advantage over a linear spectral system, a linear spectral enhancement system was built and

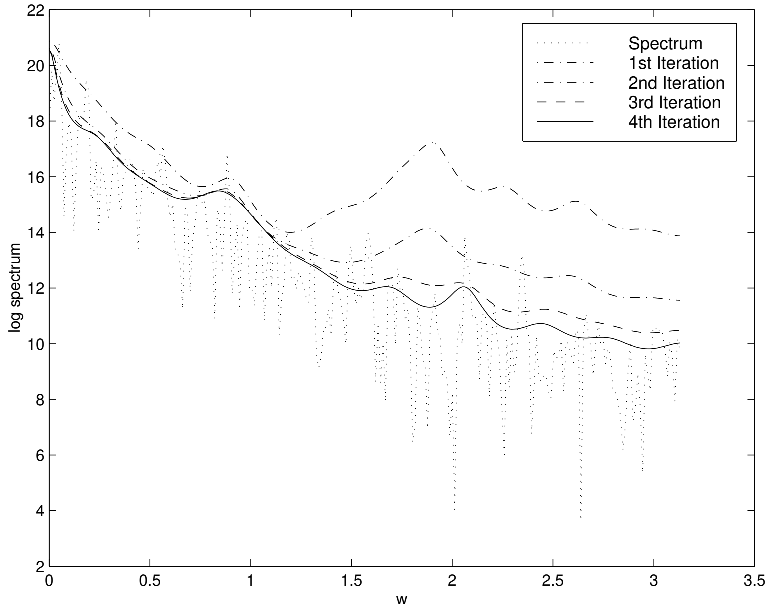

Figure 7.5: Convergence of the noise estimate for Lynx noise at 12dB.

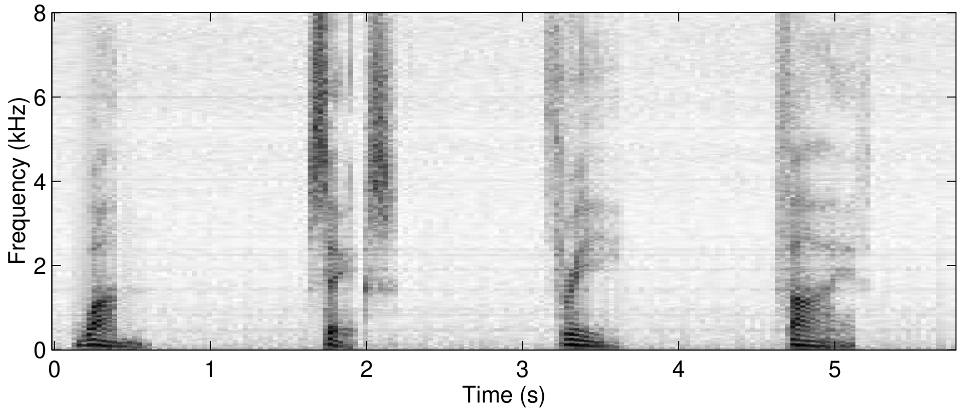

Figure 7.6: Spectrogram for speech with Lynx noise at 12dB; Enhanced using maximum likelihood noise estimation and Wiener filters.

	Distortion		% Error (D,S,I)			
SNR dB	All	Speech	MFCC Recoded		AR Compensated	
-6	0.67	0.71	28.25	(30,68,15)	26.75	(28,62,17)
0	0.42	0.44	7.50	(5,18,7)	6.50	(3,15,8)
6	0.28	0.29	2.00	(0,5,3)	1.50	(0,3,3)
12	0.18	0.19	0.25	(0,1,0)	0.00	(0,0,0)
18	0.14	0.13	0.00	(0,0,0)	0.00	(0,0,0)

Table 7.7: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimation; MMSE PSD estimation and word-based HMMs; autoregressive order 15; derived from Table E.6.

tested. This system was analogous to the AR-HMM system except a linear spectral distortion measure was used to determine the probabilities required for choosing the enhancement filters.

Specifically, a clean speech recogniser was built using data parameterised into 13 Mel spectrum bins. These features were modelled using standard HMMs with Gaussian state probabilities with diagonal covariance matrices. The digit and silence models had the same number of states as the AR-HMM system. Again two mixture components per state were used. The Mel spectrum system was then used to train non-warped AR models for the filtering stage as before.

The enhancement system was constructed as follows. The Mel spectrum models were used to determine the probability of each state given the noisy observations. The non-warped AR models were then used to filter the observations according to these probabilities. The system is analogous to that in Figure 6.5 and is shown in Figure 7.7. The number of parameters for both systems is approximately the same since the Mel spectrum system has twice the number of parameters per mixture component.

Table 7.8 summarises the distortion measures and recognition results for the linear system. A MMSE PSD estimator was used to determine the enhanced spectrum.

Comparison of Tables 7.6 and 7.8 shows that the two systems produce comparable results. However the results for the linear system are significantly worse than the AR-HMM results. Possibly the linear system's results could be improved using full covariance matrices but this would be at considerable computational expense. Thus in this domain it appears that there is some advantage in using the distortion measure provided by an AR-HMM system.

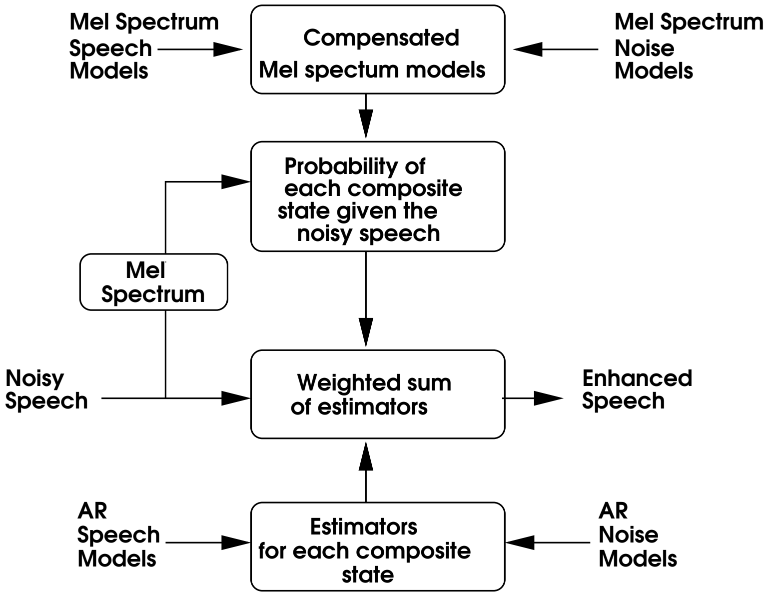

Figure 7.7: A linear spectral enhancement system; The probability of each composite state is given using a linear distortion measure.

	Distortion		% Error (D,S,I)	
SNR dB	All	Speech	MFCC Recoded	
-6	0.61	0.89	43.75	(66,58,51)
0	0.50	0.72	25.25	(69,32,0)
6	0.28	0.37	0.50	(1,1,0)
12	0.22	0.29	0.00	(0,0,0)
18	0.20	0.22	0.00	(0,0,0)

Table 7.8: Distortions and word error rates for corrupted speech enhanced adaptively using linear spectral models; derived from Table E.8.

7.10 General Speech Models

The work so far has demonstrated that good enhancement can be achieved using maximum likelihood noise parameter reestimation and word-based HMMs. In this section, a system based on simpler models is studied.

The system is similar to the original enhancement system proposed by Ephraim. Instead of using word or phone based HMMs, a general, multiple mixture speech model was trained. Here each mixture component models sounds with similar spectral characteristics. Although there is less prior information in this system, it is still possible to perform effective enhancement since ultimately it is the filter used that is important.

The HMM topology for the initial experiments consisted of a two-state model with the first state modelling speech using 128 mixture components and the second state modelling silence using a single mixture component. Transitions between the states were freely allowed.

Both warped and non-warped models were required for the enhancement systems as before. The warped models were initialised using single-pass retraining on a MFCC general speech system. The speech state of this MFCC system was trained using K-means clustering of MFCC vectors of size 30 on speech data as described in (Seymour 1996). The silence state was initialised on labelled silences. The MFCC system was used for initialisation in order to take advantage of its superior distortion measure.

After initialisation, the AR HMMs were reestimated using Baum-Welch reestimation. Single pass retraining was then used to train the non-warped AR models from the warped AR models as before. The same clean word-based MFCC models used previously were used to recognise the MFCC parameters calculated from the enhanced spectra.

In the previous experiments, Viterbi alignment was used to obtain the most likely speech and noise states corresponding to each frame. In this section, the forward-backward equations (e.g. (Rabiner & Juang 1993)) were used instead to calculate the likelihood of each mixture component of the compensated model. This was then used to construct a weighted sum of estimators for enhancement. Also, when maximum likelihood estimates of the noise are made, a weighted sum of estimators is used.

The experiments here are mostly concerned with the maximum likelihood noise parameter estimation scheme since this has been shown to be superior. However, since the likelihood of each compensated mixture component is available, the weighted recognised silence noise estimation scheme described in Section 5.1.1 is studied here for the first time.

7.10.1 Noise Estimation Using Weighted Recognised Silences

Table 7.9 summarises the results for a 128-mixture system. Here the probability of an observation being in the 'silence' state was used to weight the

	Distortion		% Error (D,S,I)	
SNR	All	Speech	MFCC	
dB			Recoded	
-6	1.02	0.83	89.25	(345,10,0)
0	0.80	0.59	83.75	(318,17,0)
6	0.53	0.38	56.50	(137,79,10)
12	0.32	0.23	25.25	(36,53,12)
18	0.16	0.13	2.50	(3,7,0)

Table 7.9: Distortions and word error rates for corrupted speech enhanced adaptively using weighted silences to estimate the noise; MMSE PSD estimation and general HMMs; 128 mixture components; derived from Table E.9.

	Distortion		% Error (D,S,I)	
SNR	All	Speech	MFCC	
dB			Recoded	
-6	0.90	0.79	85.75	(332,11,0)
0	0.67	0.56	78.75	(274,41,0)
6	0.43	0.37	51.50	(93,103,10)
12	0.25	0.22	15.75	(16,44,3)
18	0.15	0.13	2.50	(3,6,1)

Table 7.10: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimation; MMSE PSD estimation and general HMMs; 128 mixture components; derived from Table E.10.

noise estimates. The results in this table show that this system is inferior to any of the enhancement systems studied so far. This was confirmed by listening tests.

7.10.2 Maximum Likelihood Noise Parameter Estimation

Table 7.10 summarises the results for a 128-mixture system using maximum likelihood noise parameter estimates. These results are significantly better than those for the weighted silence technique above. However they are still inferior to the word-based system despite having a comparable number of parameters. Thus it can be concluded that some performance is sacrificed by the use of simpler models.

Figure 7.8 shows the enhanced spectrogram for the usual test segment of speech enhanced using this system with Wiener filters at the output stage. Listening tests showed some residual noise was present at 12dB. The residual

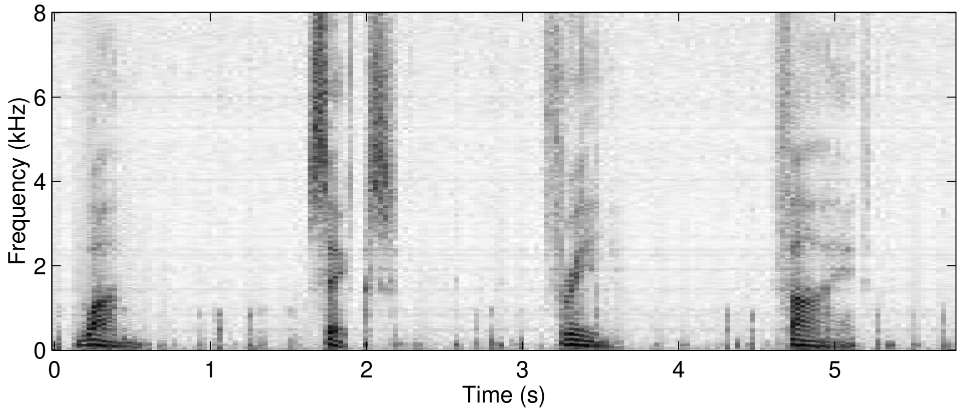

Figure 7.8: Spectrogram for speech with Lynx noise at 12dB; Enhanced using maximum likelihood noise parameter estimation and Wiener filters; 128 mixture components; general models.

noise is annoying in that it constantly fluctuates in the silence regions. This is because the filters used for enhancement can vary widely from frame to frame. This phenomenon was rarely observed for word-based systems since in this case, the language model combined with Viterbi alignment resulted in the same filter being used for many consecutive frames.

Increasing the Number of Mixture Components

The above experiment was repeated for a 256-mixture system. The results are summarised in Table 7.11. These are significantly better than the 128-mixture system results yet still do not approach the results of word-based systems. Listening tests showed a slight perceptual improvement over the 128-mixture system with less residual noise. The enhanced spectrogram for the usual test segment is shown as Figure 7.9.

Including Temporal Information

In (Seymour 1996) it was noted that a general speech model could be improved by the addition of temporal information. Specifically, rather than model the clean speech by a single state several states were used.

To test this, a 33 state general speech model was trained. The first 32 states contained 4 mixture components each and modelled the speech. The final state contained a single mixture component and modelled silence as before. The model was trained by continually splitting mixture components from an initial single mixture system.

Table 7.12 shows the performance for this system. The results are significantly better than both the 128 and 256 mixture component systems.

	Distortion		% Error (D,S,I)	
SNR dB	All	Speech	MFCC Recoded	
-6	0.86	0.81	87.50	(332,18,0)
0	0.64	0.57	76.00	(238,59,7)
6	0.41	0.37	48.00	(79,104,9)
12	0.23	0.22	10.00	(8,28,4)
18	0.15	0.14	2.00	(1,6,1)

Table 7.11: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimation to estimate the noise; MMSE PSD estimation and general HMMs; 256 mixture components; derived from Table E.11.

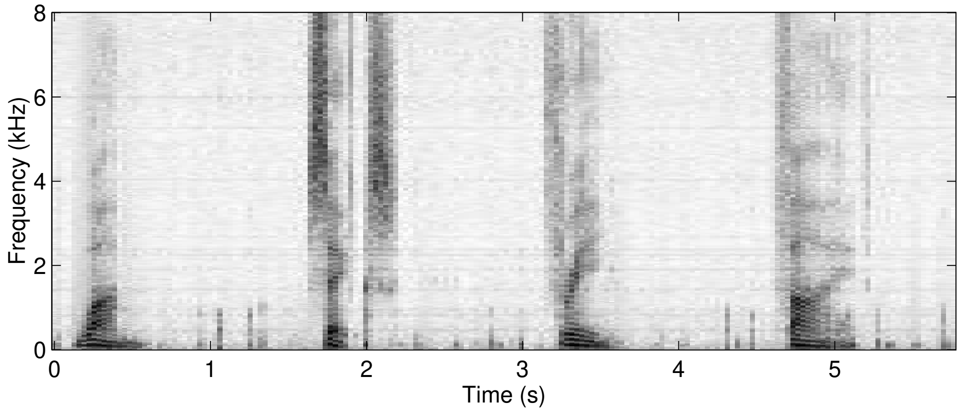

Figure 7.9: Spectrogram for speech with Lynx noise at 12dB; Enhanced using maximum likelihood noise parameter estimation and Wiener filters; 256 mixture components; general models.

	Distortion		% Error (D,S,I)	
SNR dB	All	Speech	MFCC Recoded	
-6	0.85	0.75	81.00	(182,96,46)
0	0.57	0.52	76.00	(95,119,38)
6	0.34	0.35	48.00	(26,70,27)
12	0.22	0.22	10.00	(4,24,13)
18	0.15	0.14	2.00	(1,6,0)

Table 7.12: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimation; MMSE PSD estimation and general HMMs; 32x4 mixture components; derived from Table 7.12.

The enhanced speech also had less residual noise. Thus there appears to be some advantage in adding temporal information.

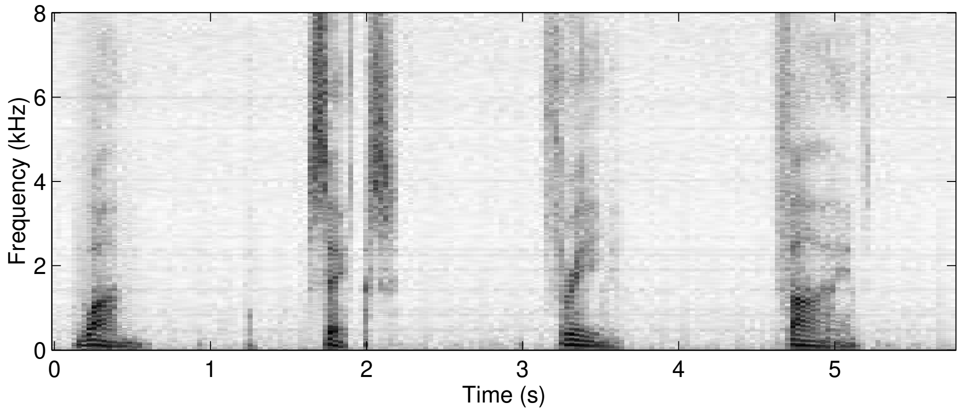

Figure 7.10: Spectrogram for speech with Lynx noise at 12dB; Enhanced using maximum likelihood noise parameter estimation and Wiener filters; 32x4 mixture components; general models.

