Model-Based Speech Enhancement

Philip John Harding

A thesis submitted for the Degree of
Doctor of Philosophy

University of East Anglia
School of Computing Sciences

July 2013

Abstract

A method of speech enhancement is developed that reconstructs clean speech from a set of acoustic features using a harmonic plus noise model of speech. This is a significant departure from traditional filtering-based methods of speech enhancement. A major challenge with this approach is to estimate accurately the acoustic features (voicing, fundamental frequency, spectral envelope and phase) from noisy speech. This is achieved using maximum a-posteriori (MAP) estimation methods that operate on the noisy speech. In each case a prior model of the relationship between the noisy speech features and the estimated acoustic feature is required. These models are approximated using speaker-independent GMMs of the clean speech features that are adapted to speaker-dependent models using MAP adaptation and for noise using the Unscented Transform.

Objective results are presented to optimise the proposed system and a set of subjective tests compare the approach with traditional enhancement methods. Three-way listening tests examining signal quality, background noise intrusiveness and overall quality show the proposed system to be highly robust to noise, performing significantly better than conventional methods of enhancement in terms of background noise intrusiveness. However, the proposed method is shown to reduce signal quality, with overall quality measured to be roughly equivalent to that of the Wiener filter.

Acknowledgements

First, thanks go to Dr. Ben Milner for his excellent supervision throughout my time at the University of East Anglia. This work would not have been possible without his support and advice and I owe him a lot for that. I would also like to thank Prof. Stephen Cox for his advice during my PhD as well as his leadership of the Speech, Language and Audio Processing group.

Thanks also go to everyone in the speech lab, past and present, not only for their excellent technical advice, but for making my time at UEA particularly enjoyable. I will miss, and certainly not forget, the time I spent here.

I owe a lot to Connie, as well as my family, and am grateful for all their love and support. Without it I would not have got this far.

Finally, I would also like to acknowledge and thank the UEA and School of Computing Sciences for funding me through my PhD, and also extend my thanks to my examiners, Dr. Barry Theobald and Dr. Mike Brookes, for their comments and suggestions which have no doubt improved the quality of this thesis.

Contents

List of Publications

1. Harding, P and Milner, B. (2011). Speech enhancement by reconstruction from cleaned acoustic features. In *Proceedings of Interspeech*, pages 1189-1192

2. Harding, P and Milner, B. (2012). Enhancing Speech by Reconstruction from Robust Acoustic Features. In *Proceedings of Interspeech*

3. Harding, P and Milner, B. (2012). On the use of Machine Learning Methods for Speech and Voicing Classification. In *Proceedings of Interspeech*

List of Abbreviations

ASR	Automatic Speech Recognition
CMOS	Comparative Mean Opinion Score
DCT	Discrete Cosine Transform
$f0$	Fundamental Frequency
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HNM	Harmonic plus Noise Model
HTK	Hidden Markov Model Toolkit
LFCC	Linear-Frequency Cepstral Coefficients
LPC	Linear Predictive Coding
LSF	Line Spectral Frequencies
MAP	Maximum *a-posteriori*
MFCC	Mel-Frequency Cepstral Coefficients
ML	Machine Learning
MLLR	Maximum Likelihood Linear Regression
MLP	Multilayer Perceptron
MOS	Mean Opinion Score
OLA	Overlap and Add
PCA	Principal Component Analysis
PDF	Probability Density Function
PMC	Parallel Model Combination
SMC	Serial Model Combination
SPLICE	Stereo-based Piecewise Linear Compensation for Environments
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
STSA	Short-Time Spectral Amplitude
SVM	Support Vector Machine
UT	Unscented Transform
VAD	Voice Activity Detection
VC	Voicing Classification
VTS	Vector Taylor Series

List of Figures

List of Tables

Chapter 1

Introduction

This chapter describes the problem of speech enhancement and introduces the proposed method of speech enhancement by reconstruction. The chapter begins by describing the effect of noise on speech and the constraints on single-channel audio-only speech enhancement. The structure of the thesis is then described.

Contents

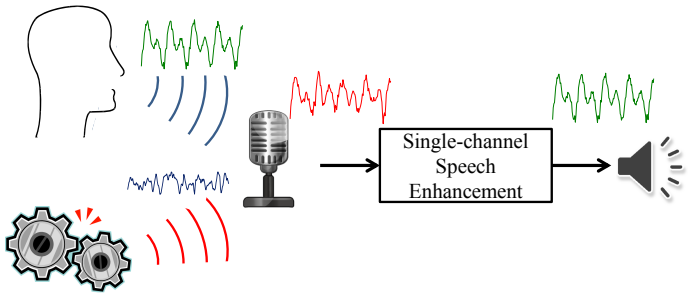


Figure 1.1: Single-channel audio-only speech enhancement

1.1 Introduction

Speech enhancement is the process of removing the effect of noise from speech recorded in noisy environments. Noise has two main effects on the perception of speech. First, the perceived quality of the signal is deteriorated whilst second, the intelligibility of the speech may also be reduced. The joint effect of these two degradations is to increase listener fatigue and, in some cases, to reduce the amount of information which may be successfully conveyed.

In this work a novel method of audio-only single-channel speech enhancement is described. The only available information about the speech is therefore the monaural noisy audio signal as illustrated in Figure 1.1. This is a more challenging problem than multi-channel speech enhancement where stereo (or higher dimensional) signals are available which contain signals from additional microphones or even video cameras for audio-visual speech enhancement as illustrated in Figures 1.2 and 1.3 respectively. In the case of audio-only multichannel speech enhancement the position of the speaker and noise source may be identified to enable better source separation [Meyer and Simmer, 1997], whilst in the case of audio-visual speech enhancement facial features such as the position of the lips and other visible articulators, which are not dependent on SNR, may be tracked to provide further information about the speech [Almajai and Milner, 2009]. From this point forward all techniques are described in the context of audio-only single-channel speech enhancement.

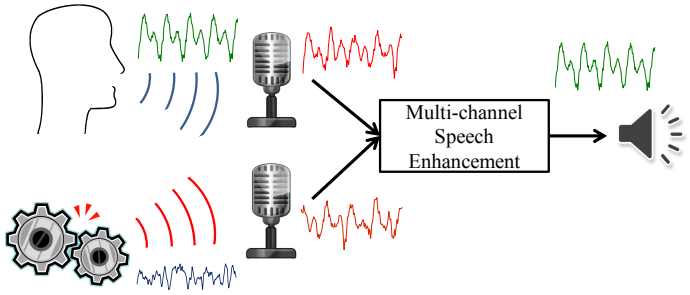


Figure 1.2: Multi-channel audio-only speech enhancement

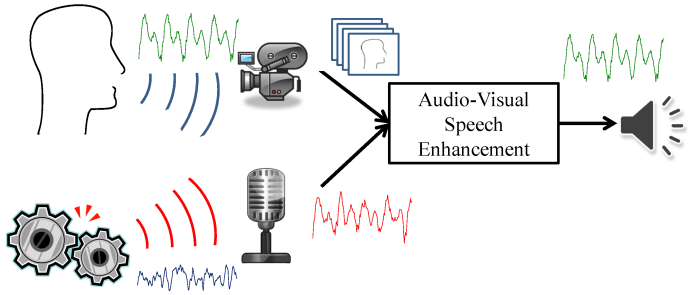


Figure 1.3: Audio-visual speech enhancement

1.1.1 Problem definition

We wish to remove the detrimental effects of the noise whilst preserving the underlying speech signal by estimating the clean speech signal, $x(m)$, from the noisy speech, $y(m)$. The noise is assumed to be additive and so the noisy speech signal, $y(m)$, can be described in terms of the clean signal, $x(m)$, and the noise signal, $n(m)$ as:

$$y(m) = x(m) + n(m). \tag{1.1}$$

An intuitive approach to noise remove is therefore to subtract an estimate of the noise from the noisy signal. Noise estimation is inherently challenging, with accurate estimation of the noise impossible. Undesirable effects occur when inaccurate estimates of the noise are subtracted from the noisy signal, and these can be grouped into two categories: underestimation and overestimation. First, in the

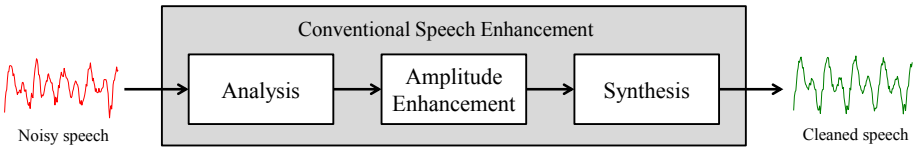


Figure 1.4: Flowchart of conventional methods of speech enhancement

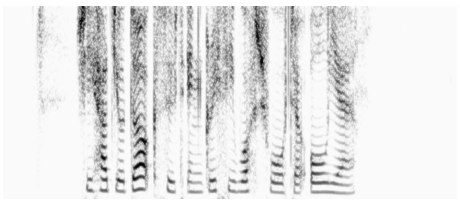
case of underestimation, some of the noise will remain in the signal after enhancement. Second, overestimation of the noise may cause the speech signal to also be suppressed resulting in speech distortion which may further reduce the intelligibility of the speech [Loizou and Kim, 2011].

There are many alternative functions of noise removal, from conventional filtering techniques to binary time-frequency masks and subspace methods. The operation of these methods is illustrated in Figure 1.4 and they are described in more detail in Chapter 2. Whilst these methods have been shown to be effective in relatively low levels of stationary noise, performance reduces in non-stationary noises [Loizou, 2007]. This is largely due to the noise estimation process not accurately tracking the noise and so time varying and impulsive noises often remain in the enhanced signal. The effect of this is shown in Figure 1.5 where log MMSE, one of the best performing methods of speech enhancement, is used to enhance an utterance of female speech with white noise (Figure 1.5(d)) and babble noise (Figure 1.5(e)), both at 5dB SNR.

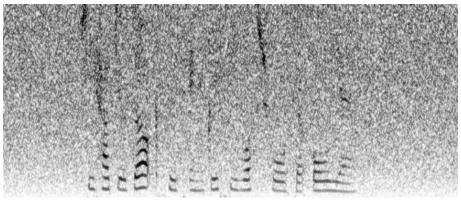
In the case of white noise the noise has been underestimated causing a considerable amount of residual noise to remain in the signal, similar to the original noise. When the speech is affected by babble noise the enhanced signal contains artifacts known as 'musical noise'. These artifacts are visible as isolated regions of noise across time and frequency which are audible as annoying 'musical' tones and are caused by inaccuracies in noise tracking.

1.1.2 Proposed method

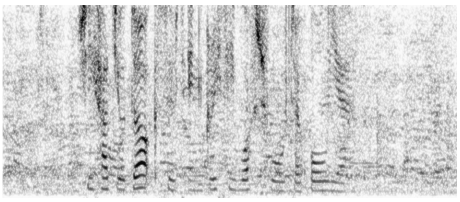
The method of speech enhancement described in this thesis takes an approach of speech enhancement by reconstruction. By reconstructing speech using an appropri-



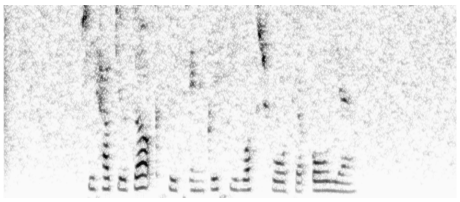
(a) Clean



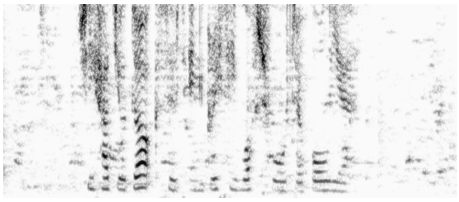
(b) White noise



(c) Babble noise



(d) White noise (log MMSE)



(e) Babble noise (log MMSE)

Figure 1.5: Narrowband spectrograms showing an utterance in a.) clean conditions, b.) white noise at 5dB SNR, c.) babble noise at 5dB SNR, d.) white noise at 5dB SNR and enhanced using log MMSE and e.) babble noise at 5dB SNR and enhanced using log MMSE

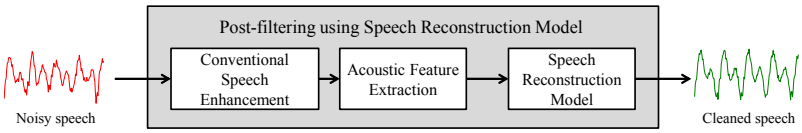


Figure 1.6: Flowchart of speech enhancement methods using speech reconstruction as a post-filter

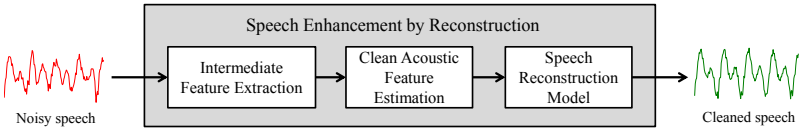


Figure 1.7: Flowchart of proposed speech enhancement by reconstruction method

ate model of reconstruction rather than filtering the noisy signal it is expected that artifacts such as musical noise will be eliminated as they will not be reconstructed.

The reconstruction model is driven by a set of acoustic features which must be estimated from the noisy speech. The use of speech reconstruction models in methods of speech enhancement is not a new idea, with several methods having already been developed. These existing methods are described in Chapter 2 and typically extract the acoustic features required for reconstruction from signals that have already been processed by a conventional method of speech enhancement, for example: spectral subtraction, Wiener filtering or log MMSE. This gives a three stage approach of: i.) conventional speech enhancement, ii.) acoustic feature extraction and iii.) speech reconstruction, as illustrated in Figure 1.6.

This work instead aims to estimate the acoustic features required for reconstruction directly from the noisy signal. An intermediate feature for estimation is first extracted from the noisy speech before the acoustic features required for reconstruction are estimated from this intermediate feature. The proposed system therefore takes a different three stage approach of: i.) noisy feature extraction, ii.) clean acoustic feature estimation and iii.) speech reconstruction (Figure 1.7).

By estimating acoustic features directly from the noisy speech the effect of artifacts caused by conventional methods of estimation should be avoided. This also enables a data-driven approach to acoustic feature estimation.

1.2 Thesis structure

The remainder of this thesis is divided into nine further chapters as follows:

2.) Speech Enhancement Review This chapter describes a number of existing methods of speech enhancement. These include: conventional filtering approaches, subspace methods and binary masking. A number of methods using speech reconstruction models as part of the enhancement process are also described to put the proposed system in context with existing methods.

3.) Speech Reconstruction Speech reconstruction models that may be used to reconstruct speech for this method of speech enhancement are described in this chapter. All of the considered reconstruction models are driven by a set of acoustic features and so this chapter is split into two parts: first, the reconstruction models are described and second, results from experiments measuring the correlation between the required acoustic features and parameterisations of the noisy speech are reported.

4.) Methods of Feature Estimation This chapter describes a method of acoustic feature estimation. Maximum *a-posteriori* (MAP) estimation was chosen for use in this work and relies on a prior model of the joint distribution of the noisy speech and the target acoustic feature. Methods of obtaining these distributions are therefore also described, including methods of speaker and noise adaptation.

5.) Spectral Envelope Estimation Using the method of estimation described in Chapter 4, this chapter describes the proposed system for spectral envelope estimation. Two systems are described. First, a method using a global model of speech is described before second, a method using localised models is proposed. The proposed systems are tested against the spectral amplitude estimation component of three conventional methods of speech enhancement: spectral subtraction, Wiener filtering and log MMSE.

6.) Fundamental Frequency Estimation A method of fundamental frequency ($f0$) estimation using MAP estimation is described in this chapter. Performance of the proposed system is evaluated in comparison with two conventional methods of $f0$ estimation: YIN and ETSI XAFE estimator.

7.) Voicing Classification This chapter describes a method of voicing classification. The chapter begins by reviewing a range of Machine Learning methods for the purpose of voicing classification to determine the most suitable method of data-driven classification. The most suitable method is then evaluated.

8.) Phase Estimation The final acoustic feature required for reconstruction is phase. This chapter therefore evaluates a range of phase models including: noisy signal phase, zero-phase, random-phase and minimum-phase models. Each model is evaluated in terms of the quality of reconstructed speech measured using both objective and subjective tests.

9.) Speech Enhancement System This chapter describes the proposed method of speech enhancement by reconstruction. The optimal speech reconstruction model as determined in Chapter 3 is driven by the acoustic features estimated using the methods described in Chapters 5-8 to reconstruct cleaned speech. This method is compared to conventional methods of enhancement as well as two more recent methods of reconstruction including a method of direct MFCC inversion and a model-based Wiener filter, constructed using spectral envelope estimated using the method described in Chapter 5. Performance is evaluated objectively using PESQ and subjectively using listening tests.

10.) Conclusions and Further Work The final chapter is split into two sections. The first draws conclusions about the proposed method of speech enhancement whilst the second describes how the system may be extended.

There are two appendices: Appendix A describes the datasets used in this work whilst Appendix B shows within-class correlation between clean and noisy MFCC feature vectors.

1.3 Previous Work

This thesis extends the work of Shao [2005] and Darch [2008]. Where this work has been extended, it has been appropriately cited. This work differs from the aforementioned work in several ways, including the following:

1. The method of speech reconstruction from MFCC features described by Shao [2005] was applied to the problem of speech enhancement,

2. The acoustic feature estimation techniques used by Darch [2008] were extended to use improved noise adaptation and the use of speaker-adaptation techniques was also introduced,

3. A review of machine learning methods for voicing classification was undertaken and the use of enhanced speech features was examined as an alternative to model adaptation in noisy conditions,

4. A range of phase estimation methods were applied to the reconstruction model to determine the effect of the use of the phase of the noisy speech on the quality of reconstructed speech.

Chapter 2

Speech Enhancement Review

The objective of this chapter is to put the proposed method of speech enhancement into perspective by describing existing methods of speech enhancement. First, conventional methods of speech enhancement are discussed. A general framework is described and then a number of related techniques are discussed. These include approaches based on filtering, binary masking and subspace analysis. More recently, speech reconstruction models have been applied for the purpose of speech enhancement. A number of methods of speech enhancement by reconstruction are therefore also described in this chapter. Finally, a number of methods of measuring the quality and intelligibility of processed speech are then reviewed.

Contents

2.1 Introduction

This chapter is split into two parts. The first describes a number of different approaches to speech enhancement with the aim of putting the proposed method of speech enhancement by reconstruction into perspective with existing methods, whilst the second describes methods of measuring the success of enhancement in terms of quality and intelligibility.

First, conventional methods of speech enhancement that filter out an estimate of the noise from the noisy signal are described in Section 2.2. Next, methods using binary time-frequency masks are described in Section 2.3 whilst subspace methods are described in Section 2.4. Finally, existing methods of speech enhancement by reconstruction are described in Section 2.5.

In terms of evaluation of performance, Section 2.6 describes a number subjective and objective tests used to measure the quality and intelligibility of enhanced speech.

2.2 Conventional Methods of Speech Enhancement

Conventional methods of speech enhancement are defined as those that use a filter to remove an estimate of the noise from the noisy speech to give an estimate of the noise-free speech. These methods typically take an approach of analysis followed synthesis. Before synthesis the signal parameters are modified to reduce the effect of noise to give an analysis-enhancement-synthesis approach. These methods typically focus on enhancing spectral amplitudes and so are also known as short-time spectral amplitude (STSA) methods. The three steps of such an approach can be broadly described as follows:

Analysis Utterances are processed on a frame-by-frame basis. Frames are typically 10-30ms in duration and so within each frame the signal may be assumed stationary. Due to limitations of the discrete Fourier transform (DFT) frames are windowed using a Hamming or Hann window. Frames are therefore usually

also overlapped to avoid aliasing in the modulation domain, with an overlap of 75% required to avoid aliasing completely.. Given a frame of noisy speech a window is applied and the DFT taken as:

$$Y(k) = \sum_{m=0}^{N-1} w(m)y(m)e^{-j\frac{2\pi km}{N}} \qquad \text{for } 0 \leq k \leq N-1, \tag{2.1}$$

where $y(m)$ and $w(m)$ are the mth samples of the noisy speech and window respectively and $Y(k)$ is the kth frequency bin of the complex spectrum consisting of N bins. The absolute of the complex spectrum is then taken to give the magnitude spectrum, $|Y(k)|$.

Enhancement In the case of STSA methods, enhancement focuses solely on removing the effect of noise on spectral amplitudes. The effect of noise on phase is often assumed to be inaudible [Wang and Lim, 1982], whilst the noisy phase has also been shown to be optimal under certain assumptions Loizou [2007]. Clean spectral amplitudes are estimated in some optimal way using an estimate of the noise. If $|Y(k)| = f(|X(k)|, |N(k)|)$ is a function describing the relationship between spectral amplitudes of speech, $|X(k)|$, and noise, $|N(k)|$, to give noisy spectral amplitudes, $|Y(k)|$, then enhancement methods aim to derive the inverse of this function. This gives $|\hat{X(k)}| = f^{-1}(|Y(k)|, |\hat{N(k)}|)$ where $|\hat{X(k)}|$ is an estimate of the clean spectral amplitudes and $|\hat{N(k)}|$ is an estimate of the noise. There are two challenges to such an approach: i.) computing an accurate estimate of the noise and ii.) designing an appropriate function of noise removal. In most cases the function of noise removal is expressed in terms of a gain function (i.e. filter), $H(k)$, where $f(|\hat{Y(k)}|, |\hat{N}(k)) = H(k)|Y(k)|$ and $H(k)$ is computed based on the $a-priori$ and $a-posteriori$ SNRs.

Synthesis Speech frames are resynthesised by taking the inverse DFT of the complex spectrum. The modified magnitude spectrum is combined with the orig-

inal phase spectrum as:

$$\hat{X}(k) = |\hat{X}(k)|e^{j\angle Y(k)}, \tag{2.2}$$

where $\angle X(k)$ is the phase of the original signal. The inverse DFT is then computed to give the estimated waveform:

$$\hat{x}(m) = \frac{1}{N}\sum_{k=0}^{N-1} \hat{X}(k)e^{j\frac{2\pi km}{N}} \qquad \text{for } 0 \leq m \leq N-1. \tag{2.3}$$

Overlap and add (OLA) may then be used to recombine frames to give an estimate of the clean speech signal, $s(m)$:

$$s(m) = x(m)w_{ola}(R-m)+x(m)w_{ola}(2R-m) \qquad \text{for } 0 \leq m \leq R-1, \tag{2.4}$$

where $R = N/2$ for 50% overlap and $w_{ola}(m)$ is the mth sample of the OLA window.

There are several classes of noise removal function. These include: spectral subtraction, Wiener filtering, statistical-model-based methods and subspace algorithms [Loizou, 2007]. Three methods of conventional enhancement are now considered. First, spectral subtraction is described in Section 2.2.1. Next, Wiener filtering is discussed in Section 2.2.2 before statistical-model-based methods are covered in Section 2.2.3.

2.2.1 Spectral subtraction

Spectral subtraction is one the most basic methods of speech enhancement. Assuming additive noise, an estimate of the noise may be subtracted from the noisy speech to give an estimate of the clean speech. This operation is performed in the frequency domain and is typically only applied to the magnitude spectrum. This noise removal process can be implemented by applying a gain function, $H(k)$, to the

magnitude spectrum of the noisy speech:

$$|\hat{X}(k)| = H(k)|Y(k)|, \tag{2.5}$$

where the response of $H(k)$ is computed from the noisy speech and estimate of the noise as:

$$H(k) = \frac{|X(k)|}{|X(k)| + |N(k)|} = \frac{|X(k)|}{|Y(k)|} = 1 - \frac{|N(k)|}{|Y(k)|}. \tag{2.6}$$

When $H(k)$ is applied to the noisy magnitude spectrum, $|Y(k)|$, this reduces to a simple subtraction, i.e:

$$|\hat{X}(k)| = f^{-1}(|Y(k)|, |\hat{N}(k)|) = |Y(k)| - |\hat{N}(k)|. \tag{2.7}$$

Subtraction may occur in one of several domains, indexed by p, i.e.:

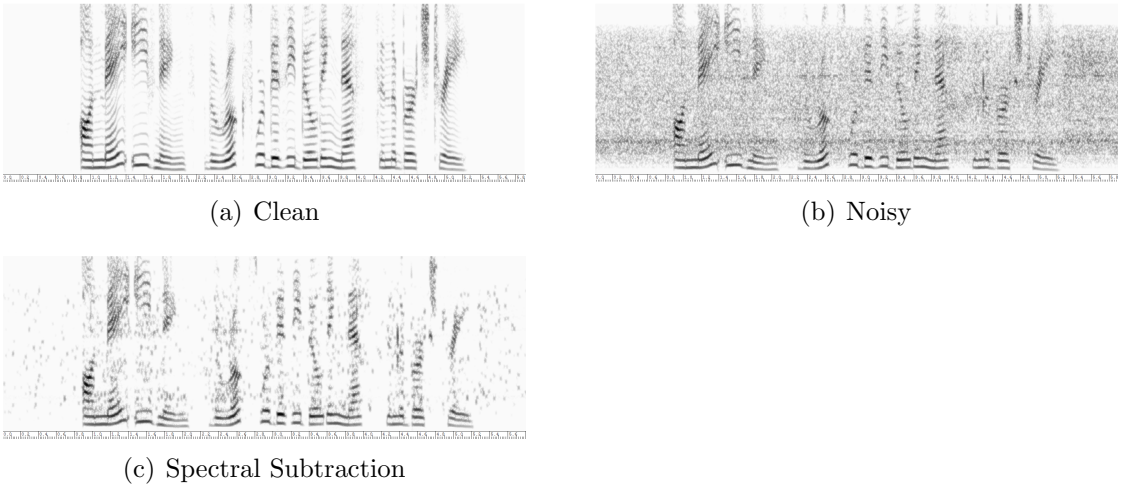
$$\sqrt[p]{|\hat{X}(k)|} = \sqrt[p]{f^{-1}(|Y(k)|^p, |\hat{N}(k)|^p)}, \tag{2.8}$$

where $p = 1$ denotes the magnitude spectrum and $p = 2$ denotes the power spectrum.

The resulting estimate of the clean speech spectrum may be negative in cases where the estimate of the noise is greater than the spectrum of the current frame. This is not valid and so half wave rectification can be applied to set negative values to zero, i.e:

$$\left|\hat{X}(k)\right| = \begin{cases} |Y(k)|^2 - |\hat{N}(k)|^2 & \text{if } |Y(k)|^2 > |\hat{N}(k)|^2 \\ 0 & \text{else} \end{cases}. \tag{2.9}$$

Whilst this approach will always give a valid magnitude spectrum half-wave rectification of the magnitude spectrum exposes random peaks causing artifacts in the reconstructed speech. The position of these peaks will vary frame-by-frame causing random tones to be heard in the enhanced signal. These tones are often known as



(a) Clean

(b) Noisy

(c) Spectral Subtraction

Figure 2.1: Narrowband spectrograms of utterance *"On May evening the rooks were busy building nests in the birch tree"* for a) clean speech, b) 10dB car noise and c) after applying spectral subtraction

'musical noise'. Figure 2.1 shows spectrograms of clean speech, noisy speech and speech enhanced by spectral subtraction to illustrate the effect of musical noise. Several alternatives to half-wave rectification have been proposed in the literature. One of these alternatives is to spectrally floor any negative spectral bins to a proportion of the noise signal estimate [Berouti et al., 1979]. The noise estimate is multiplied by an oversubtraction factor, α, and then subtracted from the noisy power spectrum. Any non-positive bins are then replaced by the noise estimate scaled by the spectral floor parameter, β:

$$\left|\hat{X}(k)\right|^2 = \begin{cases} |Y(k)|^2 - \alpha|\hat{N}(k)|^2 & \text{if } |Y(k)|^2 > (\alpha+\beta)|\hat{N}(k)|^2 \\ \beta|\hat{N}(k)|^2 & \text{else} \end{cases}. \tag{2.10}$$

This has the effect of enhancing high amplitude peaks, usually associated with speech, whilst leaving some noise in lower amplitude regions where the noise is less perceivable. The over-subtraction of the noise is intended to reduce the amplitude of broadband peaks leaving just a number of low amplitude narrowband peaks. These narrowband peaks are then masked by reintroducing a fraction of the noise estimate back on to the spectrum to fill-in the gaps between the remaining narrow-

band peaks. β controls the amount of residual noise and level of musical noise and α controlling the level of speech distortion. These parameters are typically determined either through experimentation or by forming an MMSE estimate of the optimal parameters [Sim et al., 1998]. Spectral-band or even spectral-bin level optimisation is also possible by calculating $\alpha(k)$ and $\beta(k)$ for all k.

Many tests examining both the quality and intelligibility of speech processed by various configurations of spectral subtraction-based methods have been carried out in the literature [Hu and Loizou, 2006; Vary, 1985]. Intelligibility was found to be mostly unaffected when speech enhanced using spectral subtraction was compared against noisy speech, though in some cases intelligibility was found to be slightly reduced. Overall quality and background noise intrusiveness were shown to be improved. Whilst the level of background noise can be significantly reduced, speech signal quality is shown to be slightly decreased.

2.2.2 Wiener filter

Wiener filtering is a method of conventional speech enhancement whereby the cleaned magnitude spectrum is derived based on the minimisation of the mean square error (MSE). The noise removal process is implemented as a filtering operation where the cleaned magnitude spectrum is computed as:

$$|\hat{X}(k)| = H(k)|Y(k)|, \tag{2.11}$$

where $H(k)$ is the kth component of the Wiener filter. Noise is again assumed to be additive and so $y(m) = x(m) + d(m)$ and the relationship between speech and noise in the power spectral domain is assumed to be:

$$|Y(k)|^2 = f(|X(k)|^2, |N(k)|^2) = |X(k)|^2 + |N(k)|^2. \tag{2.12}$$

The relationship between speech and noise in Equation 2.12 ignores the effect of cross-terms which are assumed to be zero on average. Section 4.5.2.1 examines this

relationship later in this thesis to determine the effect of this assumption.

One method of computing the Wiener filter is therefore:

$$H(k) = \frac{|X(k)|^2}{|X(k)|^2 + |N(k)|^2} = \frac{|X(k)|^2}{|Y(k)|^2} = 1 - \frac{|N(k)|^2}{|Y(k)|^2}. \tag{2.13}$$

This leads to the noise suppression function:

$$|\hat{X}(k)| = f^{-1}(|Y(k)|, |\hat{N}(k)|) = \left[1 - \frac{|\hat{N}(k)|^2}{|Y(k)|^2}\right] |Y(k)|. \tag{2.14}$$

Alternative methods of computing the Wiener filter values include an *a-priori* SNR based approach where the filter is given as:

$$H(k) = \frac{\xi_k}{\xi_k + 1}, \tag{2.15}$$

where ξ_k is the *a-priori* SNR of the kth frequency component and is computed as:

$$\xi_k = \frac{|X(k)|^2}{|N(k)|^2}. \tag{2.16}$$

From these equations it is clear that $H(k) \to 1$ for frequency components with high SNR, i.e. large values of ξ_k whilst $H(k) \to 0$ for low values of ξ_k. This is will result in regions of the signal with high SNR being emphasised whilst those with low SNR are attenuated. The challenge is therefore to compute the values of ξ_k. Scalart et al. [1996] proposed a method of *a-priori* SNR estimation by tracking the noise whilst several alternative methods have previously been proposed including an iterative approach by Lim and Oppenheim [1978] whilst an approach which tracked the noise using HMMs was developed by Ephraim et al. [1989]. More recently, Hadir et al. [2011] proposed the use of a model-based Wiener filter derived from log-Mel feature vectors. The feature vectors were enhanced using MMSE estimation and inverted to compute the filter response. The Mel filterbank used in the feature extraction processed caused the response of the Wiener filter to be smoothed over frequency which resulted in the fine spectral detail of the speech being retained whilst removing

the majority of the noise.

2.2.3 Statistical-model-based enhancement

Statistical-model-based methods of speech enhancement aim to derive the response of a noise suppression filter, $H(k)$, using statistical methods of estimation. There are three methods of statistical estimation commonly applied to this problem. These are: maximum likelihood (ML) estimation, minimum mean-square-error (MMSE) and maximum *a-posteriori* (MAP). Each of these methods are described in this section in the context of clean spectral amplitude estimation from noisy spectral amplitudes.

2.2.3.1 Maximum likelihood estimation

Maximum-Likelihood estimation is a widely used method of parameter estimation first applied to speech enhancement by McAulay and Malpass [1980]. Given a vector of noisy spectral amplitudes, $|\mathbf{Y}|$, we wish to estimate the most likely value of the clean spectral amplitudes, $|\mathbf{X}|$, that produced $|\mathbf{Y}|$. This is based on the assumption that whilst the relationship between $|\mathbf{X}|$ and $|\mathbf{Y}|$ is unknown, it is deterministic, i.e. not random. The most likely value of $|\mathbf{X}|$ is therefore computed by maximising the likelihood function, i.e.:

$$|\hat{\boldsymbol{X}}| = \arg\max_{|\mathbf{X}|} f(|\mathbf{Y}|; |\mathbf{X}|). \tag{2.17}$$

The maximum value is determined by differentiating the likelihood function and setting the derivative to zero. Assuming Gaussian distributions, this results in:

$$|\hat{\boldsymbol{X}}(k)| = \frac{1}{2}\left[|Y(k)| + \sqrt{|Y(k)^2| - |\hat{N}(k)^2|}\right], \tag{2.18}$$

where $|\hat{N}^2|$ is an estimate of the noise in the power spectral domain. This estimator can be expressed in terms of a filter, $H(k)$, whose frequency response is a function

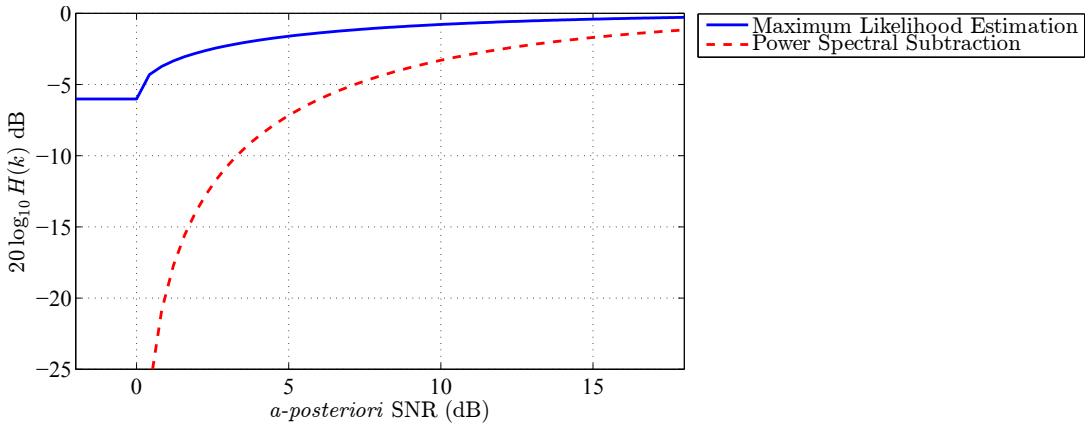


Figure 2.2: Level of noise attenuation of maximum likelihood and power spectral subtraction filters in terms of the *a-posteriori* SNR, $\gamma(k)$

of the *a-posteriori* SNR:

$$H_{ML}(k) = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{\gamma(k) - 1}{\gamma(k)}}, \tag{2.19}$$

where $\gamma(k)$ is the *a-posteriori* SNR and is computed as:

$$\gamma(k) = \frac{|Y(k)^2|}{|\hat{N}(k)^2|}. \tag{2.20}$$

Clean spectral amplitudes may then be estimated by filtering the noisy spectral amplitudes using $H(k)$:

$$|\hat{X}_{ML}(k)| = H_{ML}(k)|Y(k)|. \tag{2.21}$$

The response of the filter is now compared to the case of power spectral subtraction as a function of the *a-posteriori* SNR. The power spectral subtraction filter can be expressed in terms of $\gamma(k)$ as:

$$H_{PS}(k) = \frac{\gamma(k) - 1}{\gamma(k)}. \tag{2.22}$$

The response of $H_{PS}(k)$ and $H_{ML}(k)$ is displayed in Figure 2.2. The ML estimator is shown to attenuate very little of the noise and so is not particularly well suited

to speech enhancement. This is attributed to the lack of any prior knowledge of the speech distribution being accounted for in the process of estimation. The following methods both assume knowledge of *a-priori* distributions and as a result are shown to perform better.

2.2.3.2 Minimum mean square error

A method of estimation which minimises the mean square error (MSE) may be used to estimate the response of $H(k)$. The MMSE (minimum mean-square error) method of speech enhancement is a statistical estimation method that derives the response of the gain function using non-linear Bayesian estimation techniques. MMSE requires prior knowledge of the probability density functions (pdfs) of the speech and noise, and by taking into account this prior information the accuracy of the estimator is increased over the maximum-likelihood approach. This section begins by first describing the standard MMSE estimator. Second, a technique estimating log-spectral values, the log MMSE estimator, is covered later in the section.