7.11 Toward Real World Systems

In this section, some implementation and other issues are highlighted which show the way toward real-time, vocabulary and speaker independent implementation.

7.11.1 Approximation of $b_{\bar{x}_t m_t}(\mathbf{y}_t)$

For the relatively small NOISEX database, the computational burden is reasonable. A word-based maximum likelihood noise estimation system runs at approximately 25 times real time without pruning on a HP 9000/735 125MHz[1].

The main computational burden in the enhancement system centres around the calculation of $b_{\bar{x}_t m_t}(\mathbf{y}_t)$, the pdf of the noisy observation given state $\bar{x}_t$ and mixture component m_t. It will be recalled from Section 3.6.5 that this pdf is Gaussian with zero mean and a covariance matrix given by the sum of the corresponding speech and noise covariance matrices. Although the speech and noise processes are assumed autoregressive, this is not the case for the noisy process. Thus the covariance matrix in $b_{\bar{x}_t m_t}(\mathbf{y}_t)$ cannot be simplified such that it only depends upon $P+1$ parameters. However, since this covariance matrix has size KxK where K is the frame size, some approximations are necessary to give a tractable system.

One simplification is described in (Ephraim 1992a). Here the covariance matrix is approximated by a circulant matrix. This was the assumption made throughout this chapter so far.

An alternative approximation is suggested in (Sheikhzeheh et al. 1995). Here it is assumed that the noisy process *is* autoregressive so its pdf is simply given by the usual equation

$$b_{\bar{x}_t m_t}(\mathbf{y}_t) \approx (2\pi)^{-K/2}(\sigma^2_{\bar{x}_t m_t})^{-K/2}\exp\{-\frac{1}{2}\alpha(\sigma^{-1}_{\bar{x}_t m_t}\mathbf{y}_t; \mathbf{A}_{\bar{x}_t m_t})\}. \tag{7.1}$$

To obtain the autoregressive parameters for state $\bar{x} \equiv (x, \tilde{x})$ and mixture component m_t, it is assumed that the noisy autocorrelation function is given by the sum of the speech and noise autocorrelation functions. For each mixture component of each speech state and noise state, the average autocorrelation is stored. The autocorrelation function for each combined state is then given by the sum of these averages.

$$\mathbf{r}_{\bar{x}m} = \mathbf{r}_{xm} + \mathbf{r}_{\tilde{x}m} \tag{7.2}$$

The autoregressive parameters are then obtained from this autocorrelation function in the usual way.

This assumption results in substantial computational savings. For the NOISEX system, these are of an order of magnitude.

To investigate the effect of this approximation, a version of the maximum likelihood MMSE PSD estimation system was implemented using the approximation for $b_{\bar{x}_t m_t}(\mathbf{y}_t)$. Table 7.13 summarises the results of this experiment. Comparing these with the results in Table 7.6 it can be seen that very little performance is sacrificed for this task, particularly for the higher SNRs. The performance difference is not significant.

[1] This machine has a SPECint95 of 3.97 and a SPECfp95 of 4.61.

SNR dB	Distortion All	Distortion Speech	% Error (D,S,I) MFCC Recoded		% Error (D,S,I) AR Compensated	
-6	0.60	0.75	28.75	(26,69,20)	28.75	(26,69,20)
0	0.39	0.43	8.25	(1,14,18)	8.50	(1,14,19)
6	0.26	0.28	2.00	(0,4,4)	2.00	(0,2,6)
12	0.18	0.19	0.00	(0,0,0)	0.00	(0,0,0)
18	0.13	0.17	0.00	(0,0,0)	0.00	(0,0,0)

Table 7.13: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimation; MMSE PSD estimation and word-based HMMs; approximate $b_{\bar{x}_t m_t}(\mathbf{y}_t)$; derived from Table E.13.

7.11.2 Energy Considerations

One strength of AR-HMM systems is that they model a smoothed spectral contour. Hence the shape of the spectrum is important when distinguishing models. However, experience with MFCC systems has shown that they can benefit by the inclusion of energy information.

For clean AR-HMM recognition systems, this can be included as described in (Juang & Rabiner 1985) and implemented in Section 6.2.2. Here, the clean training observations are normalised by their autoregressive variance and the likelihood function is augmented by an extra term describing the probability of energy. Although this approach is successful, it is unfortunately not easily applicable to noisy AR-HMM recognition systems (Ephraim 1992*b*). This is because each noisy observation would need to be normalised by the variance of its corresponding clean speech autoregressive process, the determination of which is non-trivial.

A related issue is gain normalisation, the purpose of which is to correct mismatch between the clean speech training and testing data. One approach described in (Ephraim 1992*a*) and (Ephraim 1992*b*) is to normalise the clean training observations by their autoregressive variance as above, and then to match the gain of each noisy observation. At time t, the covariance matrix for the noisy observation is given by

$$S_{\bar{x}_t m_t} = g_t^2 S_{x_t m_t} + S_{\tilde{x}_t, m_t} \tag{7.3}$$

where all symbols are as previously described and g_t^2 is the gain at time t.

Although this approach is shown to improve recognition on small vocabulary tasks, it is computationally expensive and hence less applicable to large vocabulary systems. This is because the gain and hence the compensated covariance matrix $S_{\bar{x}_t m_t}$ must be computed for each observation vector. This precludes the possibility of precomputing the compensated model be-

	Distortion		% Error (D,S,I)			
SNR dB	All	Speech	MFCC Recoded		AR Compensated	
-6	0.79	1.01	48.00	(84,99,9)	42.75	(44,92,35)
0	0.56	0.72	21.00	(15,63,6)	14.50	(0,40,18)
6	0.39	0.53	6.25	(2,21,2)	15.00	(0,44,16)
12	0.28	0.35	2.00	(0,8,0)	10.00	(0,30,10)
18	0.21	0.23	0.25	(0,1,0)	1.00	(0,4,0)

Table 7.14: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimation; MMSE PSD estimation and word-based HMMs; female speaker; derived from Table E.14.

fore recognition[2]. Second, the actual process of estimating the gain contour is computationally expensive as it is iterative.

Another approach is to estimate a global gain factor as described in (Ephraim 1992*a*). This is computationally much less expensive and hence more applicable to a vocabulary independent system.

For the experiments described in this dissertation, the mismatch is small because the training and testing conditions do not vary widely given that in all cases, the testing and training data are from the same database. Therefore, no gain normalisation is used in the results presented here. It should be noted however that in a real-world system, some form of gain normalisation would be necessary and that for computational reasons, global gain normalisation would be the most applicable approach.

7.11.3 Variation of Test Speaker

So far, all the experiments have used the male speaker for testing and training. This section investigates the performance of the maximum likelihood PSD estimator enhancement scheme on a speaker which was not used to train the models. Specifically, the female speaker supplied with the NOISEX-92 database was used.

Table 7.14 shows the results for this test. The distortion measures are relative to clean female speech in order that the enhanced speech be compared to the original. Similarly, the error rates in the 'MFCC Recoded' column were obtained using MFCC models trained on the clean female speech.

It can be seen that although some performance has been sacrificed, the male models still provide reasonable enhancement on an unseen speaker.

[2] In investigative experiments with the medium vocabulary RM database, pruning did not decrease the computational load to an acceptable value.

Listening tests though showed that the enhanced speech had a not inconsiderable amount of residual noise in the speech portions and a distinct tendency to sound male. This phenomenon was more apparent at lower SNRs.

This is hardly surprising since the models used for filtering are speaker dependent. Unless it is desirable to transform a speaker's voice, it is therefore preferable to use speech models trained on many speakers in order that general filters be formed.

7.12 Summary

This chapter has investigated the performance of the enhancement algorithms developed in Chapter 5 and 6 on a small vocabulary speaker dependent task. Enhancement performance was evaluated objectively using the Itakura-Saito distortion measure and recognition performance. Some informal listing tests provided subjective results.

All the schemes developed were able to improve on the baseline results and provide effective enhancement. For the higher SNRs, the best performance was comparable to the best recognition performance achievable using models matched to the noise conditions.

The enhancement scheme based on maximum likelihood estimates of the noise parameters was the most effective. The enhancement provided by this technique was superior to that achieved using noise estimates from recognised silences as recognition errors in the latter technique due to poor noise estimates led to a poor choice of enhancement filters. The maximum likelihood scheme was also found to be superior to a similar scheme based on linear spectral models. Thus the distortion measure used in the AR-HMM framework provides some advantage.

Other conclusions from the work in this chapter are as follows. Using Wiener filters to perform the enhancement results in an enhanced signal which is perceptually pleasing but which may not necessarily be as useful as a front-end to a clean speech recogniser. This is because a Wiener filter gives the MMSE of the time domain signal yet recognition performance is influenced by errors in the short time spectrum. Conversely, a MMSE PSD estimator provides better performance as a front-end to a recogniser yet is less perceptually pleasing.

General speech models were also investigated as an alternative to the word-based models used in the earlier experiments. These were found to be inferior. Including a greater number of parameters or temporal information gave slight improvements in performance.

The chapter concluded with a discussion focusing on the changes necessary to implement a large vocabulary speaker independent system. Further experiments showed that a simplification of the compensated model did not adversely affect performance. The maximum likelihood enhancement technique was also shown to perform reasonably well on an unseen speaker although the enhanced speech exhibited some distortion.

Chapter 8

Evaluation on the Resource Management Database

In order to investigate the applicability of the enhancement schemes to a vocabulary independent system, experiments were conducted using the Resource Management (RM) Database (Price et al. 1988). The main focus of work in this chapter was to build a medium vocabulary speaker independent enhancement system.

8.1 The Resource Management Database

The RM database is suitable for medium vocabulary continuous speech experiments. It contains speech relating to a naval resource management task originally used by the DARPA speech community for benchmark tests. The total vocabulary size is 992 words.

The database is divided into speaker dependent and independent sections. Both sections contain speech sampled at 16kHz. The speaker dependent section contains speech from 8 speakers. There are 600 training utterances and 100 test utterances per speaker.

In the speaker independent section, the training data consists of 100 speakers with 40 utterances per speaker. The testing data is divided into four sets each consisting of 300 utterances comprising 30 sentences spoken by 10 speakers.

8.2 The Effect of Perceptual Frequency

In order to ascertain whether the results in Chapter 6 extended to a system with greater variance, clean speech recognition experiments were conducted using the RM database. To the author's knowledge, these experiments represent some of, if not the first, attempts to use AR-HMMs for anything but a simple recognition task. This is partly because this style of HMM was popular before the computing resources required for large scale recognition experiments were widely available.

In (Juang & Rabiner 1985), an isolated digits task was studied. Ephraim also investigated digit recognition and additionally recognition of the E-Set in (Ephraim 1992*b*). AR-HMMs have also been used for a speaker recognition task (Tishby 1991).

8.2.1 Speaker Dependent Experiments

Before moving on to a speaker independent system, experiments were conducted using speaker 'bef0_3' of the speaker dependent database. Multiple mixture 3 emitting state word-internal triphone-clustered HMMs were trained for this task using the RM Toolkit provided with HTK V1.5 (Young et al. 1993). The language model used was the word-pair grammar supplied with the database. A more complete description of the RM Toolkit can be found in (Woodland & Young 1993).

Both MFCC-based and AR-HMM-based systems were trained. The frame rate and frame size were 10ms and 32ms respectively.

The features for the MFCC-based system consisted of 12 cepstral coefficients plus one energy coefficient. Pre-emphasis was not used. A system with and without delta coefficients was trained. Diagonal covariance matrices were used.

The AR-HMM system had an autoregressive process of order 20. A system with and without perceptual frequency was trained. The clean autocorrelation coefficients were normalised by their autoregressive variance so that energy information was removed. The triphones were clustered using the Euclidean distance between the average autocorrelation coefficients for each state as the distance measure.

Word error rates for both the AR-HMM and MFCC-based systems are shown in Table 8.1. It can been seen from these results that again recognition has been improved using the bilinear transform. However, the performance of the AR-HMM systems fall far short of the performance achievable using a MFCC-based system. It appears then that for this system, while the use of a perceptual frequency scale does improve recognition, it does not completely explain the difference between the information provided by the non-warped AR-HMM system and the MFCC-no-delta system.

8.2.2 Discussion

It was seen in Section 6.2.2 that the incorporation of energy information as an extra term in the AR-HMM likelihood function could decrease the error rate. Therefore the same procedure was followed here. It was found though that only slight decreases in recognition error of the order of 1-2% absolute could be achieved by the use of this technique. Hence energy information

Model	Number Mixture Components	% Error (D,S,I)
AR no warping	3	23.6 (14,103,76)
	4	22.0 (12,83,85)
AR with warping	3	19.0 (19,77,59)
	4	17.8 (15,80,51)
MFCC no deltas	3	9.2 (13,40,22)
	4	7.8 (12,32,20)
MFCC with deltas	3	7.7 (11,29,23)
	4	6.9 (8,26,22)

Table 8.1: RM clean speech recognition error rates for speaker bef0_3. AR-HMMs with and without frequency warping are compared to MFCC systems with and without delta parameters.

does not account for the difference between the AR-HMM and MFCC-no-delta systems as was the case for the earlier experiments in Section 6.2.2.

The major source of deviation between the two systems is the distortion measure used. As discussed in Section 6.1, both systems essentially use spectral ratios to compare an utterance to trained templates. However, the MFCC-based system uses a more powerful distortion measure. While the AR-HMM system treats errors in any part of the spectrum equally, the MFCC-based system effectively weights the error in each part of the cepstrum by the trained variance of each estimate.

A point to note is that this is done at the expense of twice the number of system parameters. That is, the MFCC-based system has both a mean and a variance for each state whereas the AR-HMM system has a mean and a single variance per state. However as shown in Table 8.2, increasing the number of AR-HMM system parameters by increasing the number of mixture components does not solve this problem. This latter result was also observed in (Juang & Rabiner 1985) on a simpler task and essentially the same conclusion was reached.

8.2.3 Speaker Independent Systems

Speaker independent speech recognition is significantly harder than the speaker dependent task. State-of-the-art results are typically achieved using MFCC feature vectors augmented with both delta and delta-delta parameters (Young et al. 1993). Therefore, given the speaker dependent results in the previous section, it is unlikely that a medium vocabulary speaker independent AR-HMM system can be built without further extensions to the

Model	Number Mixture Components	% Error (D,S,I)
AR with warping	4	17.8 (15,80,51)
	5	16.1 (15,70,47)
	6	15.4 (17,60,49)

Table 8.2: RM clean speech recognition error rates for speaker bef0_3 showing dependence on the number of mixture components for an AR-HMM system with frequency warping.

AR-HMM paradigm. Even if a solution could be found to the problem of the single variance in the AR-HMM system, a further extension would be needed to incorporate delta information since this plays an even greater role in speaker independent systems.

Some attempts were made to build an AR-HMM speaker independent system but the results were not encouraging. A typical error rate for a 4 mixture component system using perceptual frequency was over 40%.

8.3 Enhancement Experiments

The enhancement experiments investigated the feasibility of vocabulary and speaker independent enhancement systems based on AR-HMMs.

Preliminary experiments with the speaker dependent triphone models indicated that a system incorporating language model information was not feasible. As in the previous word-based HMM experiments, Viterbi alignment was used to select the most likely filter to enhance each frame. However, because the performance of the clean recognition system was poor, gross errors often resulted when choosing this filter with a detrimental effect on the enhancement. For the digits system studied in the previous chapter, the recognition rate was extremely good due to the simplicity of the task. Therefore, the extra information provided by the language models helped the enhancement. For systems with more variance however, this extra information is a disadvantage.

Therefore a system based on a much simpler general speech mixture model was studied. This system does not use a language model. A similar non-adaptive enhancement system based on compensated MFCC models without delta parameters has been shown to perform well on this task (Seymour 1996).

8.3.1 Enhancement System

The enhancement system was modelled closely on the general speech systems studied in Section 7.10. In particular, the models were trained in the same manner. The first three utterances for each training speaker were used as training data similar to (Seymour 1996) in order to avoid overtraining of the models. Only the maximum likelihood noise parameter estimation scheme was investigated since this gave the best performance.

As before, warped AR-HMMs were used to calculate the probabilities required and non-warped models were used for enhancement. The approximation of $b_{\bar{x}_t m_t}(\mathbf{y}_t)$ described in Section 7.11.1 was used to calculate the compensated models to make the system computationally tractable.

In this section, the noise models were initialised from the frame of the test utterance with the minimum power. This was found to give enhanced speech which was perceptually superior to that from a system which initialised the noise using all of the test frames. A small amount of recognition performance was sacrificed by this initialisation technique (about 1% absolute on the 512 mixture system at 18dB).

8.3.2 Evaluation Methods

The recognition error rate was the main figure of merit for these experiments since it is difficult for a human assessment to be made on such a large dataset. MFCC parameters were derived directly from the enhanced spectrum as previously and tested using a clean speech MFCC recognition system.

The clean recognition system was trained using the RM Toolkit as a template. The feature vectors used contained 13 cepstral coefficients including the 0th coefficient augmented with delta and delta-delta coefficients. These were the first 13 coefficients returned from a MFCC analysis of order 24. The data was pre-emphasised by the filter $H(z) = 1 - 0.97z^{-1}$. The features were modelled using diagonal covariance matrices.