The first stage of MMSE estimation is to form an appropriate expression of the mean-square error (MSE), i.e.:

$$e = E\left[(|\hat{X}(k)| - |X(k)|)^2\right]. \tag{2.23}$$

In the Bayesian approach the expectation is performed with respect to the joint pdf of the clean and noisy magnitude spectra and so the Bayesian MSE, B_{MSE} is defined as:

$$B_{MSE}(|\hat{X}(k)|) = \int\int (|\hat{X}(k)| - |X(k)|)^2 f(\mathbf{Y}, |X(k)|) d\mathbf{Y} d|X(k)|. \tag{2.24}$$

This function is minimised by differentiation and so the MMSE estimate of $|X(k)|$, $|\hat{X}(k)|$, is given as:

$$|\hat{X}(k)| = \int |X(k)| f(|X(k)|\,|\mathbf{Y}) d|X(k)|, \tag{2.25}$$

where $|\hat{X}(k)|$ is shown to depend on every coefficient of $\mathbf{Y}$ and the posterior pdf of $|X(k)|$ is given as:

$$f(|X(k)|\,|Y(k)) = \frac{f(Y(k)|\,|X(k)|)f(|X(k)|)}{f(Y(k))}. \tag{2.26}$$

By assuming statistical independence between coefficients the Bayesian MSE estimator can be simplified to:

$$|\hat{X}(k)| = \int x_k f(x_k|Y(k))dx_k = \frac{\int_0^\infty x_k f(Y(k)|x_k)f(x_k)dx_k}{\int_0^\infty f(Y(k)|x_k)f(x_k)dx_k}. \tag{2.27}$$

Whilst the MMSE estimator may be used to compute estimates of the clean speech magnitude spectrum it has no basis in the human listening process. The human ear has a logarithmic response to sound intensity and so an MMSE approach to estimation of the log-magnitude spectrum was therefore proposed by [Ephraim and Malah, 1985]. In this approach the MSE is defined as:

$$e_{log} = E\left[(\log(|\hat{X}(k)|) - \log(|X(k)|))^2\right] \tag{2.28}$$

and so the log MMSE estimator is:

$$\log(|\hat{X}|) = E[\log(|X(k)|)|Y(k)] \tag{2.29}$$

and so the estimate of the clean speech magnitude spectrum, $|\hat{X}|$, is computed as:

$$|\hat{X}| = \exp(E[\log(|X(k)|)|Y(k)]). \tag{2.30}$$

The gain function of the log MMSE estimator, $H(k)$ can then be proven to be:

$$H(k) = \frac{\xi(k)}{\xi(k)+1}\exp\left(\frac{1}{2}\int_{v(k)}^\infty \frac{e^{-t}}{t}dt\right), \tag{2.31}$$

where $\xi(k)$ is the *a-priori* SNR of the kth frequency bin and is computed as:

$$\xi(k) = \frac{|X(k)^2|}{|\hat{N}(k)^2|}. \tag{2.32}$$

$|X(k)^2|$ and $|\hat{N}(k)^2|$ are the power spectral values of the clean speech and noise, respectively. $v(k)$ is defined as:

$$v(k) = \frac{\xi(k)}{1+\xi(k)}\gamma(k), \tag{2.33}$$

where $\gamma(k)$ is the *a-posteriori* SNR defined as:

$$\gamma(k) = \frac{|Y(k)^2|}{|\hat{N}(k)^2|}. \tag{2.34}$$

The noise suppression filter, $H(k)$, may then be applied to the magnitude spectrum of the noise speech in the normal way, i.e.:

$$|\hat{X}(k)| = f^{-1}(|Y(k)|, \xi(k), \gamma(k)) = H(k)|Y(k)|. \tag{2.35}$$

The log MMSE filter is therefore applied as follows:

Step 1: Analysis Compute DFT coefficients of noisy speech

Step 2: Parameter estimation Estimate the *a-priori* and *a-posteriori* SNRs. The *a-posteriori* SNR is computed as $\gamma(k) = \frac{|Y(k)^2|}{|\hat{N}(k)^2|}$ whilst the *a-priori* SNR, $\xi(k)$, is computed using the method described by Ephraim and Malah [1984].

Step 3: Enhancement Compute the response of the filter $H(k)$ using $\gamma(k)$ and $\xi(k)$ and apply the filter to the magnitude spectrum of the noisy speech as $|\hat{X}(k)| = H(k)|Y(k)|$.

Step 4: Synthesis Combine $|\hat{X}(k)|$ with the phase of the noisy speech to give a modified complex spectrum and resynthesise speech signal using inverse DFT.

This approach has a significant advantage over spectral subtraction and Wiener

filtering. Speech enhanced using the log MMSE estimator was observed to contain significantly fewer artifacts (musical noise) compared to the ML estimator [Ephraim and Malah, 1985]. The reasons for this were attributed by Cappé [1994] to the effect of suppression as a function of the *a-priori* SNR. The *a-priori* SNR contributes most to noise suppression with the *a-posteriori* having relatively little influence. The ML estimator is a function only of the *a-posteriori* SNR and so attenuates relatively little of the noise which results in the musical noise.

2.2.3.3 Maximum *a-posteriori* estimation

The MMSE estimator is the mean of the *a-posteriori* pdf. If the *a-posteriori* pdf cannot be evaluated in closed form then it may be more appropriate to instead maximise this distribution to give the maximum *a-posteriori* (MAP) estimator, i.e.:

$$|\hat{X}(k)| = \arg\max_{|X(k)|} f\left(|X(k)||Y(k)\right). \tag{2.36}$$

In the case that the *a-posteriori* distribution is Gaussian the maximum (MAP) and the mean (MMSE) are identical and so the MAP and MMSE estimators are equal. Loizou [2007] gives more details regarding the MAP estimator for the cases where the pdf is non-Gaussian. In the case of spectral amplitude estimation the *a-posteriori* pdf is usually assumed Gaussian and so the MAP approach is not described.

2.3 Binary Time-Frequency Masking

Time-frequency masking-based methods of speech enhancement use a mask to remove the effect of noise from speech. Masks are time-frequency matrices of scaling factors and are applied to the spectrogram of the noisy speech as an element-wise multiplication as:

$$|\hat{X}(j,k)| = M(j,k)|Y(j,k)|, \tag{2.37}$$

where $M(j,k)$ is the value of the mask at the kth frequency of the jth frame of speech and $0 \leq M(j,k) \leq 1$. When $M(j,k)$ is allowed to take any value between 0 and 1 it is known as a 'soft-decision' mask and can be seen to be equivalent to the conventional filtering methods previously described. Alternatively, $M(j,k)$ can be applied as a binary mask where:

$$M(j,k) = \begin{cases} 1 & \text{if } \textit{speech} \\ 0 & \text{else} \end{cases} . \tag{2.38}$$

This has the effect of removing regions of non-speech whilst retaining spectral components related only to the speech. After application of this binary mask the retained speech amplitudes will still be affected by the noise, however this method has been found to be effective at increasing the intelligibility of speech [Kim et al., 2009]. The ideal binary mask is used by many as a benchmark for optimal performance of this method. The ideal binary mask is computed by measuring the *a-priori* SNR at each time-frequency component and setting a cut off at the point where the noise is more powerful than the speech:

$$M(j,k) = \begin{cases} 1 & \text{if } 10\log_{10}(\frac{|X(j,k)|^2}{|N(j,k)|^2}) > 0 \\ 0 & \text{else} \end{cases} , \tag{2.39}$$

where $10\log_{10}(\frac{|X(j,k)|^2}{|N(j,k)|^2})$ is the instantaneous *a-priori* SNR in decibels. The *a-priori* SNR is often unknown and so must be estimated. Ephraim and Malah [1984] proposed a method of estimation using a gain function and *a-posteriori* SNR:

$$\hat{\xi(j,k)} = \alpha \frac{(H(j-1,k)|Y(j-1,k)|)^2}{|\hat{N}(j-1,k)|^2} + (1-\alpha)\max(\gamma(j,k)-1,0), \tag{2.40}$$

where $\alpha = 0.98$ and $H(j-1,k)$ is a gain function as defined earlier. $\gamma(j,k)$ is the *a-posteriori* SNR as defined in Equation 2.34 whilst $|Y(j-1,k)|$ and $|N(j-1,k)|$ are the magnitude spectra of the previous frames of the noisy speech and the estimate of the noise. The frequency response of the gain function (or filter)

may be determined using any one of the previously defined estimators, i.e. Spectral Subtraction, Wiener, ML, MMSE, logMMSE or MAP. Such a method of estimating the *a-priori* SNR clearly relies on an accurate estimate of the noise as well the chosen gain function. Hu and Loizou [2008] therefore tested a range of gain functions and noise estimators in order to determine the optimal configuration. Performance of the MMSE-based methods was found to be best with either the VAD-based or MCRA2 noise estimators [Loizou, 2007].

Performance of the best method of estimating the binary mask as determined by Hu and Loizou [2008] was tested by Jensen and Hendriks [2011] in terms of objective quality as measured using PESQ (Section 2.6.2.3), and objective intelligibility as measured using STOI (Section 2.6.4). The binary mask method was compared against Ephraim and Malah's MMSE spectral estimator described in the previous section [Ephraim and Malah, 1984]. The MMSE spectral estimator was shown to improve the quality of speech versus the noisy speech in terms of PESQ results whilst the binary mask reduced the quality of speech. In terms of intelligibility the binary mask was shown to improve performance relative to the noisy speech but was still outperformed by the MMSE spectral estimator.

2.4 Subspace Enhancement

The methods of speech enhancement described in the previous sections have assumed that the effect of noise on speech can be removed by filtering the signal in some way to remove an estimate of the noise. Subspace methods of speech enhancement take a different approach in assuming that speech occupies a small subspace of the overall space of the noisy speech, whilst white noise occupies the entire space. By identifying and removing the subspace that is exclusively occupied by the noise and resynthesising the modified frames the effect of the noise should be removed. In practise however the noise also affects the space occupied by the speech and so further processing is required to completely remove the noise. Typically, subspace

methods of enhancement take a three stage approach of i.) separating the subspaces of the noise and clean speech plus noise subspaces, ii.) removing the noise subspace and iii.) post-processing the clean speech plus noise subspace to remove the effect of noise from the clean speech. The second stage removes the effect of the noise without modification of the speech signal. Despite this, the post-processing stage has been found to be important for improved removal of the noise, however this often introduces speech distortions due to modification of the subspace occupied by the speech signal [Hermus and Wambacq, 2006].

A method of subspace enhancement proposed by Hu and Loizou [2003] is now described. This method of enhancement assumes additive noise where the noisy signal is defined as $y(m) = x(m) + n(m)$. A frame-based approach is taken whereby each frame is of sufficiently short duration so the signal may be assumed stationary. A linear model of x is defined as:

$$\mathbf{x} = \mathbf{\Psi} \cdot \mathbf{s}, \tag{2.41}$$

where $\mathbf{\Psi}$ is a rank deficient $K \times M$ matrix with rank M where $M < K$ and $\mathbf{s}$ is $M \times 1$. $\mathbf{\Psi}$ must be rank deficient to allow the separation of the subspaces occupied by the speech and by the noise [Hermus and Wambacq, 2006]. A linear estimator may be computed from this linear model [Loizou, 2007], of the form:

$$\hat{\boldsymbol{x}} = \mathbf{H} \cdot \mathbf{y}, \tag{2.42}$$

where the optimal estimator, $\mathbf{H}$, is defined as:

$$\mathbf{H} = \mathbf{R_x}(\mathbf{R_x} + \mu\mathbf{R_n})^{-1}, \tag{2.43}$$

where $\mathbf{R_x}$ represents the covariance matrix of the clean speech and $\mathbf{R_n}$ represents the covariance matrix of the noise. $\mathbf{R_n}$ may be estimated from non-speech portions of the signal, however $\mathbf{R_x}$ is not available and so an alternative approach of estimating $\mathbf{H}$ must be taken.

A matrix $\mathbf{\Sigma}$ is defined as:

$$\mathbf{\Sigma} = \mathbf{R_n^{-1}} \mathbf{R_y} - I, \tag{2.44}$$

where I is the identity matrix and $\mathbf{R_y}$ is the covariance matrix of the noisy speech. The eigenvalue and eigenvectors of $\mathbf{\Sigma}$ are then computed using eigenvector decomposition (EVD) to give the relationship:

$$\mathbf{\Sigma V} = \mathbf{V \Lambda_x}, \tag{2.45}$$

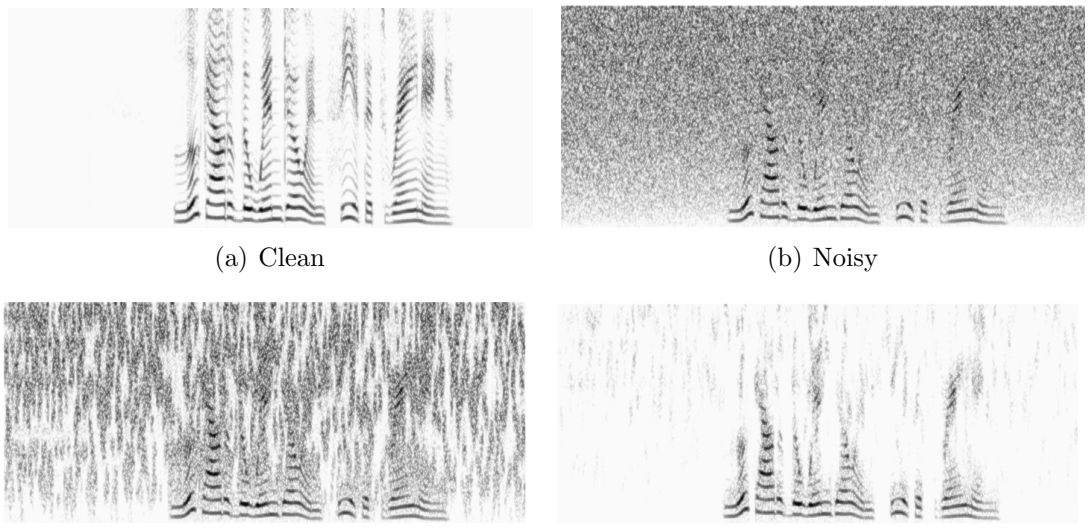
where $\mathbf{V}$ denotes the eigenvectors of $\mathbf{\Sigma}$ and $\mathbf{\Lambda_x}$ the eigenvalues. The noise subspace is nulled by setting the non-positive eigenvalues to zero based on the assumption that the signal is represented by the largest eigenvalues. The signal may then be resynthesised by defining the estimator, $\mathbf{H}$, as:

$$\mathbf{H} = \mathbf{V}^{-T} \mathbf{G} \mathbf{V}^{T}, \tag{2.46}$$

where $\mathbf{G}$ is a $K \times K$ matrix with diagonal elements:

$$G(k,k) = \begin{cases} 1 & \text{for } \Lambda(k,k) > 0 \\ 0 & \text{else} \end{cases} \qquad \text{for } k = 1 \ldots K. \tag{2.47}$$

Applying this estimator to the noisy speech using Equation 2.42 allows resynthesis of a modified speech signal exclusive of noise subspace. The resynthesised subspace, the speech plus noise subspace, is still be affected by noise and so further processing is usually required for good quality speech. Removal of the effect of noise from the speech plus noise subspace can be achieved using one of the filters described in the previous section. In this case, the Wiener filter is used to process the speech plus



(a) Clean (b) Noisy

(c) Noise subspace nulling (d) Noise subspace nulling plus filtering

Figure 2.3: Narrowband spectrograms showing: a.) clean speech, b.) noisy speech (white noise at 5dB SNR), c.) the effect of subspace nulling on the noisy speech and d.) the effect of subspace nulling followed by filtering

noise subspace as:

$$G(k,k) = \begin{cases} \frac{\Lambda(k,k)}{\Lambda(k,k)+\mu} & \text{for } \Lambda(k,k) > 0 \\ 0 \quad \text{else} \end{cases} \qquad \text{for } k = 1 \ldots K, \tag{2.48}$$

where μ is the Lagrange multiplier with values:

$$\mu = \begin{cases} \mu_0 - (SNR_{dB})/s \quad \text{for } -5 < SNR_{dB} < 20 \\ 1 \quad \text{for } SNR_{dB} \geq 20 \\ 5 \quad \text{for } SNR_{dB} \leq -5 \end{cases} \tag{2.49}$$

determined by Hu and Loizou [2003] where $\mu_0 = 4.2$ and $s = 6.25$. The optimal linear estimator, $\mathbf{H}$, can then be computed as per Equation 2.46 and subsequently applied as per Equation 2.42. This simultaneously nulls the noise subspace (zero diagonal elements of $\mathbf{G}$) whilst attenuating the noise in the speech plus noise subspace.

The effect of nulling the noise subspace and subsequently filtering the speech plus noise subspace is now illustrated in Figure 2.3. White noise was added to an utterance of speech spoken by a female speaker at an SNR of 5dB. The utterance *"Look out of the window and see if it's raining"* was used. Figure 2.3(a) shows the narrowband spectrogram of the clean speech whilst Figure 2.3(b) shows the noisy speech. The effect of nulling the subspace and retaining the signal plus noise subspace is shown in Figure 2.3(c). Some of the noise has been removed, however a large proportion of the noise remains. This is attributed to the effect of the noise retained in the speech plus noise subspace. Finally, the effect of nulling the subspace and then filtering the speech plus noise subspace is shown in Figure 2.3(d). Almost all of the noise has been removed in this case though some musical noise remains in the signal. High frequency, low SNR components of the signal are clearly missing in the enhanced signal whilst some inter-harmonic noise is also present however this is relatively low in amplitude and therefore likely to be masked by the harmonics.

2.5 Speech Enhancement by Reconstruction

Reconstruction model-based methods of speech enhancement operate similarly to the conventional analysis-enhancement-synthesis approach. Instead of directly resynthesising speech through the use of an inverse Fourier transform speech is reconstructed using a reconstruction model driven by a set of acoustic features. This gives a three-stage approach of enhancement of: i.) acoustic feature extraction (analysis), ii.) acoustic feature enhancement and iii.) speech reconstruction using enhanced acoustic features (synthesis). Alternatively, speech reconstruction models can be used as a post-processing stage to reduce the effect of artifacts caused by conventional enhancement, e.g. musical noise. This gives an approach of: i.) conventional speech enhancement, ii.) acoustic feature extraction, iii.) speech reconstruction.

Speech reconstruction models were primarily developed for the purposes of channel coding and speech modification but have several attractive properties that make

them suitable for speech enhancement. The main benefit of using a speech reconstruction model as opposed to direct resynthesis is the constraints applied to the reconstructed signal. Models of reconstruction are designed to only reconstruct components of the signal that relate to the speech and so artifacts, including musical noise, that result from inaccurate spectral envelope estimation are not reconstructed. This also leads to some noise being inherently removed by the reconstruction model in a similar fashion to previous filtering-based approaches such as those by Hanson et al. [1983] and Nehorai and Porat [1986] where an adaptive comb-filter was adaptively adjusted to follow the harmonics of the speech.

We first consider the the application of speech reconstruction models as a post-filter for conventional methods of speech enhancement. The earliest known application of such an approach was proposed by Kang and Fransen [1989] who used spectral subtraction as a noise reduction stage before reconstructing speech using the LPC vocoder. Later, Guilmin et al. [1999] published a similar method this time using Wiener filtering for noise reduction. These methods were shown to be effective with the speech reconstruction models reducing the effect of artifacts caused by the conventional methods of noise reduction. More recently Zavarehei et al. [2007] developed a method of post-processing conventionally enhanced speech by reconstructing regions of the speech spectra distorted by noise reduction. This is achieved through the use of the Harmonic plus Noise (HNM) reconstruction model to reconstruct damaged harmonics [Stylianou, 2001]. The HNM reconstructs speech as a sum of harmonic sinusoids modulated by amplitude and frequency and offset for relative phase:

$$s(m) = \sum_{l=1}^{L} A(lf_0)\cos(2\pi l f_0 m + \theta(lf_0)) + n(m), \tag{2.50}$$

where $s(m)$ is the mth sample of the reconstructed signal, L is the number of harmonics and $A(lf_0)$ is the value of the spectral envelope sampled at the lth harmonic where f_0 is fundamental frequency. Finally, $\boldsymbol{\theta}$ represents the phase spectrum and $n(m)$ is filtered noise. This structure ensures only speech energy is reconstructed

in voiced frames. The method of post-filtering developed by Zavarehei et al. [2007] tracks harmonic amplitudes and frequencies ($A(lf_0)$ and lf_0, respectively) and recovers missing or damaged harmonic components through the use of a codebook trained on uncorrupted clean speech. An approach of corrupted speech reconstruction was also taken by Krini and Schmidt [2009] for the purpose of removing noise from speech recorded from in-car environments. In such environments low frequency harmonics are often subject to much lower SNRs than high frequency components due to engine and wind noise. A conventional speech enhancement method is first applied. Spectral envelope is then extracted and smoothed using an IIR filter. A codebook is used to enhance low-SNR regions of speech. In the case of voiced speech the signal is then reconstructed at harmonic frequencies through the use of an inverse Fourier transform to give a reconstruction model similar to that of the HNM.

The HNM reconstruction model has also been successfully applied as a method of enhancement by directly reconstructing speech rather than as a method of post-filtering. Typically, a method of conventional enhancement is used for spectral envelope enhancement before the reconstruction model is directly applied for resynthesis. An example of such an approach was proposed by Jensen and Hansen [2001] where the acoustic features required for reconstruction were estimated through an iterative process of Wiener filtering for noise reduction and an analysis stage of updating acoustic features. A similar approach was proposed by Moharir et al. [2002] who used spectral subtraction to pre-process spectral envelope before reconstruction. More recently, Chen et al. [2012] applied a more advanced framework for acoustic feature estimation. The HNM was again used for reconstruction. Fundamental frequency and voicing were estimated from a pre-cleaned speech signal whilst spectral envelope was estimated through the use of a method of time-frequency tracking and modification of LSFs extracted from the pre-cleaned speech signal. In all cases significant noise reduction was achieved with no musical noise present in the reconstructed signal, though listening tests performed by Chen et al. [2012] showed some degree of signal distortion to have been introduced.

2.6 Measuring Performance

The evaluation of speech signals plays an important role in this work. Noise degrades the speech signal and we wish to reduce this degradation. To measure the effectiveness of speech enhancement techniques we therefore require a method of measuring the severity of the degradation between the original clean speech and the noisy signal and also between the original and enhanced signals.

Evaluation methods can be categorised as measuring either speech quality or intelligibility. A speech signal may be free of noise and of good 'quality' but be unintelligible whilst the introduction of noise or other processing distortion may reduce the quality of the speech but still be fully intelligible. These categories are further subdivided into objective methods and subjective methods. Subjective methods use human listeners who are presented with a range of utterances and asked to respond to a series of questions relating to the quality or intelligibility of the signals. Subjective measurement of performance is expensive and time consuming, with many listeners required for accurate results. Instead, objective measures are designed to emulate subjective tests with the use of digital signal processing (DSP).

Ideally, methods of objectively measuring quality and intelligibility will have high correlation with subjective results, however this is not always the case. We therefore examine a range of subjective and objective methods of evaluation. This work is based on the comprehensive review of methods carried out by Loizou [2007].

2.6.1 Subjective quality measures

The ultimate objective of this method of speech enhancement is to improve the quality of processed speech whilst retaining intelligibility by removing the effect of noise. The quality of speech is ultimately determined by the users of the system and so subjective evaluation is of particular importance. Subjective quality experiments are conducted as listening tests in which a range of listeners are asked to rate utterances based on one, or a number, of performance or preference metrics. There

Table 2.1: Comparison Category Rating (CCR) rating scale

Category	Score
Much Better	+3
Better	+2
Slightly Better	+1
About the Same	0
Slightly Worse	-1
Worse	-2
Much Worse	-3

are many types of listening test and these can be categorised as either relative preference methods or absolute category rating methods [Loizou, 2007].

2.6.1.1 Relative preference methods

Relative preference methods measure the relative quality of speech. Users are asked to compare processed utterances to either reference utterances or those processed using alternative methods. The isopreference test was one of the earliest methods of relative performance measurement [Munson and Karlin, 1962]. In this system 'Transmission Preference Units' (TPU) were used to rate the quality of processed speech compared to ideal conditions, i.e. speech recorded in clean conditions with no processing distortion. A similar approach, the comparative mean opinion score (CMOS), was standardised by ITU [1996] as P.830. In this method users are presented with two utterances and asked to compare them based on a comparison category rating (CCR). This rating system consists of seven categories, ranging from 'much better' to 'much worse', which are listed in Table 2.1. Such testing answers the question of which method is preferable, and in some cases by how much, but does not answer the question as to *why* this is the case.

2.6.1.2 Absolute category rating methods

Absolute category rating (ACR) methods are designed to determine the overall quality of utterances measured in isolation. Unlike relative performance measures,

Table 2.2: Mean Opinion Score (MOS) rating scale

Category	Score
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

ACR methods typically measure a range of signal properties to determine why one method may be preferred over another. There are three main methods of ACR: mean opinion score (MOS), diagnostic acceptability measure (DAM) and the ITU P.835 3-point MOS (3MOS) test. These three methods are now summarised.

MOS The MOS test requires that listeners hear each utterance in isolation and are asked to rate it on a five-point scale as listed in Table 2.2. Results for each condition are then averaged to form the mean opinion score. Scores are normalised through the use of a training stage. Listeners first hear a number of utterances which are judged to relate to the extremes of the scale as well as the middle point. Only then are listeners asked to rate the utterances which contribute to the final results. This training phase is particularly important as it ensures listeners are aware of what constitutes a 'good' utterance and a 'bad' utterance.

DAM The MOS test, like the relative performance measures, provides a rating of quality but does not give any insight as to *why* those ratings were given. The DAM therefore asks listeners to rate each utterance across 22 categories to give a multi-dimensional result describing more accurately how the signal is perceived [Voiers, 1977]. Table 2.3 displays the scales used in DAM tests. In each case the listener is asked to rate the utterance in terms of a particular property, i.e. as part of the tests listeners will be asked to rate the signal in terms of how 'rasping' or 'distant' it sounds on a scale of 0 to 100. Clearly such testing has the potential to provide fine-grained evaluation of speech signals though this comes with a significant disadvantage. In order to obtain reliable

Table 2.3: Diagnostic Acceptability Measure (DAM) rating scales

Parametric scales		
Name	Abbreviation	Descriptor
Signal	SF	Fluttering, bubbling
	SH	Distant, thin
	SD	Rasping, crackling
	SL	Muffled, smothered
	SI	Irregular, interrupted
	SN	Nasal, whining
	TSQ	Total signal quality
Background	BN	Hissing, rushing
	BB	Buzzing, humming
	BF	Chirping, bubbling
	BR	Rumbling, thumping
	TBQ	Total background quality
Metametric scales		
	I	Intelligibility
	P	Pleasantness
Isometric scales		
	A	Acceptability
	CA	Composite acceptability

results a large number of experienced listeners are required, with each test taking a considerable amount of the listeners' time. This makes such testing very expensive.

3MOS The 3-way MOS test is an extension of the standard MOS test and was standardised by the ITU as P.835. 3MOS testing splits the standard MOS test into three separate scales which measure background intrusiveness (BAK), signal distortion (SIG) and overall quality (OVL). Listeners hear each utterance three times and are asked to use a different scale each time. The overall quality is measured as per the standard MOS scale whilst background and signal quality are rated on the five-point scales displayed in Tables 2.4 and 2.5. This allows the contribution of background noise and signal distortion to overall quality to be directly measured at considerably less expense than using DAM tests.

Table 2.4: Background intrusiveness rating scale (BAK)

Category	Score
Not noticeable	5
Somewhat noticeable	4
Noticeable but not intrusive	3
Fairly conspicuous, somewhat intrusive	2
Very conspicuous, very intrusive	1

Table 2.5: Signal quality rating scale (SIG)

Category	Score
Very natural, no degradation	5
Fairly natural, little degradation	4
Somewhat natural, somewhat degraded	3
Fairly unnatural fairly degraded	2
Very unnatural, very degraded	1

2.6.2 Objective quality measures

Listening tests are expensive to run and so for practical evaluation of methods it would be beneficial to have a method of approximating subjective quality through the use of objective measurements. The aim of objective evaluation is therefore to maximise the correlation between subjective and objective measurements. A number of objective quality measures have been developed, most of which are based on simple difference measures between signals in either the time or frequency domain. This is effective at measuring the effect of noise on speech (i.e. the SNR), however it is not necessarily optimal to measure all types of signal distortion in this way. In particular, when measuring the quality of reconstructed speech it is important to take into account that the waveform of reconstructed speech can vary significantly from the original signal whilst remaining perceptually similar to the original signal due to small variations in fundamental frequency and phase. Objective quality measures based on simple difference measurements are therefore expected to be unlikely to give reliable results when measuring reconstructed speech.

In this section a number of objective quality measures are evaluated. These include: segmental signal to noise ratio (SNR), log likelihood ratio (LLR) and percep-

tual evaluation of speech quality (PESQ). The most appropriate method will provide high correlation with listening test results for both clean and degraded speech and also be robust to imperceivable changes in the signal caused by reconstruction.

2.6.2.1 Segmental signal to noise ratio (SNR)

Segmental SNR is one of the most basic objective measures and is based on a simple mathematical difference. The speech signal is split into frames and the SNR of each frame is measured in either the time or frequency domain. The overall rating is then calculated as the mean of the SNR of all frames as defined in Equation 2.51, where M is the number of frames, N is the number of samples in the original signal, x, and $\hat{x}$ is the processed signal [Loizou, 2007].

$$SNRseg = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}(x(n)-\hat{x}(n))^2}. \tag{2.51}$$

This method provides an accurate measure of the SNR for traditional enhancement methods, however due to the simplicity of the distance measure the signals must be perfectly aligned in time and phase. Speech reconstruction models rely on fundamental frequency estimates that are not always accurate and in some methods the phase is replaced entirely and so it is expected that the segmental SNR measure will provide particularly poor results when used to measure reconstructed speech. The segmental SNR also does not take into account any perceptual properties of speech and so methods such as HNM that synthesise only perceptually important components, i.e. harmonics, are also likely to give poor results even if perfectly accurate fundamental frequency and phase estimates are used.

2.6.2.2 Log likelihood ratio (LLR)

The LLR measure is based on an LPC representation of the spectral envelope. All-pole models of the clean and processed speech are constructed and a distance measure is formed as per Equation 2.52, where $\mathbf{R}_x$ is the autocorrelation matrix of

the original signal, $\mathbf{a}_x$ is the vector of LPC coefficients of the original signal and $\bar{\mathbf{a}}_{\hat{x}}$ is the vector of LPC coefficients of the processed signal [Loizou, 2007].

$$d_{LLR}(\mathbf{a}_x, \bar{\mathbf{a}}_{\hat{x}}) = \log \frac{\bar{\mathbf{a}}_{\hat{x}}^T \mathbf{R}_x \bar{\mathbf{a}}_{\hat{x}}}{\mathbf{a}_x^T \mathbf{R}_x \mathbf{a}_x}. \tag{2.52}$$

An alternative form of this measure is shown in equation 2.53, where A_x and $\bar{A}_{\hat{x}}$ are the spectral representations of $\mathbf{a}$ and $\bar{\mathbf{a}}_{\hat{x}}$. In this form it can be considered similar to the frequency domain segmental SNR measure (Section 2.6.2.1) in that it is a simple measure of the differences between spectral envelopes.

$$d_{LLR}(\mathbf{a}_x, \bar{\mathbf{a}}_{\hat{x}}) = \log \left(1 + \int_{-\pi}^{\pi} \left| \frac{A_x(\omega) - \bar{A}_{\hat{x}}(\omega)}{A_x(\omega)} \right|^2 d\omega \right). \tag{2.53}$$

When $A_x(\omega)$ is large, spectral differences will result in a higher score, penalising differences in such areas. This is perceptually advantageous as these high amplitude regions are typically located around formant locations, suggesting any differences in formant locations or amplitudes will be penalised heavily.

All reconstruction models aim to preserve the spectral envelope and so it is predicted that LLR scores will correlate well with subjective results. As LLR is based on the spectral envelope, as opposed to the magnitude or power spectrum, small differences in fundamental frequency and phase are unlikely to have a significant effect on results, though it is still important to ensure that utterances are perfectly time-aligned.

2.6.2.3 Perceptual evaluation of speech quality (PESQ)

PESQ is an objective measure designed to overcome some of the issues encountered by previously developed measures. It is designed primarily for voice over IP (VoIP) applications where the signal could be affected by packet loss, delay and codec distortion [Loizou, 2007].

Before a distance measurement is calculated the input signals are normalised and

time-aligned in a pre-processing stage to overcome issues caused by delay and non-matching gains. Signals are also filtered using an Intermediate Reference System (IRS) filter to model the frequency response of a standard telephone handset.

After pre-processing, the signals are compared using a perceptually motivated distance measure. The signals are first split into a number of 32 msec frames. The power spectrum of each frame has a bark scale filterbank consisting of 42 bands applied and then a loudness spectrum is produced after further frequency and gain equalisation stages. A simple difference between the signals is then calculated. Unlike most other objective measures, positive and negative differences are treated differently. Negative differences relate to components being added to the signal, for example noise. Positive differences would suggest that the signal has been attenuated in some way. Additive noise sources are seen to be more of an audible nuisance than distortion caused by attenuation because of masking effects in the human hearing process. This is true of the sinusoidal model, where inter-harmonic regions are completely discarded with little audible difference.

The disturbance values calculated from the differences between the signals are then used to form a single score by producing an average disturbance value per frame and then linearly combining frame scores to produce an overall disturbance. The overall disturbance score is then scaled to within the range of 1.0 and 4.5 to produce a score which can be compared to MOS listening test results.

It is expected that PESQ will provide a good correlation with subjective tests when rating reconstructed speech. Typically, objective measures penalise missing frequency components, however, it is possible that the reduced weighting applied to attenuated components will produce a rating that correlates well with subjective results. In previous tests PESQ has been shown to rate speech processed by noise suppression algorithms lower than subjective tests [Ditech Networks, 2007]. For these reasons it will be particularly interesting to evaluate how PESQ performs against MOS listening tests in both clean and noisy conditions when speech is synthesised using a speech reconstruction model.

2.6.3 Subjective intelligibility measures

The objective of speech is communication and so the processing of speech should not reduce the understanding or distort the message in any way. The measurement of intelligibility is therefore an important aspect of performance evaluation with the most valuable results coming from end-user evaluation of intelligibility. A number of tests have been designed to measure intelligibility and these include: nonsense syllable tests, word tests, phonetically balanced word tests, rhyming word tests and sentence tests. In each case listeners are asked to listen to a word, or sequence of words, and identify what they heard [Loizou, 2007]. These tests are described as follows:

Nonsense syllable tests The earliest form of intelligibility tests were proposed by Fletcher and Steinberg [1930]. Nonsense words constructed using three phones in the format consonant-vowel-consonant (/C-V-C/) were read to listeners who were asked to identify what was spoken. Later, Miller and Nicely [1955] refined the test to use only consonants that most often occur in fluent speech. These consonants were also corrupted at varying levels of noise before being presented to listeners. Nonsense syllable tests measure performance of speech enhancement algorithms in terms of their ability to process individual phonemes but do not provide a realistic measure of intelligibility in real-world scenarios.

Word tests There are two main categories of word test. First, phonetically balanced word tests use a carefully selected list of words from which to measure intelligibility. Egan [1948] constructed 20 lists of 50 common English monosyllabic words. Each list is designed to be of equal difficulty, phonetic content and phonemic distribution (i.e. phonetically balanced). The careful selection of words is important to achieve a useful measure of performance: if the test is too easy results will suffer from the 'floor effect' whereby all tests score 100% intelligibility whilst at the other end of the scale if the test is too difficult

all tests will score 0% [Loizou, 2007]. Rhyming word tests were proposed by Fairbanks [1958] as an alternative to phonetically balanced word tests. Early variants of this method of testing were similar to nonsense syllable tests in that words were chosen that matched the /C-V-C/ format with listeners asked to identify the first consonant only given the remaining letters. Given the example 'dot', the listener would be given the letters '_ot' and asked to identify the first consonant. Alternative rhyming words for this example include *cot, got, hot, not, rot.* The Diagnostic Rhyme Test (DRT) was developed as a refinement to such tests and forms the basis of intelligibility testing of most recent algorithms [Voiers, 1983].

Sentence tests Word tests are useful to identify intelligibility in isolated cases but do not take into account the contextual information available in conversational speech. Sentence tests are designed to measure intelligibility for conversational speech using carefully structured sentences. Examples of sentence tests include the Speech Perception in Noise (SPIN) [Kalikow et al., 1977] tests and Hearing in Noise Test (HINT)[Nilsson et al., 1994].

2.6.4 Objective intelligibility measures

Subjective measurement of intelligibility is expensive and time consuming and so it is useful to estimate intelligibility through the use of objective intelligibility tests. The purpose of objective intelligibility testing is to predict the intelligibility of an utterance automatically. A number of methods have been proposed and are summarised as follows:

Articulation index (AI) The AI was one of the first methods of automatic prediction of intelligibility and was developed to quantify speech intelligibility over telephone networks [French and Steinberg, 1947]. Later, this method was adapted to predict the intelligibility of speech for patients with hearing loss [Kryter, 1962]. This method works by measuring signal intensity relative

to background sound levels, i.e. the SNR. The SNR is measured across twenty frequency bands and weighted based on the degree to which that frequency band is expected to contribute to intelligibility.

Speech intelligibility index (SII) This measure evolved from the AI and was standardised as ANSI S3.5-1997 [ANSI, 1997]. SII is calculated in approximately the same way as AI but is more flexible in terms of the number of frequency bands which comprise the overall measurement as well as a number of correction factors used to correct for effects such as spectral masking.

Speech transmission index (STI) Measurement of the STI takes into account a number of distortions which can affect intelligibility. These include: speech level, frequency response of the channel, non-linear distortions (i.e. waveform clipping), background noise, echos and reverberations [Steeneken and Houtgast, 1980]. STI predicts the likelihood of utterances being comprehended in terms of syllables, words and sentences on a numerical scale between 0 and 1. A reference scale was introduced by Barnett and Knight [1996] which categorises these ratings into a five-point scale ranging from 'bad' to 'excellent'. A number of other intelligibility measures based on the STI have been developed including those by Rhebergen and Versfeld [2005] and Kates and Arehart [2005].

Short-time objective intelligibility measure (STOI) STOI, developed by Taal et al. [2010] predicts intellibility based on the measurement of 15 frequency bands in a similar way to the AI and SII measures. The signal to distortion ratio (SDR) of each frequency band is measured. This requires access to the original utterance. The intelligibility of each frequency band is then computed as an estimate of the linear-correlation coefficient between clean and modified speech. A weighted average of these estimates is then taken to form the overall measurement of intelligibility, ranging from 0 (unintelligible) to 1 (fully intelligible). This method was found to correlate strongly with subjective listening tests with a correlation of $R = 0.95$ reported in the original paper.

2.6.5 Summary

Objective quality and intelligibility measures have been described as being designed to simulate the results of subjective tests. These algorithms are often based on models of the human auditory processes, however perfect correlation between objective measures and subjective measures has not yet been achieved. As such, subjective evaluation will always be an important tool for measuring performance. Subjective performance evaluation is more time consuming than objective measurement and so in this work objective performance measures will be used for system development whilst performance of the overall system will be measured subjectively.

In terms of objective evaluation, LLR (Section 2.6.2.2 was shown to measure spectral envelope distortion and so will be used to measure spectral envelope estimation accuracy whilst PESQ will be used to predict the performance of the overall system due to its high correlation with subjective MOS testing. A 3-way MOS test (Section 2.6.1.2) will then be used to measure overall performance subjectively.

Chapter 3

Speech Reconstruction

In this chapter a range of speech reconstruction models are examined with the objective of finding a suitable method of reconstruction for the proposed method of speech enhancement. The chapter begins by looking at the process of speech production to identify the properties of the signal which must be preserved for high quality, intelligible, speech reconstruction. A number of speech reconstruction models and methods of encoding the required acoustic features are described, whilst results of experiments which are used to determine the optimal speech reconstruction model and feature configuration are presented.

Contents

3.1 Introduction

Developing a high quality method of speech reconstruction is important in this method of speech enhancement as the overall quality of speech enhanced using the system is directly linked to the quality of the reconstruction model as well as the accuracy of the estimated features. The reconstruction model used in this work must therefore have two attributes: firstly, the model must reconstruct high quality speech from a set of acoustic features and, secondly, these acoustic features must be obtainable from noisy speech.

Many methods of speech reconstruction exist, with most aiming to directly model the human speech production process and have been developed primarily for the purpose of speech coding [Spanias, 1994] and speech synthesis [Macon and Clements, 1996]. This chapter therefore starts with a description of this process in Section 3.2 which highlights not only the process itself but also the challenges in developing a good model of reconstruction.

Next, four of the most widely used reconstruction models are reviewed in Section 3.3. These are: LPC vocoder [Kondoz, 2004], sinusoidal model [McAulay and Quatieri, 1986], HNM [Stylianou, 2001] and STRAIGHT [Kawahara et al., 1999]. Each model is evaluated in terms of the quality of reconstructed speech and the acoustic features required for reconstruction. Ultimately, the acoustic features of clean speech will be estimated from features obtained from noisy speech. This process will limit the amount of information that may be reliably obtained and thus the trade-off between overall quality and feature complexity is also considered.

Common to all speech reconstruction methods is the requirement for spectral amplitudes. Linear predictive coding (LPC) coefficients are commonly used to model the spectral envelope in many speech encoding and transmission applications such as VoIP. Whilst LPC coefficients have been proven to be effective in clean conditions, they are not robust to noise and so a number of alternative features are also considered. These include: spectrum-based features (i.e. magnitude and power

spectrum), alternative LPC-based methods such as line spectral frequencies (LSF) and filterbanks.

Finally, results of experiments determining the most suitable speech reconstruction model and spectral feature configuration are presented in Section 3.5. The reconstruction model and feature configuration chosen form the basis of all future feature estimation and speech reconstruction in this work.

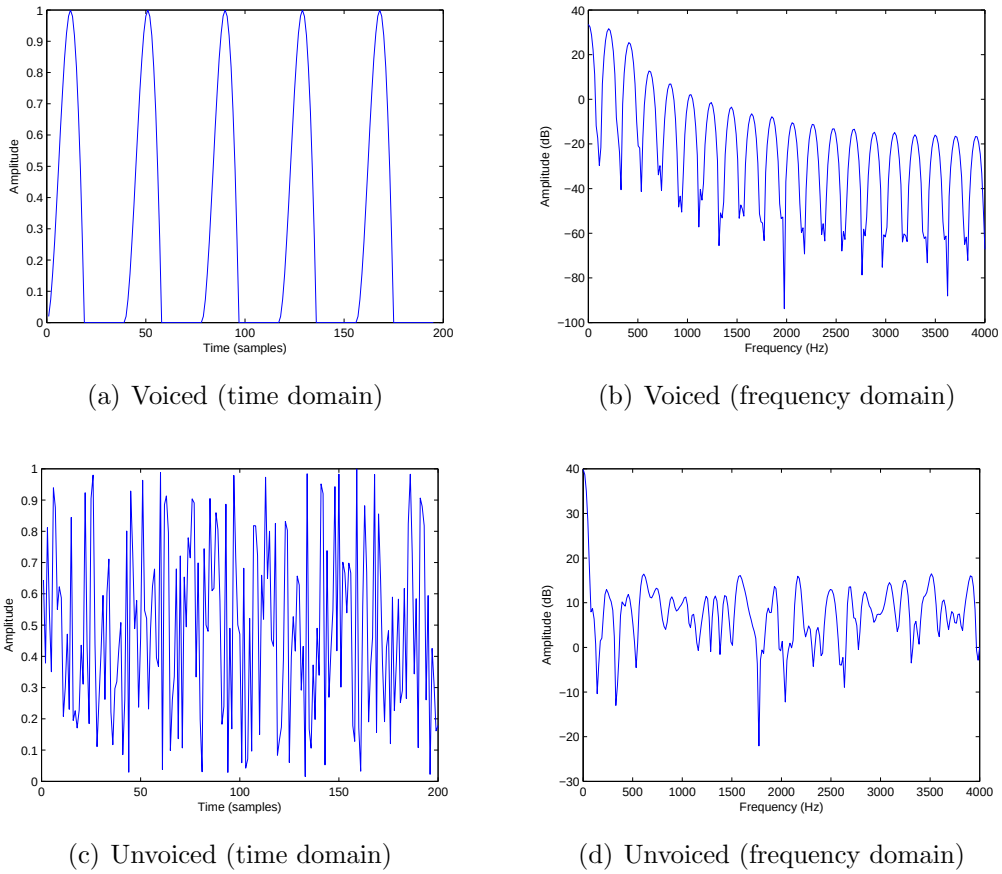
3.2 Speech Production Process

This section describes the human speech production process, the process which we aim to model for the purpose of speech reconstruction. It is therefore important to understand not only the process itself, but also the features that must be extracted from the original signal to efficiently and accurately reconstruct the original speech without loss of quality or intelligibility.

The speech production process can be split into two components: excitation from the lungs and vocal folds, and filtering by the vocal tract. This model of speech production is commonly known as the source/filter model.

We first consider the excitation component of the process, which begins at the lungs. The primary purpose of the lungs is to oxygenate blood, but a by-product of this process is exhalation which pushes air through the larynx. The larynx is composed of a number of muscles, ligaments and cartilage and is used to control the vocal folds which are positioned across the larynx.

There are three states of the vocal folds which dictate the type of sound that can be produced. When the folds are in the breathing state, air flows freely past the folds and no sound is generated, whilst in the voicing state the folds are moved closer together and vary in tension along with the build up and release of pressure caused by the restricted airflow. This variation in tension and pressure cause the folds to open and close periodically to give a buzz-like excitation to the vocal tract. The rate at which the vocal folds open and close defines the fundamental frequency



(a) Voiced (time domain)

(b) Voiced (frequency domain)

(c) Unvoiced (time domain)

(d) Unvoiced (frequency domain)

Figure 3.1: Illustration of excitation signals in the time and frequency domains, where: a.) voiced excitation signal in the time domain, b.) voiced excitation signal in the frequency domain, c.) unvoiced excitation in the time domain and d.) unvoiced excitation in the frequency domain

of the excitation. The final possible state is the unvoiced state, where the vocal folds are moved closer together but are not varied in tension. The tongue is then used to constrict the airflow to create turbulent airflow on exhalation to give a noise-like excitation [Loizou, 2007]. Figure 3.1 shows the difference between voiced and unvoiced excitation signals in both the time and frequency domains.

Figure 3.1(a) shows a time-domain plot of a synthetic voiced excitation signal. The signal is clearly periodic, which is also shown in the frequency domain plot of the signal in Figure 3.1(b). The frequency domain plot shows a clear harmonic structure, i.e. there are peaks at the fundamental frequency and integer multiples of the fundamental. This is in contrast to the spectrum of the unvoiced excitation

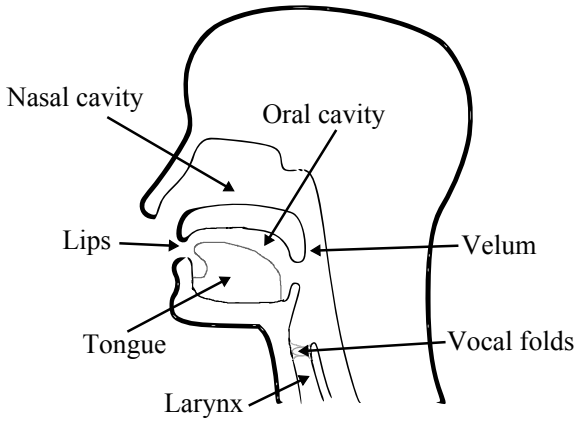


Figure 3.2: Cross section illustrating the human vocal system. Figure adapted from Liesenborgs [2000]

signal shown in Figure 3.1(d) where no clear structure is apparent due to the noise-like properties of the signal as also shown in the time domain in Figure 3.1(c).

Next, we consider the 'filtering' stage of the speech production process caused by the vocal tract. The vocal tract is made up of the oral and nasal cavities (Figure 3.2). These two cavities are linked by the velum, which controls whether air passes through the nasal cavity. The size and shape of the vocal tract is varied with the position of the articulators, namely the tongue, teeth, lips and jaws. These changes in size and shape spectrally shapes the airflow passed through from the larynx when it resonates with the natural frequencies of the vocal tract. This resonance forms the formant structures observed in speech signals as shown in Figure 3.3.

The magnitude spectrum in Figure 3.3 can be seen to contain both source and filter information, with source information being represented by the harmonic peaks at approximately 270Hz intervals and filter information represented by the overall shape of the spectrum. The spectral envelope encodes only the filter information and can be seen to follow the shape of the spectrum, ignoring any source information.

The first formant, F1, is generally considered to be affected by changes in the size of the mouth opening, with small mouth openings having low frequency first formants. The second formant, F2, is affected by the oral cavity and changes of the position of the lips and tongue. The position and intensity of the third formant is

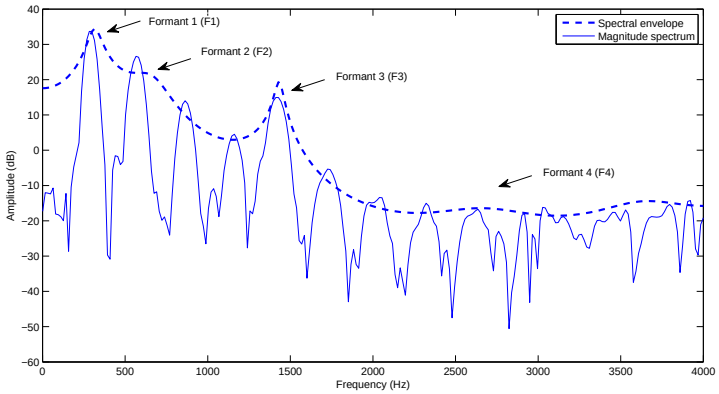


Figure 3.3: Spectrum illustrating speech formants of voiced phoneme /U/ as spoken by a female speaker

affected by constriction of the oral cavity.

Continuous speech consists of a wide range of sounds of varying intensities, duration and spectral characteristics. The types of sounds produced can be classified based on the state of the vocal folds and the size and shape of the vocal tract. These classifications are: vowels, nasals, plosives, fricatives, approximants and affricates. The following list describes the characteristics of each type of sound and how they are produced:

Vowels are produced when the vocal folds are in the fully voiced state and the vocal tract is fully open, i.e. there is no further build up of air pressure past the vocal folds.

Nasals are sounds that are produced when air is diverted through the nasal cavity when the velum opens. Phonemes such as /m/, /n/ and /ng/ are examples of nasal sounds.

Plosives are produced by a build up and sudden release of pressure within the vocal tract. Plosives can be voiced or unvoiced. Examples of voiced plosives include /p/, /t/ and /k/ while /b/, /d/ and /g/ are all unvoiced.

Fricatives are produced by passing the excitation airflow through a narrow constriction in the vocal tract. Examples of unvoiced fricatives include /f/, /s/

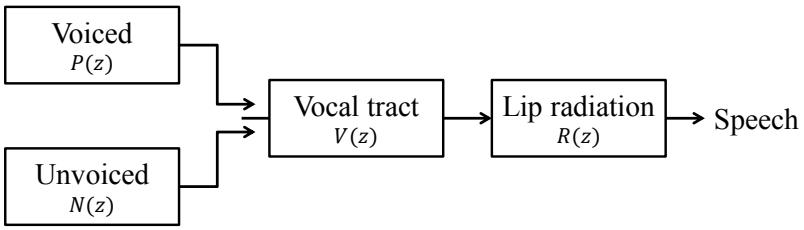


Figure 3.4: Block diagram of the independent source/filter model of speech production (Reproduced from Loizou [2007]

and /sh/, whilst /v/ and /z/ are two examples of voiced fricatives.

Approximants are similar to fricatives in that turbulence is produced by articulators in the vocal tract constricting airflow. Approximants differ from fricatives in the degree of turbulence; the articulators restrict less of the airflow and so less turbulence is caused. /w/ is an example of an unvoiced approximant.

Affricates are a combination of plosives and fricatives. /ch/ is an example of an unvoiced affricate and /j/ is an example of a voiced fricative.

Speech signals clearly have a wide range of characteristics, however the nature of the production process allows effective modelling of the signals. Separating the excitation from the filtering allows us to easily model voiced and unvoiced segments of speech, whilst the spectral envelope may be defined in terms of a finite number of formant locations and bandwidths.

A typical implementation of the source/filter model is illustrated in Figure 3.4 and can be seen to directly model the voicing state and shape of the vocal tract. The source (excitation) is modelled as either a pulse train, P, for voiced excitation or white noise, N, for unvoiced excitation. A switch then selects from either of these two types of excitation based on the voicing required.

The size and shape of the vocal tract is then modelled using a digital filter, V. An appropriate filter is constructed, modelling the speech formants, and is used to shape the excitation signal. Finally a further filter, R, is used to model the radiation of sound from the lips. This is usually of the form $R(z) = 1 - z^{-1}$ to give a 6dB/octave high-pass boost [Loizou, 2007].

In the Z-domain, the output signal, X, can be represented as a linear combination of these stages, i.e:

$$X(z) = P(z)V(z)R(z) \tag{3.1}$$

for voiced speech and:

$$X(z) = N(z)V(z)R(z) \tag{3.2}$$

for unvoiced speech. The most direct implementation of this theoretical model is the LPC vocoder as described in Section 3.3.1, though other methods of reconstruction are also considered.

Whilst the properties of speech signals have been presented as being constrained by a relatively simple process there are still a number of challenges in realising a good model of speech analysis and synthesis (reconstruction). In terms of speech analysis, separating the source and filter components of the signal is still a challenging task, whilst in the synthesis (reconstruction) stage accurately modelling these components introduces additional challenges.

At the analysis stage, errors in fundamental frequency, phase and spectral amplitude estimation can all contribute to a reduction in the perceptual quality of artificially reproduced speech whilst at the synthesis stage the quality of the excitation signal is also critical. Some implementations of the source/filter model assume a simple Dirac delta impulse for excitation whilst the actual excitation signal is somewhat more complex (Figure 3.1(a)).

The following sections examine a number of speech reconstruction models and highlights the ways in which they address the issues identified in this section to produce high quality reproductions of existing speech signals.

3.3 Speech Reconstruction Models

This section evaluates four reconstruction models in terms of their suitability for use in the proposed speech enhancement system. The methods considered are: the LPC vocoder, sinusoidal model, HNM and STRAIGHT. All of these models reconstruct speech from a set of acoustic features in an analysis/synthesis process closely related to the human speech production process (Section 3.2). First, a set of acoustic features related to the excitation signal and state of the vocal tract are obtained in the analysis stage. Later, in the synthesis stage, these features are used to reconstruct the speech signal. This is typically a frame-based approach, with frames durations of 10-30ms being typical due to assumptions of stationarity which may be made.

The quality of speech produced by this method of speech enhancement is directly linked to the reconstruction model as well as the accuracy of the associated acoustic features. Ideally, speech quality should not be reduced by the process of reconstruction as the optimal performance of this model-based speech enhancement method is bounded by the quality of the reconstruction model. The ideal model for this work will therefore reconstruct high quality speech, indistinguishable from the original, from a minimal set of easily obtainable parameters.

This section begins by describing the LPC vocoder in Section 3.3.1 before moving on to the HNM in Section 3.3.3. Finally, STRAIGHT is described in Section 3.3.4. Methods of spectral feature extraction are examined in Section 3.4.

3.3.1 LPC vocoder

The LPC vocoder is a method of speech reconstruction closely related to the source/filter model of speech production. Each frame is reconstructed from two components; a filter that models the response of the vocal tract, and an excitation signal which is either an impulse train for voiced speech or white noise for unvoiced speech [Kondoz, 2004]. Figure 3.5 shows the processes of analysis and synthesis for the vocoder.

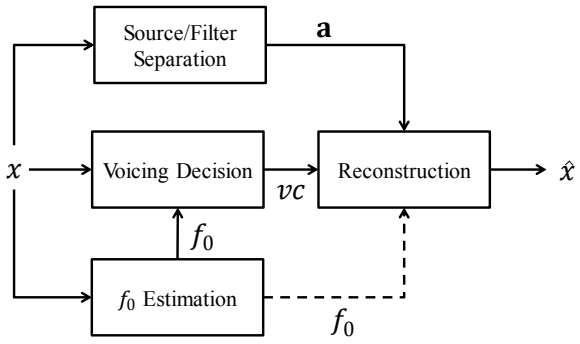


Figure 3.5: Analysis and synthesis processes of the LPC vocoder

The processes in this figure can be seen to relate directly to the source/filter model of speech production shown in Figure 3.4. Each block of the analysis and synthesis stages is examined in the remainder of this section.

Source/Filter separation One of the key challenges of this model is the derivation of the filter coefficients which best separate the response of the vocal tract from the excitation. The aim of this process is to find the Pth order linear predictor coefficients $\mathbf{a} = [1, a_2, \ldots, a_{P+1}]$ which represent the signal as:

$$\hat{x}(m) = \sum_{i=1}^{P} a_i x(m-i), \tag{3.3}$$

where $x(m)$ is the mth sample of the original signal and $\hat{x}(m)$ is the predicted signal. The coefficients are selected as the values which minimise the error $e(m) = x(m) - \hat{x}(m)$.

The values of $\mathbf{a}$ are typically derived based on optimising the root mean square (RMS) error, also known as the *autocorrelation criterion*. In this method the expected value of the squared error, $\mathbb{E}[e^2(n)]$ is minimised. This gives the series of equations to be optimised as:

$$\sum_{i=1}^{P} a_i R_x(j-i) = -R_x(j) \quad 1 \leq j \leq P, \tag{3.4}$$

where $\mathbf{R}_x$ is the autocorrelation of the original signal. These normal equations,

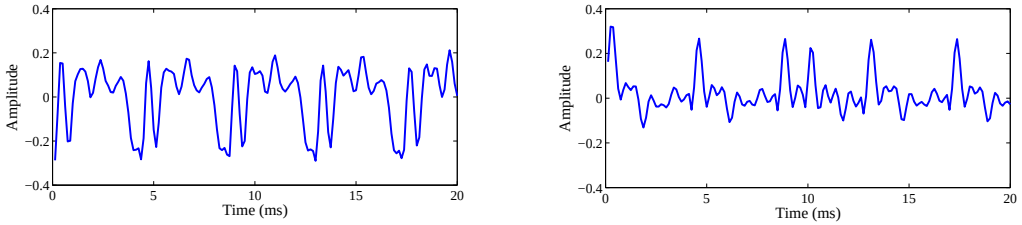
known as the *Yule-Walker* equations can be efficiently solved using *Levinson-Durbin* recursion [Nagarajan and Sankar, 1998].

Whilst this method works well for clean speech it is not robust to noise [Tierney, 1980]. The autocorrelation function is significantly affected by additive noise which degrades the quality of fit of the predictor coefficients, with more coefficients required to fit the speech and noise. Many methods of noise robustness have been developed for LPC encoding, most of which rely on conventional-style noise estimation and filtering in the time or autocorrelation domain [Tierney, 1980; Kang and Fransen, 1989; Lim, 1978]. The focus of this work is to move away from such frame-based noise estimation methods and so this work will use the standard model of predictor coefficient estimation.

Voicing classification and fundamental frequency estimation A different excitation signal is used for the LPC vocoder, based on the voicing of the frame. As such, a method of voicing classification is required. This can be achieved at the same time as fundamental frequency ($f0$) estimation. Frames with a value of $f0$ attributed to them are synthesised as voiced speech whilst all other frames of unvoiced speech or silence are synthesised as unvoiced. Methods of voicing classification and $f0$ estimation are described in Chapters 6 and 7 respectively.

Reconstruction Speech is synthesised by exciting a filter, constructed with the previously calculated predictor coefficients, with an excitation signal. In voiced speech a simple Dirac delta function is used with impulses spaced at $\frac{1}{f0}F_s$ sample intervals where F_s is the sampling rate. This is a simple model of the true excitation signal as in Figure 3.1(a). White noise to model the turbulent airflow in the unvoiced speech production process (Figure 3.1(c)).

Next, we examine the case of a single frame of voiced speech. Figure 3.6 shows the original time domain frame (Figure 3.6(a)) and the resynthesised frame using a 10th order LPC filter (Figure 3.6(b)).



(a) Original (b) 10th order LPC

Figure 3.6: Comparison of voiced frames of: a.) original and b.) reconstructed speech using 10th order LPC filter in the time domain

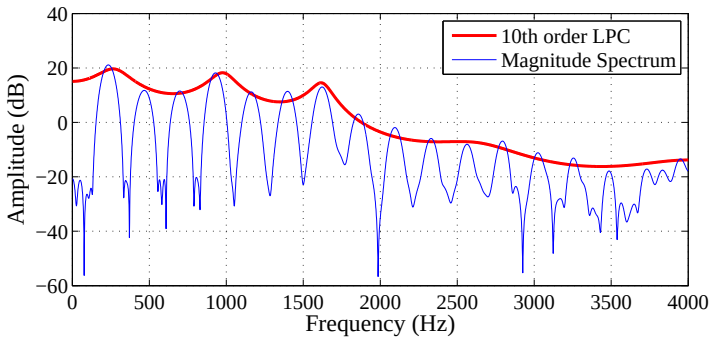


Figure 3.7: LPC frequency domain response of voiced frame

There are clear differences visible between the two time domain plots. This is thought to have occurred due to the minimum phase reconstruction resulting from the excitation signal. Figure 3.7 shows the frequency response of the LPC filter of the same frame compared to the magnitude spectrum.

The frequency response of the LPC filter is shown to accurately capture the spectral envelope information with the first three formants being easily visible. Finally, Figure 3.8 compares the magnitude spectra of the original frame with that of the reconstructed frame.

The two spectra are shown to be very similar with the largest differences found in higher frequency regions where small errors in $f0$ cause some harmonic positions to be shifted slightly in the reconstructed signal.

In summary, the LPC vocoder has been shown to provide a reasonably accurate reconstruction of clean speech signals. Despite this, two concerns remain regarding

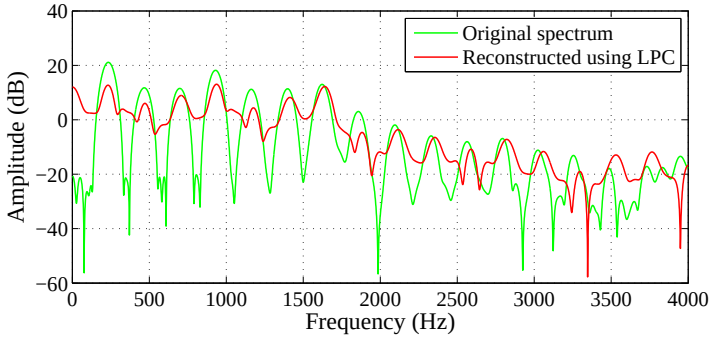


Figure 3.8: Comparison of voiced frames of original and reconstructed speech using 10th order LPC filter in the frequency domain

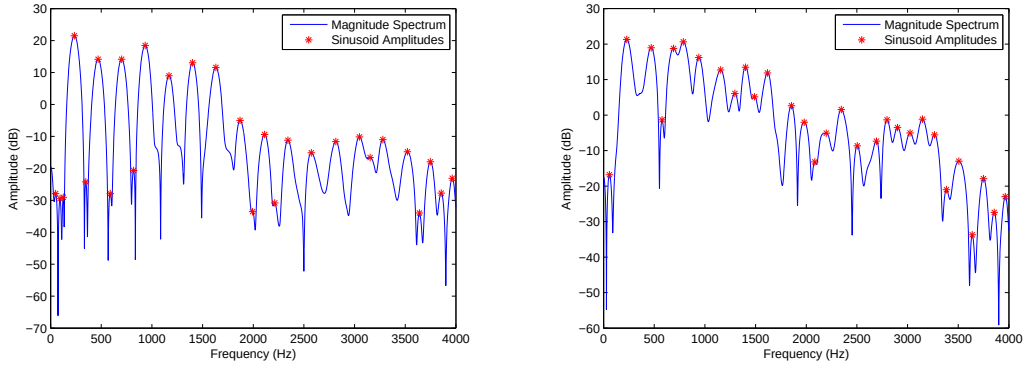
its use in this work. Firstly, the excitation signal, and resulting minimum phase assumption, may degrade the quality of reconstructed speech. Secondly, it is widely reported that this method is not inherently robust to noise which may cause further difficulties [Sambur and Jayant, 1976; Tierney, 1980]. The results of experiments comparing the LPC vocoder with other methods of speech reconstruction are presented in Section 3.5.

3.3.2 Sinusoidal model

The sinusoidal model can be considered to be an enhancement of the source/filter vocoder. Instead of using a time-domain impulse train to excite a filter representing the vocal tract response, voiced speech is reconstructed by synthesising a set of sinusoids relating to the original speech signal. In Fourier analysis, any signal may be reconstructed using a sufficient number of sinusoids [Oppenheim et al., 1989]. The sinusoidal model therefore reconstructs speech using a set of L sinusoids with amplitudes, a_l, frequencies, f_l, and phase offsets,θ_l:

$$s(m) = \sum_{l=1}^{L} a_l \cos(2\pi f_l m + \theta_l), \tag{3.5}$$

where a is computed by sampling the speech spectral envelope estimate, $A(f)$, at the required frequencies, i.e. $a_l = A(f_l)$. Sinusoid frequencies and phase-offset values are



(a) Clean (b) 0dB

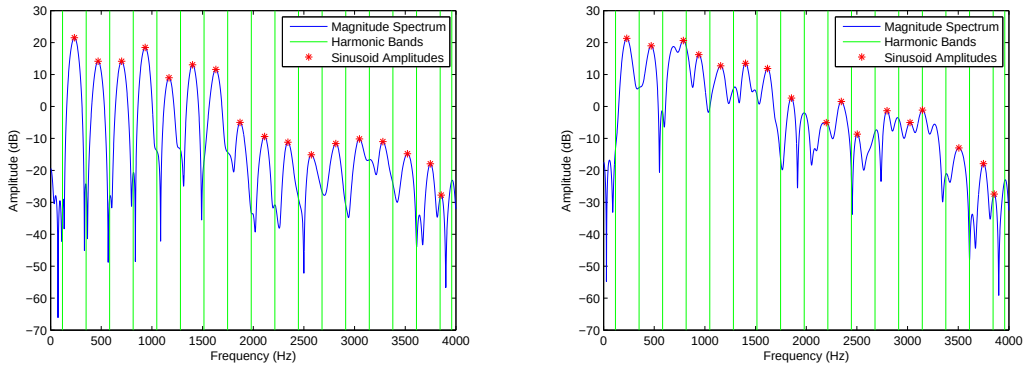
Figure 3.9: Spectra illustrating the processes of peak picking directly from the magnitude spectrum with a.) no noise and b.) in car noise at 0dB SNR

determined using one of the peak-selection methods described later in this section.

Various methods of selecting the sinusoids to use for synthesis exist, including peak-picking directly from the magnitude spectrum and harmonic sampling of the spectral envelope [Jensen and Hansen, 2001]. In clean conditions, it is possible to pick peaks from the magnitude or power spectrum to reconstruct a near-perfect representation of the original speech.

Figure 3.9 illustrates the process of peak-picking from the magnitude spectrum in clean and noisy conditions. In clean conditions, the positions of harmonics are clear. In noisy conditions many peaks exist around harmonics, in some cases masking the position of the true speech signal. If every peak were to be selected, a significant amount of noise would be reconstructed alongside the speech. In an attempt to avoid this issue the spectrum can be divided into harmonic bands and the largest peak selected from each band as illustrated in Figure 3.10. This is based on the assumption that the speech harmonic will always be the highest energy component in the band, an assumption that is clearly only valid where the local SNR is greater than $0dB$.

Using harmonic bands to select peaks is shown to be effective at selecting peaks relating to the harmonics, however in some cases only a spectral envelope is available.



(a) Clean (b) 0dB

Figure 3.10: Spectra illustrating peak picking from the magnitude spectrum using harmonic bands with a.) no noise and b.) car noise at 0dB SNR

The spectral envelope does not contain any source information and so a different strategy must be employed. Based on the excitation signal, voiced speech may be assumed to be harmonic and so, given an estimate of the fundamental frequency, amplitudes may be sampled from the spectral envelope at integer multiples of the fundamental frequency. This technique leads to a variant of the sinusoidal model known as the harmonic plus noise model (HNM).

3.3.3 Harmonic plus noise model (HNM)

The harmonic plus noise model (HNM) is a variant of the sinusoidal model. The sinusoidal model reconstructs the speech signal as a sum of modulated sinusoids [Quatieri and McAulay, 2002]. The problem which remains is how to select the sinusoids for reconstruction. Several methods of sinusoid selection were described in Section 3.3.2 though none were particularly robust to noise. By exploiting the harmonic structure of voiced speech the HNM improves on the sinusoidal model by adding constraints to the sinusoid frequencies. Techniques which may improve the quality of reconstructed speech are also described. A method of emphasising formant locations is described in Section 3.3.3.2 followed by a method of sub-frame synthesis in Section 3.3.3.3 which is used to improve harmonic trajectories. Finally, overlap and

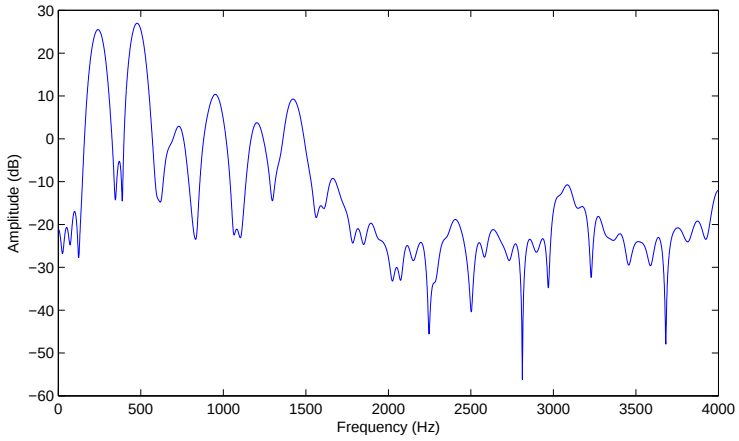


Figure 3.11: Spectrum of mixed-excitation frame with voiced/unvoiced transition at approximately 1.6kHz

add, a method of frame combination used to smooth frame transitions is described in Section 3.3.3.4.

3.3.3.1 HNM reconstruction

When using the HNM each frame is assumed to be either voiced, unvoiced or of mixed voicing. In voiced frames the sinusoid frequencies are assumed to have a harmonic relationship to the fundamental frequency, f_0, i.e. $f_l = lf_0$. Unvoiced frames are reconstructed using noise filtered by a filter derived using the spectral envelope estimate whilst the harmonic component of the equation is set to zero. Frames with mixed voicing are reconstructed as voiced up to a threshold frequency and unvoiced at all remaining frequencies. This results in speech being reconstructed as:

$$s(m) = \sum_{l=1}^{L} A(lf_0)\cos(2\pi lf_0 m + \theta(lf_0)) + n(m). \tag{3.6}$$

Figure 3.11 shows the magnitude spectrum of a frame of clean speech which has been classified as a voiced fricative, i.e. it has mixed voicing.

In the magnitude spectrum shown in Figure 3.11 there is a clear harmonic struc-

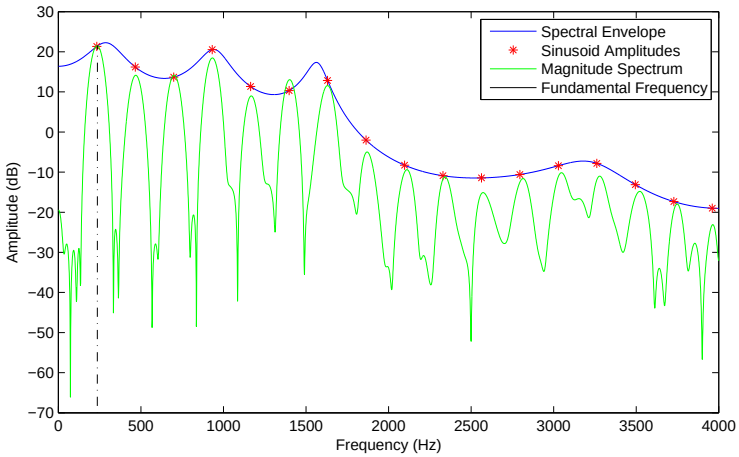


Figure 3.12: Illustration of sampling of sinusoid amplitude parameters using spectral envelope and estimate of the fundamental frequency

ture up to approximately 1.6kHz. Beyond this point the spectrum consists of noise from the unvoiced component of the excitation. Whilst it is possible to estimate a voiced/unvoiced boundary from the original speech signal, in noisy conditions this estimation is unreliable and so an empirically determined fixed value is used in some applications [Sorin and Ramabadran, 2003]. In this work a binary classification of voiced/non-voiced is used and so frames are synthesised as being either completely voiced or unvoiced.

Figure 3.12 shows the first step in the process of synthesising a frame of speech from the spectral envelope using HNM.