A non-standard frame rate and frame size of 16ms and 32ms respectively were used as in previous enhancement experiments. A 5 mixture component/state triphone-clustered system was built. Each triphone was modelled by a 3-state left-to-right HMM.

The non-standard frame rate affects the modelling of short phones by increasing the minimum duration. This problem was alleviated by the introduction of a skip state into each triphone model. The frame rate also affects the period of time used to calculate the delta and delta-delta coefficients. This effect was not considered since the baseline performance of the modified system was adequate.

8.3.3 Formation of Noise Corrupted Data

Since the RM database contains clean speech only, noise was added to the test sets in order to conduct enhancement experiments. Specifically, a random segment of Lynx noise was taken from the NOISEX-92 database, scaled appropriately and added to each utterance.

Two noise conditions were considered corresponding to the attenuation of the Lynx noise by 20dB and 12dB respectively. The NIST utility *wavmd* was used to estimate the SNR of the corrupted utterances.

wavmd estimates the SNR without prior knowledge of the noise statistics. Therefore, the SNR values calculated are not exact. Figure 8.1 shows SNRs calculated using *wavmd* on the NOISEX-92 database verses the quoted SNRs. It can be seen that for this noise, *wavmd* tends to overestimate the SNR, especially when the SNR is low. This should thus be borne in mind when comparing results from this chapter with those in Chapter 7.

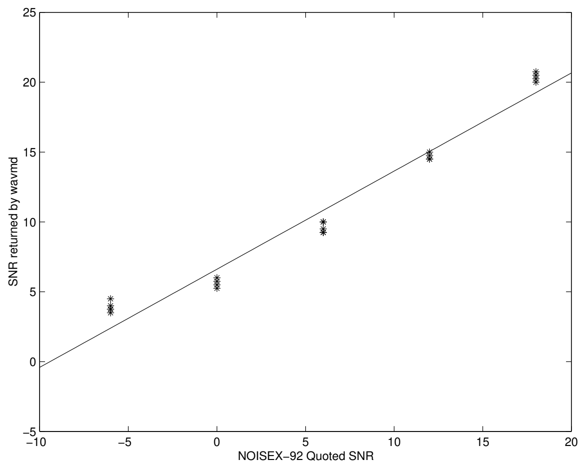

Figure 8.1: Performance of *wavmd* on Lynx Noise in NOISEX-92 with a straight line fitted to the data. Note that *wavmd* tends to overestimate the SNR.

Table 8.3 shows the SNRs returned by *wavmd* on the test sets for the two noise conditions. These will henceforth be referred to by their average SNRs 18dB and 12dB respectively.

8.3.4 Baseline Performance

Table 8.4 shows the word error rates for the clean and noisy speech on the four test sets. The clean baseline is worse than state-of-the-art performance on this database because of the decreased frame rate as discussed. It can be seen that the addition of noise has a substantial effect on the error rate.

Noise Attenuation	SNR (dB)				
	Feb89	Oct89	Feb91	Sep92	Avg.
Clean	48.7	48.5	48.5	49.4	48.8
20dB	17.7	17.9	18.2	19.0	18.2
12dB	11.1	11.5	11.4	12.5	11.6

Table 8.3: SNR values returned by *wavmd* on the RM Database with Lynx noise added at various attenuations.

Noise	% Error (D,S,I)				
	Feb89	Oct89	Feb91	Sep92	Avg.
Clean	6.3 (66,86,9)	7.3 (73,115,9)	5.9 (47,86,13)	11.0 (99,157,25)	7.6
Lynx 18dB	38.9 (212,691,92)	30.4 (182,550,85)	35.8 (150,646,92)	43.1 (207,755,141)	37.0
Lynx 12dB	80.4 (498,1268,163)	81.0 (556,1490,131)	77.7 (498,1268,163)	85.2 (561,1473,147)	81.1

Table 8.4: Baseline results for the RM database speaker independent test sets for clean speech and speech corrupted using Lynx noise at various attenuations.

Noise	% Error (D,S,I)				
	Feb89	Oct89	Feb91	Sep92	Avg.
Lynx 18dB	16.7 (139,266,23)	14.8 (139,242,16)	14.1 (107,227,16)	21.0 (163,333,41)	16.7
Lynx 12dB	40.8 (360,654,31)	31.3 (296,508,35)	34.0 (268,544,33)	40.4 (319,653,61)	36.6

Table 8.5: Results for the RM speaker independent test sets tested using matched models for various noise conditions.

Nr. Mixes	% Error (D,S,I)				
	Feb89	Oct89	Feb91	Sep92	Avg.
128	23.6 (117,419,69)	20.1 (114,359,66)	21.7 (87,377,74)	27.6 (131,470,104)	23.2
256	18.7 (106,320,52)	16.5 (117,275,50)	18.9 (82,323,64)	23.9 (133,402,76)	19.5
512	18.1 (110,301,52)	15.5 (119,253,43)	18.1 (80,301,68)	24.2 (132,412,76)	19.0

Table 8.6: Enhancement results for the RM speaker independent test sets for Lynx noise at 18dB SNR. Speech enhanced using general speech models with varying numbers of mixture components.

Table 8.5 shows the word error rates achievable when the training and testing conditions are matched. The matched MFCC models were obtained by adding Lynx noise to the training set and then using this data to train models using single pass retraining. These results give an indication of the best performance achievable by any enhancement system.

8.3.5 Enhancement Performance

Experiments were first conducted on the test files at 18dB. Table 8.6 shows the word error rates for the described enhancement system for various numbers of mixture components in the models. It can be seen that a substantial improvement has been made on the baseline performance. The performance improves as the number of mixture components increases although only a slight, non-significant improvement is noticed for the 512-mixture system over the 256-mixture system.

The error rates for speech at 12dB enhanced using a 512 mixture component system are shown in Table 8.7. Again the baseline performance has been substantially improved upon. From these two test conditions, it ap-

Nr.	% Error (D,S,I)				
Mixes	Feb89	Oct89	Feb91	Sep92	Avg.
512	42.8 (154,752,189)	35.7 (138,628,193)	37.9 (95,657,189)	46.7 (158,804,232)	40.8

Table 8.7: Enhancement results for the RM speaker independent test sets for Lynx noise at 12dB SNR. Speech enhanced using general speech models with 512 mixture components.

pears that the improvement gained by the enhancement technique halves the error rate. However this performance is significantly worse than the matched model results given in Table 8.5 suggesting that there is a modelling deficiency.

Inclusion of Temporal Information

In Section 7.10 and (Seymour 1996), it is seen that including temporal information improves the performance of general speech models. Therefore a 32-state, 16 mixture component/state model was implemented similar to (Seymour 1996).

The results of this experiment for the two noise conditions are shown in Table 8.8. The results at 18dB are not significantly different to the 512 mixture component system. The 12dB results are significantly worse than the 12dB 512 mixture component system.

Thus there appears to be no advantage to using more than one speech state. One possible reason for the inferior results may be the different training procedures for the models since the 32x16 system was formed by continually splitting a AR-HMM system whereas the 512 system was initialised from single pass retraining on a MFCC system. It seems that the superior distortion measure of the MFCC system provides some advantage for initialisation.

Perceptual Analysis

Informal listening tests showed that the enhanced speech contained some residual noise. This was quite considerable and annoying for the speech at 12dB. Seymour reported much less noise in (Seymour 1996). Since in this previous work, the filters were chosen according to MFCC probabilities, it seems likely that the inferior modelling ability of AR-HMMs is causing the increased distortion.

Figures 8.2 to 8.6 show the clean, noisy and enhanced speech for the first sentence for speaker 'alk0_3'. The enhancement was performed using

SNR	% Error (D,S,I)				
dB	Feb89	Oct89	Feb91	Sep92	Avg.
18	18.0 (119,289,52)	16.4 (115,278,48)	18.0 (82,312,53)	24.5 (137,401,90)	19.2
12	46.0 (145,819,214)	38.9 (127,703,213)	42.3 (102,711,238)	48.5 (152,852,236)	43.9

Table 8.8: Enhancement results for the RM speaker independent test sets for Lynx noise at various SNRs. Speech enhanced using general speech models with 32-state, 16 mixture component/state models.

the 512-mixture system. The text of this sentence is 'WHEN WILL THE PERSONNEL CASUALTY REPORT FROM THE YORKTOWN BE RESOLVED'.

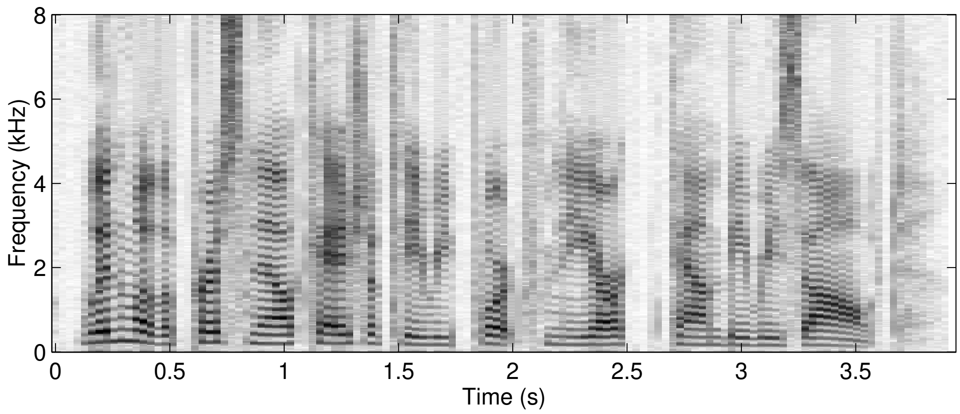

Figure 8.2: Clean speech spectrogram for the first sentence for speaker alk0_3.

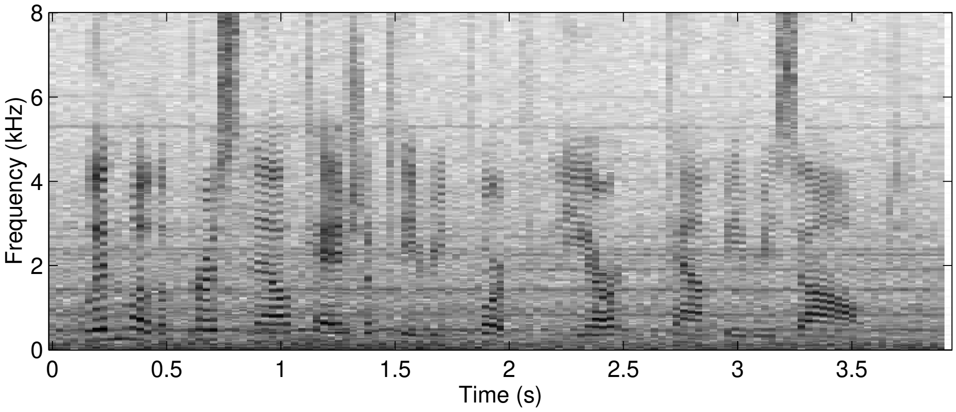

Figure 8.3: Speech corrupted by Lynx noise at 12dB; first sentence for speaker alk0_3.

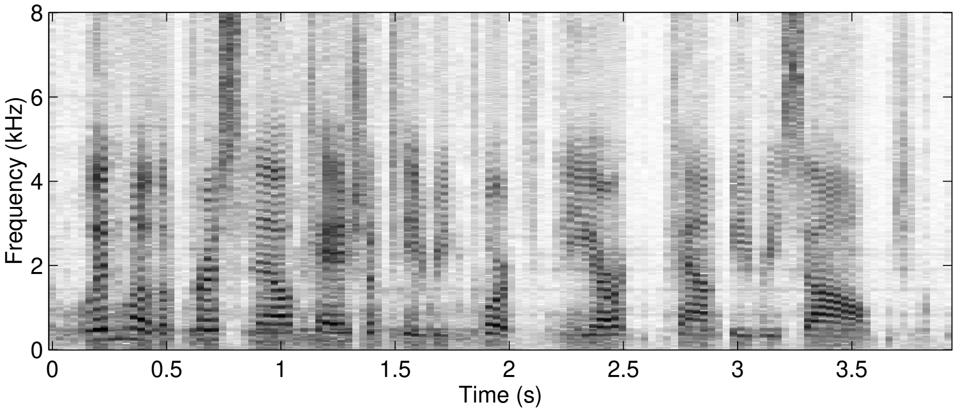

Figure 8.4: Speech corrupted by Lynx noise at 12dB enhanced using Wiener filters formed from 512 mixture component models; first sentence for speaker alk0_3.

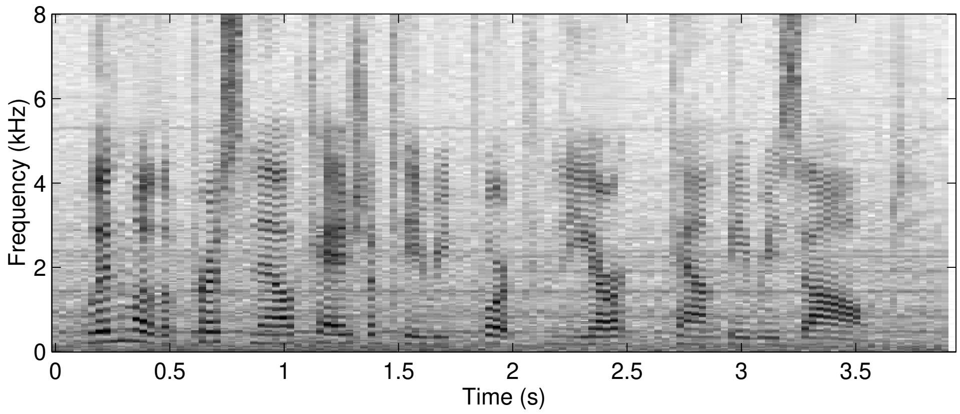

Figure 8.5: Speech corrupted by Lynx noise at 18dB; first sentence for speaker alk0_3.

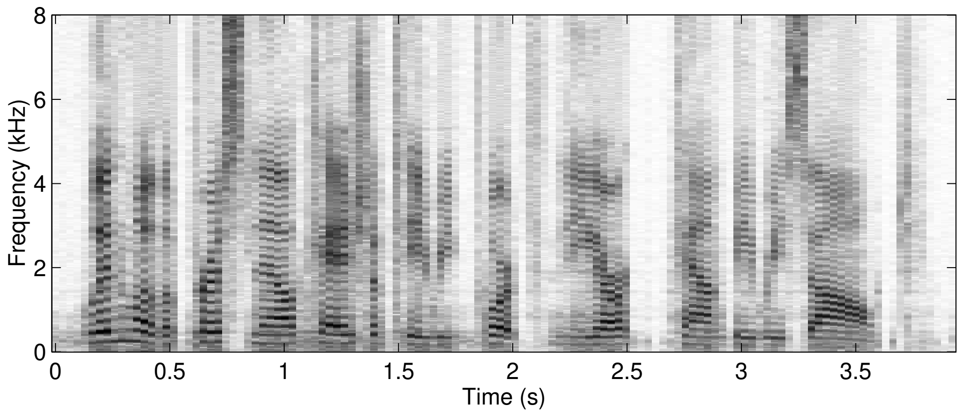

Figure 8.6: Speech corrupted by Lynx noise at 18dB enhanced using Wiener filters formed from 512 mixture component models; first sentence for speaker alk0_3.

8.4 Summary

This chapter has investigated the performance of AR-HMM recognition and enhancement systems on a medium vocabulary database.

The first section investigated whether the addition of perceptual frequency to AR-HMMs improved the modelling power of these systems. Experiments on clean speaker dependent data showed that a relative reduction in error of 19% was possible.

The improvement did not however bring the AR-HMM results up to the same level as a MFCC system without delta parameters as had been observed in earlier experiments. This was because of the larger variance in the RM task. In AR-HMM systems, the variance of the spectrum is modelled using a single variance per mixture component whereas MFCC systems are able to model the variance of each cepstral coefficient.

Since the recognition results were not encouraging for clean speaker dependent data, it was not surprising that speaker independent results were poor. Here, a MFCC system requires delta-delta coefficients in addition to delta coefficients in order to achieve good performance. Thus even if the AR-HMM variance modelling problem can be solved, further additions to the AR-HMM paradigm would be necessary to deal with this more difficult task.

The second section investigated medium vocabulary speaker independent enhancement. Since a medium vocabulary AR-HMM system could not be built for the reasons outlined above, enhancement based on general speech models was studied. Only the maximum likelihood noise reestimation scheme was investigated. Performance evaluation was performed in terms of recognition accuracy of enhanced speech. Informal listening tests were also conducted.

The experiments showed that substantial improvements over baseline results could the made using the enhancement technique at noise levels of 18dB and 12dB. Perceptually, the enhanced speech was inferior to the speech generated using speaker dependent models in the previous chapter. However, reductions in interfering noise were achieved.

Chapter 9

Conclusions and Future Work

The work in this dissertation has aimed to enhance speech corrupted by additive noise in an unknown noisy environment when only one microphone is available.

Speech enhancement has wide application in the field of speech processing. It is used to improve the perceptual aspects of speech to increase intelligibility and decrease listener fatigue. In addition, it is useful for cleaning up corrupted speech prior to input to coding and recognition systems in order to improve their performance. However, many existing enhancement techniques are unable to adapt to unknown noise. This provides the primary motivation for the work in this dissertation.

The techniques developed are based on an enhancement system by Ephraim (Ephraim 1992*a*) which models speech and noise statistics using AR-HMMs. These statistics must be trained using *a priori* information. Given these models, very effective enhancement can be achieved.