Sinusoid amplitudes, $\mathbf{a}$, are estimated from the spectral envelope at multiples of the fundamental frequency. In this case $f_0 \approx 240Hz$. Minor sampling errors are observed in high frequency regions where slight errors in the fundamental frequency estimate are amplified due to the multiplicative process of calculating the harmonic sampling points, $\mathbf{f}$. The phase, $\boldsymbol{\theta}$, is sampled using the same technique of harmonic sampling. Finally, voicing estimates are made using a voicing estimation method.

Reconstruction begins by generating a number of sinusoids at harmonic frequencies, $\mathbf{f}$, and then applying amplitude modulation and phase offsets using the previously sampled parameters $\mathbf{a}$ and $\boldsymbol{\theta}$. These sinusoids are then summed to give the

reconstructed frame. This is beneficial for the purpose of spectral envelope modelling as a more coarse spectral envelope may be used for reconstruction as no source information is required for peak selection as per the sinusoidal model.

3.3.3.2 Formant emphasis

When sinusoid amplitudes are sampled from the spectral envelope the resulting speech can sometimes sound muffled due to over-smoothing of formants. To compensate for this a method of formant enhancement is used which post-filters the reconstructed frame to sharpen formants and thus improve speech quality [Chen and Gersho, 1987; Quatieri and McAulay, 2002]. The process of formant filtering consists of three stages:

Transformation of spectral envelope to LPC domain Spectral envelope is first transformed to the LPC domain by first applying an inverse Fourier transform to the power spectrum of the spectral envelope to obtain the autocorrelation vector. Levinson-Durbin recursion is then applied to the autocorrelation values to give LPC predictor coefficients, **a** [Kondoz, 2004].

Parameter modification LPC coefficients are transformed to the Z-domain to give a pole-zero representation of the filter:

$$H(z) = K\frac{\prod_n^N (z - Z(n))}{\prod_n^N (z - P(n))}, \tag{3.7}$$

where K is the gain of the filter, N is the order of the filter, Z represents filter zeros and P represents the filter poles. Pole values are modified to give new pole and zero values, with modified pole values computed as:

$$\mathbf{P}' = p\mathbf{P} \tag{3.8}$$

and modified zero values computed as:

$$\mathbf{Z}' = z\mathbf{P}. \tag{3.9}$$

p and z are tunable parameters which control the extent to which formants are modified with the values $p = 0.95$ and $z = 0.85$ found to offer best performance [Kleijn and Paliwal, 1995]. These pole and zero values are then converted back to LPC coefficients using the original gain value, K.

Filtering The filter is applied to the reconstructed waveform in the time-domain in the standard way [Kondoz, 2004].

3.3.3.3 Sub-frame reconstruction

During periods of rapid change in fundamental frequency step changes in harmonic frequencies occur between frames. These discontinuities cause a slight degradation in the quality of reconstructed speech. Sub-frame reconstruction is therefore used to interpolate $f0$ values between frames to provide smoother harmonic transitions.

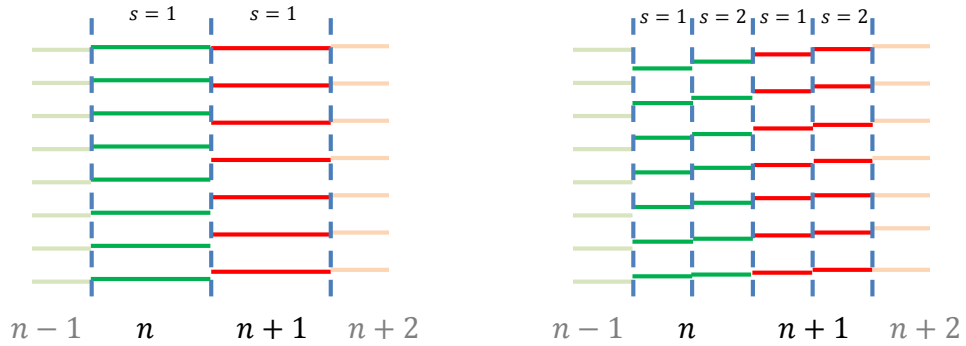
Each frame is split into S subframes. $f0$ is varied across each subframe with amplitude and phase values resampled based on the new harmonic positions.

The fundamental frequency of each frame is derived by linearly interpolating between the current and next frame parameters as:

$$f0_s = f0(n) + \frac{s(f0_{n+1} - f0_n)}{S}, \tag{3.10}$$

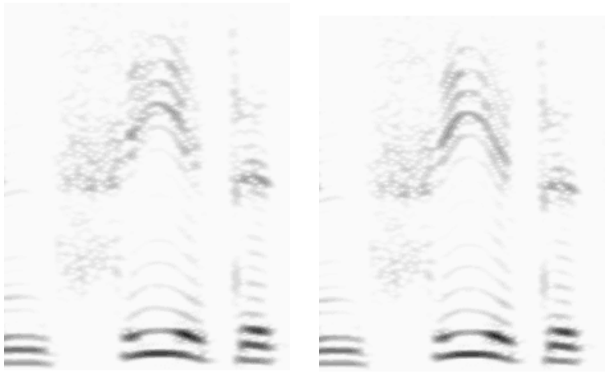
where s is the subframe number of a total S subframes and n is the current frame index.

Figure 3.13(a) illustrates the step change in harmonics between frames where fundamental frequency is changing rapidly. At lower frequencies the discontinuities are less visible, however at high frequencies clear differences exist between harmonics with the number of total harmonics in the frames also varying. The effect of



(a) Standard ($S = 1$) (b) Subframe ($S = 2$)

Figure 3.13: Illustration of step change between harmonic frequencies in periods of rapid f_0 change



(a) Standard ($S = 1$) (b) Subframe ($S = 4$)

Figure 3.14: Narrowband spectrograms of reconstructions of the utterance *"and see if it"* comparing a.) standard HNM reconstruction and b.) subframe reconstruction

subframe reconstruction with $S = 2$ is illustrated in Figure 3.13(b). Here, the step change between frames is reduced due to the linear interpolation of the fundamental frequency and subsequent resampling of harmonic frequencies.

Figure 3.14 now compares spectrograms of the utterance *"and see if it"* for both standard reconstruction and sub-frame reconstruction using $S = 4$. At the beginning and end of the utterance there is very little change in f_0 and therefore no significant differences between the spectrograms. In the centre of the utterance, for the change between /i:/ and /I/, there are significant changes in f_0. This results in harmonic confusion in the case of $S = 1$ which appears as noise-like segments on the spectrogram. In the case of $S = 4$ the harmonic structure is much clearer due

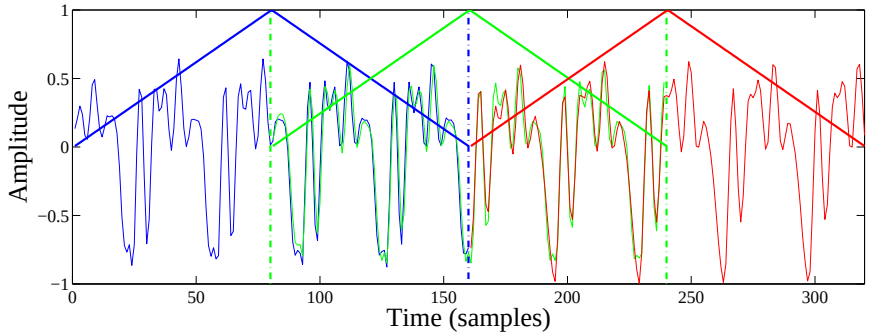


Figure 3.15: Illustration of the overlap and add process in the time domain showing overlapping windows

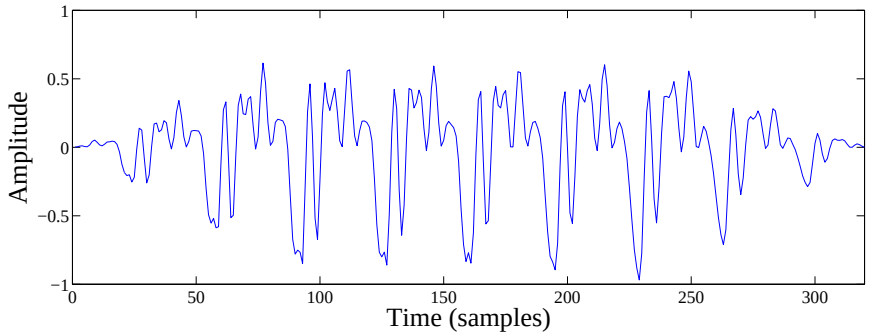


Figure 3.16: Reconstructed signal using overlap and add

to the reduced step-change in harmonic frequencies between subframes.

3.3.3.4 Overlap and add

In this work the HNM is used with a frame rate of 100fps and a frame duration of 20ms. This gives a 50% overlap which must be accounted for at reconstruction. Whilst it is possible to reconstruct using no overlap, overlapping frames significantly reduces the effect of any phase discontinuity between frames. Figure 3.15 illustrates this process of overlap and add. First, each frame is windowed using a triangular window. Next, frames are overlapped by 50% and then finally added together to give a weighted average of the combination of frames. Figure 3.16 shows the result of the overlap and add process shown in Figure 3.15. The reconstructed signal tapers to zero at either end of the signal which allows the segment to be joined with other frames. At the beginning and end of the utterance half-triangular windows are used

to maintain signal amplitudes.

3.3.3.5 Summary

The HNM has attractive properties compared to both the source/filter vocoder and sinusoidal model. Firstly, when compared to the source/filter vocoder, the lack of an explicit excitation signal reduces the complexity of the model and gives a potential increase in quality; as there is no excitation signal required, errors associated with reproducing an excitation signal are not possible. When compared the original sinusoidal model, the simplifying assumption that sinusoid amplitudes are harmonically related to the fundamental frequency vastly reduces the complexity of the analysis stage. This harmonic assumption also provides the model with an inherent robustness to noise on the assumption that a good estimate of f_0 is available.

3.3.4 STRAIGHT

STRAIGHT is a channel vocoder that was designed to allow real-time manipulation of speech parameters, such as the fundamental frequency of a speaker [Kawahara et al., 1999]. For effective parameter modification a complete separation of the source and filter components is desirable. The conventional LPC vocoder separates source and filter information into a time-domain excitation signal and a vocal tract filter. However, perfect separation is not always possible as traces of periodicity remain in the filter (Section 3.3.1). STRAIGHT takes an alternate approach of estimating the 'spectral surface' which describes the vocal tract filter whilst source information is extracted independently using a fundamental frequency estimator.

Spectral surface (filter) estimation It is assumed that voiced frames give a partial sampling of the spectral surface at harmonic intervals. Short analysis windows result in high time resolution analysis of this surface but low frequency resolution, whilst longer windows increase frequency resolution at the expense of time resolution. Equally high resolution in both dimensions is re-

quired to accurately reconstruct the spectral surface. Kawahara et al. [1999] proposed a windowing function that varies with $f0$ in order to reduce the amount of periodicity in the resulting frequency domain analysis. Remaining periodic variations in time and frequency are then smoothed based on a partial representation of the spectral surface using a second-order cardinal B-spline function. Alternatively, a set of phase insensitive windows may be used to produce a spectrogram free of periodic variations.

Source estimation A fundamental frequency estimation system based on instantaneous frequency was proposed as part of STRAIGHT. High resolution in time is required for smooth parameter modifications and so the method focuses on high performance in clean conditions rather than noise robustness.

Speech is reconstructed using the standard vocoder driven by two acoustic features: the high resolution spectral surface and the fundamental frequency (Section 3.3.1). Phase is synthesised using a minimum phase model. Whilst STRAIGHT is demonstrated to reconstruct speech of high quality the high resolution of the features required for reconstruction may make the model unsuitable for this work. High resolution parameter estimation is possible in clean conditions, however with the addition of noise this estimation is expected to become difficult due to the masking effect of the noise.

3.4 Spectral Features

Applications such as speech compression [Kondoz, 2004] and speech transmission (i.e. VoIP) extract acoustic features from an existing signal with the aim of minimising the amount of data which needs to be transmitted whilst retaining the important information relating to the signal. In other applications, such as text-to-speech (TTS) synthesis, acoustic features may be generated or selected from codebooks [Stylianou, 2001]. In this work the aim is to reconstruct speech from 'cleaned' acoustic features, estimated from those extracted from noisy speech.

Common to all reconstruction models is the requirement of the spectral envelope of the signal. The aim is therefore to identify a method of encoding the spectral envelope that preserves the information carried in the signal, i.e. the formant locations and bandwidths. This section therefore examines the use of feature vectors to model the spectral envelope.

Whilst our ultimate goal is a set of spectral amplitudes for use in our speech reconstruction model, working in an alternative feature domain for the spectral enhancement process can provide a number of advantages. Feature extraction processes typically incorporate a range of filters and transforms which give a more perceptually relevant representation of the signal in a more compact form. For example, a considerable amount of redundancy exists in the magnitude and power spectra of speech; the value of $A(k)$ will be highly correlated with $A(k-1)$ and $A(k+1)$ and so it should be possible to significantly reduce the dimensionality of the feature compared to using the full spectrum, reducing the complexity of feature modelling.

In this section we consider a range of methods in terms of coding efficiency and resulting speech quality. An ideal method will be robust to noise and provide a compact representation, i.e. $M \ll N_{fft}$, where M is the feature size. In doing so it is important to retain the same level of quality as when using the original high resolution magnitude spectrum.

Spectral features considered in this work include spectrum-based features (i.e. magnitude spectrum and power spectrum) in Section 3.4.1, linear and Mel-spaced filterbanks in Section 3.4.2 and finally linear predictive coding (LPC) based methods including line spectral frequencies (LSF) in Section 3.4.3.

3.4.1 Spectrum-based features

Spectrum-based features consist of the magnitude or power spectra of the signal. First, the signal is split into short frames and then a window applied. Windows such

as the Hamming or Hann windows taper the signal at frame boundaries in order to reduce spectral leakage around peaks in the spectrum, though this comes at the cost of broadening those spectral peaks. The design of the window is essentially a trade off between spectral leakage and width of spectral peaks. The Hamming window is used for this application as it provides the best balance of these factors for speech processing. The Hamming window is defined as:

$$w(m) = 0.54 - 0.46\cos\left(\frac{2\pi m}{N-1}\right), \tag{3.11}$$

where $w(m)$ is the mth sample of the window. This window is applied to the signal as:

$$x(m) = x(m)w(m). \tag{3.12}$$

A DFT is then applied to give the complex spectrum of the signal from which the absolute value is taken to give the magnitude spectrum which is defined as $|\mathbf{X}|$. Only the first half of the spectrum is retained due to the mirroring of the spectrum caused by the Nyquist frequency. Optionally, this spectrum may be raised to the power of p, i.e. $|\mathbf{X}^p|$, where $p = 2$ gives the power spectrum.

Spectrum-based features have the attractive property of the process being fully invertible with no loss of information. This comes at the cost of features with high dimensionality; for no loss of information the spectrum must be at least as long as the number of samples in the time-domain frame.

3.4.2 Filterbank-based features

This section compares and contrasts two filterbank-based features: linear-spaced cepstral coefficients (LFCCs) and Mel-spaced cepstral coefficients (MFCCs), by examining the processes of feature extraction (Section 3.4.2.1) and inversion from feature vector to spectral envelope (Section 3.4.2.2).

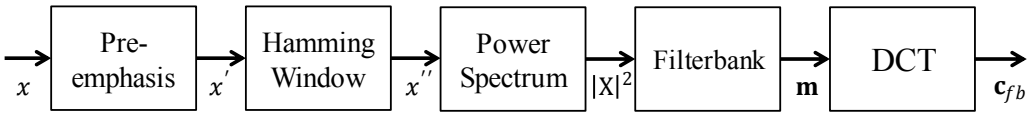


Figure 3.17: Flowchart of LFCC/MFCC feature extraction process

3.4.2.1 Feature extraction

This section explains the full process of extracting LFCCs and MFCCs from a signal using a process based on the ETSI Aurora standard [Sorin and Ramabadran, 2003]. Each stage operates on a single frame of speech, typically between 10-30ms in duration and overlapping by 50% with adjacent frames. Figure 3.17 shows the feature extraction process. Each stage is explained in further detail in the following parts of this section.

Log energy Each frame has an energy coefficient associated with it and is calculated as:

$$E = \sum_{m=0}^{N-1} x(m)^2, \tag{3.13}$$

where $x(m)$ is the mth sample of the current frame and N is the total number of samples in the current frame. The log of energy parameter, E, is then taken and thresholded as:

$$lnE = \begin{cases} \log(E) & \text{if } E \geq E_{thresh} \\ \log(E_{thresh}) & \text{else} \end{cases}, \tag{3.14}$$

where $E_{thresh} = \exp(-50)$.

Windowing The frame is then Hamming windowed:

$$x'(m) = x(m)w(m), \tag{3.15}$$

where $x'(m)$ is the nth sample of the Hamming windowed signal.

Fourier transform After the pre-processing stages the signal is transformed into the frequency domain using a discrete Fourier transform (DFT):

$$X(k) = \sum_{m=0}^{N-1} x'(m) e^{-\frac{2\pi i}{N} km} \qquad k = 0, \ldots, N-1. \tag{3.16}$$

The power spectrum is derived from the Fourier transformed signal as:

$$|X(k)|^2 . \tag{3.17}$$

Filterbank Next, a filterbank is applied to the power spectrum. LFCCs use a set of linearly spaced triangular filters which provide equal weighting across frequency. This is beneficial for accurate signal reconstruction, however the human ear does not have a linear frequency response. MFCCs therefore use a set of Mel-spaced filters, a psycho-acoustic scale used to compensate for the non-linear frequency response of human hearing. The Mel scale is defined as:

$$\text{Mel}(f) = 2595 \cdot \log_{10}(1 + \frac{f}{700}), \tag{3.18}$$

by [O'Shaughnessy, 1987]. Figure 3.18 shows the relation between linear frequency and the Mel scale. The result of this scaling is that more filterbank channels are present in the perceptually most important regions of the speech signal.

The filterbank matrix is calculated as a set of basis functions, each corresponding to a filterbank channel. The filterbanks consist of N triangular filters with centre frequencies spaced at either linear intervals or Mel-spaced intervals. The start and end points of each filter correspond to the centre frequencies of the previous and next filters respectively. Figure 3.19 compares the linear and Mel-scale filterbank matrices.

The filterbank channels are applied to the power spectrum as a matrix multi-

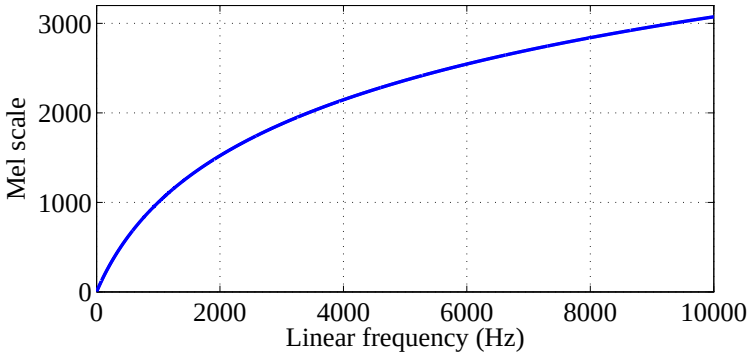
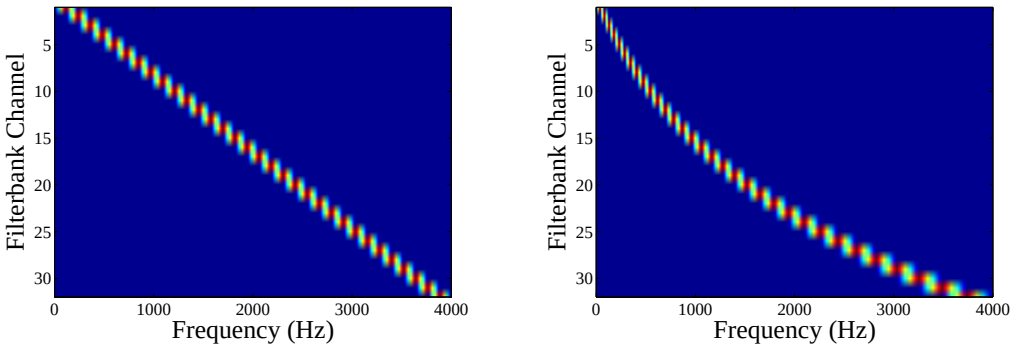


Figure 3.18: Relationship between linear and Mel frequency scales



(a) Linear Filterbank (b) Mel Filterbank

Figure 3.19: Visual representation of a.) linear and b.) Mel-spaced filterbank matrices

plication,

$$\mathbf{m} = |\mathbf{X}|^2 \times \mathbf{W} \tag{3.19}$$

giving our vector of N filterbank channels, $\mathbf{m}$.

Cepstral transformation A DCT is applied to the output of the filterbank to transform the feature to the cepstral domain. MFCCs and LFCCs can therefore be viewed as an extension to the standard cepstral coefficients which are calculated by taking the magnitude spectrum of the log of the magnitude spectrum, i.e,

$$c(l) = |DFT(\log|X(f)|)|\,, \tag{3.20}$$

where $c(l)$ is the lth cepstral coefficient and $|X(f)|$ is the magnitude spectrum of the signal. LFCCs are very closely related to cepstral coefficients, the main difference being the application of a set of linearly-spaced filterbank channels to the spectrum before transformation to the cepstral domain. MFCCs further differ by using Mel-spaced filterbank channels to apply a perceptual weighting to the spectrum.

The DCT represents the filterbank channels as a set of cosine basis functions. If $\mathbf{c}_{fb}$ is our cepstral feature, $c_{fb}(0)$ represents the overall energy of the signal (i.e. the DC level) whilst $c_{fb}(1)$ represent spectral slope. High order coefficients represent finer detail which may be discarded to smooth the spectral representation. This transform can therefore be seen to both decorrelate the feature space and allow for efficient source separation; two properties which are believed to be beneficial for later feature estimation and reconstruction.

Firstly, decorrelating the feature space allows for more efficient modelling of the feature space. The value of each filterbank channel will be correlated with neighbouring filterbanks channels due to the spectral relationship between channels. Decorrelation of the feature space allows speech recognition systems

to use GMMs with diagonal co-variance to model the feature space without a large drop in performance over full-covariance systems due to minimal interaction between feature coefficients [Gales, 1998].

Secondly, the DCT allows efficient separation of source and filter information. The low-order cosine basis functions model course detail such as energy and spectral slope whilst higher order basis functions model more fine (high quefrency) detail. This fine detail can be seen to consist of the source information and therefore discarding higher order DCT coefficients should leave the spectral envelope of the signal. This effect is examined further in Section 3.4.2.2.

LFCCs are therefore the DCT of the logarithm of the linearly space filterbank channels whilst MFCCs are the DCT of the logarithm of the Mel-spaced filterbank channels. First, the logarithm is applied and results floored at $M_{thresh} = -10$:

$$\mathbf{m}_{log}(k) = \begin{cases} \log[\mathbf{m}(k)] & \text{if } \log[\mathbf{m}(k)] > M_{thresh} \\ M_{thresh} & \text{else} \end{cases}. \tag{3.21}$$

Next, the log-filterbank channels are transformed to the cepstral domain:

$$\mathbf{c}_{fb} = \mathbf{m} \times \mathbf{C}, \tag{3.22}$$

where $\mathbf{c}_{fb}$ is the cepstral feature vector and $\mathbf{C}$ is the matrix of DCT basis functions.

3.4.2.2 Feature inversion

As described in Section 3.4.2, LFCC and MFCC feature extraction consist of a number of stages including several lossy operations. Whilst the effect of these operations are not fully recoverable, it is still possible to form an estimate of the spectral envelope. Figure 3.20 illustrates the process of feature inversion. It should be noted that the stages are not inverted in the exact reverse order of the feature extraction

Figure 3.20: Flowchart of LFCC/MFCC feature inversion process

process.

Cepstral equalisation The inversion process begins by subtracting the cepstral-domain impulse response of the feature extraction process from the feature vector to equalise for the filterbank bandwidths in the cepstral domain. This is particularly beneficial as the impulse response may be pre-computed and stored for later use. The first stage is therefore to obtain the equalised cepstral vector and is computed as:

$$\mathbf{c}'_{fb} = \mathbf{c}_{fb} - \mathbf{h}, \tag{3.23}$$

where $\mathbf{h}$ is the impulse response of the feature extraction process and $\mathbf{c}'_{fb}$ is the equalised cepstrum.

Inverse DCT and exponential The DCT and logarithm are inverted as:

$$|\hat{\mathbf{X}}'|^2 = \exp(\mathbf{c}'_{fb} \times \mathbf{C}^{-1}) \tag{3.24}$$

to give a sparse power spectrum with points at filterbank centre frequencies.

Interpolation The estimate of the complete magnitude spectrum is formed by linearly interpolating between filterbank centre points and then finally performing a square-root to transform from power to magnitude spectrum.

The relationship between two filterbank channels, $n-1$ and n, with centre frequencies f_{n-1} and f_n and the interpolated point at frequency f can be described as:

$$\frac{|\hat{X}|^2(f) - |\hat{X}'|^2(f_{n-1})}{f - f_{n-1}} = \frac{|\hat{X}'|^2(f_n) - |\hat{X}'|^2(f_{n-1})}{f_n - f_{n-1}}. \tag{3.25}$$

Solving the equation for $|\hat{\mathbf{X}}|^2$ gives:

$$|\hat{X}|^2(f) = |\hat{X}'|^2(f_{n-1}) + \frac{(f - f_{n-1})|\hat{X}'|^2(f_n) - (f - f_{n-1})|\hat{X}'|^2(f_{n-1})}{f_n - f_{n-1}}. \quad (3.26)$$

Square root Finally, the estimated magnitude spectrum is given as:

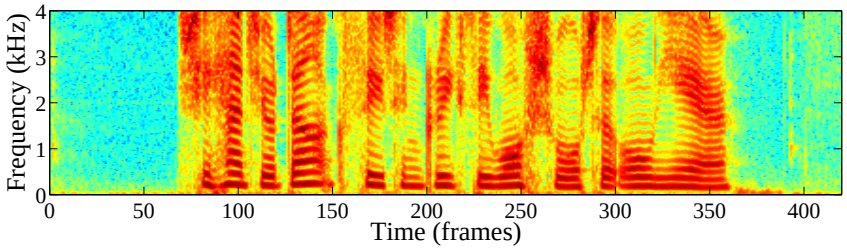
$$|\hat{\mathbf{X}}| = \sqrt{|\hat{\mathbf{X}}|^2}. \quad (3.27)$$

3.4.2.3 Effect of feature inversion

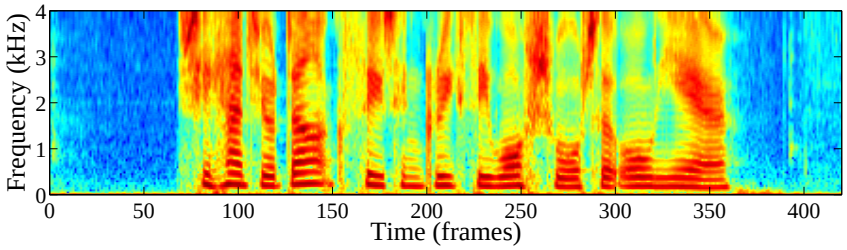
Next, we examine the effect of both feature inversion and removal of high-order DCT coefficients. In Figure 3.21 64 Mel-spaced filterbank channels were extracted from the utterance *"She had your dark suit in greasy wash water all year"* spoken by a female speaker.

The harmonic structure of the original utterance in Figure 3.21(a) is clear across the entire frequency range. Comparing the spectrogram of the original utterance to the recovered magnitude spectrum of a 64-channel MFCC vector in Figure 3.21(b) shows some spectral smearing in high and mid-frequency regions with considerable spectral detail remaining between 0-1kHz. This is attributed to Mel-spacing of the filterbank channels placing more channels in this range than at higher frequencies.

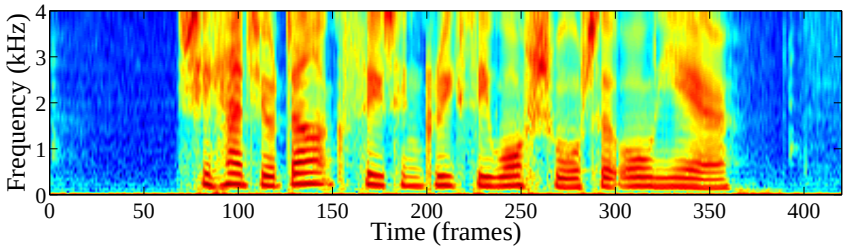
Reducing the number of retained DCT coefficients to 32 in Figure 3.21(c) increases the amount of spectral smearing whilst retaining significant source information. Retaining only the first 16 coefficients (Figure 3.21(d)) begins to reduce the amount of source information in the spectrum whilst formants begin to become clearer. Finally, when only the first 8 coefficients are retained as in Figure 3.21(e) the formant structure is easily visible with no source information remaining. The effect of feature extraction and inversion on reconstructed speech quality, including the effect of reducing the number of DCT coefficients, is examined further in Section 3.5.2.



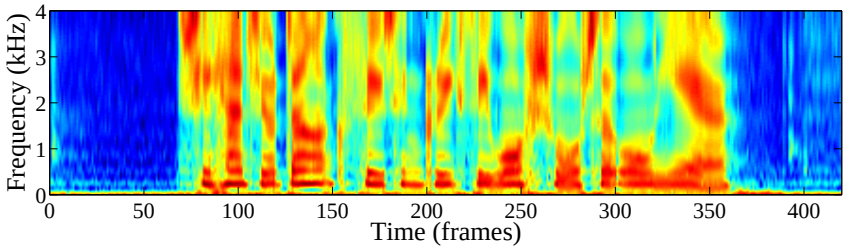
(a) Original Spectrum



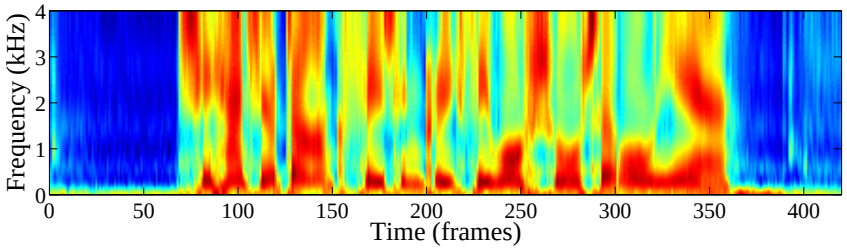
(b) 64 coefficients retained from 64 point transform



(c) First 32 coefficients retained from 64 point transform



(d) First 16 coefficients retained from 64 point transform



(e) First 8 coefficients retained from 64 point transform

Figure 3.21: Effect of feature inversion on the log-magnitude spectrum with varying number of DCT coefficients retained

3.4.3 LPC-based features

LPC coefficients are a common feature in speech reconstruction applications due to their use in the LPC vocoder. This section examines line spectral frequencies (LSF) as an alternative to the standard LPC coefficients. As described in Section 3.3.1, LPC coefficients model the spectral envelope of a signal as an all-pole filter. LSFs are a transform of LPC coefficients which have several attractive properties such as a smaller sensitivity to quantisation noise, and therefore estimation, and a more direct link to formant positions and bandwidths.

3.4.3.1 Line spectral frequencies (LSF)

In the z-domain LPC coefficients have roots anywhere within the unit circle. LSFs decompose the LPC coefficient polynomial:

$$\hat{x}(m) = \sum_{i=1}^{P} a_i x(m-i), \tag{3.28}$$

into the sum filter:

$$P(z) = A(z) + z^{-p+1}A(z^{-1}), \tag{3.29}$$

and difference filter:

$$Q(z) = A(z) - z^{-p+1}A(z^{-1}), \tag{3.30}$$

where both $P(z)$ and $Q(z)$ have roots directly on the unit circle and correspond to the vocal tract with the glottis closed and open respectively. This representation is useful as LSFs can be shown to correspond to formant locations and bandwidths [Itakura, 1975].

3.4.3.2 Feature inversion

Spectral values can be obtained directly from LSF parameters. A filter is constructed, equivalent to the LPC filter, i.e:

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 + 1/2\left[(P(z) - 1) + (Q(z) - 1)\right]}, \tag{3.31}$$

where A is the LPC coefficients, P are the filter poles and Q are the filter zeros in the Z-domain. The spectral envelope is therefore the frequency response of this filter Kondoz [2004].

3.5 Results

This section presents results of a series of experiments used to determine the most suitable speech reconstruction model and also the best feature configuration to encode the spectral envelope.

First, results of experiments that objectively measure the optimal speech quality of a range of speech reconstruction methods are presented in Section 3.5.1. Once the most appropriate model has been chosen and the required acoustic features are known, Section 3.5.2 determines the optimal acoustic feature configuration in terms of encoding the spectral envelope. To determine the effect this has on the quality of reconstructed speech experiments measuring spectral distortion and objective speech quality are carried out using a range of feature configurations. Finally, the correlation between feature vectors extracted from clean and noisy speech is examined in Section 3.5.3.1 to determine the most suitable feature in terms of clean feature estimation.

Table 3.1: Objective quality, as measured using PESQ and LLR, of 100 utterances from different speakers reconstructed using STRAIGHT, LPC vocoder and HNM.

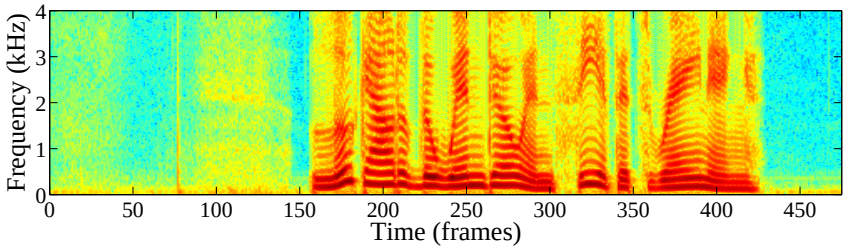
Method	PESQ	LLR
STRAIGHT	3.66	0.39
LPC vocoder	3.01	0.22
HNM	3.51	0.20

3.5.1 Quality of speech reconstruction

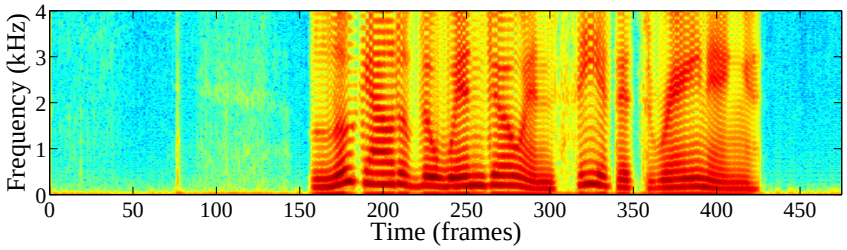
This section examines the quality of speech reconstructed using the models considered in this work. The results presented focus on clean speech reconstruction. Fundamental frequency was estimated from laryngograph recordings taken at the original time of recording whilst all other acoustic features were obtained directly from clean speech. Three of the four described reconstruction models are tested: LPC vocoder, HNM and STRAIGHT. The sinusoidal model was not included as the HNM is essentially an enhancement of the sinusoidal model. In each case female speech from the NuanceCatherine dataset was reconstructed and performance measured using objective quality measures. Utterances were originally recorded using a 16kHz sample rate and downsampled to 8kHz. 246 utterances, with an average duration of ≈ 4 seconds, were reconstructed using each reconstruction model.

Figure 3.22 shows narrowband spectrograms of the utterance *"Look out of the window and see if it's raining"* from the original recording and reconstructed using each of the reconstruction models. On visual inspection HNM and STRAIGHT reconstruct speech true to the original. Whilst the LPC vocoder clearly reconstructs the harmonics of voiced frames and captures most of the formant structure there are still considerable differences between the reconstructed speech and the original, primarily in inter-formant regions.