Working in the AR-HMM domain has three advantages for enhancement of additive noise. The first is that the feature vectors used are linearly combinable. This is important when forming a compensated system to model the corrupted speech and also when forming maximum likelihood estimates of unknown parameters. The second advantage is that the distortion measure used to compare features to templates is the Itakura-Saito distortion measure. This is more effective than a linear spectral distortion measure which would be the metric used if a linear spectral HMM system was built. Finally, the computational requirements of AR-HMMs are reasonable.

It should be noted that AR-HMMs have been neglected in recent years. Thus a further reason for working in this domain was to investigate the performance of such systems on difficult tasks which were computationally infeasible when these models were first proposed.

The work in this dissertation extends (Ephraim 1992*a*) by estimating the noise statistics directly from the signal to be enhanced. Two main approaches were developed. The first considers estimating the noise from detected pauses. The AR-HMM framework was used for the pause detection. The second approach uses maximum likelihood parameter estimation to estimate the noise statistics given a compensated AR-HMM model of the noisy speech.

Additional work in this dissertation investigates improving the performance of AR-HMM systems. Here perceptual frequency is incorporated using the bilinear transform. This extension can be used in AR-HMM recognition and enhancement systems.

9.1 Summary of Results

The enhancement schemes developed were evaluated using the NOISEX-92 and RM databases providing information about performance on small vocabulary speaker dependent and medium vocabulary speaker independent tasks respectively. The addition of perceptual frequency to AR-HMM systems was also evaluated on these databases with additional small vocabulary speaker independent experiments performed using the ISOLET database.

The performance of the enhancement schemes was evaluated using distortion measures and recognition results obtained using features directly parameterised from the enhanced spectra. Informal listening tests were also conducted. The perceptual frequency extension was evaluated using recognition scores.

The main result of this dissertation is that maximum likelihood parameter estimation can be used in the AR-HMM domain to provide effective speech enhancement in unknown stationary additive noise. The scheme showed substantial improvement over baseline results. Perceptual improvement was also observed. Performance was shown to be effective on a variety of stationary noises at various SNRs on several tasks of varying complexity.

For the small vocabulary task, it was found that the performance of the enhancement system which estimated the noise statistics from recognised pauses was significantly worse than the maximum likelihood system. In this domain, the performance of a linear spectral estimator was also significantly worse. Further experiments showed that word-based models were superior to general speech models but that temporal information could improve the latter.

The addition of perceptual frequency to AR-HMM systems was shown to improve clean recognition performance substantially on both small and medium vocabulary tasks. The results for the former system were comparable to a MFCC system without delta parameters. This was not the case for the medium vocabulary system. This was because AR-HMMs are less able to model systems with increased variance since they only have a single variance parameter per mixture component, whereas MFCC systems have a variance parameter for each cepstral coefficient in each mixture component.

AR-HMMs are not at present able to incorporate delta features. Since these have a substantial impact as the complexity of a recognition task increases, it was not possible to build a medium vocabulary speaker independent AR-HMM recognition system. Therefore, the medium vocabulary speaker independent enhancement tests were performed using general speech models rather than word or phone based models.

These tests showed that effective enhancement could be performed on a larger system at several noise levels for the Lynx helicopter noise. Substantial improvements over baseline results were obtained with the error rate decreasing by an absolute factor of approximately 50%.

Perceptually, the enhanced speech was best for the speaker dependent system. As more speakers were introduced, the quality of filters available for enhancement decreased as did the probability of choosing an appropriate filter.

9.2 Future Directions

Future research directions can be divided into two categories: improvements to the enhancement algorithm and improvements to AR-HMMs.

One extension to the enhancement algorithm would be to investigate sequential techniques of noise parameter estimation. These approaches estimate the unknown parameters to maximise the likelihood at each time slice and thus allow adaptation to slowly varying noise.

In general, the performance of the algorithms in non-stationary noise was not studied. However, the maximum likelihood algorithm is able to enhance speech corrupted by noises which can be modelled by multi-state HMMs. Therefore, it would be interesting to investigate how well the algorithm performs in such circumstances.

Another path for investigation is an examination of the method used to obtain the maximum likelihood estimates of the noise parameters. At present, this is achieved using the expectation maximisation algorithm which is known to be slow and to often converge to a local likelihood maximum. Other gradient descent methods may lead to more optimal or less computationally expensive algorithms.

Careful investigation of AR-HMM systems has led to an improvement in their performance by the incorporation of perceptual frequency. Further extensions are necessary however to improve the performance of these models on large vocabulary systems.

Comparison with MFCC systems has suggested that the inclusion of more variance information and delta parameters should be the primary areas of investigation. There are also unresolved issues in the areas of inclusion of energy information and clustering techniques.

Appendix A

Maximum Likelihood Estimation of Noise Model Parameters

In this appendix, maximum likelihood estimation of the noise parameters is described within an autoregressive HMM framework. The technique is similar to that described in (Rose et al. 1994) but differs in that AR-HMMs are used instead of mixtures of Gaussians to model the speech and noise. This affects the type of parameters reestimated. In (Rose et al. 1994), the means and covariances of Gaussian mixture models are estimated whereas here, the estimated parameters are the transition probabilities and the autocorrelation function corresponding to each HMM state. Also, in (Rose et al. 1994) the speech statistics are unknown whereas here the noise statistics are unknown.

The notation used in this appendix is identical to that used in Chapter 3. To simplify the equations, it is assumed that there is a single mixture component per AR-HMM state. The extension to the multiple mixture case is straightforward.

A.1 Probability Density Functions

The pdfs describing the clean speech and noise processes are given by Equations A.1 and A.2 below respectively.

$$p(\mathbf{S}|\lambda_s) = \sum_{\mathbf{X}} a_{x_0 x_1} \prod_{t=1}^{T} a_{x_t x_{t+1}} b_{x_t}(\mathbf{s}_t|\lambda_s) \tag{A.1}$$

$$p(\mathbf{D}|\lambda_d) = \sum_{\tilde{\mathbf{X}}} a_{\tilde{x}_0 \tilde{x}_1} \prod_{t=1}^{T} a_{\tilde{x}_t \tilde{x}_{t+1}} b_{\tilde{x}_t}(\mathbf{d}_t|\lambda_d) \tag{A.2}$$

Here $\mathbf{S}$ is a sequence of K dimensional clean speech observations, $\mathbf{X}$ is a sequence of clean speech states, $a_{x_t x_{t+1}}$ is the transition probability from state x_t to state x_{t+1} and $b_{x_t}(\mathbf{s}_t|\lambda_s)$ is the pdf of the output vector $\mathbf{s}_t$ from the state x_t. The sum over $\mathbf{X}$ represents the summation over all possible clean speech state sequences. Similarly $\mathbf{D}$ is a sequence of noise observations and $\tilde{\mathbf{X}}$ is a sequence of noise states. λ_s and λ_d refer generally to the speech and noise model parameters.

The pdfs $b_{x_t}(\mathbf{s}_t|\lambda_s)$ and $b_{\tilde{x}_t}(\mathbf{d}_t|\lambda_d)$ are assumed Gaussian with zero mean and covariance matrices Σ_{x_t} and $\Sigma_{\tilde{x}_t}$ respectively. Later it will be assumed that the speech and noise models are autoregressive. However, the initial part of the analysis holds for general covariance matrices.

Since additive noise is being considered, the speech and noise at time t, $\mathbf{s}_t$ and $\mathbf{d}_t$, are related to the noisy observation $\mathbf{y}_t$ by

$$\mathbf{y}_t = \mathbf{s}_t + \mathbf{d}_t. \tag{A.3}$$

The likelihood of the noisy speech observation is given by

$$P(\mathbf{Y}|\lambda) = \sum_{\mathbf{X}} \sum_{\tilde{\mathbf{X}}} \int \int_{\mathbf{C}} P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}}|\lambda) d\mathbf{S} d\mathbf{D} \tag{A.4}$$

where

$$P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}}|\lambda) = a_{x_0 x_1} a_{\tilde{x}_0 \tilde{x}_1} \prod_{t=1}^{T} a_{x_t x_{t+1}} b_{x_t}(\mathbf{s}_t) a_{\tilde{x}_t \tilde{x}_{t+1}} b_{\tilde{x}_t}(\mathbf{d}_t). \tag{A.5}$$

Here $\int \int_{\mathbf{C}}$ is taken over the contour $\mathbf{C}_t$ defined by the relationship $f(\mathbf{s}_t, \mathbf{d}_t) = \mathbf{y}_t$ (Equation A.3).

A.2 Auxiliary Function

The method of maximum likelihood parameter estimation chooses unknown parameters to maximise the likelihood of the observed data. Thus in this case, Equation A.4 must be maximised with respect to the unknown parameters. The parameters of interest in this case are the noise parameters in Equation A.2. These are the noise transition probabilities and the noise autoregressive parameters for each noise state.

No closed form solution exists to find the maximum likelihood estimates of these parameters from Equation A.4. However, the Expectation-Maximisation algorithm can be used to iteratively find a solution. This algorithm is based on the use of an auxiliary Q function introduced in (Baum et al. 1970). There it is shown that maximising the auxiliary or Q function with respect to an unknown parameter produces a new model λ' such that $P(\mathbf{Y}|\lambda') \geq P(\mathbf{Y}|\lambda)$. The Q function is defined for this problem as

$$\begin{aligned} Q(\lambda, \lambda') &= E\{\log P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}}|\lambda')\} \qquad \text{(A.6)} \\ &= \sum_{\mathbf{X}} \sum_{\tilde{\mathbf{X}}} \int \int_{\mathbf{C}} P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}}|\lambda) \log P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}}|\lambda') d\mathbf{S} d\mathbf{D}. \end{aligned}$$

Substituting from Equation A.5 leads to

$$
\begin{aligned}
&Q(\lambda, \lambda') \\
&= \sum_{\mathbf{X}} \sum_{\tilde{\mathbf{X}}} \int \int_{\mathbf{C}} P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}} | \lambda) \\
&\quad \cdot \log \left[a'_{x_0 x_1} a'_{\tilde{x}_0 \tilde{x}_1} \prod_{t=1}^{T} a'_{x_t x_{t+1}} b'_{x_t}(\mathbf{s}_t) a'_{\tilde{x}_t \tilde{x}_{t+1}} b'_{\tilde{x}_t}(\mathbf{d}_t) \right] d\mathbf{S} d\mathbf{D} \\
&= \sum_{\mathbf{X}} \sum_{\tilde{\mathbf{X}}} \int \int_{\mathbf{C}} P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}} | \lambda) \\
&\quad \cdot \sum_{\forall x_\tau} \sum_{\forall x_{\tau+1}} \sum_{\forall \tilde{x}_\tau} \sum_{\forall \tilde{x}_{\tau+1}} n_0(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}, \mathbf{X}, \tilde{\mathbf{X}}) \\
&\quad \cdot \log \left[a'_{x_\tau x_{\tau+1}} a'_{\tilde{x}_\tau \tilde{x}_{\tau+1}} \right] d\mathbf{S} d\mathbf{D} \\
&\quad + \sum_{t=1}^{T} \sum_{\mathbf{X}} \sum_{\tilde{\mathbf{X}}} \int \int_{\mathbf{C}} P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}} | \lambda) \\
&\quad \cdot \sum_{\forall x_\tau} \sum_{\forall x_{\tau+1}} \sum_{\forall \tilde{x}_\tau} \sum_{\forall \tilde{x}_{\tau+1}} n_t(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}, \mathbf{X}, \tilde{\mathbf{X}}) \\
&\quad \cdot \log \left[a'_{x_\tau x_{\tau+1}} b'_{x_\tau}(\mathbf{s}_t) a'_{\tilde{x}_\tau \tilde{x}_{\tau+1}} b'_{\tilde{x}_\tau}(\mathbf{d}_t) \right] d\mathbf{S} d\mathbf{D} \qquad \text{(A.7)}
\end{aligned}
$$

where $n_t(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}, \mathbf{X}, \tilde{\mathbf{X}})$ is the counting function defined by

$$
n_t(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}) = \begin{cases} 1 & \text{if } x_t = x_\tau, x_{t+1} = x_{\tau+1}, \tilde{x}_t = \tilde{x}_\tau, \tilde{x}_{t+1} = \tilde{x}_{\tau+1} \\ 0 & \text{otherwise} \end{cases} . \qquad \text{(A.8)}
$$

By defining $\xi_t(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1})$ as

$$
\xi_t(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}) = \sum_{\mathbf{X}} \sum_{\tilde{\mathbf{X}}} n_t(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}, \mathbf{X}, \tilde{\mathbf{X}}) P(\mathbf{S}, \mathbf{D}, \mathbf{X}, \tilde{\mathbf{X}} | \lambda) \qquad \text{(A.9)}
$$

Equation A.7 can be further simplified to

$$
\begin{aligned}
Q(\lambda, \lambda') &= \sum_{\forall x_\tau} \sum_{\forall x_{\tau+1}} \sum_{\forall \tilde{x}_\tau} \sum_{\forall \tilde{x}_{\tau+1}} \int \int_{\mathbf{C}} \xi_0(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}) \\
&\quad \cdot \log \left[a'_{x_\tau x_{\tau+1}} a'_{\tilde{x}_\tau \tilde{x}_{\tau+1}} \right] d\mathbf{S} d\mathbf{D} \\
&\quad + \sum_{t=1}^{T} \sum_{\forall x_\tau} \sum_{\forall x_{\tau+1}} \sum_{\forall \tilde{x}_\tau} \sum_{\forall \tilde{x}_{\tau+1}} \int \int_{\mathbf{C}} \xi_t(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}) \\
&\quad \cdot \log \left[a'_{x_\tau x_{\tau+1}} b'_{x_\tau}(\mathbf{s}_t) a'_{\tilde{x}_\tau \tilde{x}_{\tau+1}} b'_{\tilde{x}_\tau}(\mathbf{d}_t) \right] d\mathbf{S} d\mathbf{D} \qquad \text{(A.10)}
\end{aligned}
$$

A.3 Maximum Likelihood Reestimation

Now consider maximisation of Equation A.10 with respect to $a'_{\tilde{x}_\tau \tilde{x}_{\tau+1}}$ and $b'_{\tilde{x}_\tau}(\mathbf{d}_t)$, the required noise model parameters.

The reestimation of $a_{\tilde{x}_\tau \tilde{x}_{\tau+1}}$ is fairly straightforward in that it follows almost exactly the derivation in (Rose et al. 1994). The only difference is that here a *transition* probability is reestimated whereas in (Rose et al. 1994) a *mixture weight* probability is reestimated. This does not affect the derivation a great deal. The reestimation formula is

$$a'_{\tilde{x}_\tau \tilde{x}_{\tau+1}} = \frac{\sum_{t=0}^{T} p(\tilde{x}_t = \tilde{x}_\tau, \tilde{x}_{t+1} = \tilde{x}_{\tau+1}, \mathbf{Y}|\lambda)}{\sum_{t=0}^{T} p(\tilde{x}_t = \tilde{x}_\tau, \mathbf{Y}|\lambda)}. \tag{A.11}$$

Here $p(\tilde{x}_t = \tilde{x}_\tau, \tilde{x}_{t+1} = \tilde{x}_{\tau+1}, \mathbf{Y}|\lambda)$ is the joint likelihood of state x_τ at time t and state $x_{\tau+1}$ at time $t+1$ and the observation sequence $\mathbf{Y}$. $p(\tilde{x}_t = \tilde{x}_\tau, \mathbf{Y}|\lambda)$ is the joint likelihood of state x_τ at time t and the observation sequence $\mathbf{Y}$.