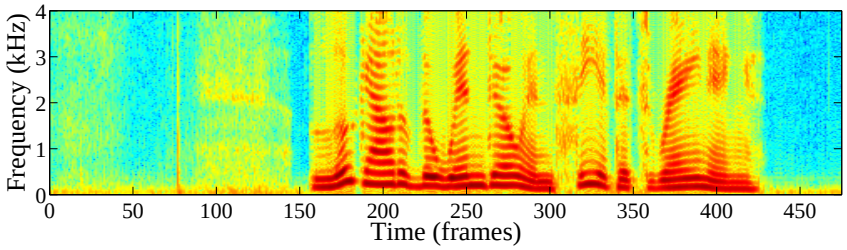
The result of measuring the quality of reconstructed speech using objective measures are presented in Table 3.1. PESQ and LLR were used to measure objective speech quality and spectral distortion respectively. PESQ shows speech recon-



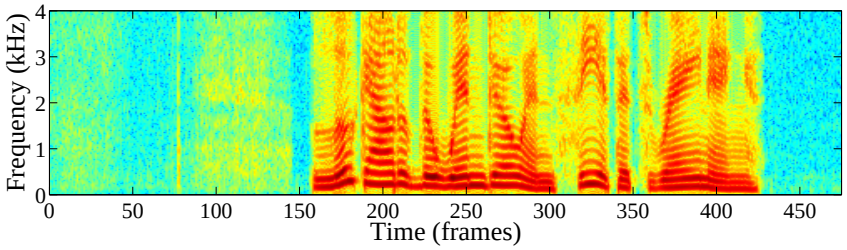
(a) Original utterance



(b) LPC vocoder



(c) HNM



(d) STRAIGHT

Figure 3.22: Comparison of narrowband spectrograms of a.) clean speech and speech reconstructed using: b.) LPC vocoder, c.) HNM and d.) STRAIGHT

structed using STRAIGHT to have the highest quality. This is despite significantly different results in terms of spectral envelope distortion as measured by LLR. This can be described by the objectives of STRAIGHT as a method of speech modification rather than straight-forward speech reconstruction. Using STRAIGHT, the spectral envelope is warped based on the estimate of the fundamental frequency. Changes in fundamental frequency are reflected in altered formant positions, thus affecting LLR results whilst maintaining high perceptual quality when errors in fundamental frequency are low.

Considering now the differences between the HNM and vocoder, a much smaller difference between LLR results is observed. This is due to neither method altering the spectral envelope during the reconstruction process. Much larger differences are observed when considering perceptual quality. Speech reconstructed using the vocoder is measured to be 0.5 MOS points worse than with HNM and 0.65 MOS points worse than with STRAIGHT. This is attributed to the use of an artificial phase offset by the source/filter vocoder which causes the model to produce 'robotic' sounding speech. In addition, the Dirac delta function used to model the vocal tract excitation in the production of voiced speech is an overly simplistic interpretation of the excitation signal which causes a 'buzzing' sound in the reconstructed speech.

An ideal speech reconstruction method for the purposes of this application will produce high quality speech, preferably transparent to the original speech in terms of observed quality, and will be driven by a minimal set of easily-obtainable acoustic features. STRAIGHT clearly meets the first criterion, however the fine resolution in both frequency and time makes this method unsuitable. In acoustic environments with no background noise estimates of these high resolution parameters are relatively easy to obtain, however the addition of noise makes this task prohibitively difficult. Instead, HNM is considered to be the most suitable reconstruction model. PESQ results have shown speech reconstructed using HNM to be within 0.15 MOS points of STRAIGHT whilst the required model parameters are substantially lower resolution and should therefore be easier to estimate.

3.5.2 Acoustic feature configuration

Speech is reconstructed using a reconstruction model driven by a set of acoustic features. In this work we are interested in estimating clean acoustic features from those extracted from noisy speech. In this case it is potentially beneficial to parameterise speech frames and to then subsequently estimate the required acoustic features from this new parameterisation. In this section the aim is to therefore determine the most suitable method of parameterisation. Several methods of parameterisation are tested in this section and include: spectrum-based features, LPC-based features and linear and Mel-spaced filterbanks. From now on 'feature' is used to refer to the parameterised speech frame whilst the term 'acoustic features' is used to refer to the features used to drive the reconstruction model.

The optimal feature configuration is a careful balance of maximising overall speech quality and acoustic feature correlation and, where possible, minimising number of elements which comprise the feature, M. These properties are all closely related. The quality of reconstructed speech is directly related to the quality of the acoustic features used to drive the reconstruction model. These features will ultimately be estimated directly from the features extracted from noisy speech and so the feature must contain sufficient information relating to the acoustic features, i.e. there must be a sufficient degree of correlation between acoustic features and feature vectors extracted from the noisy speech. To improve the efficiency of estimation the feature will ideally also be compact; a significant amount of redundancy typically exists in the complex spectra of speech and many methods of parameterisation aim to reduce this through compression. Whilst this is usually beneficial, the level of compression must be managed in order not to lose important information.

Previously, the HNM was determined to be the most appropriate model of reconstruction. This model is driven by the following acoustic features: spectral envelope, fundamental frequency, voicing classification and phase. Experiments in this section aim to determine the optimal feature for encoding this information by reconstructing speech using spectal envelope values determined from a range of parameterisations

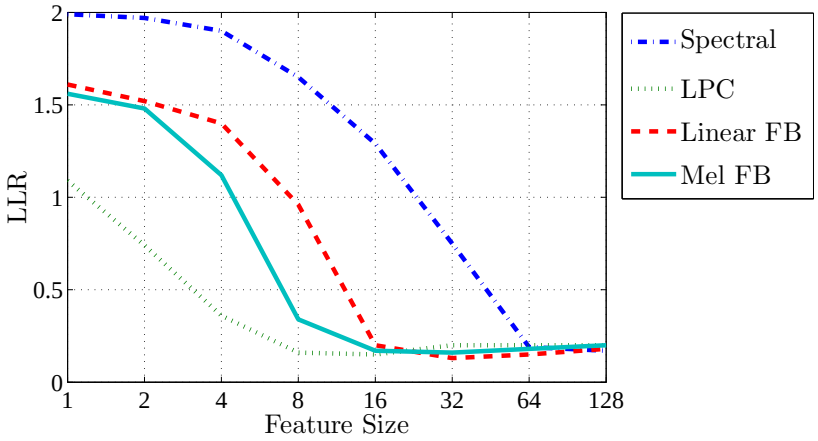


Figure 3.23: Mean LLR of 246 utterances of female speech reconstructed using the HNM using spectral amplitudes estimated from spectral features, LPC, linear-spaced filterbanks and Mel-spaced filterbanks

of the speech. All other acoustic features are obtained directly from the clean speech as per Section 3.5.1. This allows us to determine directly the effect that compressing the spectral envelope has on the quality of reconstructed speech. Later chapters examine the effect of obtaining such estimates from features extracted from noisy speech in terms of spectral envelope, fundamental frequency, voicing classification and phase.

Speech is reconstructed using the same data as in Section 3.5.1, that is 246 utterances of female speech sampled at 8kHz recorded in clean conditions. Considering first the effect of feature size on spectral envelope distortion, Figure 3.23 displays results of reconstructing clean speech using the HNM driven by spectral envelope estimated from a range of feature configurations. In each case the number of coefficients are varied for each feature type to determine the effect this has on spectral distortion. Spectrum-based features are clearly sensitive to vector size with speech reconstructed from feature vectors made up of less than 64 coefficients suffering from significant distortion. On the other hand, LPC based methods are shown to perform very well even with a minimal number of coefficients, though it should be noted that as LLR is based on an LPC representation of the spectral envelope results may be biased towards this class of feature. An interesting comparison can

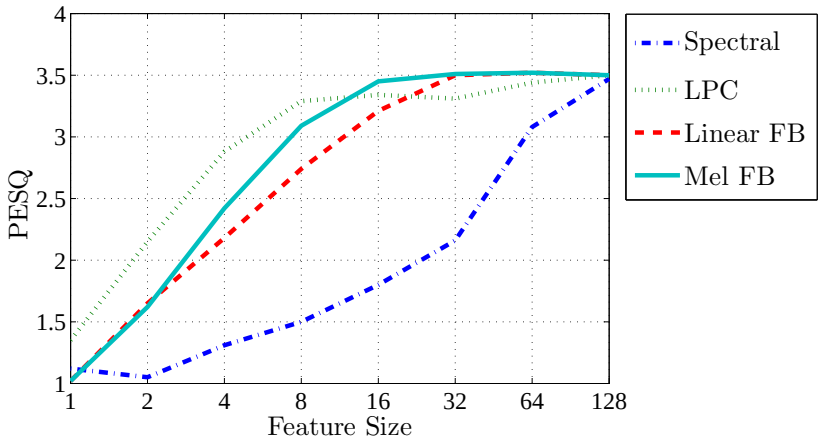


Figure 3.24: Mean PESQ of 246 utterances of female speech reconstructed using the HNM using spectral amplitudes estimated from spectral features, LPC, linear-spaced filterbanks and Mel-spaced filterbanks

be made between linear and Mel-spaced filterbanks. Mel-spaced filterbanks perform considerably better in situations where there are relatively few channels ($M < 16$), however when there are a larger number of channels results are roughly equal. This is attributed to the much higher density of filterbank channels in low frequency bands which make up the majority of the high-energy detail in voiced speech. As the number of filterbank channels rises this benefit reduces as the density of linear filterbank channels approaches that of the Mel-spaced feature.

In terms of overall speech quality, Figure 3.24 shows results taken from the same experiment, this time measured using PESQ. These results show a relatively high degree of correlation between spectral envelope distortion and overall objective quality. The two notable differences between the LLR and PESQ results concern spectrum and LPC-based features. LLR results are biased by the use of LPC in the measurement process. PESQ ratings show the difference in quality between LPC and the two filterbank methods to be much smaller, with LPC not realising the full potential of the reconstruction model until reaching 128 coefficients. Spectral features are again shown to be ineffective at lower feature sizes, though this time performance degrades significantly when $M < 128$.

Based on the results presented in this section a Mel-spaced filterbank-based fea-

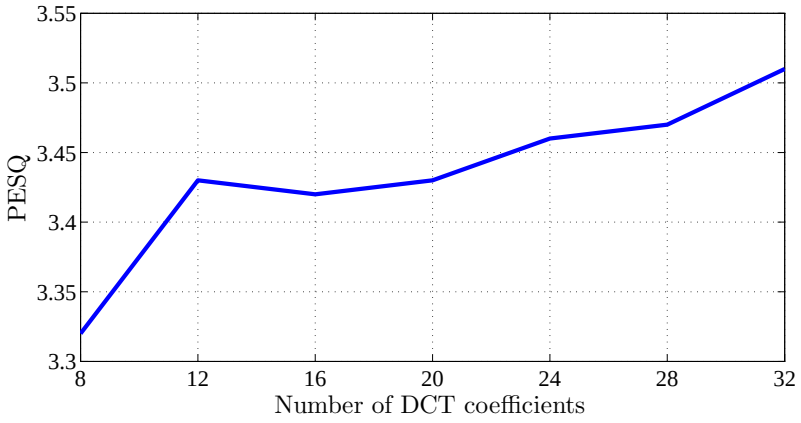


Figure 3.25: Effect of discarding DCT coefficients from 32-dimensional MFCC feature vectors on speech quality as measured using PESQ

ture has been chosen for use in the remainder of this work. The Mel-spaced filterbank has been shown to offer low levels of spectral distortion and high perceptual quality at low feature sizes. The most appropriate feature vector size was determined to be $M = 32$ as this is the point at which the speech reconstruction model becomes the limiting factor.

A final variable that may be optimised when using the MFCC feature is the number of DCT components that are retained. High-order coefficients may be discarded to further reduce the size of the feature vector, smoothing the spectral representation in the process. This could allow for more efficient estimation at the enhancement stage. Figure 3.25 shows the effect of discarding higher order coefficients on objective speech quality. Mel-filterbank features with $M = 32$ coefficients were extracted from clean speech and then transformed to the cepstral domain through the use of the DCT. High order coefficients were then discarded to give the new feature size. As expected, speech quality degrades as the number of DCT coefficients is reduced. Whilst a reduction in quality can be expected, the relatively steep drop off in quality was decided to be too much to out-weigh any benefits in compression and so the full set of 32 DCT coefficients was retained.

3.5.3 Acoustic feature correlation

This section presents results of measuring the correlation between intermediate feature vectors (i.e. MFCC, LPC, etc.) and acoustic features, namely: spectral envelope and $f0$ and voicing class. Correlation is measured using multiple linear regression. A linear model is built which describes the relationship between the independent variable, $\mathbf{x}$, (intermediate feature) and θ, the dependent variable (acoustic feature). The jth element of the acoustic feature, $\theta(j)$, may then be represented in terms of the intermediate feature vector, $\mathbf{x}$, and a set of $P+1$ regression coefficients, $\mathbf{b}$, as follows:

$$\theta(j) = b_{j,0} + b_{j,1}x(1) + b_{j,2}x(2) + \cdots + b_{j,P}x(P) + \epsilon, \tag{3.32}$$

where ϵ represents the modelling error. Regression coefficients are computed as described by Chatterjee and Hadi [1986]t. These coefficients are then used to predict the value of $\hat{\theta}(j)$, from $\mathbf{x}$. The correlation can then be measured in terms of the R^2 which is defined as:

$$R(j)^2 = 1 - \frac{\sum_i \left(\theta_i(j) - \hat{\theta}_i(j)\right)^2}{\sum_i \left(\theta_i(j) - \bar{\theta}_i(j)\right)^2}, \tag{3.33}$$

where $\bar{\theta}_i(j)$ is the mean of the jth coefficient of θ and i relates to the frame number. This term was calculated for every coefficient and then averaged to give a global R term:

$$R = \sqrt{\frac{1}{J}\sum_{j=1}^{J} R(j)^2}. \tag{3.34}$$

R^2 is the proportion of variance in the clean speech that can be accounted for by knowing the noisy speech and so higher values of R are preferred.

The correlation between clean and noisy intermediate features is first examined for the purpose of spectral envelope estimation in Section 3.5.3.1 before correlation

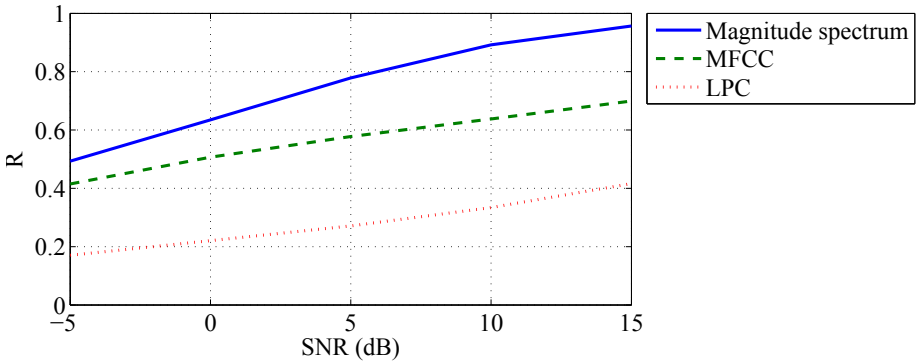


Figure 3.26: Correlation between clean and noisy feature vectors extracted from 30 minutes of female speech using: magnitude spectrum, MFCC and LPC features.

between intermediate features and $f0$ is investigated in Section 3.5.3.2. Finally, voicing class correlation is examined in Section 3.5.3.3.

3.5.3.1 Spectral envelope correlation

For an accurate estimate of spectral envelope to be computed sufficient information relating to the clean speech features must be contained within the noisy features. This is measured in this section using multiple linear regression. Results are presented in terms of the correlation coefficient, R.

First, feature correlation across a range of noise levels is examined. White noise was added to 30 minutes of clean speech from ten speakers at a range of SNRs. The correlation between intermediate feature vectors extracted from clean and noisy speech was then measured for each SNR. Three intermediate features were considered: LPC, MFCC and magnitude spectrum. Feature sizes determined as optimal in Section 3.4 were used, that is 16 coefficents for LPC, 32 for MFCC and 128 for spectral features. Figure 3.26 shows how feature correlation is affected by SNR. LPC offers worst performance with very little correlation across all noise levels. Magnitude spectra offer best performance with very high correlation especially at high SNR. Whilst MFCCs do not offer best performance they are still deemed to be the most appropriate feature for efficient enhancement as they offer reasonable correlation and are the most easily modelled.

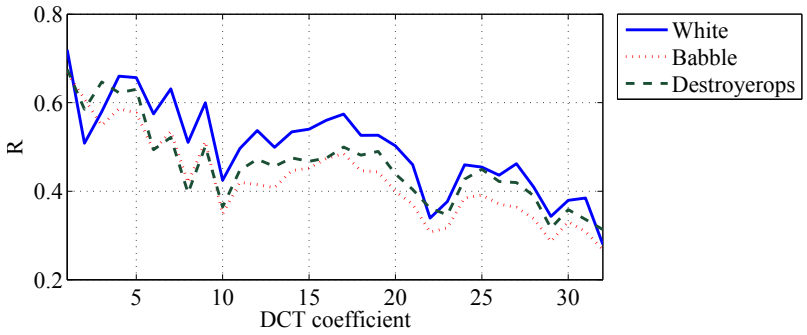


Figure 3.27: Individual feature correlation between clean and noisy MFCCs extracted from 30 minutes of female speech using white, babble and destroyerops noise at 0dB SNR

Figure 3.27 examines mean correlation across each coefficient. Most correlation is likely to exist between low-order coefficients which represent coarse spectral detail whilst the fine detail represented by high-order coefficients expected to be low. Three different types of noise are considered: white, babble and destroyerops (Appendix A). All noises were mixed with clean speech at 0dB SNR. Spectral envelope information is contained in low-quefrency coefficients and so it is not unexpected that correlation is shown to be inversely proportional to quefrency. Looking in more detail, there are two regions that have higher than expected correlation; in the mid-range a fairly broad range of coefficients display high correlation whilst a group of higher order coefficients also have higher than expected correlation. This high quefrency region can be seen as representing the harmonic structure of the frame.

All experiments so far have examined the global correlation of speech. In this work a localised approach to estimation is proposed (Section 5.3). It is therefore also useful to examine the within class correlation. Appendix B shows within class correlations for MFCCs in white noise at 0dB SNR. A phoneme level classification scheme using forced-alignment decoding was used to segment the data. Correlation profiles are shown to be similar across phonemes within the same articulation class confirming the validity of the articulation class scheme for estimation. By comparing each phoneme-specific plot to the global results in Figure 3.27 it is also possible to predict which phonemes will be most accurately estimated. Table 3.2 displays the average correlation for each phoneme as well as the proportion of the training data

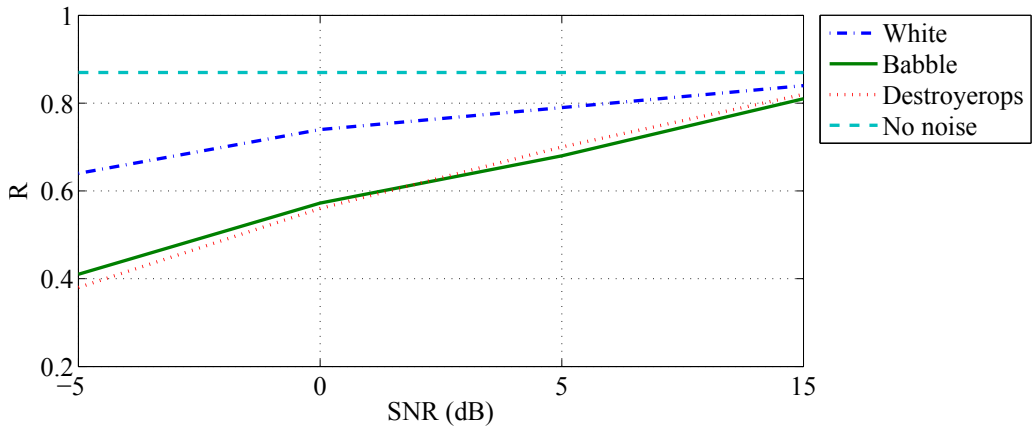


Figure 3.28: Mean overall correlation between MFCC features extracted from noisy speech and $f0$ in white, babble and destroyerops noises at varying SNRs

which makes up each class.

Correlation is clearly higher for diphthongs, liquids, monophthongs, R-coloured vowels and semi-vowels whilst silence, affricates, fricatives, nasals and stops all have lower than average correlation. This is attributed to the local SNR of each class; classes exhibiting higher correlation are typically those with more energy and therefore higher SNR, assuming constant noise energy. Unvoiced phonemes also have less defined structure due to the noise-like excitation which is likely to reduce correlation.

3.5.3.2 Fundamental frequency correlation

In terms of $f0$ estimation from MFCCs it is important to determine the level of correlation that exists between $f0$ and MFCC features.

Correlation in various levels of noise is measured. MFCCs were extracted from noisy speech at a range of SNRs whilst reference $f0$ was used. MFCCs with 32 coefficients are used as per spectral envelope estimation and were extracted using a frame width of 20ms at 100fps. 30 minutes of multi-speaker speech was used for testing. Figure 3.28 shows correlation as measured using multiple linear regression. Correlation between $f0$ and clean speech is measured to be very high with a value of $R = 0.87$ showing that just over 75% of the variance of $f0$ is represented by the clean speech. Correlation clearly reduces in noisy conditions, though relatively little

Table 3.2: Mean phoneme correlation in white noise at 0dB SNR

Acoustic Class	Phoneme	R	Dataset %
Global	*	0.51	100
Silence	sil	0.09	23.48
Affricates	ch	0.33	0.58
	jh	0.36	0.68
Diphthongs	aw	0.71	0.69
	ay	0.71	1.97
	ey	0.70	2.10
	ow	0.71	1.27
	oy	0.67	0.29
Fricatives	dh	0.47	1.31
	f	0.26	1.91
	hh	0.48	0.98
	s	0.15	6.09
	sh	0.40	1.50
	th	0.24	0.54
	v	0.41	1.38
	z	0.25	3.34
	zh	0.48	0.09
Liquids	l	0.60	2.53
	r	0.66	2.31
Monophthongs	aa	0.71	1.04
	ae	0.67	1.81
	ah	0.70	1.09
	ao	0.68	1.55
	ax	0.56	3.13
	eh	0.69	2.03
	ih	0.58	4.29
	iy	0.63	1.57
	uh	0.60	0.17
	uw	0.62	1.06
Nasals	m	0.53	2.02
	n	0.49	4.98
	ng	0.50	0.90
R-coloured vowels	ea	0.73	0.36
	er	0.71	0.86
	ia	0.64	0.63
	ua	0.68	0.13
Semi-vowels	w	0.62	1.11
	oh	0.68	1.31
	y	0.63	0.94
Stops	b	0.34	1.24
	d	0.30	2.51
	g	0.35	0.65
	k	0.18	3.53
	p	0.23	2.39
	t	0.18	5.65

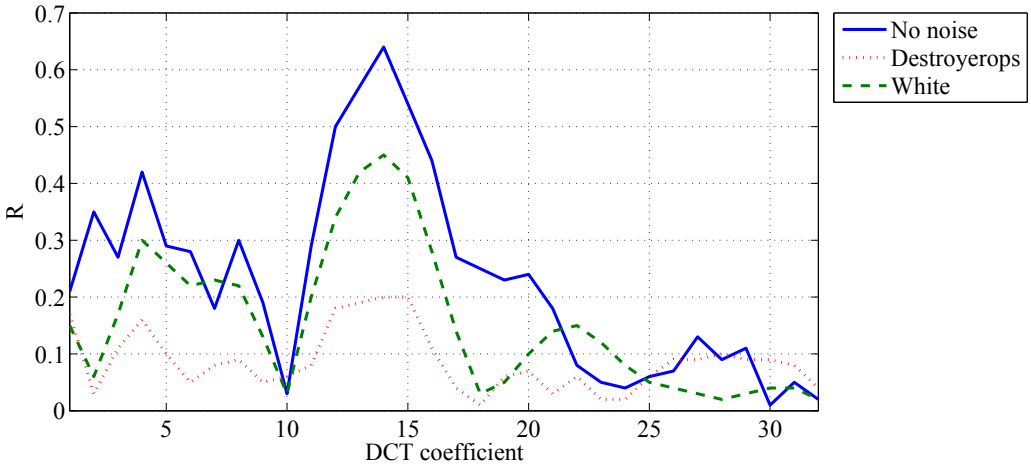


Figure 3.29: Individual coefficient correlation between $f0$ and MFCCs with no noise, white noise and destroyerops noise

effect is measured at 15dB SNR across all noises. Significant differences are noted at very low SNR between stationary and non-stationary noise types. Correlation in white noise is shown to degrade to $R = 0.64$ at -5dB whilst the non-stationary noises, namely babble and destroyerops noise, are shown to reduce correlation to as little as 0.41 and 0.38 respectively.

Figure 3.29 therefore now examines coefficient-level correlation in the cepstral domain. Three conditions are tested: no noise, white noise at 0dB SNR and destroyerops noise at 0dB SNR. All data was taken from the same female speaker from the NuanceCatherine dataset. Periodicity in the frequency domain is represented as peaks in the cepstral domain and so there should be relatively strong correlation in a small range of cepstral coefficients relating to the fundamental frequency and its harmonics. A clear peak in correlation is visible around the 14th coefficient which relates to this harmonic structure. In clean conditions $R = 0.64$, reducing to $R = 0.45$ when white noise is added at 0dB SNR. Correlation is further reduced when destroyerops noise is added in place of the white noise at the same SNR resulting in a value $R = 0.20$.

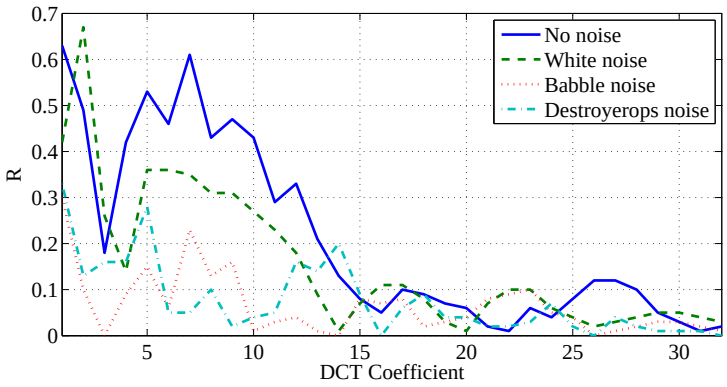


Figure 3.30: Correlation between voicing class and MFCCs in white noise, babble noise and destroyerops noise at 0dB SNR

3.5.3.3 Voicing class correlation

This section examines the correlation between MFCCs and voicing class. The HNM reconstruction model requires frames to be classified based on a binary classification problem, i.e. does a frame contain voiced speech or not. Features must therefore contain sufficient voicing information for such a classification to be possible. The main discriminating features between frames of voiced and unvoiced speech can be seen to be the periodicity of voiced frames and the energy distribution across the spectrum; the majority of unvoiced energy is typically found in higher frequency regions whilst voiced energy is typically focused in the lower frequency regions. These properties efficiently captured by the MFCCs used for classification, $\mathbf{c}$. $c(0)$ can be seen to model overall frame energy whilst $c(1)$ models spectral tilt. Source information, and therefore the periodicity of the frame, is captured in mid-order coefficients whilst high-order coefficients model fine detail. As we are only interested in distinguishing between voiced speech and 'not' voiced speech (i.e. unvoiced/non-speech), it is clear that the information required to accurately model the discriminating features will be represented by a limited selection of feature coefficients.

Figure 3.30 shows the correlation between the voicing class and each feature coefficient as measured using multiple linear regression. Higher values of R represent

a stronger correlation between coefficient and voicing class. Features were extracted from speech from five male and five female speakers in clean conditions as well as in white, babble and destroyerops noises at 0dB SNR. Correlation is shown to be relatively high in clean conditions with noise shown to reduce correlation, especially in the case of babble and destroyerops which are both examples of non-stationary noises. Based on these results it is clear that the majority of useful information exists between $c(0)$ and $c(15)$. This result is reinforced by the spectrogram plots in Figure 3.21 which show that considerable source information remains in MFCC feature vectors when as few as 16 coefficients are retained from a 64-point transform. We may therefore consider reducing the dimensionality by discarding high-order coefficients. This should increase modelling efficiency by removing some of the finer spectral detail which has been shown to be unimportant for voicing classification.

3.6 Summary

In this chapter several speech reconstruction models were considered for their suitability for use in the proposed speech enhancement system. In addition to considering the complexity of the models, the HNM was determined to be the most suitable model through objective tests of overall speech quality and spectral distortion. Whilst STRAIGHT offers the best overall speech quality, this comes at the expense of a considerably more complex system with much higher feature requirements. This would be highly likely to hinder any attempts at feature enhancement due to the very high resolution requirements in both time and frequency. In terms of feature selection, MFCCs were determined to be the most effective method of modelling the spectral envelope. This decision was made after a series of objective tests including assessment of objective speech quality and spectral distortion.

The quality of speech reconstructed by the HNM as well as the low spectral distortion and compact representation of the spectral envelope given by the MFCCs gives a good platform on which to base the proposed speech enhancement system.

Chapter 4

Methods of Feature Estimation

This chapter examines methods of acoustic feature estimation. This method of speech enhancement reconstructs speech using a reconstruction model driven by a set of acoustic features. These acoustic features must be estimated from noisy speech and so methods of robust estimation are examined. The chapter begins by reviewing related methods of robust feature estimation. Maximum *a-posteriori* estimation is identified as a suitable method of estimation and described in detail. The remainder of the chapter focuses on improving the robustness of the estimation models through the use of stereo trained models and model adaptation.

Contents

4.1 Introduction

The HNM was selected in Chapter 3 as a suitable method of reconstruction for the proposed method of speech enhancement. The model is driven by a set of acoustic features, namely: spectral envelope, fundamental frequency, voicing and phase. MFCCs were identified as a suitable method of encoding speech frames and so a suitable method of estimating acoustic features of clean speech from MFCC features extracted from noisy speech is required. Such acoustic feature estimation is not a new idea, with several such applications developed over the last twenty years for the purposes of robust automatic speech recognition (ASR) [Deng et al., 2000] and speech reconstruction [Chazan et al., 2000]. This chapter therefore begins by reviewing existing literature on the topic of acoustic feature estimation in Section 4.2. MAP estimation is identified as a suitable method of robust feature estimation and so the technique is described in Section 4.3. One key challenge in such a system is how to obtain the model of joint density. Two methods are considered. First, stereo training data may be used to train models directly (Section 4.4). Stereo-trained systems build robustness into the model by training on data from the same speaker and noise as the target environment. Alternatively, models trained on a variety of speakers in clean conditions may be adapted to the target speaker and environment using methods of adaptation as described in Section 4.5.

4.2 Feature Estimation Review

Acoustic feature estimation has roots in both straight-forward acoustic feature estimation and robust ASR. In the literature there are several examples of acoustic feature estimation in both clean and noisy conditions, notably fundamental frequency and formant estimation, whilst methods of feature and model compensation developed for robust ASR are also of use for this method of speech enhancement.

The section begins by examining related work on acoustic feature estimation in Section 4.2.1. The performance of model-based estimation is known to reduce when

the feature domain is not matched to the model domain. A significant amount of work has been done in the field of robust ASR on reducing such mismatches in terms of both speaker and noise. Methods developed for use in robust ASR applications are therefore examined in Section 4.2.2.

4.2.1 Acoustic feature estimation

Starting with case of estimating acoustic features from clean speech, a significant amount of correlation was measured between clean speech and fundamental frequency and formants by Darch et al. [2007]. Droppo and Acero [1998] were the first to propose such a scheme for fundamental frequency estimation. Using a feature vector based on the concept of 'predictable energy' it was demonstrated that fundamental frequency may be estimated from clean speech using MAP estimation. Later, Tabrikian et al. [2004] presented a similar method of fundamental frequency estimation where it was shown that MAP estimation-based methods can provide robust estimates in very low SNRs (-15dB). In addition, Milner et al. [2005] successfully estimated formant frequencies and voicing classification from relatively low dimensional clean speech MFCCs using MAP estimation. Whilst estimating such acoustic features from clean speech is useful for applications such as playback in distributed speech recognition (DSR) systems, in this application the acoustic features must be estimated from noisy speech. In the case of DSR speech is reconstructed using a reconstruction model driven by a set of acoustic features estimated from the MFCC originally extracted for speech recognition [Chazan et al., 2000; Milner and Shao, 2006]. The proposed method of speech enhancement requires that these acoustic features are estimated from MFCCs extracted from noisy speech.

There are two approaches to estimating acoustic features from noisy speech. These can be categorised as either model-based or feature-based. Model-based techniques aim to adapt model parameters to the domain of the noisy features whilst feature-based methods compensate features to match the model domain. A review of both techniques for the purpose of $f0$, voicing and formant estimation was car-

ried out by Milner et al. [2008]. In this review spectral subtraction was used as the method of feature compensation whilst the log-normal approximation described by Gales and Young [1993] was used as the method of model adaptation. Experiments with no noise compensation and matched training/testing environments were also performed. Overall, the matched system was shown to offer best performance whilst the uncompensated systems performed worst. The model adapted system performed best out of the two compensation systems though in the majority of cases performance was not significantly different to using feature-based compensation using spectral subtraction.

Whilst good results were achieved by Milner et al. [2008] using spectral subtraction and log-normal model adaptation, the limitations of these techniques have been widely discussed with respect to robust ASR [Hu and Huo, 2006; Shinohara and Akamine, 2009; Li et al., 2012]. Techniques developed for robust ASR are therefore examined in Section 4.2.2.

4.2.2 Feature estimation for robust ASR

ASR recognition accuracy is known to fall with the introduction of noise or changes in speaker. A wide range of techniques used to increase robustness have been developed and as per acoustic feature estimation these can be broadly categorised as either feature-based or model-based. The objective of the feature-based methods is to transform the feature space to the domain of the clean speech to allow the use of existing clean-trained acoustic models. There are two main ways in which this can be achieved. Feature compensation techniques aim to directly process features to improve robustness and are discussed in Section 4.2.2.1 whilst feature estimation methods use estimation techniques to directly estimate cleaned features from noisy features and are discussed in Section 4.2.2.2. The objective of model-adaptation techniques is to adapt the model-domain to the feature-domain and are discussed in Section 4.2.2.3.

4.2.2.1 Feature compensation

This section discusses methods of feature compensation. Feature normalisation, as described by Viikki and Laurila [1998], is a method of feature compensation and can improve recognition results by normalising for variations in mean and variance resulting from the addition of noise. An extension of this method includes smoothing by an ARMA filter leading to a method known as MVA (Mean, Variance and ARMA filtering) [Chen et al., 2005]. Alternatively, conventional speech enhancement methods can also be applied for feature enhancement. These methods may be applied in the time domain and then features extracted from the modified signal or adapted for use in the feature domain itself. Popular methods of enhancement in this category include spectral subtraction and Wiener filtering [Vaseghi and Milner, 1997] as well as log MMSE [Yu et al., 2008]. These methods have been shown to offer significant reductions in recognition error, with improvements of over 70% in word error rate typically observed relative to uncompensated systems [Viikki and Laurila, 1998] but suffer from the same issues as when applied for speech enhancement. A significant issue with conventional enhancement methods is the introduction of artifacts known as 'musical noise' (Section 2.2). Conventional techniques typically work by filtering out an estimate of the noise, which is assumed to be stationary. Subsequently, if the noise is non-stationary over, or under, filtering of the noise occurs which results in erroneous peaks in the spectrum. In addition, *a-priori* knowledge of clean speech is typically not incorporated into such techniques and so there is no control over whether the output is even a valid speech frame. These shortcomings motivated the development of a scheme of feature *estimation* rather than feature compensation which will be described next.

4.2.2.2 Feature estimation

The motivation behind feature estimation rather than feature compensation is that, given prior knowledge of the relationship between clean and noisy features, clean features may be estimated directly from noisy features. Such a method is described

by Deng et al. [2000]. The technique developed was named SPLICE (stereo piecewise linear compensation for environments) and built upon work by Acero and Stern [1990] whereby quantised clean speech vectors were obtained from noisy feature vectors through the use of a codebook and vector quantisation (VQ) which was itself closely related to earlier work by Neumeyer and Weintraub [1994].

Instead of directly estimating the noise, SPLICE models the relationship between clean and noisy speech by training on stereo data consisting of time aligned vectors of clean and noisy speech, i.e. $\mathbf{z} = [\mathbf{y}, \mathbf{x}]$, where $\mathbf{x}$ are MFCCs extracted from clean speech. $\mathbf{y}$ represents MFCCs extracted from the same utterances after noise has mixed with the clean speech. The system is therefore able to model relatively non-stationary noise assuming the noise is well represented in the training data. This method of model-based estimation is also beneficial as the model includes an inherent model of clean speech. Given observations of noisy speech, the model of clean and noisy speech can be used to estimate the corresponding clean feature vectors using MAP estimation (Section 4.3). These estimated feature vectors may subsequently be used as the front end of an unadapted ASR system.

SPLICE is shown to perform very well for the task of robust ASR, outperforming spectral subtraction-based systems and even those using matched train/test conditions [Deng et al., 2001]. Several further systems have been developed based on SPLICE such as those described by Cui et al. [2008] and Afify et al. [2007]. All of these systems are related by the requirement of a model of the joint distribution of clean and noisy speech, with slightly different approaches taken in terms of feature estimation. Clearly the reliance on stereo training data is the limiting factor and so various methods exist to increase the flexibility of such methods. For example, Stouten et al. [2003] and Deng et al. [2004] describe techniques whereby VTS is used to construct the required joint distribution from separate models of clean and noisy speech.

4.2.2.3 Model adaptation

Model-based methods of achieving robust ASR are now considered. The method described by Gales and Young [1993] was originally developed for ASR rather than acoustic feature estimation. Whilst results of this method are shown to be good there are several other methods of adaptation which have been shown to offer better performance. PMC [Gales and Young, 1995], MAP [Gauvain and Lee, 1994], MLLR [Gales and Woodland, 1996], VTS [Acero et al., 2000] and most recently, the Unscented Transform (UT) [Li et al., 2010] have all been demonstrated to provide superior performance with UT offering best performance based on results published by Shinohara and Akamine [2009]. Such methods will therefore be considered for the purpose of acoustic feature estimation for this method of speech enhancement.

4.3 Maximum *a-posteriori* Estimation

Maximum *a-posteriori* (MAP) estimation is a method of Bayesian estimation used to obtain point-estimates of an unobserved quantity based on empirical data. MAP estimation is closely related to maximum likelihood (ML) estimation, differing in the optimisation objective by incorporating the prior distribution of the unobserved quantity. In the case of this work, acoustic features, which include clean spectral envelope and fundamental frequency, are estimated from feature vectors obtained from noisy speech. It is preferred for use over ML estimation as the priors required for estimation may be obtained from the training data. MAP estimation has been proven effective for acoustic feature estimation in existing works, i.e. Darch et al. [2006]; Hadir et al. [2011]; Lotter and Vary [2005]; Deng et al. [2001]. This section begins by providing a general definition of MAP estimation in Section 4.3.1. This method of estimation requires a model of the joint density of the feature and unobserved quantity and so Gaussian mixture models (GMMs) are examined in Section 4.3.2. MAP estimation using GMMs is then described in Section 4.3.3.

4.3.1 General definition

Given suitable information about the joint distributions of the unobserved quantity, $\boldsymbol{\theta}$, and our observations, $\mathbf{x}$, we can compute an estimate of $\boldsymbol{\theta}$ given $\mathbf{x}$. The *likelihood function*, $f(\mathbf{x}|\boldsymbol{\theta})$, gives the probability of $\mathbf{x}$ when the unobserved quantity is equal to $\boldsymbol{\theta}$. Finding the value of $\boldsymbol{\theta}$ which maximises the output of this function gives us the *maximum likelihood* estimate of $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} f(\mathbf{x}|\boldsymbol{\theta}). \tag{4.1}$$

Assuming now that the prior distribution over $\boldsymbol{\theta}$, g, is available allows $\boldsymbol{\theta}$ to be treated as a Bayesian random variable. Applying Bayes' theorem gives the posterior distribution:

$$f(\boldsymbol{\theta}|\mathbf{x}) = \frac{f(\mathbf{x}|\boldsymbol{\theta})\, g(\boldsymbol{\theta})}{\displaystyle\int_{\vartheta \in \boldsymbol{\theta}} f(\mathbf{x}|\vartheta)\, g(\vartheta)\, d\vartheta}. \tag{4.2}$$

The MAP estimate of $\boldsymbol{\theta}$ given $\mathbf{x}$ is therefore defined as:

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{x}) = \arg\max_{\boldsymbol{\theta}} \frac{f(\mathbf{x}|\boldsymbol{\theta})\, g(\boldsymbol{\theta})}{\displaystyle\int_{\vartheta \in \boldsymbol{\theta}} f(x|\vartheta)\, g(\vartheta)\, d\vartheta} = \arg\max_{\theta} f(\mathbf{x}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) \tag{4.3}$$

which is equivalent to the mode of the posterior distribution. The denominator of the posterior probability does not depend on $\boldsymbol{\theta}$ and so can be dropped from the derivation. Comparing the ML and MAP estimates, given in Equations 4.1 and 4.3 respectively, shows the MAP estimate to be equal to the ML estimate when the prior, g, is a constant function.

4.3.2 Gaussian mixture models

Mixture models may be used to model arbitrary distributions by combining ('mixing') a set of distributions. Each distribution models a *sub-population* of an overall population. Gaussian mixture models (GMMs) are a popular form of mixture model and use a mixture of Gaussian distributions to model an overall population. GMMs are structured as follows:

Mixture priors The prior probability, α_k, defines the proportion of the total data belonging to the kth sub-population (and therefore also kth mixture component) and is computed as:

$$\alpha_k = \frac{N_k}{N_p}, \tag{4.4}$$

where N_k is the number of data points which comprise the kth sub-population and N_p is the total size of the population.

Mean vector The mean of a mixture component defines the centre point of the distribution and is computed as the sample mean of the data of the kth sub-population and is computed as:

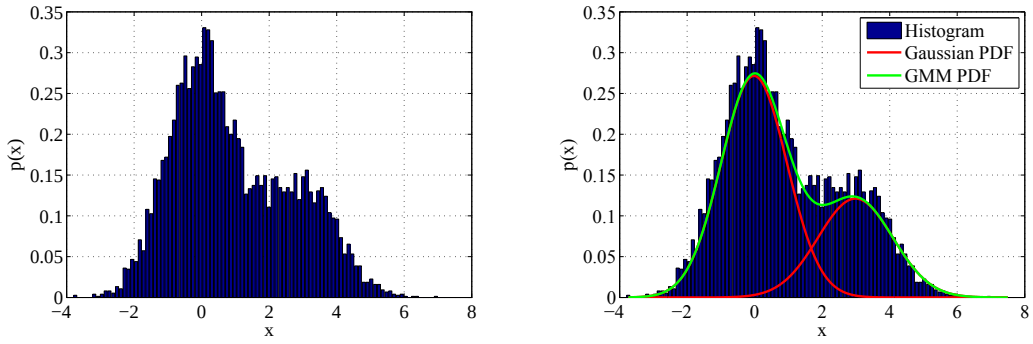
$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N_k} x_k(n), \tag{4.5}$$

where $x_k(n)$ is the nth data point of the population $\mathbf{x}$ assigned to cluster k.

Covariance matrices Covariances are computed for each mixture component as:

$$\mathbf{\Sigma}_k = \frac{1}{N_k - 1} \sum_{n}^{N_k} (x_k(n) - \mu_k)(x_k(n) - \mu_k). \tag{4.6}$$

In the case of multi-dimensional data $x(n)$ becomes a vector and all operations are subsequently performed element-wise. Sub-populations can be represented as



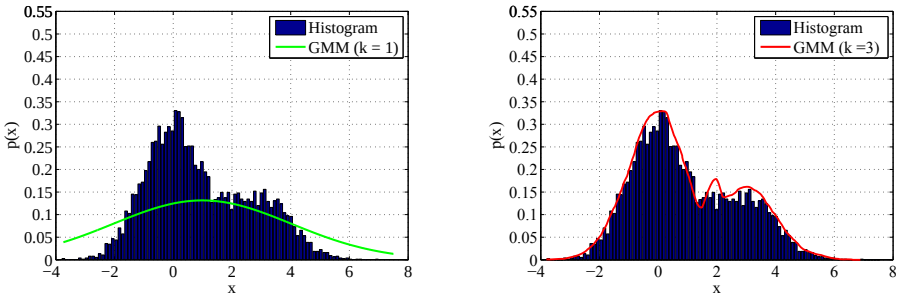
(a) Data (b) GMM mixture components

Figure 4.1: Example of using a GMM with 2 mixture components to model the distribution of artificially generated bimodal data

a clustering of the data. These sub-populations may therefore be estimated using any clustering algorithm. In this work the Linde-Buzo-Grey (LBG) variant of the K-means clustering algorithm is used to estimate model parameters [Linde et al., 1980]. This is an iterative process which successively forms new centroids by splitting existing clusters. Clusters chosen by the K-means algorithm may be further refined through the use of Expectation Maximisation (EM) [Dempster et al., 1977]. In preliminary tests this method was not found to improve performance and so was not used in the final system.

To illustrate how GMMs can be used to model multi-modal distributions an example of modelling the distribution of a 1-dimensional data series is displayed in Figure 4.1. Figure 4.1(a) is a histogram of the data series where the distribution is shown to be clearly non-Gaussian. Figure 4.1(b) illustrates how this distribution is modelled using a GMM. First, two Gaussian distributions are fitted to the data. The PDFs of these Gaussian distribution are then scaled by the mixture priors and summed to give the PDF of the GMM.

A key issue with mixture models is the appropriate selection of the number of mixture components, k. Increasing k allows more detailed distributions to be modelled. If insufficient data is available this results in over-fitting where the model of the distribution is biased to the available data samples rather than the distribution



(a) $k = 1$ (b) $k = 3$

Figure 4.2: Example of over-fitting and under-fitting distributions using GMM

of the overall population. The opposite of this problem is under-fitting whereby k is set to be too low and subsequently is unable to model the characteristics of the distribution. k is usually determined empirically and so it is important to ensure training data is representative of the overall data and that test data is not drawn from the training data. Figure 4.2 illustrates the problems of under-fitting and over-fitting the GMM PDF. In Figure 4.2(a) $k = 1$ and so a single Gaussian is used to model the data. The fit is clearly very poor as neither mode has been modelled. In terms of over-fitting Figure 4.2(b) shows the PDF of a GMM with $k = 3$. In this case there is too much detail in the PDF which does not accurately reflect the true distribution.

4.3.3 MAP using Gaussian distributions

Now that the general approach of MAP estimation has been defined in Section 4.3.1, we now look at applying the technique given an appropriate model of our feature distributions. Section 4.3.2 has shown GMMs to be effective at modelling the distributions of speech feature vectors and so we will examine the case of MAP using Gaussian distributions in this section.

In keeping with the notation used in the general case, we define an augmented

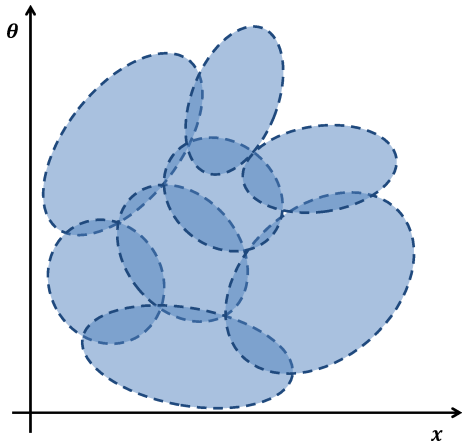


Figure 4.3: Visual representation of GMM mixture components

feature vector, $\mathbf{y}$, as:

$$\mathbf{y} = [\mathbf{x}, \boldsymbol{\theta}]^T . \tag{4.7}$$

The joint density of the augmented feature vector is then modelled as:

$$f(\mathbf{y}) = \Phi(\mathbf{y}) = \sum_{k=1}^{K} \alpha_k \phi_k(\mathbf{y}) = \sum_{k=1}^{K} \alpha_k \mathcal{N}\left(\mathbf{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{4.8}$$

where $\Phi(\mathbf{y})$ is a GMM comprising K mixture components which localise the joint density of $\mathbf{y}$, with α_k representing the prior probability of the kth mixture component, $\phi_k(\mathbf{y})$. $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ denote the mean and covariance of the joint vector within the kth mixture component where

$$\boldsymbol{\mu}_k^y = \begin{bmatrix} \boldsymbol{\mu}_k^x \\ \boldsymbol{\mu}_k^\theta \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_k^y = \begin{bmatrix} \boldsymbol{\Sigma}_k^{xx} & \boldsymbol{\Sigma}_k^{x\theta} \\ \boldsymbol{\Sigma}_k^{\theta x} & \boldsymbol{\Sigma}_k^{\theta\theta} \end{bmatrix} . \tag{4.9}$$

The mean vector comprises $\mathbf{x}$ and $\boldsymbol{\theta}$ mean vectors whilst the covariance matrix consists of $\mathbf{x}$ and $\boldsymbol{\theta}$ covariance matrices as well as the cross-covariances, $\boldsymbol{\Sigma}_k^{\theta x}$ and $\boldsymbol{\Sigma}_k^{x\theta}$. Figure 4.3 gives a visual representation of the GMM. Each mixture component is shown as an oval described by the covariance matrix of the component and centred on the mean.

The MAP estimate of θ is therefore defined as:

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} \left(f(\boldsymbol{\theta}|\mathbf{x}, \Phi) \right), \tag{4.10}$$

with the MAP estimate for the kth GMM mixture component defined as:

$$\hat{\boldsymbol{\theta}}^{k}_{MAP} = \arg\max_{\boldsymbol{\theta}} \left(f(\boldsymbol{\theta}|\mathbf{x}, \phi_k) \right). \tag{4.11}$$

The posterior probability of $\boldsymbol{\theta}$ given $\mathbf{x}$ and the kth GMM mixture component, ϕ_k, is defined as:

$$\begin{aligned} f(\boldsymbol{\theta}|\mathbf{x}, \phi_k) = & \\ & \frac{1}{(2\pi)^{D/2}(\boldsymbol{\Sigma}^{\mathbf{y}}_k)^{1/2}} \exp\{-0.5[\boldsymbol{\theta} - \boldsymbol{\mu}^{\boldsymbol{\theta}}_k - \boldsymbol{\Sigma}^{\boldsymbol{\theta}\mathbf{x}}_k(\boldsymbol{\Sigma}^{\mathbf{xx}}_k)^{-1}(\mathbf{x} - \boldsymbol{\mu}^{\mathbf{x}}_k)]^T \\ & \qquad [\boldsymbol{\Sigma}^{\boldsymbol{\theta}\boldsymbol{\theta}}_k - \boldsymbol{\Sigma}^{\boldsymbol{\theta}\mathbf{x}}_k(\boldsymbol{\Sigma}^{\mathbf{xx}}_k)^{-1}\boldsymbol{\Sigma}^{\mathbf{x}\boldsymbol{\theta}}_k]^{-1} \\ & \qquad [\boldsymbol{\theta} - \boldsymbol{\mu}^{\boldsymbol{\theta}}_k - \boldsymbol{\Sigma}^{\boldsymbol{\theta}\mathbf{x}}_k(\boldsymbol{\Sigma}^{\mathbf{xx}}_k)^{-1}(\mathbf{x} - \boldsymbol{\mu}^{\mathbf{x}}_k)]\} \end{aligned}, \tag{4.12}$$

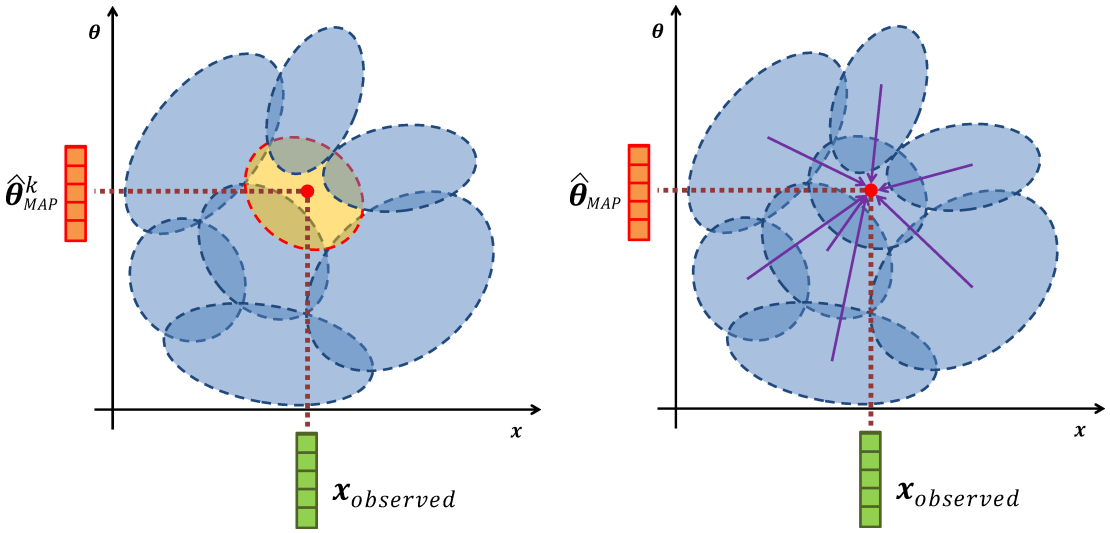
where D is the dimensionality of the augmented feature vector, $\mathbf{y}$, and exp operates element-wise. To maximise this function the derivative is found and set to zero, i.e:

$$\frac{d}{d\boldsymbol{\theta}} f(\boldsymbol{\theta}|\mathbf{x}, \phi_k) = 0. \tag{4.13}$$

In this case the point where this occurs is equivalent to setting the exponential term to zero. This term is then substituted into Equation 4.11 to give the MAP estimate of $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}^{k}_{MAP} = \arg\max_{\boldsymbol{\theta}} \left(f(\boldsymbol{\theta}|\mathbf{x}, \phi_k) \right) = \boldsymbol{\mu}^{\boldsymbol{\theta}}_k - \Sigma^{\boldsymbol{\theta}\mathbf{x}}_k(\Sigma^{\mathbf{xx}}_k)^{-1}(\mathbf{x} - \boldsymbol{\mu}^{\mathbf{x}}_k). \tag{4.14}$$

The process of obtaining the MAP estimation from the kth mixture component is illustrated in Figure 4.4(a). The observed vector, $\mathbf{x}_{observed}$, falls within the distribution described by the highlighted mixture component. The means and covariance of



(a) Single mixture component (b) Weighted average

Figure 4.4: Illustration of obtaining the MAP estimate from a single mixture component (a) and as a weighted average of all mixture components (b)

that particular component are then used to compute an estimate of the unobserved vector $\hat{\boldsymbol{\theta}}^k_{MAP}$. In this case, where $\mathbf{x}$ and $\boldsymbol{\theta}$ are assumed jointly Gaussian, the MAP estimator is equal to the minimum mean square error (MMSE) estimator [Gallager, 2008].

There are now two options available to obtain the final estimate of $\boldsymbol{\theta}$. Firstly, an estimate can be made from a single mixture component as per SPLICE, described by Deng et al. [2001]. In this case the estimate from the most likely mixture component was used, with k determined as:

$$\hat{k} = \arg\max_k \alpha_k f(\mathbf{x}|\phi^{\mathbf{x}}_k). \tag{4.15}$$

Alternatively the estimates from each mixture component in the joint density can be combined. The most straightforward way of doing this is by weighting each estimate by the posterior probability, i.e. the probability of $\mathbf{x}$ belonging to the kth mixture component. Figure 4.4(b) illustrates this process. Whilst the observed vector falls within the Gaussian described by a single mixture component, a weighted

average of all mixture components is used:

$$\hat{\boldsymbol{\theta}}_{MAP} = \sum_{k=1}^{K} h_k(\mathbf{x}) \arg\max_{\boldsymbol{\theta}} \left(f(\boldsymbol{\theta}|\mathbf{x}, \phi_k) \right), \tag{4.16}$$

where the posterior probability of $\mathbf{x}$, $h_k(\mathbf{x})$, is defined as:

$$h_k(\mathbf{x}) = \frac{\alpha_k f(\mathbf{x}|\phi_k^{\mathbf{x}})}{\sum_{k=1}^{K} \alpha_k f(\mathbf{x}|\phi_k^{\mathbf{x}})}. \tag{4.17}$$

$f(\mathbf{x}|\phi_k)$ is the marginalised distribution of $\mathbf{x}$. The weighted estimate of θ is therefore computed as:

$$\hat{\boldsymbol{\theta}}_{MAP} = \sum_{k=1}^{K} h_k(\mathbf{x}) \arg\max_{\boldsymbol{\theta}} \left(f(\boldsymbol{\theta}|\mathbf{x}, \phi_k) \right) = \sum_{k=1}^{K} h_k(\mathbf{x}) \left(\boldsymbol{\mu}_k^{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_k^{\theta\mathbf{x}} (\boldsymbol{\Sigma}_k^{\mathbf{xx}})^{-1} (\mathbf{x} - \boldsymbol{\mu}_k^{\mathbf{x}}) \right). \tag{4.18}$$

Alternate methods of combining individual component estimates also exist. One such example is described by Boucheron and Leon [2012] whereby weights are determined using a novel mapping matrix. In this work the posterior probabilities will be used to combine individual estimates.

4.4 Model Training using Stereo Data

One method of obtaining the joint distribution of noisy speech and an unknown parameter, $\boldsymbol{\theta}$, is to use stereo data for model training. This method assumes that all parameters, including the noise type and SNR is known at the training stage to obtain the model joint density directly. This approach was used in the SPLICE feature enhancement method as described by Deng et al. [2001]. In SPLICE, cepstral features of clean speech are estimated from cepstral features of noisy speech. The model is therefore trained on an augmented feature vector consisting noisy speech

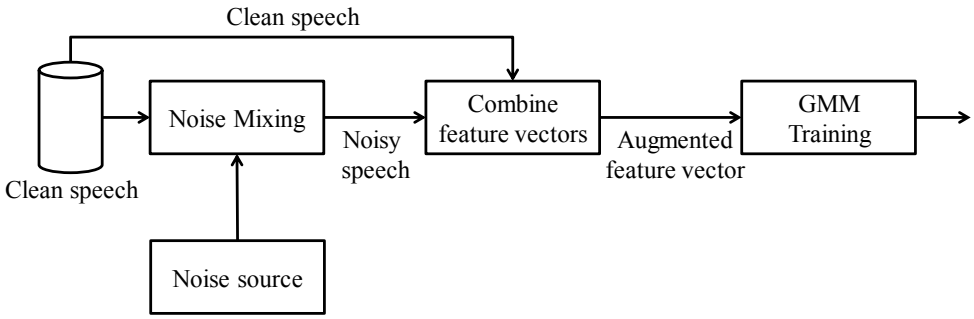


Figure 4.5: Flowchart illustrating the process of training a joint density model of clean and noisy speech using stereo training data

and clean speech cepstral feature vectors with the each clean speech frame aligned with the corresponding noisy frame of noisy speech.

Figure 4.5 illustrates the process of model training in SPLICE. The noise source can either be a recording of a noise environment or a system which generates random samples from a given noise distribution, with actual noise recordings being preferable [Deng et al., 2001]. These noise samples are then mixed with clean speech at the target SNR to give noisy speech which matches the target environment. Whilst it is possible to use a cache of clean speech and an estimate of the noise at runtime, this method would require significant resources in terms of both processor and memory and so a more direct way of obtaining model parameters without completely retraining the resultant GMM is desirable.

4.5 Model Adaptation

It has been shown in Section 4.3 that an estimate of an unobserved feature can be made using MAP estimation given a GMM of the joint feature distribution. This process is known to perform best when the environment and speaker in the operating environment are matched to the environment and speaker in the training process [Deng et al., 2000]. Stereo training (Section 4.4) is clearly not practical in all cases and so this section examines methods of adapting a 'generic' model to specific speakers and environments. These methods assume that a universal background

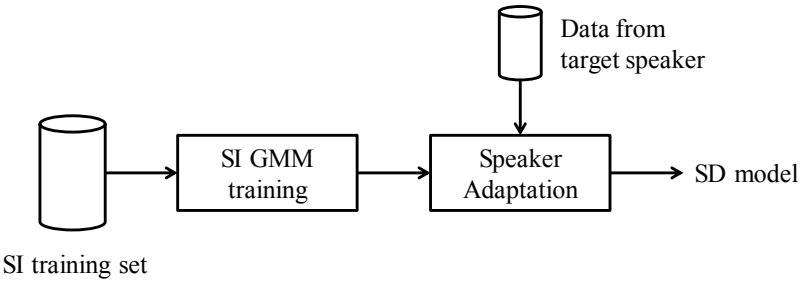


Figure 4.6: Illustration of the process of applying speaker adaptation to a speaker-independent GMM

model (UBM), a GMM trained on a large range of speakers in an environment with no noise, is available. Small amounts of data representing the speaker and noise are then used to adapt the model this speaker independent, environment independent model to a speaker-dependent, environment-dependent model. Methods of speaker adaptation are discussed in Section 4.5.1 followed by methods of noise adaptation in Section 4.5.2.

4.5.1 Speaker adaptation

This section examines two methods of speaker adaptation, namely: maximum likelihood linear regression (MLLR) and maximum *a-posteriori* (MAP) adaptation. Both methods were developed for ASR and rely on an existing speaker-independent model and a small amount of data from the target speaker [Leggetter and Woodland, 1995; Gauvain and Lee, 1994]. Figure 4.6 shows the general approach to obtaining speaker-dependent models from speaker-independent models. MLLR is described in Section 4.5.1.1 whilst MAP adaptation is described in Section 4.5.1.2.

4.5.1.1 Maximum likelihood linear regression (MLLR)

Originally designed for HMM-GMM based speech recognition systems, maximum likelihood linear regression (MLLR) is widely used as a method of adapting systems to compensate for variations in either speaker or environment (or both). Several variations of this technique exist, including systems for mean-only adaptation, mean

and covariance adaptation and modified systems for use in situations where either minimal adaptation data or low processing and memory resources are available. Speech recognition systems are known to perform better using speaker-dependent models, however sufficient speaker-dependent data is not always available to train such models and so a method of adaptation is desirable.

The objective of MLLR is to compute transformations of a speaker independent system based on new observations which represent the target speaker and/or environment. Given a sufficient amount of data performance should tend towards the speaker-dependent case. These transformations are designed to maximise the likelihood of the adaptation data [Leggetter and Woodland, 1995].

In the case of standard MLLR, for mixture component k, means, $\boldsymbol{\mu}_k$, are adapted as:

$$\hat{\boldsymbol{\mu}}_k^{sd} = \mathbf{A}_k \boldsymbol{\mu}_k^{si} + \mathbf{b}_k, \tag{4.19}$$

where $\mathbf{A}$ is the transformation matrix and $\mathbf{b}$ representing the bias vector. sd denotes the speaker dependent source domain and si denotes the speaker-independent target domain. Originally, MLLR was designed as a mean-only adaptation method, however it was later extended for mean and covariance adaptation, with covariances, $\boldsymbol{\Sigma}_k$, adapted as:

$$\hat{\boldsymbol{\Sigma}}_k^{sd} = \mathbf{H}_k \boldsymbol{\Sigma}_k^{si} \mathbf{H}_k, \tag{4.20}$$

where $\mathbf{H}$ is the covariance transformation matrix. Mean-only adaptation achieves increases in performance of between 13-17% over speaker-independent systems while introducing the covariance transformations increases performance by a further 2-7% [Gales and Woodland, 1996].

The transformations given in Equations 4.19 and 4.20 assume that sufficient adaptation is available to compute transforms for each mixture component. In cases where minimal adaptation data is available mixture components can be grouped

together and share transformation matrices [Leggetter and Woodland, 1995; Gales and Woodland, 1996]. Furthermore, Digalakis et al. [1995] proposed the same transformation matrix could be used for both mean and covariance adaptation to give an adaptation process known as constrained MLLR (CMLLR), i.e:

$$\hat{\boldsymbol{\mu}}_k^{sd} = \mathbf{H}_k(\boldsymbol{\mu}_k^{si} - \mathbf{b}_k), \tag{4.21}$$

and

$$\hat{\boldsymbol{\Sigma}}_k^{sd} = \mathbf{H}_k \boldsymbol{\Sigma}_k^{si} \mathbf{H}_k. \tag{4.22}$$

Despite these measures to increase the robustness to small amounts of speaker adaptation data, performance can still suffer. In extreme cases performance can fall below that of the unadapted speaker-independent case when insufficient adaptation data exists [Woodland, 2001].

4.5.1.2 MAP adaptation

MAP adaptation is an unsupervised method of speaker adaptation which updates speaker-independent model parameters using the sufficient statistics of a number of observations from the target speaker.

Assuming an appropriate form of the prior distribution, $g(\gamma)$, of the target parameter, γ, is available MAP estimation may be used to estimate the updated model parameters, i.e.

$$\hat{\theta}_{MAP} = \arg\max_{\gamma} f(x|\gamma)g(\gamma). \tag{4.23}$$

As demonstrated in Section 4.3 this is equivalent to setting the parameters to the mode of the posterior distribution, $f(x|\gamma)g(\gamma)$. A Gaussian prior of finite dimension does not exist for this case and so an alternative approach is usually used to obtain the distribution. This approach is described by Gauvain and Lee [1994].

Model parameters may then be updated by applying the Expectation-Maximisation (EM) algorithm. Given a GMM trained on speaker-independent data, Φ_{SI}, and a set of observation vectors from the target speaker, $\mathbf{Y} = \mathbf{y}_1, \ldots, \mathbf{y}_N$, the expectation stage (E-step) of the EM algorithm is used to compute the sufficient statistics of the model. This begins by determining the mixture components to which each vector belongs in the existing model, i.e. calculating:

$$f(k|\mathbf{y}_n, \Phi_{SI}) = \frac{\alpha_k(\mathbf{y}_n) f(\mathbf{y}_n|\phi_{SI}^k)}{\sum_{j=1}^{K} \alpha_j(\mathbf{y}_n) f(\mathbf{y}_n|\phi_{SI}^j)}. \tag{4.24}$$

Next, the sufficient statistics of the model are computed for each mixture component across target speaker observations:

$$\eta_k = \sum_{n=1}^{N} f(k|\mathbf{y}_n, \Phi_{SI}), \tag{4.25}$$

$$E_k(\mathbf{y}) = \sum_{n=1}^{N} f(k|\mathbf{y}_n, \Phi_{SI})\mathbf{y}_n, \tag{4.26}$$

$$E_k(\mathbf{y}^2) = \sum_{n=1}^{N} f(k|\mathbf{y}_n, \Phi_{SI})\mathbf{y}_n^2. \tag{4.27}$$

To update the model parameters the maximisation stage (M-step) of the EM procedure is applied, i.e:

$$\hat{\alpha}_k = \frac{\alpha_k - 1 + \eta_k}{\sum_{k=1}^{K} \alpha_k - 1 + \eta_k}, \tag{4.28}$$

$$\hat{\mu}_k = \frac{\tau^{\alpha}\mu_k + E_k(\mathbf{y})}{\tau^{\alpha} + \eta_k}, \tag{4.29}$$

$$\hat{\Sigma}_k = \frac{(\tau^{\Sigma} - 1)\Sigma_k + \tau^{\mu}(\hat{\mu}_k - \mu_k)^2 + (E_k(\mathbf{y}^2) - 2\hat{\mu}_k E_k(\mathbf{y}))\eta_k + \hat{\mu}_k^2}{\tau^{\Sigma} - 1 + \eta_k}, \tag{4.30}$$

where τ^{μ} and τ^{Σ} are the tunable parameters which weight the updates of the means and covariances respectively and $\hat{\alpha}_k$, $\hat{\mu}_k$ and $\hat{\Sigma}_k$ are the modified priors, means and covariances for the kth mixture component of the model [Huang et al., 2001]. τ^{μ} and τ^{Σ} can be seen as 'confidence parameters' and dictate by how much the adaptation data influences the update of the model parameters. This can be set as proportional to the amount or quality of adaptation data available, though typically $\tau^{\mu} = \tau^{\Sigma} = 12$ [Reynolds et al., 2000].

The E-step and M-step of the EM algorithm are iterated, with the updated model parameters from the M-step used to feed into the next iteration of the E-step, until either convergence or a pre-defined number of iterations have been completed.

The use of MAP for speaker adaptation is advantageous in several ways. Firstly, due to the use of a prior distribution of the parameters less adaptation data is required to obtain robust estimates of the model parameters compared to MLLR. This is partially due to the local approach of the technique; only mixture components represented in the adaptation data will be updated. Finally, given sufficient training data the MAP estimate converges to the ML estimate as the prior tends to a constant function [Woodland, 2001].

4.5.2 Noise adaptation

The noise adaptation methods described in this section follow the parallel model combination (PMC) framework whereby models of clean speech and noise are combined to give a model of $\mathbf{z}$. Figure 4.7 illustrates the process, i.e. a model of the

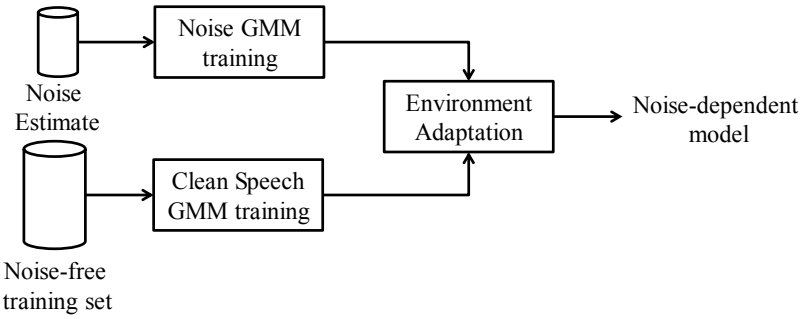


Figure 4.7: Illustration of the process of applying noise adaptation to a GMM of clean speech

clean speech:

$$\Phi_{\mathbf{x}} = \sum_{k=1}^{K} \alpha_k^{\mathbf{x}} \phi_k^{\mathbf{x}}(\mathbf{x}) = \sum_{k=1}^{K} \alpha_k^{\mathbf{x}} \mathcal{N}\left(\mathbf{x}, \boldsymbol{\mu}_k^{\mathbf{x}}, \boldsymbol{\Sigma}_k^{\mathbf{x}}\right) \tag{4.31}$$

is combined with a model of the noise:

$$\Phi_{\mathbf{n}} = \phi_k^{\mathbf{n}}(\mathbf{n}) = \mathcal{N}\left(\mathbf{n}, \boldsymbol{\mu}_k^{\mathbf{n}}, \boldsymbol{\Sigma}_k^{\mathbf{n}}\right) \tag{4.32}$$

to give a model of the joint distribution:

$$\Phi_{\mathbf{z}} = \sum_{k=1}^{K} \alpha_k^{\mathbf{z}} \phi_k^{\mathbf{z}}(\mathbf{z}) = \sum_{k=1}^{K} \alpha_k^{\mathbf{z}} \mathcal{N}\left(\mathbf{z}, \boldsymbol{\mu}_k^{\mathbf{z}}, \boldsymbol{\Sigma}_k^{\mathbf{z}}\right). \tag{4.33}$$

Models of clean speech are easily available using widely available speech corpora such as the WSJ, WSJCAM0 and NuanceCatherine datasets (Appendix A). The problem of noise adaptation can therefore be split into three components. Firstly, noise statistics must be obtained in a reliable fashion. Next, a mismatch function which defines the relationship between noisy speech, clean speech and the noise is required. Finally, the clean speech and noise models must be combined using the mismatch function. In the case of MFCCs this is a non-linear function and so an appropriate method of approximating the new model statistics is required; during feature extraction the clean speech and noise are transformed by the application of a logarithm. Whilst appropriate mismatch functions are defined in Section 4.5.2.1

these functions cannot be directly applied to the model parameters as they relate to the geometric mean rather than the arithmetic mean and so:

$$\boldsymbol{\mu}_k^{\mathbf{y}} \neq \mathbf{C} f(\mathbf{C^{-1}}\boldsymbol{\mu}_k^{\mathbf{x}}, \mathbf{C}^{-1}\boldsymbol{\mu}^{\mathbf{n}}, \alpha). \tag{4.34}$$

The first part of this section focuses on mismatch functions in Section 4.5.2.1 before moving on to methods of noise estimation in Section 4.5.2.5. Next, methods of model applying these noise estimates to adapt clean-trained models using the mismatch functions are discussed. Two methods of adaptation are covered: Vector Taylor Series (VTS) in Section 4.5.2.2 and the Unscented Transform (UT) in Section 4.5.2.3. These transforms assume the noise may be modelled as a Gaussian distribution and so methods of extending these transforms to handle non-Gaussian noises are proposed in Section 4.5.2.4.

4.5.2.1 Mismatch functions

The purpose of a mismatch function is to model the interaction between background noise and clean speech. The aim therefore is to find an appropriate form of the function $\mathbf{y} = f(\mathbf{x}, \mathbf{n})$ where $\mathbf{x}$, $\mathbf{n}$ and $\mathbf{y}$ are clean speech, noise and noisy speech in the MFCC domain.

Assuming additive noise in the time domain the effect of noise on speech can be represented as:

$$y(m) = x(m) + n(m), \tag{4.35}$$

where $x(m)$ is the mth sample of clean speech, $n(m)$ is noise and $y(m)$ is the resulting noisy speech. In the magnitude domain a similar relationship is observed:

$$|Y(k)| = |X(k) + N(k)|. \tag{4.36}$$

This work considers the use of MFCCs to represent frames of speech and so a simple

way of representing this relationship in the MFCC domain is to assume the same relationship as in the magnitude spectrum and perform the addition in that domain, i.e:

$$\mathbf{y} = f(\mathbf{x}, \mathbf{n}) = \mathbf{C} \log \left(\exp(\mathbf{C}^{-1}\mathbf{x}) + \exp(\mathbf{C}^{-1}\mathbf{n}) \right), \tag{4.37}$$

where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{n}$ are clean speech, noisy speech and noise in the MFCC domain and $\mathbf{C}$ is the cepstral transformation matrix.