The reestimation of the state-dependent noise statistics for each state is less straightforward. Let $Q_{b_{\tilde{x}_\tau}}$ be the terms of Q which depend on the $b'_{\tilde{x}_\tau}(\mathbf{d}_t)$. Thus

$$Q_{b_{\tilde{x}_\tau}}(\lambda, \lambda') = \sum_{t=1}^{T} \sum_{\forall x_\tau} \sum_{\forall x_{\tau+1}} \sum_{\forall \tilde{x}_\tau} \sum_{\forall \tilde{x}_{\tau+1}} \int \int_{\mathbf{C}} \xi_t(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}) \log b'_{\tilde{x}_\tau}(\mathbf{d}_t) d\mathbf{S} d\mathbf{D}. \tag{A.12}$$

By defining $\gamma_t(x_\tau, \tilde{x}_\tau)$ as

$$\gamma_t(x_\tau, \tilde{x}_\tau) = \sum_{\forall x_{\tau+1}} \sum_{\forall \tilde{x}_{\tau+1}} \xi_t(x_\tau, x_{\tau+1}, \tilde{x}_\tau, \tilde{x}_{\tau+1}) \tag{A.13}$$

and dropping the dependence on τ since it is no longer necessary to make this explicit, Equation A.12 can be rewritten

$$Q_{b_{\tilde{x}}}(\lambda, \lambda') = \sum_{t=1}^{T} \sum_{\forall x} \sum_{\forall \tilde{x}} \int \int_{\mathbf{C}} \gamma_t(x, \tilde{x}) \log b'_{\tilde{x}}(\mathbf{d}_t) d\mathbf{S} d\mathbf{D}. \tag{A.14}$$

Assuming that $b'_{\tilde{x}}$ describes an autoregressive Gaussian process,

$$b'_{\tilde{x}}(\mathbf{d}_t) = (2\pi)^{-K/2} (\sigma^{2'}_{\tilde{x}})^{-K/2} \exp\left\{ -\frac{1}{2} \alpha(\sigma^{-1'}_{\tilde{x}} \mathbf{d}_t; \mathbf{A}'_{\tilde{x}}) \right\} \tag{A.15}$$

where $\sigma^{2'}_{\tilde{x}}$ and $\mathbf{A}'_{\tilde{x}}$ are the variance and autoregressive filter parameters of noise state $\tilde{x}$ and K is the frame size (Juang 1984). Here $\alpha(\sigma^{-1'}_{\tilde{x}} \mathbf{d}_t; \mathbf{A}'_{\tilde{x}})$ is given by

$$\alpha(\sigma^{-1'}_{\tilde{x}} \mathbf{d}_t; \mathbf{A}'_{\tilde{x}}) = r'_{A_{\tilde{x}}}(0) \frac{r_{\mathbf{d_t}}(0)}{\sigma^{2'}_{\tilde{x}}} + 2 \sum_{i=1}^{P} r'_{A_{\tilde{x}}}(i) \frac{r_{\mathbf{d_t}}(i)}{\sigma^{2'}_{\tilde{x}}} \tag{A.16}$$

where $r_{\mathbf{d_t}}(i)$ and $r'_{A_{\tilde{x}}}(i)$ are the autocorrelation functions of the noise and the noise autoregressive filter parameters respectively and P is the order of the autoregressive process. Substituting Equation A.15 into Equation A.14 yields

$$Q_{b_{\tilde{x}}}(\lambda, \lambda') = \tag{A.17}$$
$$\sum_{t=1}^{T}\sum_{\forall x}\sum_{\forall \tilde{x}}\int\int_{\mathbf{C}} \gamma_t(x,\tilde{x})\left[-\frac{K}{2}\log(2\pi) - \frac{K}{2}\log(\sigma_{\tilde{x}}^{2\prime}) - \frac{1}{2}\alpha(\sigma_{\tilde{x}}^{-1\prime}\mathbf{d}_t; \mathbf{A}'_{\tilde{x}})\right] d\mathbf{S}d\mathbf{D}.$$

Now, consider maximisation of $Q_{b_{\tilde{x}}}(\lambda, \lambda')$ with respect to a particular set of autoregressive parameters $\mathbf{A}'$ for a particular state $\tilde{x}$. The required parameters will minimise the function

$$F_{\tilde{x}}(\lambda, \lambda') = \sum_{t=1}^{T}\sum_{\forall x}\int\int_{\mathbf{C}} \gamma_t(x,\tilde{x})\alpha(\sigma_{\tilde{x}}^{-1\prime}\mathbf{d}_t; \mathbf{A}'_{\tilde{x}})d\mathbf{S}d\mathbf{D} \tag{A.18}$$

or equivalently, by substituting from Equation A.16,

$$F_{\tilde{x}}(\lambda, \lambda') = \sum_{t=1}^{T}\sum_{\forall x}\int\int_{\mathbf{C}} \gamma_t(x,\tilde{x})\left[\frac{r'_{A_{\tilde{x}}}(0)}{\sigma_{\tilde{x}}^{2\prime}} r_{\mathbf{d_t}}(0) + 2\sum_{i=1}^{P}\frac{r'_{A_{\tilde{x}}}(i)}{\sigma_{\tilde{x}}^{2\prime}} r_{\mathbf{d_t}}(i)\right] d\mathbf{S}d\mathbf{D}. \tag{A.19}$$

Now consider each of the terms in the integral separately. In particular, consider $\int\int_{\mathbf{C}} \gamma_t(x,\tilde{x}) r_{\mathbf{d_t}}(i) d\mathbf{S}d\mathbf{D}$. From (Rose et al. 1994), it is seen that

$$\begin{aligned}
&\int\int_{\mathbf{C}} \gamma_t(x,\tilde{x}) r_{\mathbf{d_t}}(i) d\mathbf{S}d\mathbf{D} \\
&= \prod_{\tau\neq t}^{T} p(\mathbf{y}_\tau|\lambda)\int\int_{C} r_{\mathbf{d_t}}(i) p(\mathbf{s}_t, \mathbf{d}_t, x_t = x, \tilde{x}_t = \tilde{x}|\lambda) d\mathbf{s}_t d\mathbf{d}_t \\
&= \frac{p(\mathbf{Y}|\lambda)}{p(\mathbf{y}_t)}\int\int_{C} r_{\mathbf{d_t}}(i) p(\mathbf{s}_t, \mathbf{d}_t|x_t = x, \tilde{x}_t = \tilde{x}, \lambda) p(x_t = x, \tilde{x}_t = \tilde{x}|\lambda) d\mathbf{s}_t d\mathbf{d}_t \\
&= p(\mathbf{Y}|\lambda) p(x_t = x, \tilde{x}_t = \tilde{x}|\lambda) E\{r_{\mathbf{d_t}}(i)|\mathbf{y}_t, x_t = x, \tilde{x}_t = \tilde{x}, \lambda\} \\
&= p(\mathbf{Y}, x_t = x, \tilde{x}_t = \tilde{x}|\lambda) E\{r_{\mathbf{d_t}}(i)|\mathbf{y}_t, x_t = x, \tilde{x}_t = \tilde{x}, \lambda\}.
\end{aligned} \tag{A.20}$$

Equation A.19 is thus rewritten as

$$\begin{aligned}
&F_{\tilde{x}}(\lambda, \lambda') \\
&= \frac{r'_{A_{\tilde{x}}}(0)}{\sigma_{\tilde{x}}^{2\prime}}\left[\sum_{t=1}^{T}\sum_{\forall x} p(\mathbf{Y}, x_t = x, \tilde{x}_t = \tilde{x}|\lambda) E\{r_{\mathbf{d_t}}(0)|\mathbf{y}_t, x_t = x, \tilde{x}_t = \tilde{x}, \lambda\}\right] \\
&\quad +2\sum_{i=1}^{P}\frac{r'_{A_{\tilde{x}}}(i)}{\sigma_{\tilde{x}}^{2\prime}}\left[\sum_{t=1}^{T}\sum_{\forall x} p(\mathbf{Y}, x_t = x, \tilde{x}_t = \tilde{x}|\lambda) E\{r_{\mathbf{d_t}}(i)|\mathbf{y}_t, x_t = x, \tilde{x}_t = \tilde{x}, \lambda\}\right] \\
&= \frac{r'_{A_{\tilde{x}}}(0)}{\sigma_{\tilde{x}}^{2\prime}} r'_{\tilde{x}}(0) + 2\sum_{i=1}^{P}\frac{r'_{A_{\tilde{x}}}(i)}{\sigma_{\tilde{x}}^{2\prime}} r'_{\tilde{x}}(i)
\end{aligned} \tag{A.21}$$

where

$$r'_{\tilde{x}}(i) = \sum_{t=1}^{T} \sum_{\forall x} p(\mathbf{Y}, x_t = x, \tilde{x}_t = \tilde{x}|\lambda) E\{r_{\mathbf{d_t}}(i)|\mathbf{y}_t, x_t = x, \tilde{x}_t = \tilde{x}, \lambda\}. \tag{A.22}$$

Thus by using this reestimated autocorrelation function, reestimates of the noise autoregressive parameters for noise state $\tilde{x}$, $\mathbf{A}'_{\tilde{x}}$, can then be found in the usual way (e.g. the autocorrelation method in (Deller et al. 1993)). These will then minimise the desired function in Equation A.21.

To obtain $\sigma^{2\prime}_{\tilde{x}}$, consider maximisation of $H_{\tilde{x}}(\lambda, \lambda')$ with respect to $\sigma^{2\prime}_{\tilde{x}}$ where the optimal values $\mathbf{A}'_{\tilde{x}}$ and $r'_{\tilde{x}}$ have been substituted.

$$H_{\tilde{x}}(\lambda, \lambda') = \tag{A.23}$$
$$\sum_{t=1}^{T} \sum_{\forall x} \sum_{\forall \tilde{x}} \int \int_{\mathbf{C}} \gamma_t(x, \tilde{x}) \left[-\frac{K}{2} \log(\sigma^{2\prime}_{\tilde{x}}) - \frac{1}{2} \alpha(\sigma^{-1\prime}_{\tilde{x}} \mathbf{d}_t; \mathbf{A}'_{\tilde{x}}) \right] d\mathbf{S} d\mathbf{D}$$

Taking the derivative of Equation A.23 with respect to $\sigma^{2\prime}_{\tilde{x}}$ yields

$$\sigma^{2\prime}_{\tilde{x}} = \frac{r'_{A_{\tilde{x}}}(0) r'_{\tilde{x}}(0) + 2\sum_{i=1}^{P} r'_{A_{\tilde{x}}}(i) r'_{\tilde{x}}(i)}{K \sum_{t=1}^{T} \sum_{\forall x} p(\mathbf{Y}, x_t = x, \tilde{x}_t = \tilde{x}|\lambda)} \tag{A.24}$$

It should be noted that the same solution for $\mathbf{A}'_{\tilde{x}}$ and $\sigma^{2\prime}_{\tilde{x}}$ will be obtained even if $r'_{\tilde{x}}(i)$ is scaled by an arbitrary scaling factor. Therefore it is convenient to write

$$\mathbf{r}'_{\tilde{x}} = \frac{\sum_{t=1}^{T} \sum_{\forall x} p(\mathbf{Y}, x_t{=}x, \tilde{x}_t{=}\tilde{x}|\lambda) E\{\mathbf{r}_{\tilde{x}}|\mathbf{y}_t, x_t{=}x, \tilde{x}_t{=}\tilde{x}, \lambda\}}{\sum_{t=1}^{T} \sum_{\forall x} p(\mathbf{Y}, x_t{=}x, \tilde{x}_t{=}\tilde{x}|\mathbf{y}_t, \lambda)} \tag{A.25}$$

and thus

$$\sigma^{2\prime}_{\tilde{x}} = \frac{r'_{A_{\tilde{x}}}(0) r'_{\tilde{x}}(0) + 2\sum_{i=1}^{P} r'_{A_{\tilde{x}}}(i) r'_{\tilde{x}}(i)}{K}. \tag{A.26}$$

$p(\mathbf{Y}, x_t{=}x, \tilde{x}_t{=}\tilde{x}|\lambda)$ is the joint likelihood of states x_τ and $\tilde{x}_\tau$ at time t and the observation sequence $\mathbf{Y}$.

Appendix B

Digital Warping of Spectra Using the Bilinear Transform

This appendix summarises relevant parts of (Oppenheim & Johnson 1972) in which the representation of a continuous signal by a digital sequence is described. Of most interest is the result that several sequences can represent the same continuous signal and that a simple conversion exists between these two sequences. This conversion can be used to implement a non-linear warping on the digital frequency axis.

Consider first the digital representation $\{f_n\}$ of the continuous signal $f(t)$.

$$f(t) = \sum_{n=-\infty}^{+\infty} f_n \phi_n(t) \tag{B.1}$$

Here $\phi_n(t)$ is a continuous-time function. It can be shown that for continuous convolution to be mapped to discrete convolution, $\phi_n(t)$ must satisfy

$$\Phi_n(s) = [\Phi_1(s)]^n \tag{B.2}$$

where $\Phi_n(s)$ is the Laplace transform $\{\phi_n(t)\}$.

A common example of a choice of $\{f_n\}$ and $\phi_n(t)$ is periodic sampling. Choosing

$$f_n = Tf(nT) \tag{B.3}$$

where T is the sampling period and

$$\phi_n(t) = \sin \frac{\pi}{T}(t - nT)/\pi(t - nT) \tag{B.4}$$

satisfies Equations B.1 and B.2. Therefore $f(t)$ is faithfully represented by $\{f_n\}$ while the property of convolution is preserved.

Other choices of $\{f_n\}$ and $\phi_n(t)$ are possible however. Consider an alternative representation

$$f(t) = \sum_{k=-\infty}^{+\infty} g_k \lambda_k(t) \tag{B.5}$$

where again

$$\Lambda_k(s) = [\Lambda_1(s)]^k. \tag{B.6}$$

It is possible to map between the sequences $\{f_n\}$ and $\{g_k\}$ as follows.

Assuming that $\{\phi_n(t)\}$ is complete, each $\lambda_k(t)$ can be expanded in terms of $\{\phi_n(t)\}$. Thus

$$\lambda_k(t) = \sum_{n=-\infty}^{+\infty} \psi_{k,n}\phi_n(t). \tag{B.7}$$

Hence the sequences $\{f_n\}$ and $\{g_k\}$ are related by

$$f_n = \sum_{k=-\infty}^{+\infty} g_k\psi_{k,n}. \tag{B.8}$$

Again it can be shown that for the mapping from $f(t)$ to f_n and from $f(t)$ to g_k to preserve convolution, $\{\Psi_k(z)\}$ must satisfy the relation

$$\Psi_k(z) = [\Psi_1(z)]^k \tag{B.9}$$

where $\{\Psi_k(z)\}$ is the z-transform of $\psi_{k,n}$.

Given this relation, the z-transform of f_n can be written as

$$\begin{aligned}
F(z) &= \sum_{n=-\infty}^{+\infty} f_n z^{-n} && \text{(B.10)}\\
&= \sum_{n=-\infty}^{+\infty}\sum_{k=-\infty}^{+\infty} g_k\psi_{k,n}z^{-n} && \text{(B.11)}\\
&= \sum_{k=-\infty}^{+\infty} g_k \sum_{n=-\infty}^{+\infty} \psi_{k,n}z^{-n} && \text{(B.12)}\\
&= \sum_{k=-\infty}^{+\infty} g_k\Psi_k(z) && \text{(B.13)}\\
&= \sum_{k=-\infty}^{+\infty} g_k[\Psi_1(z)]^k && \text{(B.14)}\\
&\triangleq G(\hat{z}) && \text{(B.15)}\\
& && \text{(B.16)}
\end{aligned}$$

where

$$\hat{z} = [\Psi_1(z)]^{-1} \triangleq m(z). \tag{B.17}$$

Thus by a suitable choice of $m(z)$ it is possible to transform the sequence $\{f_n\}$ to a new sequence $\{g_k\}$ which represents the same continuous time function and preserves convolution. The z-transforms of these sequences will be related by

$$F(z) = G[m(z)] = G(\hat{z}). \tag{B.18}$$

Therefore, to obtain the sequence $\{g_k\}$ with Fourier transform equal to the Fourier transform of $\{f_n\}$ but on a warped frequency scale, $m(z)$ is chosen to satisfy

$$e^{j\hat{\Omega}} = m[e^{j\Omega}] \tag{B.19}$$

so that the unit circle in the z plane is mapped to the unit circle in the $\hat{z}$ plane and the angular frequency is mapped according to

$$\hat{\Omega} = \theta(\Omega). \tag{B.20}$$

(Taking the Fourier transform of a sequence is equivalent to evaluating the z-transform at $e^{j\Omega}$ i.e. on the unit circle.)

A useful choice for $m(z)$ is

$$\hat{z} = m(z) = \frac{1 - az^{-1}}{z^{-1} - a}. \tag{B.21}$$

In this case

$$\hat{\Omega} = \theta(\Omega) = \arctan\left[\frac{(1-a^2)\sin\Omega}{(1+a^2)\cos\Omega - 2a}\right]. \tag{B.22}$$

The parameter a determines the amount of warping.