Increased flexibility in the mismatch function can be achieved by introducing a weighting term which relates to the domain in which the clean speech and noise are added, i.e:

$$\mathbf{y} = f(\mathbf{x}, \mathbf{n}) = \frac{1}{\gamma}\mathbf{C} \log \left(\exp(\mathbf{C}^{-1}\gamma\mathbf{x}) + \exp(\mathbf{C}^{-1}\gamma\mathbf{n}) \right), \tag{4.38}$$

where $\gamma = 1$ represents the power domain and $\gamma = 0.5$ the magnitude domain. Whilst this parameter allows tuning of the function to obtain the best result, it is not precise for power spectral features. Examining the effect of noise on the computation of power spectral features shows an additional term to be present which is not handled in the previously described functions. This is shown in the following transform from magnitude to power spectral features:

$$\begin{aligned} |Y|^2 &= (|X + N|)^2 && (4.39)\\ &= (|X + N|) \times (|X + N|) && (4.40)\\ &= |X|^2 + |N|^2 + 2\Re(XN) && (4.41)\\ &= |X|^2 + |N|^2 + 2|X||N| \cos(\theta), && (4.42) \end{aligned}$$

where θ denotes the phase between clean speech and the noise. Applying the Mel filterbank matrix to each term to give Mel-spaced filterbank features, i.e. $\bar{X} = W|X|^2$, $\bar{N} = W|N|^2$ and $\bar{Y} = W|Y|^2$ allows the mismatch function to be expressed

as:

$$\bar{Y} = \bar{X} + \bar{N} + 2\boldsymbol{\alpha}\sqrt{\bar{X}\bar{N}}, \tag{4.43}$$

where $\boldsymbol{\alpha}$ represents the vector which describes the phase difference between the clean speech and noise and is defined as:

$$\boldsymbol{\alpha}(m) = \frac{\sum_{k=1}^{K} W(m,k)\cos(\theta(k))|X(k)||N(k)|}{\sqrt{\bar{X}(k)\bar{N}(k)}}, \tag{4.44}$$

where $\boldsymbol{\alpha}(m)$ is the phase factor relating to the mth filterbank channel and k indexes power spectral bins. Applying a logarithm to the phase-sensitive mismatch function yields:

$$\log(\bar{Y}) = \log(\bar{X} + \bar{N} + 2\alpha\sqrt{\bar{X}\bar{N}}). \tag{4.45}$$

In most speech enhancement applications it is assumed that $E[\boldsymbol{\alpha}] = 0$, and indeed some studies have aimed to model this term as a Gaussian distribution with zero mean [Droppo et al., 2002; Deng et al., 2004] with some success. Due to the nonlinear transform (logarithm) in Equation 4.45 the pdf of $\boldsymbol{\alpha}$ is also modified in a nonlinear fashion, with the resulting pdf being demonstrated to be non-Gaussian empirically by Faubel et al. [2008].

Defining $\bar{\mathbf{x}} = \log(\bar{X})$, $\bar{\mathbf{n}} = \log(\bar{N})$ and $\bar{\mathbf{y}} = \log(\bar{Y})$ allows the Equation 4.45 to be rewritten as:

$$\bar{\mathbf{y}} = f(\bar{\mathbf{x}}, \bar{\mathbf{n}}, \boldsymbol{\alpha}) = \bar{\mathbf{x}} + \log(1 + \exp^{\bar{\mathbf{n}}-\bar{\mathbf{x}}} + 2\boldsymbol{\alpha}\sqrt{\exp^{\bar{\mathbf{n}}-\bar{\mathbf{x}}}}) \tag{4.46}$$

to give the phase-dependent mismatch function $f(\bar{\mathbf{x}}, \bar{\mathbf{n}}, \boldsymbol{\alpha})$. Excluding the phase term gives the standard mismatch function in the log domain, i.e:

$$\bar{\mathbf{y}} = f(\bar{\mathbf{x}}, \bar{\mathbf{n}}, \alpha = 0) = \log(\exp^{\bar{\mathbf{x}}} + \exp^{\bar{\mathbf{n}}}) \tag{4.47}$$

This mismatch function can be applied to MFCC vectors by inverting the DCT at the input stage, i.e:

$$\mathbf{y} = \mathbf{C}f(\mathbf{C}^{-1}\mathbf{x}, \mathbf{C}^{-1}\mathbf{n}, \boldsymbol{\alpha}). \tag{4.48}$$

This mismatch function may now be used to update model parameters of the GMM given an appropriate method of adaptation.

4.5.2.2 Vector Taylor series (VTS)

Vector Taylor Series (VTS) is a method of approximating the result of a non-linear function at a particular point. The function is represented as an infinite sum of terms relating to the derivatives of the function at the desired point. In reality it is not possible to calculate an infinite number of terms and so a limited number of terms is used, with relatively few terms required for a good approximation [Moreno et al., 1996].

Given an appropriate mismatch function, i.e. $\mathbf{y} = f(\mathbf{x}, \mathbf{n}, \boldsymbol{\alpha})$, $\boldsymbol{\mu}_k^{\mathbf{y}}$ can be represented as:

$$\boldsymbol{\mu}_k^{\mathbf{y}} = E\left[f(\mathbf{x}, \mathbf{n}, \boldsymbol{\alpha}) + f'(\mathbf{x}, \mathbf{n}, \boldsymbol{\alpha})(\mathbf{x} - \boldsymbol{\mu}_k^{\mathbf{x}}) + \frac{f''(\mathbf{x}, \mathbf{n}, \boldsymbol{\alpha})(\mathbf{x} - \boldsymbol{\mu}_k^{\mathbf{x}})^2}{2} + \ldots\right], \tag{4.49}$$

where $f'(\mathbf{x}, \mathbf{n}, \boldsymbol{\alpha})$ represents the first derivative of the function, i.e.:

$$f'(\mathbf{x}, \mathbf{n}, \boldsymbol{\alpha}) = \left.\frac{\partial f(\mathbf{x}, \mathbf{n}, \boldsymbol{\alpha})}{\partial \mathbf{x}}\right|_{\boldsymbol{\mu}_k^{\mathbf{x}}, \boldsymbol{\mu}_{\mathbf{n}}, \boldsymbol{\alpha}}. \tag{4.50}$$

The effect of high order components is assumed to be negligible by Acero et al. [2000] and so an approximation of the noisy speech mean can be computed as:

$$\hat{\boldsymbol{\mu}}_k^{\mathbf{y}} = E\left[f(\mathbf{x}, \mathbf{n}, \boldsymbol{\alpha}) + f'(\mathbf{x}, \mathbf{n}, \boldsymbol{\alpha})(\mathbf{x} - \boldsymbol{\mu}_k^{\mathbf{x}})\right]. \tag{4.51}$$

Given this approximation of the noisy speech mean, the means of the kth mixture

component of the joint distribution are as follows:

$$\hat{\boldsymbol{\mu}}_k^{\mathbf{z}} = [\hat{\boldsymbol{\mu}}_k^{\mathbf{y}}, \boldsymbol{\mu}_k^{\mathbf{x}}]. \tag{4.52}$$

The updated covariance matrices are therefore computed as:

$$\hat{\boldsymbol{\Sigma}}_k^{\mathbf{z}} = E\left[(\mathbf{z} - \hat{\boldsymbol{\mu}}_k^{\mathbf{z}})(\mathbf{z} - \hat{\boldsymbol{\mu}}_k^{\mathbf{z}})^T\right]. \tag{4.53}$$

4.5.2.3 Unscented transform (UT)

The Unscented Transform (UT) is a form of data-driven parallel model combination (DPMC). DPMC methods typically use Monte Carlo sampling to approximate the parameters of the noisy speech model, i.e. samples are drawn from the clean speech and noise distributions and then processed as:

$$\hat{\boldsymbol{\mu}}_k^{\mathbf{y}} = \frac{\sum_{i=1}^{I} f(\mathbf{x}_i, \mathbf{n}_i, \boldsymbol{\alpha})}{I}, \tag{4.54}$$

where $\mathbf{x}_i$ is the ith sample drawn from the clean speech distribution, $\mathcal{N}(\boldsymbol{\mu}_k^{\mathbf{x}}, \boldsymbol{\Sigma}_k^{\mathbf{x}})$, and $\mathbf{n}_i$ is the ith sample drawn from the noise distribution, $\mathcal{N}(\boldsymbol{\mu}^{\mathbf{n}}, \boldsymbol{\Sigma}^{\mathbf{n}})$. This method has the advantage of converging on the exact statistics of the noisy speech model as $I \to \infty$, however in practise this is clearly not practical. The difficulty is therefore deciding on an appropriate value of I in order to obtain a good approximation of the model parameters whilst keeping memory requirements within realistic levels. The solution is to draw samples from the distributions in a more structured way to guarantee sufficient coverage across a minimal number of samples. This is the rationale behind the Unscented Transform.

The Unscented Transform requires $2(D_{\mathbf{x}} + D_{\boldsymbol{\theta}})$ points to form a good approximation of model parameters, where $D_{\mathbf{x}}$ is the dimensionality of the clean speech feature vectors and $D_{\boldsymbol{\theta}}$ is the dimensionality of the unknown parameter [Julier and Uhlmann, 2004].

First, a model of the joint density of clean speech and the estimated parameter

is built where cross-correlation terms are assumed zero, i.e:

$$\mathbf{z} = [\mathbf{x}, \boldsymbol{\theta}]^T, \tag{4.55}$$

$$\boldsymbol{\mu}_k^{\mathbf{z}} = \begin{bmatrix} \boldsymbol{\mu}_k^{\mathbf{x}} \\ \boldsymbol{\mu}_k^{\boldsymbol{\theta}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_k^{\mathbf{z}} = \begin{bmatrix} \boldsymbol{\Sigma}_k^{\mathbf{xx}} & 0 \\ 0 & \boldsymbol{\Sigma}_k^{\boldsymbol{\theta\theta}} \end{bmatrix}. \tag{4.56}$$

A set of 'sigma' points are then generated from the distributions, as:

$$\mathbf{s}_{i,k}^{\mathbf{z}} = \begin{bmatrix} \mathbf{s}_{i,k}^{\mathbf{x}} \\ \mathbf{s}_{i,k}^{\boldsymbol{\theta}} \end{bmatrix} = \begin{cases} \boldsymbol{\mu}_k^{\mathbf{z}} + (\sqrt{(D_{\mathbf{x}} + D_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_k^{\mathbf{z}}})_i & \text{if } i = 1 \ldots D_{\mathbf{x}} + D_{\boldsymbol{\theta}} \\ \boldsymbol{\mu}_k^{\mathbf{z}} - (\sqrt{(D_{\mathbf{x}} + D_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_k^{\mathbf{z}}})_{i-(D_{\mathbf{x}}+D_{\boldsymbol{\theta}})} & \text{if } i = D_{\mathbf{x}} + D_{\boldsymbol{\theta}} + 1 \ldots 2(D_{\mathbf{x}} + D_{\boldsymbol{\theta}}). \end{cases} \tag{4.57}$$

Next, the same process is applied to the noise statistics to obtain the noise sigma points, $\mathbf{s}_k^{\mathbf{n}}$:

$$\mathbf{s}_i^{\mathbf{n}} = \begin{cases} [\mu^{\mathbf{n}}, \mathbf{0}] + (\sqrt{(D_{\mathbf{x}} + D_{\boldsymbol{\theta}})\bar{\boldsymbol{\Sigma}}^{\mathbf{n}}})_i & \text{if } i = 1 \ldots D_{\mathbf{x}} + D_{\boldsymbol{\theta}} \\ [\boldsymbol{\mu}^{\mathbf{n}}, \mathbf{0}] - (\sqrt{(D_{\mathbf{x}} + D_{\boldsymbol{\theta}})\bar{\boldsymbol{\Sigma}}^{\mathbf{n}}})_{i-(D_{\mathbf{x}}+D_{\boldsymbol{\theta}})} & \text{if } i = D_{\mathbf{x}} + D_{\boldsymbol{\theta}} + 1 \ldots 2(D_{\mathbf{x}} + D_{\boldsymbol{\theta}}) \end{cases}, \tag{4.58}$$

where $\bar{\boldsymbol{\Sigma}}^{\mathbf{n}}$ is the zero-padded noise covariance matrix:

$$\bar{\boldsymbol{\Sigma}}^{\mathbf{n}} = \begin{bmatrix} \boldsymbol{\Sigma}^{\bar{\mathbf{n}}} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{bmatrix}, \tag{4.59}$$

so that $\bar{\boldsymbol{\Sigma}}^{\mathbf{n}}$ becomes a $(D_{\mathbf{x}}+D_{\boldsymbol{\theta}})\times(D_{\mathbf{x}}+D_{\boldsymbol{\theta}})$ matrix. Whilst this results in redundant sigma points (zero padded covariance values will result in sigma points consisting of the mean value only) it is necessary to enable the estimation of vectors of differing lengths from the noisy speech vectors and has been found to have no significant

impact on performance. The noise sigma points therefore consist:

$$\mathbf{s}^{\bar{\mathbf{n}}} = [\mathbf{s}^{\mathbf{n}}, \mathbf{0}]^T. \tag{4.60}$$

The clean speech and noise sigma points are then combined with the use of the mismatch function to give sigma points of noisy speech, i.e:

$$\mathbf{s}^{\mathbf{y}}_{i,k} = \mathbf{C} f(\mathbf{C}^{-1}\mathbf{s}^{\mathbf{x}}_{i,k}, \mathbf{C}^{-1}\mathbf{s}^{\mathbf{n}}_{i}, \boldsymbol{\alpha}). \tag{4.61}$$

These are then used to replace the clean speech sigma points to form a new augmented vector of sigma points:

$$\mathbf{s}^{\bar{\mathbf{z}}}_{k} = [\mathbf{s}^{\mathbf{y}}_{k}, \mathbf{s}^{\boldsymbol{\theta}}_{k}]^T. \tag{4.62}$$

The parameters of the model of the joint density of the noisy speech and unknown parameter, $\mathbf{y}$ and $\boldsymbol{\theta}$ are then constructed for each mixture component as follows:

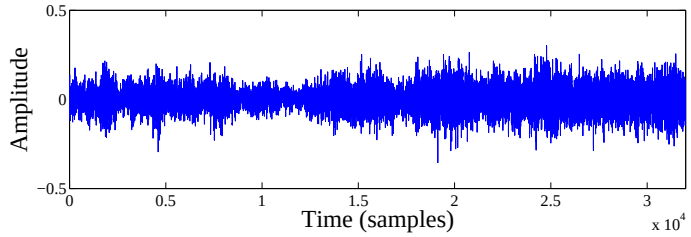
$$\boldsymbol{\mu}^{\bar{\mathbf{z}}}_{k} = \begin{bmatrix} \boldsymbol{\mu}^{\mathbf{y}}_{k} \\ \boldsymbol{\mu}^{\boldsymbol{\theta}}_{k} \end{bmatrix} = \sum_{i=1}^{2(D_{\mathbf{x}}+D_{\boldsymbol{\theta}})} \frac{\mathbf{s}^{\bar{\mathbf{z}}}_{i,k}}{2(D_{\mathbf{y}}+D_{\boldsymbol{\theta}})}, \tag{4.63}$$

$$\boldsymbol{\Sigma}^{\bar{\mathbf{z}}}_{k} = \begin{bmatrix} \boldsymbol{\Sigma}^{\mathbf{yy}}_{k} & \boldsymbol{\Sigma}^{\mathbf{y}\theta}_{k} \\ \boldsymbol{\Sigma}^{\theta\mathbf{y}}_{k} & \boldsymbol{\Sigma}^{\boldsymbol{\theta\theta}}_{k} \end{bmatrix} = \sum_{i=1}^{2(D_{\mathbf{x}}+D_{\boldsymbol{\theta}})} \frac{(\mathbf{s}^{\bar{\mathbf{z}}}_{i,k} - \boldsymbol{\mu}^{\bar{\mathbf{z}}}_{k})(\mathbf{s}^{\bar{\mathbf{z}}}_{i,k} - \boldsymbol{\mu}^{\bar{\mathbf{z}}}_{k})^T}{2(D_{\mathbf{y}}+D_{\boldsymbol{\theta}})}, \tag{4.64}$$

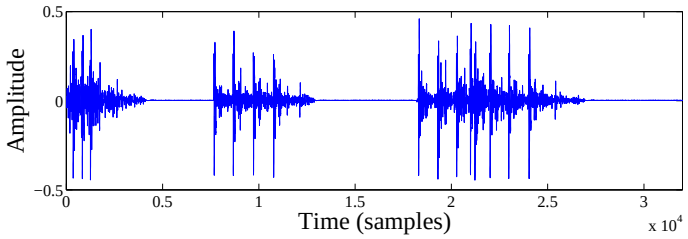
to give the following model of the joint density, $\bar{\mathbf{z}}$:

$$\Phi(\bar{\mathbf{z}}) = \sum_{k=1}^{K} \alpha_k \phi_k(\bar{\mathbf{z}}) = \sum_{k=1}^{K} \alpha_k \mathcal{N}\left(\bar{\mathbf{z}}, \boldsymbol{\mu}^{\bar{\mathbf{z}}}_{k}, \boldsymbol{\Sigma}^{\bar{\mathbf{z}}}_{k}\right). \tag{4.65}$$

For the case of noise adaptation the model priors are assumed to be equal to the priors of the clean speech model. The Unscented Transform may therefore be expressed



(a) Babble noise



(b) Machine gun noise

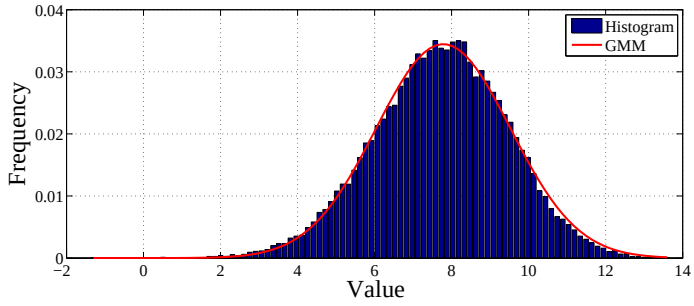
Figure 4.8: Time domain plots of noise signals comparing: a.) babble noise and b.) machine gun noise

as:

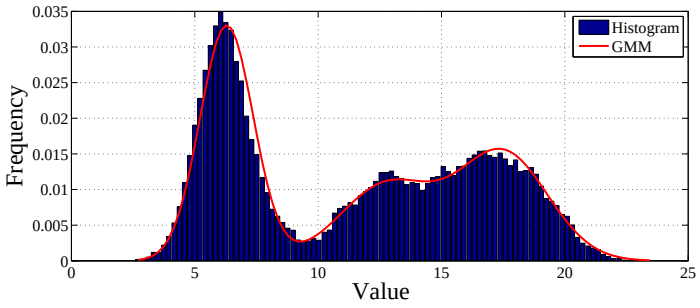
$$\Phi(\bar{\mathbf{z}}) = \upsilon(\Phi(\mathbf{x}), \Phi(\boldsymbol{\theta}), \Phi(\mathbf{n})). \tag{4.66}$$

4.5.2.4 Handling non-Gaussian noise

So far it has been assumed that the noise can be modelled as a single Guassian, i.e. $\Phi(\mathbf{n}) = \mathcal{N}(\mathbf{n}, \boldsymbol{\mu}_k^{\mathbf{n}}, \boldsymbol{\Sigma}_k^{\mathbf{n}})$. For example, Figure 4.9(a) shows the distribution of the second MFCC of babble noise to be Gaussian. In some cases, however, the noise does not conform to this assumption. Figure 4.8 compares babble noise to machine gun noise in the time domain. The machine gun noise clearly consists of significantly different periods, which can be categorised as low noise, shot and recoil. Figure 4.9(b) shows the distribution of the second MFCC of machine gun noise. The different modes are clearly visible, with the largest peak relating to the silence periods with two further peaks visible at higher coefficient values relating to the shot and recoil. Whilst a single Gaussian can be used to fit the distribution of babble noise, a GMM with $K = 3$ is required to fit the distribution of machine gun



(a) Babble noise



(b) Machine gun noise

Figure 4.9: Distributions of the zero'th MFCC for: a.) babble noise and b.) machine gun noise

noise as illustrated in Figure 4.9(b). This is clearly incompatible with the standard methods of model adaptation.

To overcome this issue a method of multiple-model adaptation is proposed. Instead of modelling the noise as a single Gaussian, a GMM of the noise with K_n mixture components is trained. Assuming the UT is used for model adaptation, the standard form of the transform

$$\Phi(\bar{\mathbf{z}}) = \upsilon\left(\Phi(\mathbf{x}), \Phi(\boldsymbol{\theta}), \Phi(\mathbf{n})\right) \tag{4.67}$$

becomes

$$\left[\Phi(\bar{\mathbf{z}})_1 \ldots \Phi(\bar{\mathbf{z}})_{K_n}\right] = \Upsilon\left(\Phi(\mathbf{x}), \Phi(\boldsymbol{\theta}), \Phi(\mathbf{n})^{(K_n)}\right), \tag{4.68}$$

where the $\Phi(\bar{\mathbf{z}})_j$ is the jth model of the joint density using the corresponding jth

noise mixture component and each model is computed in the standard way:

$$\Phi(\bar{\mathbf{z}})_j = \upsilon(\Phi(\mathbf{x}), \Phi(\boldsymbol{\theta}), \Phi(\mathbf{n})_j) \quad \text{for } j = 1 \dots K_n \ . \tag{4.69}$$

There are now several options for using these models for estimation. Two such methods are considered in this section: an HMM-based system and serial model combination.

4.5.2.4.1 HMM-based system HMMs are a popular technique for modelling stochastic processes [Vaseghi and Milner, 1997, 1995]. Given a sufficient number of observations of the noise signal it is possible to build an HMM to model the temporal and acoustic properties of the noise. The resulting HMM can then be used to decode a previously unseen sequence of noise observations. By outputting the state sequence the appropriate joint density model may then be selected for subsequent acoustic feature estimation. Such systems have already been proposed by authors including Varga and Moore [1990]; Zhao et al. [2008]; Bai [2011].

There are two challenges in the design of such a system. First, a suitable HMM topology must be designed and, second, an appropriate strategy for updating model parameters to take in to account the presence of speech in the signal must be formulated.

Considering first the topology of the HMM, there are a number of options. First, if we assume no temporal structure exists in the noise then a fully-connected ergodic topology is the most suitable model, as shown in Figure 4.10. This model allows transitions from any state to any other state at any time, including self-transitions. This type of model is useful for noises with several forms but with no significant temporal structure to the different forms, i.e. if we assume a noisy home environment one state could model the typical background noise whilst other states could model impulsive noises such as door slams.

Alternatively, if the noise source is known to have a particular temporal structure it may be modelled using a more restrictive model, for example, a circular topology

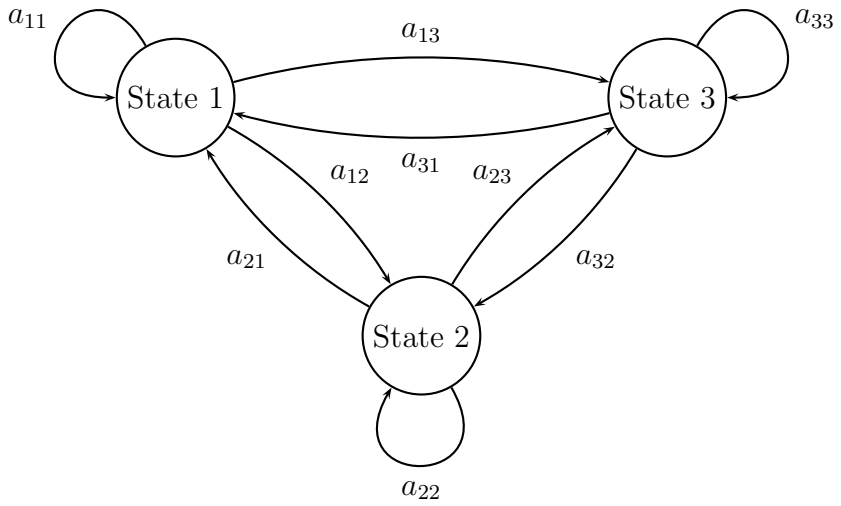


Figure 4.10: An ergodic hidden Markov model

as illustrated in Figure 4.11. In this model the system is limited to transitioning through the model from state 1 to state 3 in sequence and then looping back to state 1, with self-transitions also allowed. This type of model may be suitable for noises such as machine gun noise (Figure 4.8(b)) where periods of silence or low noise are followed by a shot and then recoil. If we allow state 1 to model the background noise, state 2 to model the shot noise and state 3 the recoil it is easy to see that such a model could be used effectively for this noise type. Figure 4.12 shows an appropriate model for machine gun noise. The model is based on the left-right model as per Figure 4.11 with an additional transition allowed between recoil and shot to prevent the need to return to silence in bursts of fire.

The next problem comes with the mismatch in the training and decoding environments. The problem is effectively the exact opposite of noisy speech recognition in that the system is trained on noise and testing observations are 'contaminated' with speech. There are two approaches to this problem. Both are analogous to methods used for noisy speech recognition. The first is model adaptation. Assuming transition probabilities remain the same it is possible to use a model adaptation method to update the acoustic models from modelling noise to noisy speech. Whilst

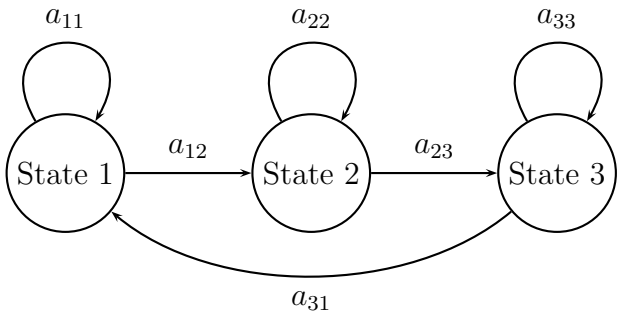


Figure 4.11: A circular hidden Markov model

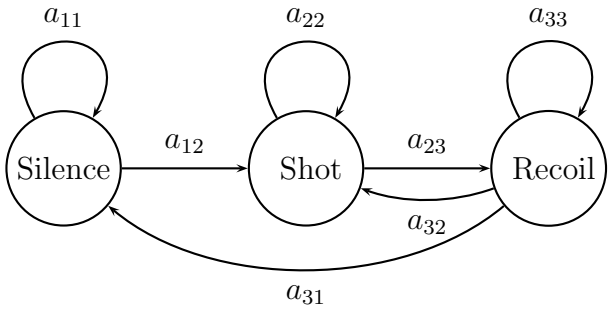


Figure 4.12: A modified left-right hidden Markov model modelling machine gun noise

this is easily possible using previously discussed methods such as VTS and UT, in most situations the speech power will to some extent mask the noise and so decoding accuracy may be adversely affected. The second option is to use an estimate of the noise as observations to the model. This is advantageous as no model parameters need updating.

Whilst this method provides a robust framework for mixture component selection, a substantial amount of information is required for good estimates of transition probabilities. Issues also remain in dynamically choosing an appropriate HMM topology for unknown noise signals.

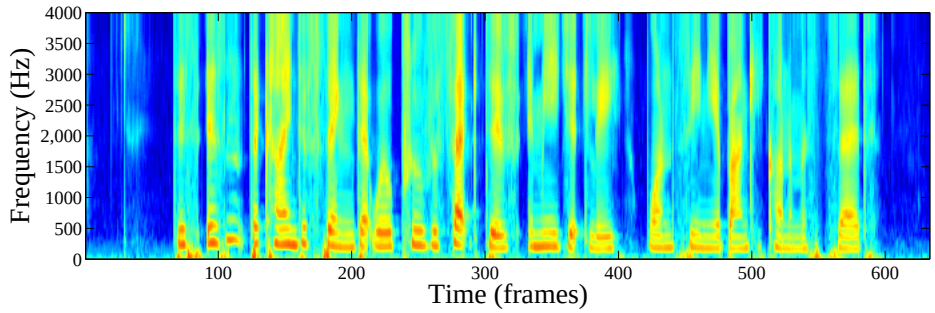
4.5.2.4.2 Serial model combination Serial model combination (SMC) is proposed as a simple method of using models created using the modified UT function, Υ. Instead of combining two models to form a model with the same dimensionality as per PMC, SMC concatenates models to give larger models in terms of the number of mixture components, K. This is equivalent to assuming that mixture components are selected independent of each other rather than being related through a Markov process as with the HMM-based method.

SMC creates a new model with $K = K_x K_n$ mixture components by incorporating all mixture components from models $\Phi(\bar{\mathbf{z}})_1$ to $\Phi(\bar{\mathbf{z}})_{K_z}$. The only modification of model parameters required is the normalisation of the mixture priors, i.e:

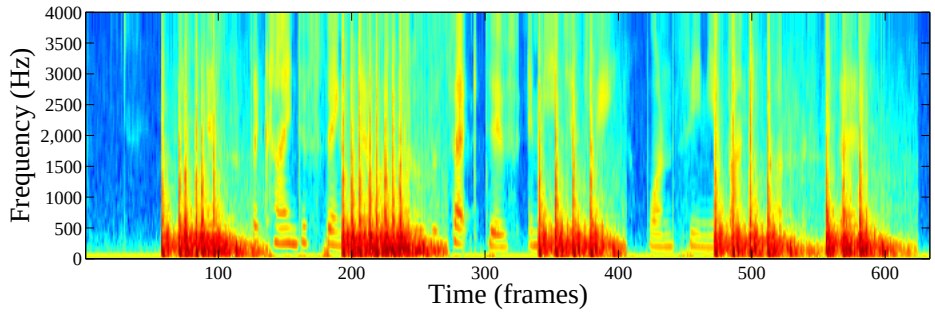
$$\alpha^{\bar{\mathbf{z}}} = \frac{\left[\alpha_1^{\bar{\mathbf{z}}} \dots \alpha_{K_n}^{\bar{\mathbf{z}}}\right]}{\sum_{j=1}^{K_z} \alpha_j^{\bar{\mathbf{z}}}}. \tag{4.70}$$

This adapted model may then be used as normal to estimate the unknown parameter. Whilst the model sizes may become prohibitively large if $\mathbf{n}$ is modelled by a large number of mixture components, in reality only a limited number of mixture components are actually required for estimation at each frame.

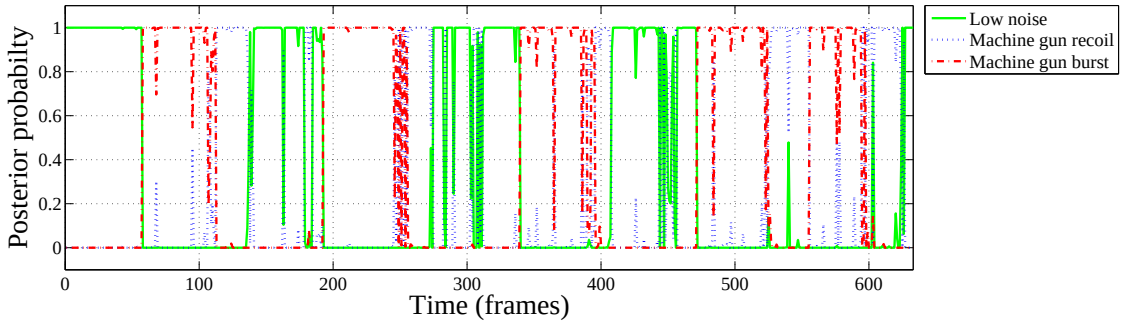
As per Equation 4.16 an estimate of θ is computed by weighting individual estimates by the posterior of $\mathbf{y}$, $h_k(\mathbf{y})$. When $h_k(\mathbf{y}) = 0$ the estimate from the kth



(a) Clean speech



(b) Noisy speech



(c) Summed noise mixture posterior probabilities

Figure 4.13: Noise type detection in machine gun noise using SMC. Spectrograms of a.) clean speech and b.) noisy speech are given for reference. c.) shows the posterior probability of each of the frames belonging to GMMs modelling: i.) low noise, ii.) machine gun recoil and iii.) machine gun burst noise

mixture component does not contribute to the overall estimate. This means that the estimate from each mixture component can be seen to be selected based on the posterior probability of that component. Figure 4.13 shows an example of this for machine gun noise.

Figure 4.13(a) shows clean speech and Figure 4.13(b) shows noisy speech. Figure 4.13(c) illustrates the mixture component selection by showing the summed posterior probabilities of the mixture components of $\Phi(\bar{\mathbf{z}})$ corresponding to the mixture components of the noise model, $\Phi(\bar{\mathbf{n}})$, i.e. low noise posteriors $= \sum_{k=1}^{K} h_k(\mathbf{y}|\Phi(\mathbf{z})_1^{(K_n)})$ and so on.

Figure 4.13 shows that in most cases all selected mixture components come from a submodel adapted by a single mixture component of the noise model. For all other mixture components the probability $f(\boldsymbol{\theta}|\mathbf{y}, \phi_k)$ need not be computed. In this example, $K_x = 256$ and $K_n = 3$ to give $K_z = 256 \times 3 = 768$. For the example utterance shown in Figure 4.13 the mean number of mixture components that contributed to the overall estimate (i.e. those with a non-zero posterior probability) was < 100.

The advantage of this algorithm is the simplicity of training and adaptation. Less parameters are required than the HMM-based method though this is due to the fundamental assumption that no useful temporal information exists in the noise signal. Whilst this is clearly not the case for certain types of noise, i.e. machine gun noise, it has been demonstrated that such noises can still be effectively modelled using this method.

4.5.2.5 Noise estimation

Mismatch functions have been described as a way of modelling the relationship between clean and noisy speech. These assume that the noise signal is known *a-priori* however this is generally not the case and so the noise must be estimated from the noisy speech. In the case of conventional speech enhancement a frame-by-frame estimate of the noise is required. A simple way of obtaining such an estimate is to apply a VAD to the noisy speech and update an estimate of the noise when no speech is detected. This is beneficial in that the estimation process is straightforward, but comes with several disadvantages. First, performance of VAD in noisy speech is typically unreliable and so some speech energy may be incorporated into the estimate of the noise. Second, such a system is only able to update the noise estimate

in periods of non-speech. There are several other approaches to this problem which aim to overcome these issues and can be categorised as: minimal-tracking, time-recursive averaging and histogram-based [Loizou, 2007]. Over time the noise is assumed to be more stationary than the speech and so in each case a number of neighbouring frames are included to form an analysis segment, the duration of which is usually set to between 400ms and 1 second. The length of the analysis window must be carefully managed to ensure that it is long enough to incorporate speech pauses and low-energy periods whilst being short enough to track variations in the noise. The motivation behind the main categories of conventional noise estimation methods are now summarised.

Minimal-tracking The assumption behind these methods is that the minimum value of a spectral bin over time will be equal to the energy of the noise. The estimate of the noise is therefore the minimum value of each spectral bin over a period of time. There are two variants of this method; the first tracks noise over the length of an analysis segment [Martin, 1994] whilst the second continually tracks the noise over the entire utterance [Doblinger, 1995]. The latter variant was found to perform better in objective and subjective testing as reported by Meyer et al. [1997].

Time-recursive averaging These methods assume that the noise has a non-uniform effect across spectral values. This assumption allows the noise estimate of each spectral bin to be updated either when the estimated local SNR is very low or, equivalently, when the probability of the bin containing speech energy is low. In the case of the SNR-based estimation proposed by Lin et al. [2003] a previous estimate of the noise is used to determine the SNR of the current frame. A smoothing function is defined based on the estimated SNR which controls the extent to which the current spectrum contributes to the overall estimate. Alternatively this smoothing function may be defined based on the speech-presence uncertainty, i.e. the probability of speech in the current spectral bin. There are many methods of computing this probability with a

technique known as minima-controlled recursive averaging (MCRA) by Cohen [2003] found to perform best [Loizou, 2007].

Histogram-based The most frequent energy value of each spectral bin can be seen to correspond to the energy of the noise. By measuring the frequency of spectral energy values the noise estimate therefore comprises the most frequent values [Hirsch and Ehrlicher, 1995].

Performance of the aforementioned methods was measured by Loizou [2007] who observed the accuracy of the noise estimates across a range of noises. No method was found to perform best overall with performance varying between noises.

The proposed methods of noise adaptation for this method of speech enhancement require knowledge of the noise statistics rather than the noise signal itself. In this case noise estimation becomes a parameter estimation problem and so an alternative approach may be taken. First, noise statistics may be obtained from estimates of the noise signal by using one of the conventional approaches. Alternatively the parameters may be estimated directly from the noisy speech. These approaches are typically based on an Expectation-Maximisation (EM) style of approach whereby an estimate of noise statistics is formed iteratively [Deng et al., 2004; Faubel and Klakow, 2010].