Appendix C

Analysis of Noise Sources

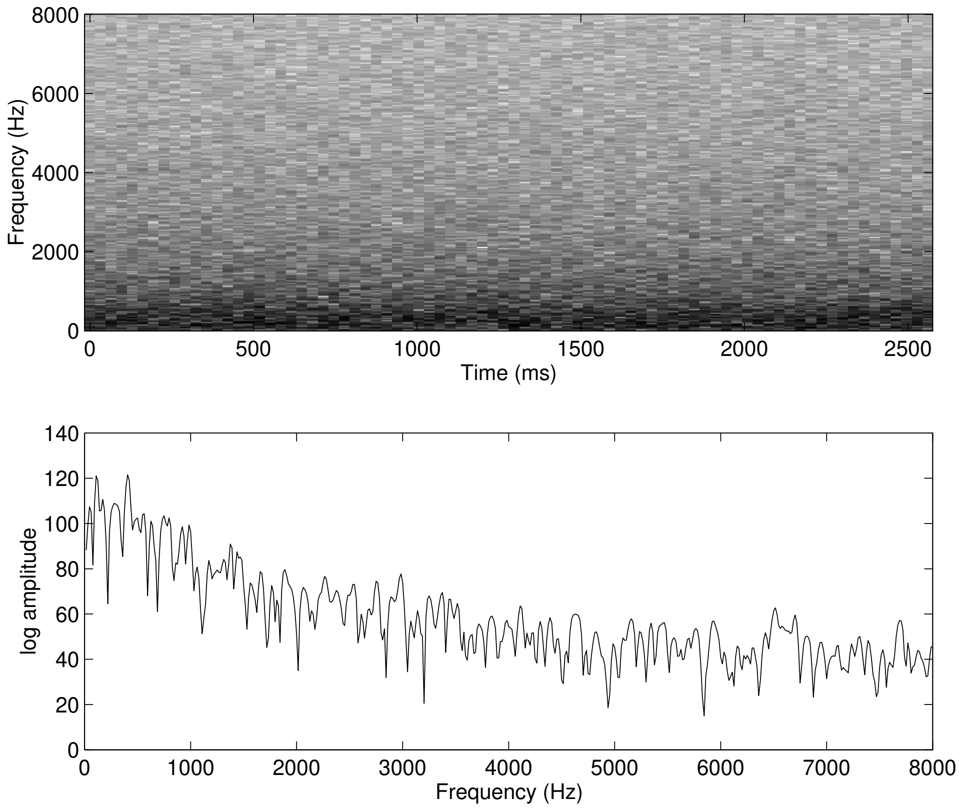

Figure C.1: Spectrogram and Typical Spectrum of Speech Noise

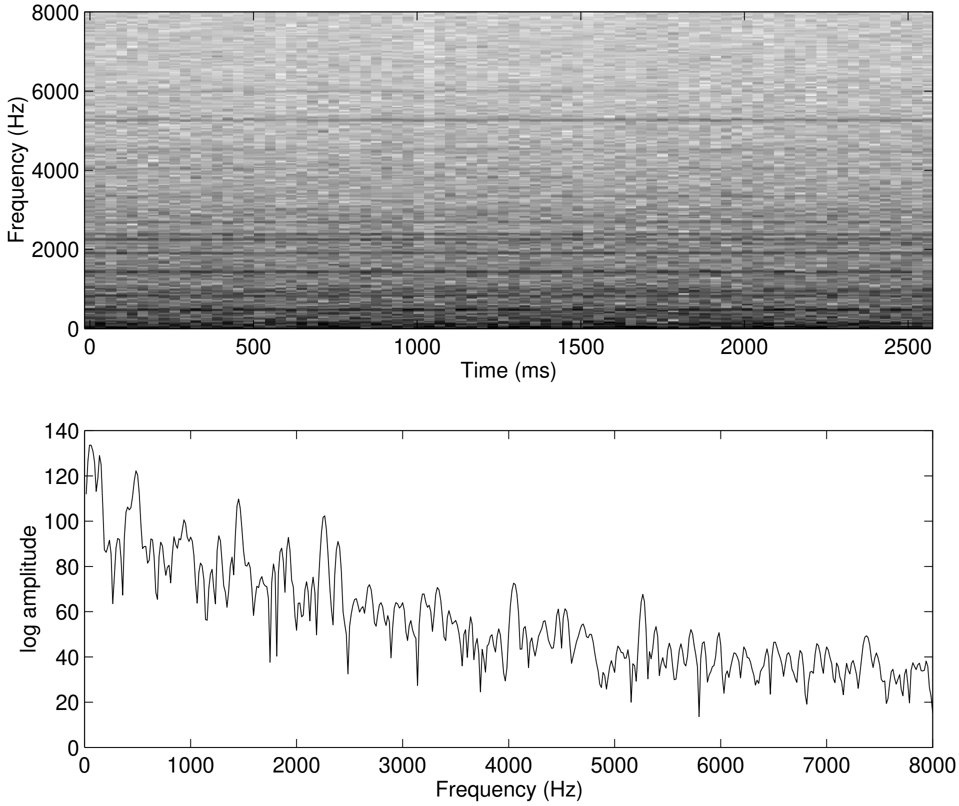

Figure C.2: Spectrogram and Typical Spectrum of Lynx Noise

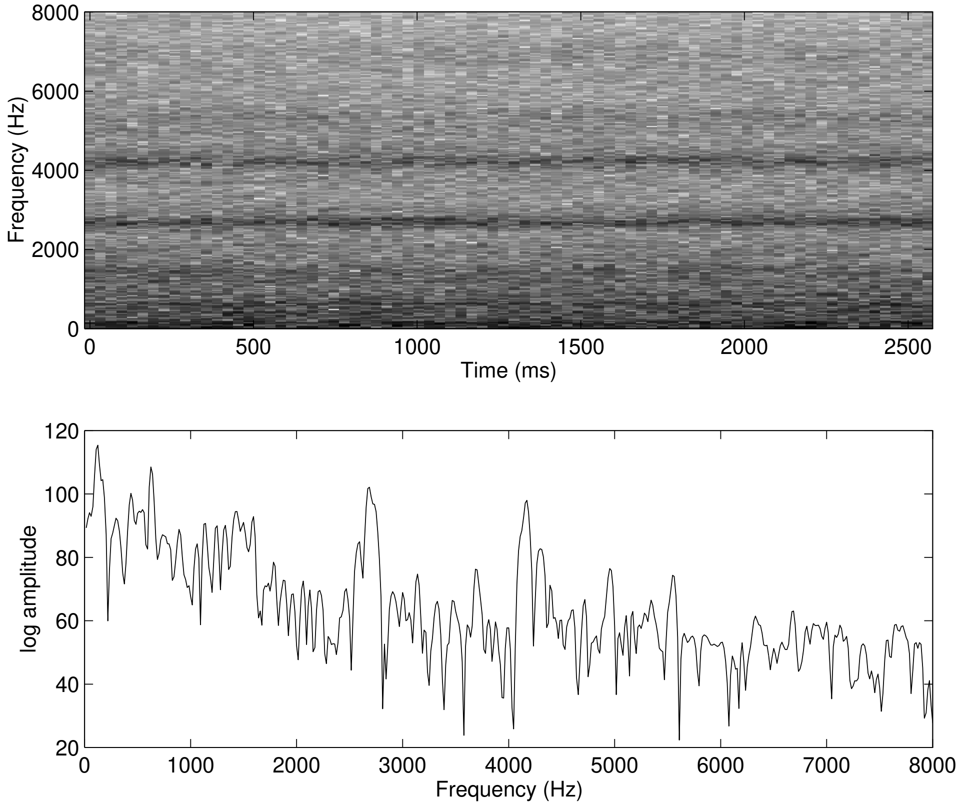

Figure C.3: Spectrogram and Typical Spectrum of F16 Noise

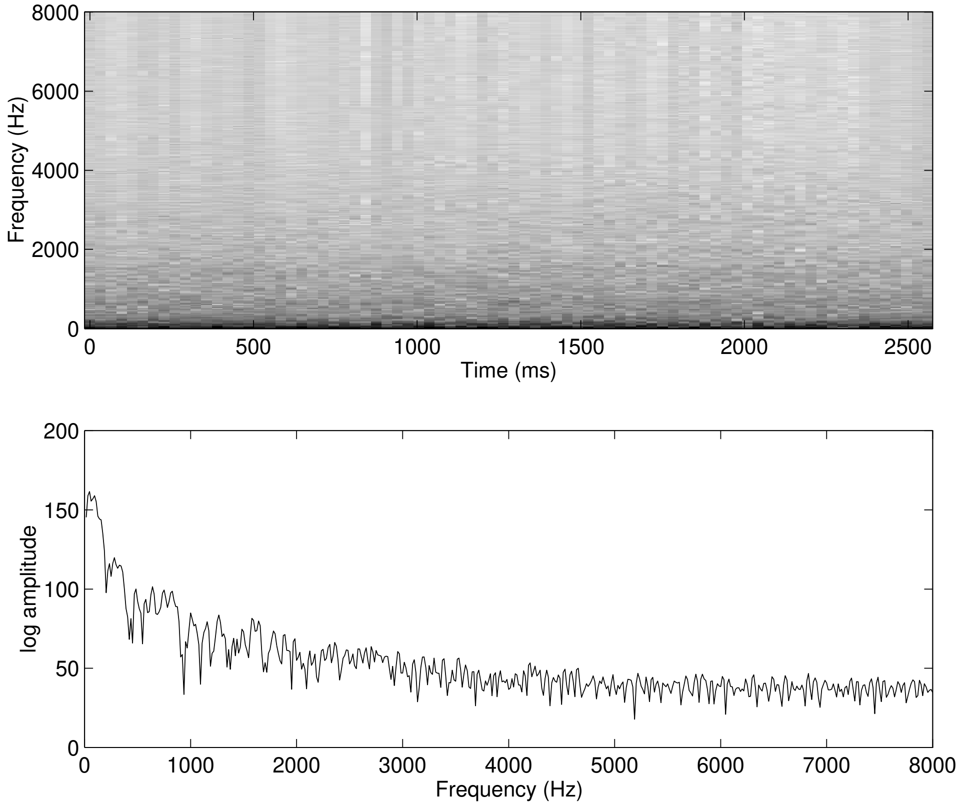

Figure C.4: Spectrogram and Typical Spectrum of Car Noise

Appendix D

Audio Compact Disc

A Compact Disc (CD) is provided with this dissertation such that the reader may assess the various algorithms studied. The CD contains samples from both the NOISEX-92 and RM databases. A detailed list of contents is given in Tables D.1, D.2 and D.3. In these tables, the enhancement scheme estimating noise from silences is denoted 'sil' and the maximum likelihood scheme is denoted 'ML'.

The NOISEX-92 examples are based on the first 20 digits in the digits test set. The correct transcription for this utterance is:

ONE, SIX, THREE, FIVE, TWO, FOUR, EIGHT, NINE, SEVEN, ZERO, SIX, THREE, SEVEN, FOUR, EIGHT, FIVE, NINE, ZERO, ONE, TWO.

The RM examples are for speaker 'alk0_3' and 'meb0_3' showing performance on a female and male speaker respectively. The correct transcriptions of the utterances are:

alk0_3	WHEN WILL THE PERSONNEL CASUALTY REPORT FROM THE YORKTOWN BE RESOLVED
meb0_3	WHO HAS TASM CAPABILITY IN BISMARK SEA.

Track	Noise	SNR dB	Algorithm	Models		Estimator
				Type	State/Mix	
1	-	∞	-	-	-	-
2	Lynx	0	-	-	-	-
3	Lynx	12	-	-	-	-
4	Lynx	0	sil	word	-	Wiener
5	Lynx	12	sil	word	-	Wiener
6	Lynx	0	sil	word	-	MMSE PSD
7	Lynx	12	sil	word	-	MMSE PSD
8	Lynx	0	ML	word	-	Wiener
9	Lynx	12	ML	word	-	Wiener
10	Lynx	0	ML	word	-	MMSE PSD
11	Lynx	12	ML	word	-	MMSE PSD
12	speech	0	-	-	-	-
13	speech	0	ML	word	-	Wiener
14	speech	12	-	-	-	-
15	speech	12	ML	word	-	Wiener
16	F16	0	-	-	-	-
17	F16	0	ML	word	-	Wiener
18	F16	12	-	-	-	-
19	F16	12	ML	word	-	Wiener
20	car	0	-	-	-	-
21	car	0	ML	word	-	Wiener
22	car	12	-	-	-	-
23	car	12	ML	word	-	Wiener
24	Lynx	0	-	-	-	-
25	Lynx	0	ML	general	1/128	Wiener
26	Lynx	0	ML	general	1/256	Wiener
27	Lynx	0	ML	general	32/4	Wiener
28	Lynx	12	-	-	-	-
29	Lynx	12	ML	general	1/128	Wiener
30	Lynx	12	ML	general	1/256	Wiener
31	Lynx	12	ML	general	32/4	Wiener

Table D.1: Contents of Part 1 of the audio Compact Disk; Male Digits

Track	Noise	SNR dB	Algorithm	Models		Estimator
				Type	State/Mix	
32	-	∞	-	-	-	-
33	Lynx	0	-	-	-	-
34	Lynx	0	ML	word	-	Wiener
35	Lynx	12	-	-	-	-
36	Lynx	12	ML	word	-	Wiener

Table D.2: Contents of Part 2 of the audio Compact Disk; Female Digits

Track	Speaker	Noise	SNR dB	Algorithm	Models		Estimator
					Type	State/Mix	
37	aem0_3	-	∞	-	-	-	-
38	aem0_3	Lynx	12	-	-	-	-
39	aem0_3	Lynx	12	ML	general	1/512	Wiener
40	aem0_3	Lynx	18	ML	-	-	-
41	aem0_3	Lynx	18	ML	general	1/512	Wiener
42	meb0_3	-	∞	-	-	-	-
43	meb0_3	Lynx	12	-	-	-	-
44	meb0_3	Lynx	12	ML	general	1/512	Wiener
45	meb0_3	Lynx	18	ML	-	-	-
46	meb0_3	Lynx	18	ML	general	1/512	Wiener

Table D.3: Contents of Part 3 of the audio Compact Disk; RM Examples

Appendix E

Full Results of Evaluation on the NOISEX-92 Database

Noise Source	SNR dB	Distortion		% Error (D,S,I)			
		Overall	Speech	MFCC Models		AR Models	
Speech	-6	1.05	1.08	95	(95,0,0)	95	(95,0,0)
	0	0.98	0.91	95	(95,0,0)	95	(95,0,0)
	6	0.87	0.69	95	(95,0,0)	95	(95,0,0)
	12	0.75	0.47	69	(18,31,20)	75	(39,28,8)
	18	0.60	0.30	49	(24,14,11)	59	(7,46,6)
Lynx	-6	1.46	1.24	95	(95,0,0)	95	(95,0,0)
	0	1.38	1.08	95	(95,0,0)	95	(95,0,0)
	6	1.27	0.87	95	(95,0,0)	95	(95,0,0)
	12	1.09	0.64	79	(43,32,4)	78	(59,17,2)
	18	0.83	0.40	64	(10,25,29)	57	(3,48,6)
F16	-6	1.19	1.52	95	(95,0,0)	95	(95,0,0)
	0	1.09	1.29	95	(95,0,0)	95	(95,0,0)
	6	0.95	0.99	95	(95,0,0)	95	(95,0,0)
	12	0.80	0.68	83	(15,51,17)	92	(20,50,13)
	18	0.64	0.42	65	(11,38,16)	67	(6,59,2)
Car	-6	1.02	0.88	95	(95,0,0)	95	(95,0,0)
	0	0.95	0.73	95	(95,0,0)	93	(93,0,0)
	6	0.86	0.56	82	(61,20,1)	90	(89,1,0)
	12	0.73	0.40	77	(12,46,19)	77	(9,65,3)
	18	0.55	0.25	40	(12,20,8)	80	(11,61,8)
∞	0	0	0	0	(0,0,0)	0	(0,0,0)

Table E.1: Distortions and word error rates for speech corrupted by various noises recognised using clean MFCC and AR models

Noise Source	SNR dB	% Error (D,S,I) MFCC Models		AR Models		Warped AR Models	
Speech	-6	22	(1,19,2)	24	(6,16,2)	16	(5,13,1)
	0	0	(0,0,0)	10	(1,9,0)	0	(0,0,0)
	6	0	(0,0,0)	6	(0,6,0)	0	(0,0,0)
	12	0	(0,0,0)	1	(0,1,0)	0	(0,0,0)
	18	0	(0,0,0)	1	(0,1,0)	0	(0,0,0)
Lynx	-6	42	(17,21,4)	45	(11,28,6)	27	(1,24,2)
	0	6	(0,6,0)	16	(1,15,0)	2	(0,2,0)
	6	0	(0,0,0)	7	(0,7,0)	0	(0,0,0)
	12	0	(0,0,0)	4	(0,4,0)	0	(0,0,0)
	18	0	(0,0,0)	1	(0,1,0)	0	(0,0,0)
F16	-6	48	(16,29,3)	70	(18,35,17)	59	(5,49,5)
	0	4	(0,4,0)	34	(3,22,9)	21	(3,16,2)
	6	1	(0,1,0)	9	(0,9,0)	3	(0,3,0)
	12	0	(0,0,0)	4	(0,4,0)	0	(0,0,0)
	18	0	(0,0,0)	1	(0,1,0)	0	(0,0,0)
Car	-6	18	(3,12,3)	52	(8,43,1)	30	(6,22,2)
	0	0	(0,0,0)	36	(1,34,1)	5	(0,5,0)
	6	0	(0,0,0)	17	(0,17,0)	0	(0,0,0)
	12	0	(0,0,0)	6	(0,6,0)	0	(0,0,0)
	18	0	(0,0,0)	2	(0,2,0)	0	(0,0,0)

Table E.2: Word error rates for corrupted speech recognised using matched MFCC and AR models

Noise Source	SNR dB	% Error (D,S,I)			
		AR Models		Warped AR Models	
Speech	-6	20	(1,15,4)	12	(1,9,2)
	0	9	(1,8,0)	0	(0,0,0)
	6	5	(0,5,0)	0	(0,0,0)
	12	1	(0,1,0)	0	(0,0,0)
	18	1	(0,1,0)	0	(0,0,0)
Lynx	-6	33	(4,24,5)	21	(3,15,3)
	0	13	(1,12,0)	2	(0,2,0)
	6	7	(0,7,0)	0	(0,0,0)
	12	3	(0,3,0)	0	(0,0,0)
	18	1	(0,1,0)	0	(0,0,0)
F16	-6	34	(8,18,8)	37	(7,25,5)
	0	21	(5,14,2)	14	(2,12,0)
	6	12	(0,12,0)	3	(0,3,0)
	12	4	(0,4,0)	0	(0,0,0)
	18	0	(0,0,0)	0	(0,0,0)
Car	-6	68	(17,43,8)	13	(3,8,2)
	0	47	(6,33,8)	3	(0,3,0)
	6	20	(3,17,0)	0	(0,0,0)
	12	5	(0,5,0)	0	(0,0,0)
	18	2	(0,2,0)	0	(0,0,0)

Table E.3: Word error rates for corrupted speech recognised using compensated AR models

Noise Source	SNR dB	Distortion All	Distortion Speech	% Error (D,S,I) MFCC Recoded		% Error (D,S,I) AR Compensated	
Speech	-6	1.56	1.21	74	(13,36,25)	71	(6,37,28)
	0	1.26	0.99	57	(35,21,1)	56	(37,18,1)
	6	0.76	0.59	27	(7,13,7)	4	(0,4,0)
	12	0.46	0.36	16	(6,6,4)	0	(0,0,0)
	18	0.41	0.31	28	(19,6,3)	0	(0,0,0)
Lynx	-6	1.36	1.08	77	(6,55,16)	79	(5,33,41)
	0	1.24	0.99	63	(33,26,4)	51	(31,19,1)
	6	0.93	0.68	37	(13,15,9)	11	(0,11,0)
	12	0.74	0.60	33	(13,17,3)	5	(0,4,1)
	18	0.58	0.46	48	(29,16,3)	0	(0,0,0)
F16	-6	1.58	1.36	81	(47,28,6)	76	(37,29,10)
	0	1.65	1.35	75	(49,21,5)	63	(41,22,0)
	6	1.22	0.86	49	(6,36,7)	35	(1,30,4)
	12	0.66	0.46	21	(1,12,8)	13	(0,12,1)
	18	0.55	0.34	34	(16,8,10)	7	(0,5,2)
Car	-6	1.50	1.26	75	(56,18,2)	71	(55,10,6)
	0	1.19	0.89	58	(33,21,4)	35	(21,14,0)
	6	0.84	0.58	44	(18,20,6)	20	(1,15,4)
	12	0.49	0.33	17	(3,9,5)	5	(0,4,1)
	18	0.52	0.42	46	(38,7,1)	0	(0,0,0)

Table E.4: Distortions and word error rates for corrupted speech enhanced adaptively using recognised silences to estimate the noise; Wiener filters and word-based HMMs

		Distortion		% Error (D,S,I)			
Noise Source	SNR dB	All	Speech	MFCC Recoded		AR Compensated	
Speech	-6	1.13	0.87	71	(6,37,28)	71	(6,37,28)
	0	0.94	0.76	56	(37,18,1)	56	(37,18,1)
	6	0.45	0.32	4	(0,4,0)	4	(0,4,0)
	12	0.24	0.18	0	(0,0,0)	0	(0,0,0)
	18	0.14	0.12	0	(0,0,0)	0	(0,0,0)
Lynx	-6	1.17	0.88	76	(2,33,41)	79	(5,33,41)
	0	0.92	0.75	51	(31,19,1)	51	(31,19,1)
	6	0.57	0.36	11	(0,11,0)	11	(0,11,0)
	12	0.35	0.23	5	(0,4,1)	5	(0,4,1)
	18	0.21	0.14	0	(0,0,0)	0	(0,0,0)
F16	-6	1.23	1.08	76	(37,29,10)	76	(37,29,10)
	0	1.17	0.96	63	(41,22,0)	63	(41,22,0)
	6	0.78	0.54	35	(1,30,4)	35	(1,30,4)
	12	0.41	0.29	13	(0,12,1)	13	(0,12,1)
	18	0.28	0.18	7	(0,5,2)	7	(0,5,2)
Car	-6	1.27	1.18	71	(55,10,6)	71	(55,10,6)
	0	0.86	0.62	35	(23,12,0)	35	(21,14,0)
	6	0.64	0.36	20	(1,15,4)	20	(1,15,4)
	12	0.37	0.20	6	(0,5,1)	5	(0,4,1)
	18	0.24	0.14	0	(0,0,0)	0	(0,0,0)

Table E.5: Distortions and word error rates for corrupted speech enhanced adaptively using recognised silences to estimate the noise; MMSE PSD estimation and word-based HMMs

Noise Source	SNR dB	Distortion All	Distortion Speech	% Error (D,S,I) MFCC Recoded		% Error (D,S,I) AR Compensated	
Speech	-6	0.42	0.48	10	(4,6,0)	11	(4,7,0)
	0	0.30	0.34	1	(1,0,0)	1	(1,0,0)
	6	0.19	0.23	0	(0,0,0)	0	(0,0,0)
	12	0.15	0.16	0	(0,0,0)	0	(0,0,0)
	18	0.12	0.12	0	(0,0,0)	0	(0,0,0)
Lynx	-6	0.74	0.72	26	(2,13,11)	26	(1,14,11)
	0	0.42	0.43	4	(0,4,0)	4	(0,4,0)
	6	0.25	0.27	0	(0,0,0)	0	(0,0,0)
	12	0.17	0.17	0	(0,0,0)	0	(0,0,0)
	18	0.12	0.11	0	(0,0,0)	0	(0,0,0)
F16	-6	0.66	0.66	36	(12,20,4)	36	(12,20,4)
	0	0.43	0.44	10	(0,6,4)	10	(0,6,4)
	6	0.27	0.28	1	(0,1,0)	1	(0,1,0)
	12	0.19	0.20	0	(0,0,0)	0	(0,0,0)
	18	0.14	0.14	0	(0,0,0)	0	(0,0,0)
Car	-6	0.70	0.87	34	(12,14,8)	34	(10,16,8)
	0	0.42	0.46	10	(0,4,6)	10	(0,4,6)
	6	0.31	0.28	1	(0,1,0)	2	(0,1,1)
	12	0.21	0.19	0	(0,0,0)	0	(0,0,0)
	18	0.14	0.12	0	(0,0,0)	0	(0,0,0)

Table E.6: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimates; MMSE PSD estimation and word-based HMMs

Noise Source	SNR dB	Distortion All	Distortion Speech	% Error (D,S,I) MFCC Recoded		% Error (D,S,I) AR Compensated	
Speech	-6	0.50	0.53	16	(5,11,1)	15	(5,10,0)
	0	0.33	0.35	2	(1,1,0)	1	(1,0,0)
	6	0.19	0.25	0	(0,0,0)	0	(0,0,0)
	12	0.15	0.17	0	(0,0,0)	0	(0,0,0)
	18	0.12	0.13	0	(0,0,0)	0	(0,0,0)
Lynx	-6	0.63	0.72	29	(4,19,6)	26	(2,18,6)
	0	0.39	0.44	4	(1,3,0)	3	(0,3,0)
	6	0.25	0.28	0	(0,0,0)	0	(0,0,0)
	12	0.17	0.17	0	(0,0,0)	0	(0,0,0)
	18	0.12	0.11	0	(0,0,0)	0	(0,0,0)
F16	-6	0.84	0.78	34	(12,17,5)	34	(12,17,5)
	0	0.54	0.51	11	(2,7,2)	11	(2,7,2)
	6	0.34	0.33	7	(0,4,3)	5	(0,2,3)
	12	0.20	0.22	1	(0,1,0)	0	(0,0,0)
	18	0.15	0.15	0	(0,0,0)	0	(0,0,0)
Car	-6	0.72	0.82	34	(9,21,4)	32	(9,17,6)
	0	0.43	0.47	13	(1,7,5)	11	(0,5,6)
	6	0.33	0.30	1	(0,1,0)	1	(0,1,0)
	12	0.21	0.21	0	(0,0,0)	0	(0,0,0)
	18	0.15	0.13	0	(0,0,0)	0	(0,0,0)

Table E.7: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimates; MMSE PSD estimation and word-based HMMs; Autoregressive order 15

		Distortion		% Error (D,S,I)	
Noise Source	SNR dB	All	Speech	MFCC Recoded	
Speech	-6	0.42	0.74	15	(1,10,4)
	0	0.39	0.63	15	(11,4,0)
	6	0.23	0.33	0	(0,0,0)
	12	0.19	0.25	0	(0,0,0)
	18	0.21	0.17	0	(0,0,0)
Lynx	-6	0.57	0.78	38	(2,14,22)
	0	0.64	0.94	48	(36,12,0)
	6	0.27	0.37	0	(0,0,0)
	12	0.21	0.27	0	(0,0,0)
	18	0.17	0.22	0	(0,0,0)
F16	-6	0.59	0.90	52	(17,21,14)
	0	0.50	0.70	24	(15,9,0)
	6	0.30	0.40	2	(1,1,0)
	12	0.22	0.30	0	(0,0,0)
	18	0.19	0.24	0	(0,0,0)
Car	-6	0.85	1.15	70	(46,13,11)
	0	0.46	0.62	14	(7,7,0)
	6	0.31	0.38	0	(0,0,0)
	12	0.26	0.32	0	(0,0,0)
	18	0.20	0.25	0	(0,0,0)

Table E.8: Distortions and word error rates for corrupted speech enhanced adaptively using linear spectral models

Noise Source	SNR dB	Distortion All	Distortion Speech	% Error (D,S,I) MFCC Recoded	
Speech	-6	1.10	0.83	91	(87,4,0)
	0	0.78	0.58	86	(84,2,0)
	6	0.44	0.35	54	(25,25,4)
	12	0.21	0.19	12	(3,8,1)
	18	0.13	0.11	0	(0,0,0)
Lynx	-6	0.99	0.78	87	(83,4,0)
	0	0.79	0.58	82	(76,6,0)
	6	0.53	0.38	53	(31,21,1)
	12	0.27	0.21	25	(4,18,3)
	18	0.14	0.12	1	(1,0,0)
F16	-6	1.05	0.88	87	(84,1,2)
	0	0.80	0.66	81	(76,5,0)
	6	0.52	0.43	64	(42,21,1)
	12	0.35	0.28	29	(18,10,1)
	18	0.19	0.15	8	(2,6,0)
Car	-6	0.95	0.81	92	(91,1,0)
	0	0.81	0.55	86	(82,4,0)
	6	0.64	0.37	55	(39,12,4)
	12	0.44	0.23	45	(11,17,7)
	18	0.17	0.14	1	(0,1,0)

Table E.9: Distortions and word error rates for corrupted speech enhanced adaptively using weighted silence to estimate the noise; MMSE PSD estimation and general speech HMMs; 128 mixture components

Noise Source	SNR dB	Distortion All	Distortion Speech	% Error (D,S,I) MFCC Recoded	
Speech	-6	0.94	0.78	91	(90,1,0)
	0	0.67	0.54	86	(77,9,0)
	6	0.38	0.34	54	(25,26,3)
	12	0.18	0.19	3	(2,1,0)
	18	0.12	0.11	0	(0,0,0)
Lynx	-6	0.92	0.74	83	(79,4,0)
	0	0.71	0.56	79	(71,8,0)
	6	0.44	0.37	52	(22,29,1)
	12	0.23	0.21	12	(2,9,1)
	18	0.13	0.12	1	(1,0,0)
F16	-6	0.86	0.80	84	(79,3,0)
	0	0.61	0.59	76	(64,12,0)
	6	0.39	0.39	48	(22,23,3)
	12	0.24	0.25	12	(2,10,0)
	18	0.16	0.15	7	(2,4,1)
Car	-6	0.86	0.84	87	(84,3,0)
	0	0.67	0.54	74	(62,12,0)
	6	0.50	0.36	52	(24,25,3)
	12	0.33	0.23	36	(10,24,2)
	18	0.17	0.13	2	(0,2,0)

Table E.10: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimates; MMSE PSD estimation and general speech HMMs; 128 mixture components

Noise Source	SNR dB	Distortion All	Distortion Speech	% Error (D,S,I) MFCC Recoded	
Speech	-6	0.91	0.78	89	(85,4,0)
	0	0.64	0.54	81	(64,16,1)
	6	0.36	0.34	54	(23,27,4)
	12	0.17	0.18	0	(0,0,0)
	18	0.12	0.12	0	(0,0,0)
Lynx	-6	0.86	0.77	86	(81,5,0)
	0	0.66	0.56	77	(61,15,1)
	6	0.41	0.37	49	(19,29,1)
	12	0.22	0.21	6	(2,4,0)
	18	0.13	0.13	2	(1,0,1)
F16	-6	0.86	0.80	85	(78,7,0)
	0	0.64	0.57	72	(59,13,0)
	6	0.41	0.37	41	(20,19,2)
	12	0.23	0.22	10	(1,8,1)
	18	0.15	0.14	5	(0,5,0)
Car	-6	0.86	0.89	90	(88,2,0)
	0	0.59	0.57	71	(54,15,2)
	6	0.39	0.37	48	(17,29,2)
	12	0.23	0.23	24	(5,16,3)
	18	0.15	0.14	1	(0,1,0)

Table E.11: Distortions and word error rates for corrupted speech enhanced adaptively using weighted silence to estimate the noise; MMSE PSD estimation and general speech HMMs; 256 mixture components

Noise Source	SNR dB	Distortion All	Distortion Speech	% Error (D,S,I) MFCC Recoded	
Speech	-6	0.86	0.65	85	(28,32,25)
	0	0.51	0.45	61	(17,24,18)
	6	0.26	0.30	22	(5,11,6)
	12	0.16	0.19	1	(0,1,0)
	18	0.12	0.13	1	(0,1,0)
Lynx	-6	0.89	0.68	75	(36,30,9)
	0	0.64	0.49	68	(17,42,9)
	6	0.35	0.35	31	(3,17,11)
	12	0.19	0.21	1	(0,1,0)
	18	0.13	0.13	0	(0,0,0)
F16	-6	0.81	0.73	77	(37,28,12)
	0	0.51	0.54	63	(26,30,7)
	6	0.34	0.38	32	(12,18,2)
	12	0.24	0.25	13	(2,10,1)
	18	0.17	0.16	5	(0,5,0)
Car	-6	0.84	0.94	87	(81,6,0)
	0	0.60	0.58	62	(35,23,4)
	6	0.42	0.37	38	(6,24,8)
	12	0.29	0.24	26	(2,12,12)
	18	0.16	0.15	1	(0,1,0)

Table E.12: Distortions and word error rates for corrupted speech enhanced adaptively using weighted silence to estimate the noise; MMSE PSD estimation and general speech HMMs; 32 States, 4 mixture components

		Distortion		% Error (D,S,I)			
Noise Source	SNR dB	All	Speech	MFCC Recoded		AR Compensated	
Speech	-6	0.40	0.53	13	(4,9,0)	13	(4,9,0)
	0	0.29	0.34	1	(1,0,0)	1	(1,0,0)
	6	0.19	0.23	0	(0,0,0)	0	(0,0,0)
	12	0.15	0.16	0	(0,0,0)	0	(0,0,0)
	18	0.12	0.12	0	(0,0,0)	0	(0,0,0)
Lynx	-6	0.54	0.59	12	(0,12,0)	12	(0,12,0)
	0	0.37	0.39	1	(0,1,0)	1	(0,1,0)
	6	0.24	0.28	2	(0,1,0)	1	(0,1,0)
	12	0.17	0.18	0	(0,0,0)	0	(0,0,0)
	18	0.12	0.11	0	(0,0,0)	0	(0,0,0)
F16	-6	0.60	0.68	31	(11,19,1)	31	(11,19,1)
	0	0.40	0.43	8	(0,4,4)	8	(0,4,4)
	6	0.26	0.29	0	(0,0,0)	0	(0,0,0)
	12	0.19	0.20	0	(0,0,0)	0	(0,0,0)
	18	0.14	0.14	0	(0,0,0)	0	(0,0,0)
Car	-6	0.87	1.20	59	(11,29,19)	59	(11,29,19)
	0	0.49	0.55	23	(0,9,14)	24	(0,9,15)
	6	0.33	0.30	6	(0,2,4)	7	(0,1,6)
	12	0.22	0.20	0	(0,0,0)	0	(0,0,0)
	18	0.14	0.13	0	(0,0,0)	0	(0,0,0)

Table E.13: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimates; MMSE PSD estimation and word-based HMMs; Approximate $b_{\bar{x}_t m_t}(\mathbf{y}_t)$

		Distortion		% Error (D,S,I)			
Noise Source	SNR dB	All	Speech	MFCC Recoded		AR Compensated	
Speech	-6	0.71	0.95	50	(22,24,4)	39	(0,31,8)
	0	0.54	0.73	16	(0,15,1)	11	(0,9,2)
	6	0.39	0.58	4	(0,4,0)	20	(0,15,5)
	12	0.27	0.36	2	(0,2,0)	12	(0,9,3)
	18	0.20	0.22	0	(0,0,0)	1	(0,1,0)
Lynx	-6	0.81	1.00	48	(21,25,2)	43	(1,24,18)
	0	0.55	0.72	28	(12,15,1)	15	(0,9,6)
	6	0.37	0.51	5	(2,3,0)	15	(0,9,6)
	12	0.27	0.33	1	(0,1,0)	11	(0,7,4)
	18	0.19	0.22	0	(0,0,0)	1	(0,1,0)
F16	-6	0.86	1.04	56	(26,28,2)	50	(38,22,0)
	0	0.60	0.77	17	(2,15,0)	6	(0,6,0)
	6	0.42	0.53	6	(0,6,0)	9	(0,8,1)
	12	0.33	0.43	4	(0,4,0)	8	(0,8,0)
	18	0.24	0.29	1	(0,1,0)	2	(0,2,0)
Car	-6	0.79	1.04	38	(15,22,1)	29	(5,15,9)
	0	0.55	0.67	23	(1,18,4)	26	(0,16,10)
	6	0.37	0.45	10	(0,8,2)	16	(0,12,4)
	12	0.26	0.29	1	(0,1,0)	9	(0,6,3)
	18	0.19	0.20	0	(0,0,0)	0	(0,0,0)

Table E.14: Distortions and word error rates for corrupted speech enhanced adaptively using maximum likelihood noise parameter estimates; MMSE PSD estimation and word-based HMMs; Tested on the female speaker

Appendix F

Analysis of Wiener Filters

As described in Section 3.1, the frequency response of the standard Wiener filter is given by

$$H(\omega) = \frac{P_s(\omega)}{P_s(\omega) + P_d(\omega)} \tag{F.1}$$

where P_s is the PSD of the speech and P_d is the PSD of the noise. Only the spectral magnitude is estimated with the noisy phase being used when the signal is reconstructed. The estimated spectral magnitude $|\hat{S}(\omega)|$ is given by

$$|\hat{S}(\omega)| = H(\omega)|Y(\omega)| \tag{F.2}$$

where $Y(\omega)$ is the Fourier transform of the noisy signal. For additive noise, this is given by

$$Y(\omega) = S(\omega) + D(\omega) \tag{F.3}$$

where $S(\omega)$ and $D(\omega)$ are the Fourier transforms of the speech and noise respectively. The PSDs of these signals can be approximated by

$$\begin{aligned} P_s(\omega) &\approx |S(\omega)|^2 && \text{(F.4)} \\ P_d(\omega) &\approx |D(\omega)|^2. && \text{(F.5)} \end{aligned}$$

Ignoring cross correlation between the speech and the noise

$$|Y(\omega)| \approx \sqrt{|S(\omega)|^2 + |D(\omega)|^2}. \tag{F.6}$$

Substituting Equations F.1, F.4, F.5 and F.6 into Equation F.2 gives

$$|\hat{S}(\omega)|^2 \approx \frac{|S(\omega)|^4}{|S(\omega)|^2 + |D(\omega)|^2}. \tag{F.7}$$

Let $|D(\omega)|^2 = K|S(\omega)|^2$ where K is a constant. Now

$$|\hat{S}(\omega)| \approx \frac{|S(\omega)|}{\sqrt{1+K}}. \tag{F.8}$$

For small K (negligible noise)

$$|\hat{S}(\omega)| \approx |S(\omega)|. \tag{F.9}$$

For large K (large noise)

$$|\hat{S}(\omega)| \to 0. \tag{F.10}$$

For intermediate K (SNR approximately zero), $|\hat{S}(\omega)|$ is given by Equation F.8.

Bibliography

Abdallah, I., Montrēsor, S. & Bauding, M. (1997), Speech signal detection in noisy environment using a local entropic criterion, *in* 'European Conference on Speech Communication and Technology', pp. 2595–2598.

Acero, A. & Stern, R. M. (1990), Environmental robustness in automatic speech recognition, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', 849–852.

Afify, M., Gong, Y. & Haton, J. (1997), A unified maximum likelihood approach to acoustic mismatch compensation: application to noisy Lombard speech recognition, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing'.

Anderson, B. & Moore, J. B. (1979), *Optimal Filtering*, Prentice Hall.

Arslan, L. & Hansen, J. H. L. (1994), Minimum cost based phoneme class detection for improved iterative speech enhancement, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. II–45–II–48.

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970), 'A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains', *Ann. Math. Statist.* **41**, 164–171.

Beattie, V. & Young, S. (1992), Hidden Markov model state-based noise cancellation, Technical report, University of Cambridge.

Boll, S. F. (1979), 'Suppression of acoustic noise in speech using spectral subtraction', *IEEE Transactions on Acoustics, Speech and Signal Processing* **2**, 113–120.

Deller, J. R., Proakis, J. G. & Hansen, J. H. L. (1993), *Discrete–Time Processing of Speech Signals*, Macmillan.

Ephraim, Y. (1992*a*), 'A Bayesian estimation approach for speech enhancement using hidden Markov models', *IEEE Transactions on Signal Processing* **40**(4), 725–735.

Ephraim, Y. (1992*b*), 'Gain-adapted hidden Markov models for recognition of clean and noisy speech', *IEEE Transactions on Signal Processing* **40**(6), 1303–1316.

Ephraim, Y. (1992*c*), 'Statistical–model–based speech enhancement systems', *Proceedings of the IEEE* **80**, 1526–1555.

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), 1109–1121.

Ephraim, Y. & Malah, D. (1985), 'Speech enhancement using a minumum mean square error log-spectral amplitude estimator', *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-33**(2), 443–445.

Ephraim, Y. & VanTrees, H. L. (1995), 'A signal subspace approach for speech enhancement', *IEEE Transactions on Speech and Audio Processing* **3**(4), 251–265.

Ephraim, Y., Malah, D. & Juang, B. H. (1989), 'On the application of hidden Markov models for enhancing noisy speech', *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**(12), 1846–1856.

Ephraim, Y., Wilpon, J. G. & Rabiner, L. R. (1987), A linear predictive front-end processor for speech recognition in noisy environments, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 31.1.1–31.1.4.

Erell, A. & Weintraub, M. (1993), 'Energy conditioned spectral estimation for recognition of noisy speech', *IEEE Transactions on Speech and Audio Processing* **1**(1), 84–89.

Erell, A. & Weintraub, M. (1994), 'Estimation of noise-corrupted speech dft-spectrum using the pitch period', *IEEE Transactions on Speech and Audio Processing* **2**(1), 1–8.

ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels (1997), Proceedings, Pont-a-Mousson, France.

Fanty, M. & Cole, R. (1990), Spoken letter recognition, *in* 'Proceedings Neural Information Processing System Conference'.

Gales, M. J. F. (1995), Model-based techniques for noise robust speech recognition, PhD thesis, University of Cambridge.

Gibson, J. D., Koo, B. & Gray, S. D. (1991), 'Filtering of colored noise for speech enhancement', *IEEE Transactions on Signal Processing* **39**(8), 1732–1741.

Gillick, L. & Cox, S. J. (1989), Some statistical issues in the comparison of speech recognition algorithms, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 532–535.

Gish, H., Chow, Y. L. & Rohlicek, J. R. (1990), Probabilitic vector mapping of noisy speech parameters for HMM word spotting, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 117–120.

Gong, Y. (1993), Base transformation for environment adaptation in continuous speech recognition, *in* 'European Conference on Speech Communication and Technology', pp. 2227–2230.

Gong, Y. (1995), 'Speech recognition in noisy environments: A survey', *Speech Communication* **16**, 261–291.

Graf, J. T. & Hubing, N. (1993), Dynamic time-warping for the enhancement of speech degraded by white Gaussian noise, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. II, pp. 339–342.

Gray, A. H., Buzo, A., Gray, R. M. & Matsuyama, Y. (1980), 'Distortion measures for speech processing', *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-28**(4), 367–376.

Hansen, J. H. L. & Clements, M. A. (1991), 'Constrained iterative speech enhancement with application to speech recognition', *IEEE Transactions on Signal Processing* **39**(4), 795–805.

Juang, B. H. (1984), 'On the hidden Markov model and dynamic time warping for speech recognition - a unified view', *AT&T Bell Laboratories Technical Journal* **63**(7), 1213–1243.

Juang, B. H. & Rabiner, L. R. (1985), 'Mixture autoregresive hidden Markov models for speech signals', *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP-33**(6), 1404–1413.

Juang, B. H. & Rabiner, L. R. (1987), Signal restoration by spectral mapping, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 6.6.1–6.6.4.

Junqua, J. C., Mak, B. & Reaves, B. (1994), 'A robust algorithm for word boundary detection in the presence of noise', *IEEE Transactions on Speech and Audio Processing* **2**(3), 406–412.

Kim, N. S. (1998), 'Non-stationery environment compensation based on sequential estimation', *IEEE Signal Processing Letters* **5**(1), 8–10.

Klatt, D. H. (1976), A digital filter-bank for spectral matching, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 699–702.

Lee, C. H. (1997), On feature and model compensation approach to robust speech recognition, *in* 'Robust Speech Recognition for Unknown Communication Channels', pp. 45–54.

Lee, K. Y. & Rheem, J. Y. (1997), A nonstationary autoregressive HMM and its application to speech enhancement, *in* 'European Conference on Speech Communication and Technology', pp. 1407–1410.

Lee, K. Y. & Shirai, K. (1996), 'Efficient recursive estimation for speech enhancement in colored noise', *IEEE Signal Processing Letters* **3**(7), 196–199.

Lee, K. Y., Lee, B. & Ann, S. (1997), 'Adaptive filtering for speech enhancement in colored noise', *IEEE Signal Processing Letters* **4**(10), 277–279.

Lee, K. Y., Lee, B., Song, I. & Yoo, J. (1996), Recursive speech enhancement using the EM algorithm with initial conditions trained by HMM's, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 621–624.

Lee, K. Y., Rheem, J. Y. & Shirai, K. (1996), Recursive estimation based on the trended hidden Markov model in speech enhancement, *in* 'IEEE Asia Pacific Conference on Circuits and Systems '96', pp. T3–PC8.1–T3–PC8.4.

Lim, J. S. & Oppenheim, A. V. (1978), 'All–pole modeling of degraded speech', *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**, 197–210.

Lim, J. S. & Oppenheim, A. V. (1979*a*), 'Enhancement and bandwidth compression of noisy speech', *Proceedings of the IEEE* **67**, 1586–1604.

Lim, J. S. & Oppenheim, A. V. (1979*b*), 'Evaluation of an adaptive comb filtering method for enhancing speech degraded by white noise addition', *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**(4), 354–358.

Little, R. J. A. & Rubin, D. B. (1987), *Statistical analysis with missing data*, John Wiley and Sons.

Logan, B. T. & Robinson, A. J. (1996), Noise estimation for enhancement and recognition within an autoregressive hidden-Markov-model framework, *in* 'Proceedings Sixth Australian International Conference on Speech Science and Technology', pp. 85–90.

Logan, B. T. & Robinson, A. J. (1997*a*), Enhancement and recognition of noisy speech within an autoregressive hidden Markov model framework using noise estimates from the noisy signal, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 843–846.

Logan, B. T. & Robinson, A. J. (1997*b*), Improving autoregressive hidden Markov model recognition accuracy using a non-linear frequency scale with application to speech enhancement, *in* '5th European Conference on Speech Communication and Technology', pp. 2103–2106.

Malah, D. & Cox, R. V. (1982), A generalized comb filtering technique for speech enhancement, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 1, pp. 160–163.

McAulay, R. J. & Malpass, M. L. (1980), 'Speech enhancement using a soft-decision noise suppression filter', *IEEE Transactions on Acoustics, Speech and Signal Processing* **ASSP–28**, 137–145.

McKinley, B. & Whipple, G. (1997), Model based speech pause detection, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 1179–1182.

Merhav, N. & Ephraim, Y. (1991), 'Maximum likelihood hidden Markov modelling using a dominant sequence of states', *IEEE Transactions on Signal Processing* **39**, 2111–2115.

Mokbel, C. & Chollet, G. (1992), Wrod recognition in the car: speech enhancement / spectral transformations, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. 2, pp. 925–928.

Moreno, P. J., Raj, B. & Stern, R. M. (1995), A vector Taylor series approach for environment-independent speech recognition, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 733–736.

Musicus, B. R. & Lim, J. S. (1979), An iterative technique for maximum likelihood parameter estimation on noisy data, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 224–227.

Nadas, A., Nahamoo, D. & Picheny, M. A. (1989), 'Speech recognition using noise-adaptive prototypes', *IEEE Transactions on Acoustics, Speech and Signal Processing* **37**(10), 1495–1502.

Oppenheim, A. V. & Johnson, D. H. (1972), 'Discrete representation of signals', *Proceedings of the IEEE* **60**(6), 681–691.

O'Shaughnessy, D. (1988), Speech enhancement using vector quantization and a formant distance measure, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 549–552.

Paliwal, K. K. (1988), 'Estimation of noise variance from the noisy ar signal and its application in speech enhancement', *IEEE Transactions on Acoustics, Speech and Signal Processing* **36**(2), 292–294.

Paliwal, K. K. & Basu, A. (1987), A speech enhancement method based on Kalman filtering, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 6.3.1–6.3.4.

Porter, J. E. & Boll, S. F. (1984), Optimal estimators for spectral restoration of noisy speech, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 18A.2.1–18A.2.4.

Price, P., Fisher, W. M., Bernstein, J. & Pallett, D. S. (1988), The DARPA 1000-word Resource Management database for continuous speech recognition, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 651–654.

Quatieri, T. F. & McAulay, R. J. (1990), Noise reduction using a soft-decision sine-wave vector quantizer, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 821–824.

Rabiner, L. R. (1989), 'A tutorial on hidden Markov models and selected applications in speech recognition', *Proceedings of the IEEE* **77**, 257–285.

Rabiner, L. R. & Juang, B. H. (1993), *Fundamentals of Speech Recognition*, Prentice–Hall.

Rabiner, L. R. & Schafer, R. W. (1978), *Digital Processing of Speech Signals*, Prentice Hall.

Raj, B., Gouvêa, E., Moreno, P. J. & Stern, R. M. (1996), Cepstral compensation by polynomial approximation for environment-independent speech recognition, *in* 'International Conference on Spoken Language Processing', pp. 2340–2343.

Ramalho, M. A. & Mammone, R. J. (1994), A new speech enhancement technique with application to speaker identification, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. I, pp. 29–32.

Rose, R. C., Hofstetter, E. M. & Reynolds, D. A. (1994), 'Integrated models of signal and background with application to speaker identification', *IEEE Transactions on Speech and Audio Processing* **2**(2), 245–257.

Saleh, G. M. K. (1996), Bayesian Inference in Speech Processing, PhD thesis, University of Cambridge.

Saleh, G. M. K. & Niranjan, M. (1998), Speech enhancement in a Bayesian framework, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing'.

Sambur, M. R. (1978), 'Adaptive noise canceling for speech signals', *IEEE Transactions on Acoustics, Speech and Signal Processing* **26**(5), 419–423.

Sankar, A. & Lee, C. H. (1995), Robust speech recognition based on stochastic matching, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 121–124.

Sankar, A. & Lee, C. H. (1996), 'A maximum-likelihood approach to stochastic matching for robust speech recognition', *IEEE Transactions on Speech and Audio Processing* **4**(3), 190–202.

Seymour, C. W. (1996), Model-Based Speech Enhancement, PhD thesis, University of Cambridge.

Sheikhzadeh, H. & Deng, L. (1994), 'Waveform-based speech recognition using hidden filter models: Parameter selection and sensitivity to power normalization', *IEEE Transactions on Speech and Audio Processing* pp. 80–89.

Sheikhzadeh, H., Sameti, H., Deng, L. & Brennan, R. L. (1994), Comparative performance of spectral subtraction and HMM-based speech enhancement with application to hearing aid design, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. I13–I16.

Sheikhzeheh, H., Brennan, R. & Sameti, H. (1995), Real-time implementation of HMM-based MMSE algorithm for speech enhancement in hearing aid applications, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', pp. 808–811.

Shikano, K. (1985), Evaluation of LPC spectral matching measures for phonetic unit recognition, Technical Report CMU-CS-86-108, Carnegie-Mellon University.

Siohan, O. & Lee, C. (1997), 'Iterative noise and channel estimation under the stochastic matching algorithm framework', *IEEE Signal Processing Letters* **4**(11), 304–306.

Smith, J. O. & Abel, J. S. (1995), The bark bilinear transform, *in* 'IEEE Workshop on Applications of Signal Processing to Audio and Acoustics'.

Strube, H. W. (1980), 'Linear prediction on a warped frequency scale', *J. Acoust. Soc. Am.* **68**(4), 1071–1076.

Tishby, N. Z. (1991), 'On the application of mixture AR hidden Markov models to text independent speaker recognition', *IEEE Transactions on Signal Processing* **39**, 563–570.

Treurniet, W. C. & Gong, Y. (1994), Noise independent speech recognition for a variety of noise types, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', Vol. I, pp. 437–440.

Van-Compernolle, D. (1989), 'Noise adaptation in a hidden Markov model speech recognition system', *Computer Speech Language* **3**(2), 151–167. 1989.

Van-Trees, H. L. (1968), *Detection, Estimation, and Modulation Theory*, John Wiley and Sons.

Varga, A. P. & Moore, R. K. (1990), Hidden Markov model decomposition of speech and noise, *in* 'Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing', 845–848.

Varga, A. P., Steeneken, H. J. M., Tomlinson, M. & Jones, D. (1992), The noisex-92 study on the effect of additive noise on automatic speech recognition, Technical report, DRA Speech Research Unit.

Wilpon, J. G. & Rabiner, L. R. (1987), 'Application of hidden Markov models to automatic endpoint detection', *Computer Speech and Language* **2**, 321–341.

Woodland, P. C. & Young, S. J. (1993), 'The HTK tied-state continuous speech recogniser', *European Conference on Speech Communication and Technology* pp. 2207–2210.

Woodland, P. C., Gales, M. J. F., Pye, D. & Young, S. J. (1997), 'Broadcast news transcription using HTK', *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing* pp. 719–722.

Young, S. J. (1996), 'A review of large-vocabulary continuous-speech recognition', *IEEE Signal Processing Magazine* pp. 45–57.

Young, S. J., Woodland, P. C. & Byrne, W. J. (1993), *HTK: Hidden Markov Model Toolkit V1.5*, Cambridge University Engineering Department Speech Group and Entropic Research Laboratories Inc.

Young, S., Jansen, J., Ollason, D. & Woodland, P. (1996), *The HTK book for HTK V2.0*, Cambridge University Technical Services Ltd. and Entropic Cambridge Research Laboratory Ltd.