The proposed method of speech enhancement is able to utilise any method of noise estimation providing the distribution of the noise may be computed. Instead of analysing overall performance based on a particular choice of noise estimation method, performance is instead measured based on the accuracy of the noise distribution. Figure 4.14 therefore examines how much of the noise signal is required for accurate parameter estimation. A 'reference' model of the noise was trained from MFCCs extracted from the entire noise signal (235 seconds). Next, models were trained on subsets of the data using Monte-Carlo sampling to select noise vectors in order to emulate the effect of noise estimation. The similarity of these models to the reference model was then measured using the Kullback-Leibler divergence [Kullback and Leibler, 1951]. Three types of noise were tested: white, babble and destroyerops,

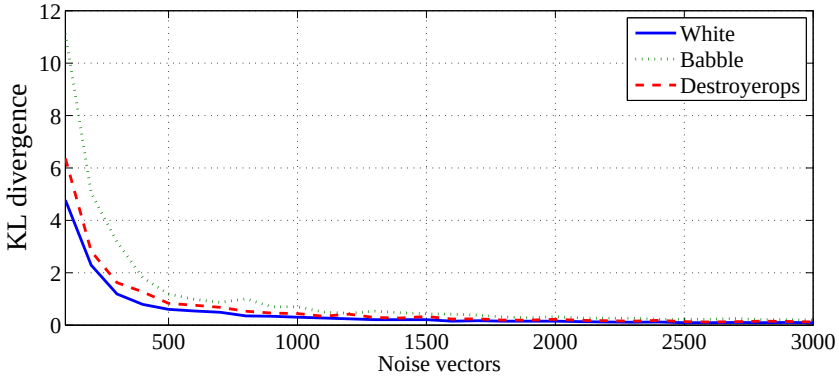


Figure 4.14: KL divergence of noise models as a function of the amount of noise used to train the model for: i.) white noise, ii.) babble noise and iii.) destroyerops noise.

all from the NOISEX dataset (Appendix A). White noise is used as an example of stationary noise whilst babble and destroyerops noises are both non-stationary. In all cases a reasonably accurate estimate of the noise is not achieved until ≈ 5 seconds of noise is available with $10 - 15$ seconds of noise required to achieve a highly accurate estimate of noise model statistics.

4.5.3 Adapting for speaker and noise

The previous sections have examined the processes of adapting a speaker-independent model of clean speech for either a specific speaker or noise condition, however in practise there is usually a mismatch in both speaker and noise and so the model must be adapted for variations in both speaker and environment. Several methods of joint speaker and noise adaptation have been developed for the purposes of robust ASR. These include unsupervised adaptation methods such as MLLR and MAP as well as more recent methods such as those by Chin et al. [2011] and Fujimoto et al. [2012] where the mismatch function is modified to incorporate speaker variability with model parameters subsequently updated using VTS. In this work a two-stage process is used. Given a model of clean speech, the first stage of adaptation is to reduce the effect of speaker variation by using either MLLR or MAP adaptation to adapt for variations in speaker characteristics. This requires a small amount of data

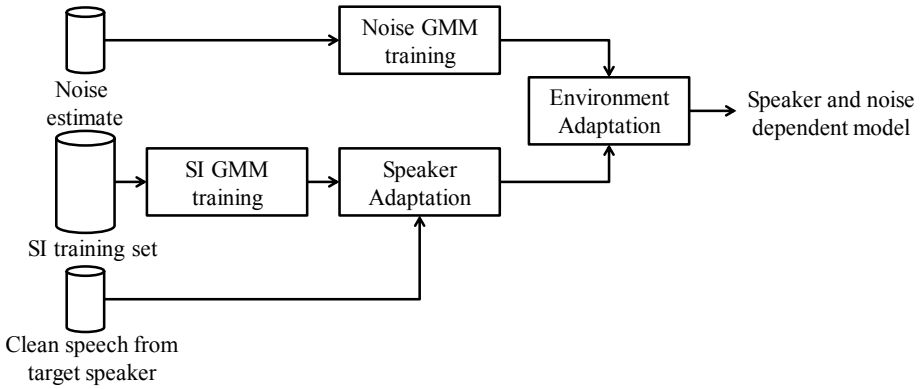


Figure 4.15: Illustration of the process of applying speaker and noise adaptation to a speaker-independent GMM trained on clean speech

from the target speaker, recorded in clean conditions. This results in a speaker-dependent model of clean speech. In the case of adapting for the environment, both VTS and UT are able to provide estimations of the cross-covariance matrices. Of these, UT has been shown to be more effective and so will be used for this method of speech enhancement [Shinohara and Akamine, 2009]. Given an estimate of the noise distribution the UT is then used to estimate the required distributions from the speaker-dependent model of clean speech. Figure 4.15 illustrates the proposed two-stage adaptation strategy.

4.6 Summary

In this chapter a framework for estimating an unknown parameter, θ, from noisy speech, $\mathbf{y}$, has been proposed. First, a method of modelling the joint distribution of $\mathbf{y}$ and θ was described and a method of using this distribution for parameter estimation, MAP estimation, subsequently defined. Methods of obtaining the required joint distribution were then examined. Whilst it is trivial to train an appropriate model with stereo training data, such data is not always available. Methods of adapting a base-model (UBM) for variations in both speaker and noise were therefore proposed.

Chapter 5

Spectral Envelope Estimation

Spectral envelope is one of the acoustic features required for speech reconstruction and so clean spectral envelope must be estimated from noisy speech. Estimation of spectral envelope is split into three stages. First, MFCC features are extracted from noisy speech. Next, clean MFCC features are estimated using MAP before finally the pseudo-inverse of the MFCC vectors is taken to give an estimate of the clean spectral envelope. The performance of the model adaptation methods used to construct the required distributions for estimation is compared to stereo-trained models. Two model configurations are considered. First, a global model of speech is used for enhancement before second, a system which models different acoustic classes with separate models is also considered.

Contents

5.1 Introduction

This work reconstructs speech using a speech reconstruction model driven by a set of acoustic features. One of these acoustic features is spectral envelope and so this chapter focuses on the estimation of this feature from noisy speech. The spectral envelope of a frame of speech represents the 'filter' in the source/filter model of speech production (Section 3.2). A significant amount of information is contained in the spectral envelope such as phonetic content, speaker age, accent and identity and this information is mostly represented in the formant locations and bandwidths [Vaseghi, 2008]. The majority of speech processing applications such as ASR therefore focus almost entirely on spectral envelope [Young et al., 2002]. As well as being an important carrier of information the spectral envelope was also shown to make a significant contribution to overall speech quality in Section 3.5 and so an accurate representation of the clean spectral envelope is important for this method of speech enhancement[1].

The ultimate objective of this work is to reconstruct clean speech from noisy speech. In terms of spectral envelope, we therefore require $|\mathbf{X}|$ given $|\mathbf{Y}|$. Noise is assumed to be additive, i.e. $|\mathbf{Y}| = |\mathbf{X} + \mathbf{N}|$ (Section 4.5.2.1). $|\mathbf{N}|$ is not known *a-priori* and so $|\mathbf{X}|$ cannot be computed directly. Instead, a popular approach of obtaining an approximation of $|\mathbf{X}|$ is to filter the noisy signal using an estimate of $|\mathbf{N}|$ which is obtained from the noisy speech (Section 2.2). This approach relies on the noise estimation process and so whilst this approach is effective for stationary noises it is generally not robust to non-stationary noises where changes in the noise spectrum are not well tracked by the noise estimator [Loizou, 2007]. Instead, this work takes a model-based approach to parameter estimation. Such a system is beneficial over conventional approaches as whilst the noise is not known *a-priori*, model-based estimation methods require only first and second order noise statistics which are more easily obtained from noisy speech than frame-by-frame noise estimates (Section 4.5.2.5).

[1]Parts of this chapter were published at Interspeech [Harding and Milner, 2011, 2012b]

The process of clean spectral envelope estimation can be divided into three components:

Feature extraction The first stage of the proposed estimation technique is the transformation of spectral envelope to an alternative domain for estimation. Whilst it is possible to form an estimate of spectral amplitudes directly from the magnitude or power spectra these representations contain a significant amount of redundant data which results in a joint feature space with very high dimensionality and intra-frame correlation which reduces modelling efficiency. Instead, an intermediate feature vector is used for estimation. In this work these features consist of MFCCs, the use of which was decided based upon the review of features in Section 3.5.2. This choice is in line with other related work [Deng et al., 2000; Darch et al., 2006; Boucheron and Leon, 2012].

Feature estimation A model of the joint-density of clean and noisy features is used to compute an estimate of clean features using MAP (Section 4.3). The model of the joint density can be acquired in a number of ways including stereo training and model adaptation (Sections 4.4 and 4.5).

Feature inversion the final stage is to invert the MFCC features to the spectral domain. The mel-filterbank matrix, $\mathbf{W}$ is applied to the magnitude spectrum as $\mathbf{W}|\mathbf{X}|$. It would therefore be reasonable to assume that this may be inverted by multiplying log-mel features by $\mathbf{W}^{-1}$. Direct inversion is not possible and so a pseudo-inverse is taken as described in Section 3.4.2.2.

Two systems based on this model-based estimation framework are proposed in this chapter. First, a method of global estimation is described in which a single model of the joint density is used for enhancement (Section 5.2). Second, a method of localised modelling is proposed in which utterances are split into acoustic classes and estimates made using class-specific models (Section 5.3). The results of experiments used to test these variants using both Gaussian and non-Gaussian noise are presented in Section 5.4.

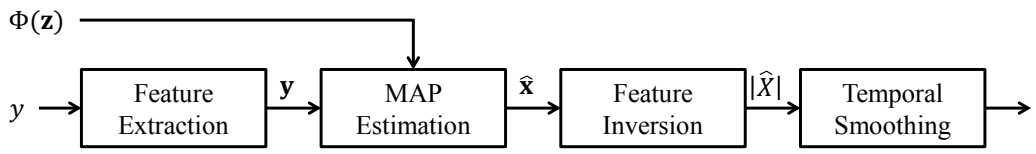


Figure 5.1: Flowchart illustrating the process of spectral amplitude estimation using a model of the joint density of clean and noisy speech features.

5.2 Global Modelling

An estimate of the clean spectral envelope may be obtained using MAP estimation with MFCCs used as an intermediate feature. MAP estimation was described in Chapter 4 as a process of estimating an unknown quantity, θ, from a model of the joint density of clean speech, $\mathbf{x}$, and θ. By substituting in the noisy speech, $\mathbf{y}$, in place of clean speech, $\mathbf{x}$, and $\mathbf{x}$ in place of $\boldsymbol{\theta}$ an estimate of the clean speech MFCCs, $\hat{\mathbf{x}}$, can be computed as:

$$\hat{\mathbf{x}} = \sum_{k=1}^{K} h_k(\mathbf{y}) \arg\max_{\mathbf{x}} \left(f(\mathbf{x}|\mathbf{y}, \phi_k) \right). \tag{5.1}$$

The process of obtaining such an estimate is illustrated in Figure 5.1. As input, the system takes a model of the joint density of MFCCs extracted from both clean and noisy speech, $\Phi(\mathbf{z})$, and, for each frame, the noisy MFCC vector, $\mathbf{y}$. $\Phi(\mathbf{z})$ models the relationship $\mathbf{z} = [\mathbf{y}, \mathbf{x}]$ where $\mathbf{x}$ are MFCCs extracted from clean speech and $\mathbf{y}$ correspond to the same frames of speech but this time in noisy conditions. The parameters of this model can be obtained in a variety of ways, including directly from stereo training data (Section 4.4) or indirectly through the use of PMC style techniques (Section 4.5).

First, MFCCs are extracted from the noisy speech to give the intermediate feature vectors. The MAP estimates of the clean features are then computed using the noisy feature vectors and joint density model as:

$$\hat{\mathbf{x}} = \sum_{k=1}^{K} h_k(\mathbf{y}) \left(\boldsymbol{\mu}_k^{\mathbf{x}} - \boldsymbol{\Sigma}_k^{\mathbf{xy}} (\boldsymbol{\Sigma}_k^{\mathbf{yy}})^{-1} (\mathbf{y} - \boldsymbol{\mu}_k^{\mathbf{y}}) \right). \tag{5.2}$$

The intermediate features are then inverted to give spectral features. No temporal information is built into the estimation model and so an additional processing block is included which uses recursive first-order averaging to smooth features across time, reducing inter-frame discontinuities, i.e:

$$|\hat{\mathbf{X}}|_i = \beta \cdot |\hat{\mathbf{X}}|_i + (1 - \beta) \cdot |\hat{\mathbf{X}}|_{i-1}, \tag{5.3}$$

where i is the frame index and all operations are element-wise. A value $\beta = 0.85$ was determined in preliminary testing.

5.3 Localised Modelling

The previous method of clean spectral envelope estimation used a single model of the joint-density of noisy and clean feature vectors. Whilst this method of estimation has been used with considerable success in other works, it is not necessarily the most effective method.

When a single model is used the assumption is made that all acoustic classes (i.e. phonemes, articulation classes etc.) can be modelled using a single, albeit multi-modal, distribution. Speech recognition systems take advantage of the distinct spectral properties of phoneme units and so it is natural to question the optimality of using a global model for enhancement. An approach using separate models for each acoustic class is therefore proposed. There are three challenges to such an approach: i.) the acoustic classes must be determined, ii.) a method of training appropriate models is required and iii.) a system of classifying feature vectors extracted from noisy speech for model selection at runtime must also be designed. These challenges are now discussed.

Acoustic classes Phoneme labels are typically used for speech recognition where a phoneme-level transcription of the utterance is ultimately required. In the case of speech enhancement a human readable transcription of the utterance

is not required and so other classifications can be considered. The acoustic classes considered in this work include phoneme and articulation classes. Phoneme classes are based on the TIMIT labelling system [Garofolo, 1993] and so comprise 41 classes including silence whilst the articulation class system uses ten classes, namely: affricates, diphthongs, fricatives, liquids, monophthongs, nasals, R-coloured vowels, semi-vowels, stops and silence.

Model training Models are trained in a two-stage process. First, assuming that the joint distribution of clean and noisy speech is again represented as $\mathbf{z} = [\mathbf{y}, \mathbf{x}]$, the dataset, Z, is first divided into M vectorpools, Ω^m, based on their acoustic class where M is the total number of acoustic classes:

$$\Omega^m = \{\mathbf{z}_i \in Z : \text{class}(\mathbf{z}_i) = m\}. \tag{5.4}$$

Training labels are based on phoneme-level transcriptions made at the time of recording (Appendix A). Articulation class transcripts were obtained by mapping phoneme classes to articulation classes and post-processing transcriptions based on this mapping. These transcriptions are not time-aligned and so time alignments are obtained using a context-independent HMM-GMM-based recognition system built using HTK [Young et al., 2002] and trained on clean speech. Recognition models are trained using iterations of the embedded Baum-Welch estimation algorithm [Baum et al., 1970]. The Viterbi algorithm is then used for forced-alignment recognition to give class boundaries. Next, GMMs are trained from each vectorpool, $\boldsymbol{\Omega}^m$, to give class-dependent models, $\Phi(\boldsymbol{\Omega}^m)$. These models are trained in the standard way as described in Section 4.3.2.

Localised estimation The final stage of estimation is to localise the region in the acoustic feature space and then make an estimation using the appropriate localised model. Models are selected on a frame-by-frame basis from classifications made from the noisy MFCC feature vectors. As per the model training

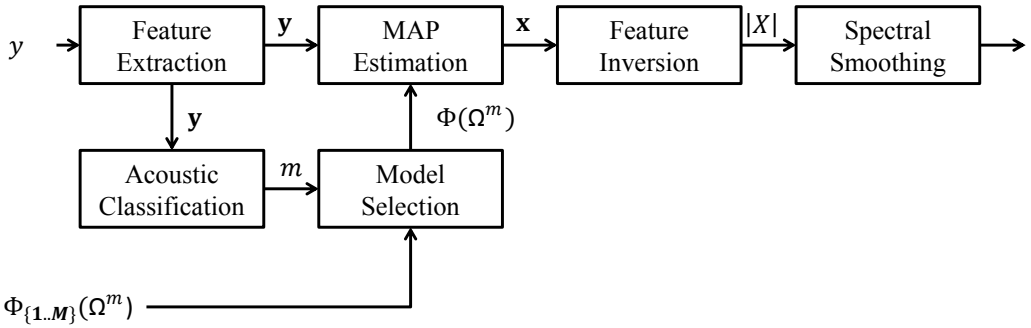


Figure 5.2: Flowchart illustrating the process of spectral amplitude estimation using a acoustic-class based models of the joint density of clean and noisy speech features.

stage, classifications are made using an HMM-GMM based recogniser. In fact, the same recognition models trained in the previous stage can be used for this purpose. No transcriptions are available at runtime and so standard Viterbi decoding is used. A basic grammar is used which allows any acoustic class to follow any other acoustic class. In more constrained tasks more specific grammars or language models may be used to give more accurate results. Such systems work well for clean data but performance rapidly deteriorates with the addition of noise. By training the recognition models on noisy data, matched to the target enhancement environment, performance can be improved [Deng et al., 2000]. This is not always practical as such data is not always available and training new models requires significant resources. Instead, either the model domain can be adapted to the target domain or the target domain can be adapted to the model domain. In terms of model adaptation, MLLR (Section 4.5.1.1) has been proven effective whilst features can be adapted using the global enhancement method *ã la* SPLICE [Deng et al., 2000].

Once appropriate models have been trained and acoustic class labels have been estimated the clean spectral envelope is estimated as illustrated in Figure 5.2. First, MFCC features are extracted from the noisy speech. These features are used for both acoustic unit classification and enhancement. Next, frames are classified according to their acoustic class, m, and the corresponding model loaded ($\Phi_m(\Omega^m)$). This

model is then used for estimation, i.e.:

$$\hat{\mathbf{x}}_i = \sum_{k=1}^{K} h_{k,m}(\mathbf{y}_i) \left(\boldsymbol{\mu}_{k,m}^{\mathbf{x}} - \boldsymbol{\Sigma}_{k,m}^{\mathbf{xy}} (\boldsymbol{\Sigma}_{k,m}^{\mathbf{yy}})^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_{k,m}^{\mathbf{y}}) \right). \tag{5.5}$$

The estimated features are then inverted and smoothed as per the process described in Section 5.2.

5.4 Results

Performance of the proposed methods of estimation are now tested. The use of the global model, described in Section 5.2, is examined in Section 5.4.1 before the localised system (Section 5.3) is tested in Section 5.4.2 in terms of model selection accuracy and overall enhancement performance. In both cases speaker-dependent, gender-dependent and speaker independent systems are examined to determine the effect of speaker variability on each system. In all cases clean spectral envelope is estimated from noisy speech. The NuanceCatherine dataset is used for speaker-dependent testing whilst speaker-independent and gender-dependent data is taken from the WSJCAM0 corpus (Appendix A). The systems are tested across three different noises: white noise, babble noise and destroyerops noise which are all taken from the NOISEX dataset. These noises are mixed with speech at four SNRs: -5, 0, 5 and 15dB. Clean spectral envelope are then estimated from MFCCs extracted from the noisy data. Performance is measured using two metrics: percentage RMS filterbank error and LLR. RMS filterbank error is used to measure the accuracy of the estimated features after the cepstral transformation has been inverted and is computed as:

$$E_{fb} = \sqrt{\frac{1}{N_c} \sum_{c=1}^{N_c} \frac{1}{N_k} \sum_{k=1}^{N_k} \left[\hat{x}_{fb}(c,k) - x_{fb}(c,k) \right]^2}, \tag{5.6}$$

where E_{fb} is the RMS error across N_k frames, $\hat{\mathbf{x}}_{fb} = \mathbf{C}^{-1}\hat{\mathbf{x}}$ and $\mathbf{x}_{fb} = \mathbf{C}^{-1}\mathbf{x}$. This metric measures performance before the pseudo-inversion of the MFCC features and so LLR (Section 2.6.2.2) is used to measure spectral envelope distortion after the MFCC features have been inverted to the spectral domain.

Three methods of obtaining the joint density models of clean and noisy MFCCs are examined. First, matched conditions are tested to determine optimal system performance. Next, two methods of model adaptation are examined. The Unscented Transform is used for noise adaptation, whilst the effect of speaker adaptation by MAP adaptation is also examined where appropriate. In all of these cases noise is assumed to fit a Gaussian distribution. Not all noises are Gaussian and so Section 4.5.2.4 identified methods of handling non-Gaussian noise. The use of such methods is examined in Section 5.4.3 to enhance speech affected by high levels machine gun noise (-20dB SNR). Finally, Section 5.5 summarises the chapter.

5.4.1 Global model

This section presents results of experiments examining the use of a global model of the joint density for spectral envelope estimation as described in Section 5.2. There are two main properties of the system which we wish to examine: first, the performance of the system across a range of noise types and levels and second, the robustness of the system to variations in gender and speaker. Optimal performance is expected when the noise and speaker are matched to the training environment but we are also interested in how results vary when models are adapted to the target conditions using small amounts of adaptation data. Results are therefore split into two parts. First, into categories in terms of speaker variability and second, within each of these categories a range of model training strategies are tested. Three speaker categories are tested, namely: speaker dependent, gender dependent and speaker independent.

The section begins by optimising estimation model parameters. Next, estimation results are presented in Sections 5.4.1.1, 5.4.1.2 and 5.4.1.3 where speaker-dependent,

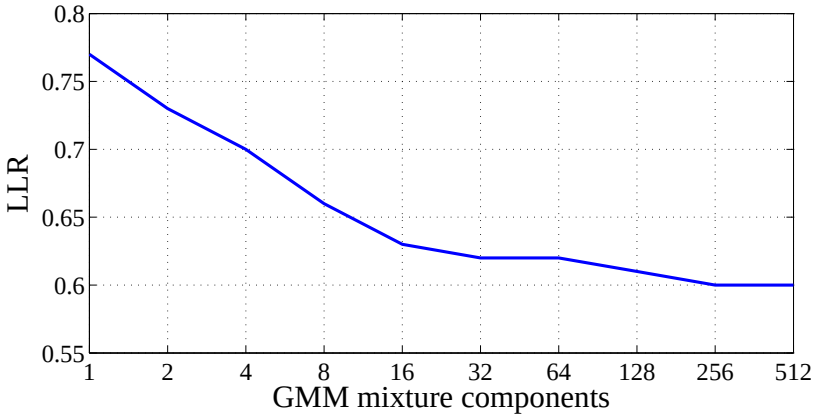


Figure 5.3: Effect of varying the number of mixture components on estimated spectral envelope error measured using LLR

gender-dependent and speaker-independent systems are tested.

For global model enhancement there are two parameters which must be optimised: the feature configuration and number of GMM mixture components. MFCCs have been chosen for feature enhancement with a feature size of 32 filterbank channels transformed to the cepstral domain with 32 DCT coefficients. This is based on previous experiments examining reconstruction quality (Section 3.5.2) and feature correlation (Section 3.5.3.1).

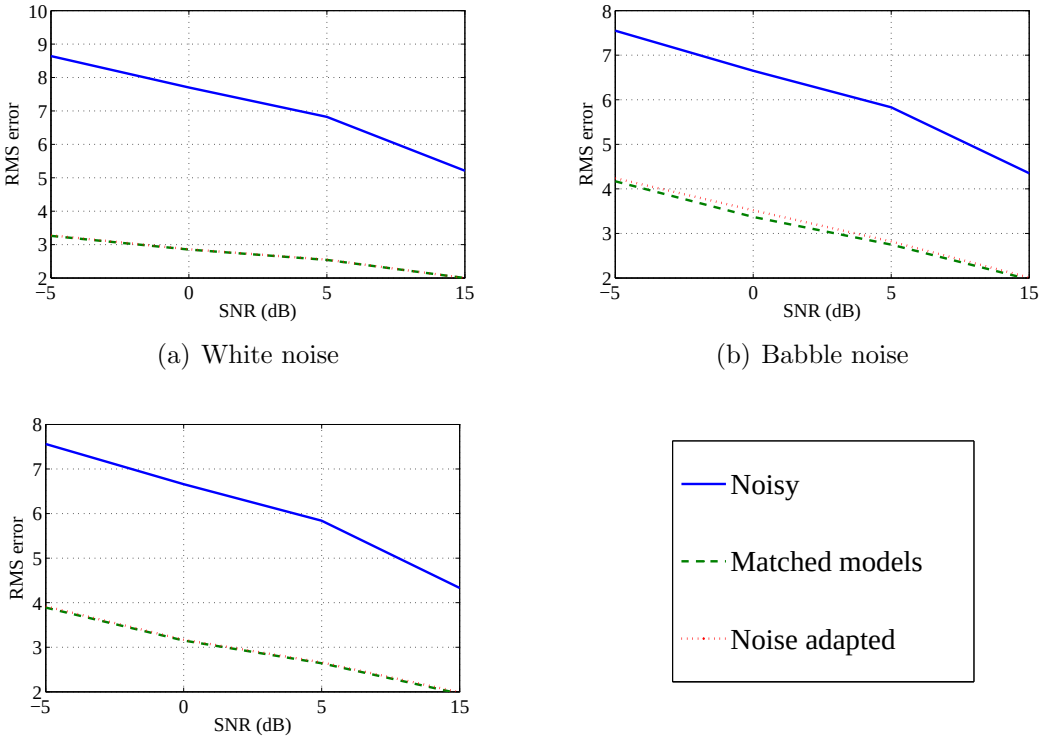
This leaves only the GMM parameters to be optimised. MAP estimation of clean MFCC vectors is reliant on the accuracy of the joint density of clean and noisy MFCC vectors which is modelled by a GMM. To establish the optimal number of mixture components in the GMM features are estimated using the proposed estimation system and the log likelihood ratio (LLR) of the inverted feature vectors is computed to measure spectral distortion. Figure 5.3 examines how estimation accuracy varies with the number of mixture components, M. GMMs were trained from speaker-dependent stereo training data in white noise at 10dB SNR. M was varied from 1 to 512. Estimation accuracy is seen to improve as M is increased until $M = 256$. Beyond 256 mixture components the level of improvement reduces and as such $M = 256$ components will be used in the GMMs. This result was also observed across other noises.

5.4.1.1 Speaker-dependent models

These results focus on speaker-dependent models. Clean spectral envelope are estimated from noisy speech across a range of noise conditions. A single female speaker from the NuanceCatherine corpus is used for training and testing purposes, with 40 minutes of data used for model training and 12 minutes of previously unseen data used for testing. Noise was mixed with clean speech at four SNRs: -5, 0, 5 and 15dB and three noises were used for testing: white, babble and destroyerops. An 8kHz sampling rate was used and MFCCs were extracted from speech using a frame width of 20ms with a 10ms overlap to give a frame rate of 100fps. Two model types for estimation are examined. A model trained on data with the same noise and SNR as the test data ('matched model') is expected to give optimal performance, whilst a model adapted from a clean-trained model of speech will also be tested. Noise adaptation is achieved using the Unscented Transform with the phase-averaged mismatch function described in Section 4.5.2.1. In the case of the noise-adapted model noise is modelled using a Gaussian distribution.

For the purposes of these tests the noise statistics are assumed to be known *a-priori* and so no estimation of these parameters takes place. This is so results are not biased towards a particular method of noise estimation and so optimal performance of the proposed method of estimation can be determined. The result of an experiment in which a method of VAD-based noise estimation is simulated is included to give an indication as to how the system would perform in real-world conditions.

Figure 5.4 shows the result of estimating clean spectral envelope from noisy speech. Performance is measured using percentage RMS filterbank error to determine the estimation accuracy of the methods. In all cases a significant improvement over the noisy data is visible with relative performance best in white noise. Relative performance is observed to be stable across SNR with improvements of $\approx 60\%$ in white noise and $\approx 53\%$ for the two non-stationary noises. Very little difference in performance is observed when noise adapted models are used as opposed to matched



(a) White noise

(b) Babble noise

(c) Destroyerops noise

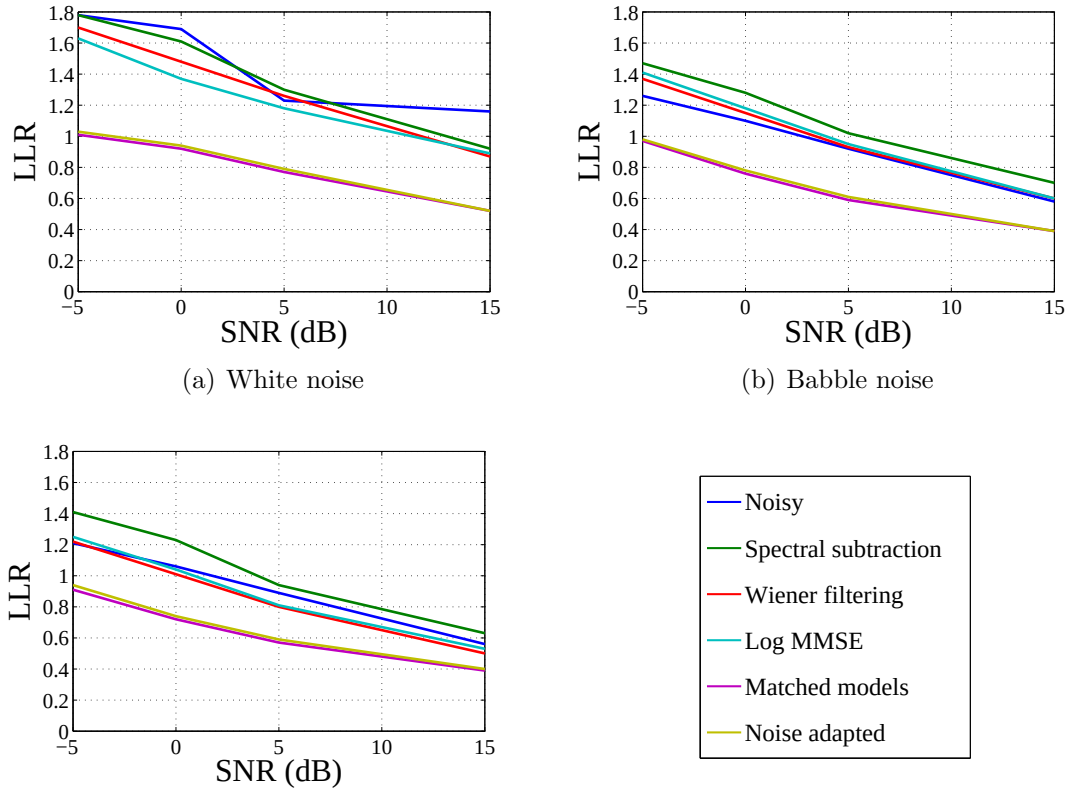
(d) Legend

Figure 5.4: RMS filterbank error of estimated spectral features of a single female speaker across noise and SNR using speaker-dependent stereo-trained (matched) and noise adapted models in a.) white noise, b.) babble noise and c.) destroyerops noise

models. A very slight increase in error is observed for babble noise, however in the case of white and destroyerops noise there are no significant differences in performance.

A number of conventional methods of speech enhancement are now compared to the proposed methods of estimation in terms of spectral envelope distortion. These include: spectral subtraction, Wiener filtering and log MMSE. This is achieved by extracting spectral envelope from the waveforms that result from conventional enhancement and measuring performance using LLR. Previously, results were measured by comparing the RMS error of the filterbank channels. This should give a good indication of final performance but such a metric does not directly measure spectral envelope distortion. Figure 5.5 therefore displays results of the same experiment but this time performance is measured using LLR to compare the accuracy of the actual spectral envelope rather than the raw filterbank channels. Using LLR, performance of noise adapted models is shown to be very similar to stereo-trained (matched) models as per the results measured using RMS error. Across all noises and SNRs the conventional methods of enhancement are shown to offer worse performance than the proposed methods. Surprisingly, in the case of babble and destroyerops noises the conventional methods are actually shown to perform worse than the unprocessed speech. This is attributed to the effect of musical noise.

Next, the effect of the accuracy of the noise model on the performance of the adapted system is measured. All experiments so far have used oracle noise models and so it is perhaps a little unsurprising that the proposed method is shown to outperform other methods. The purpose of such oracle experiments is to confirm the effectiveness of the noise adaptation process and to provide theoretical best-case results, independent of the performance of any particular noise estimation method. Figure 5.6 therefore shows the RMS filterbank error across all three noises at 0dB SNR where the amount of training data used to train the noise model has been varied. Noise samples were randomly sampled from the reference noise signal in 50ms bursts in a Monte-Carlo fashion to emulate VAD-based noise estimation. Errors



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 5.5: LLR of estimated spectral envelope, from a single female speaker, compared across noise and SNR using speaker-dependent stereo-trained (matched) and noise adapted models in a.) white noise, b.) babble noise and c.) destroyerops noise

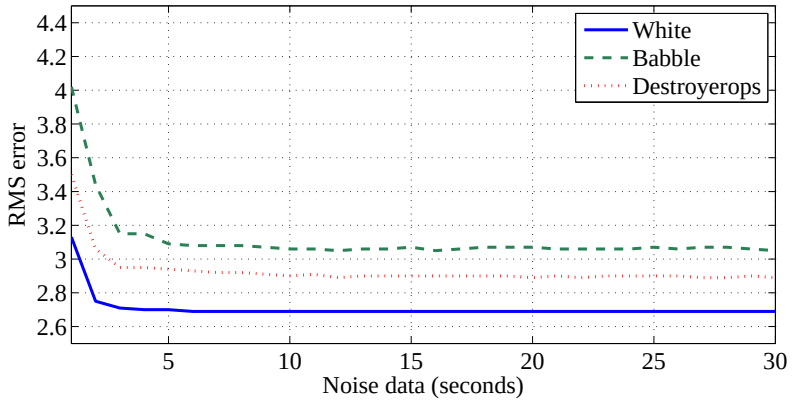
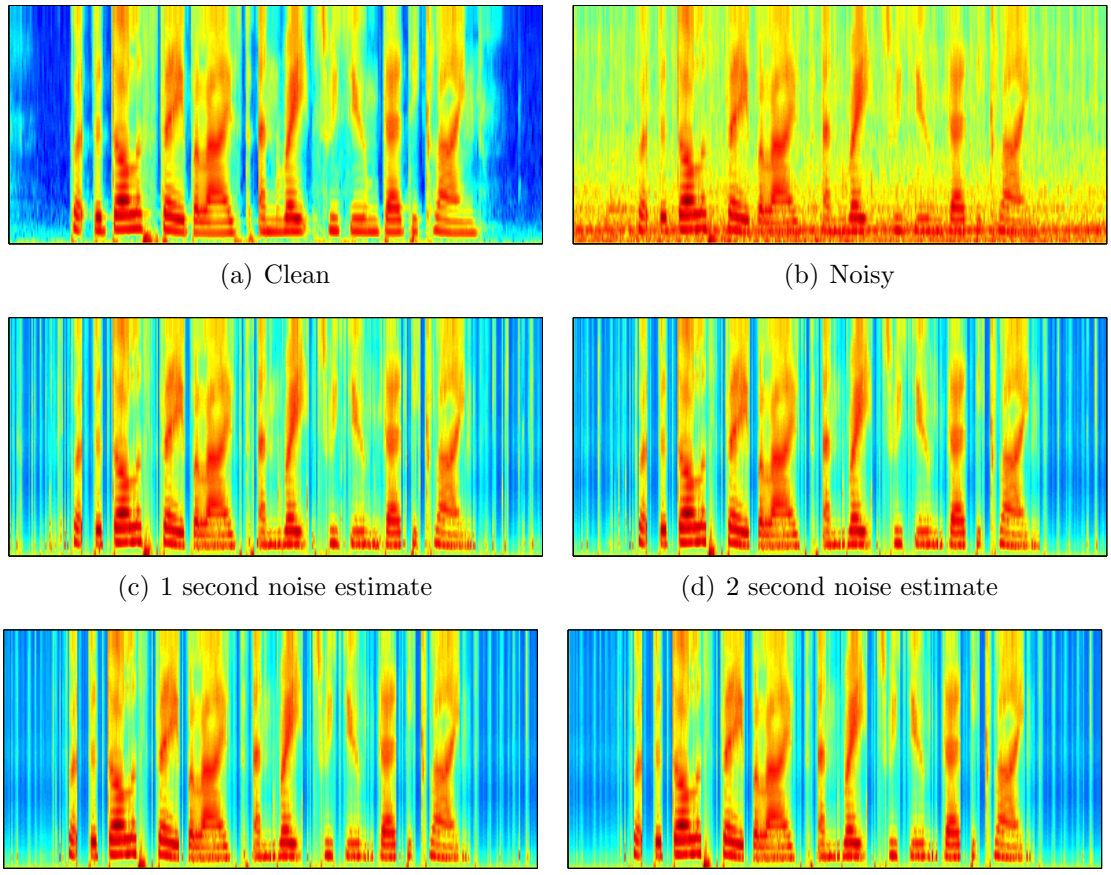


Figure 5.6: RMS filterbank error at 0dB SNR as a function of the amount of noise used to train the noise models used for adaptation in i.) white noise, ii.) babble noise and iii.) destroyerops noise



(a) Clean (b) Noisy

(c) 1 second noise estimate (d) 2 second noise estimate

(e) 5 second noise estimate (f) Entire noise signal (235 seconds)

Figure 5.7: Spectral envelope plots of speech enhanced from destroyerops noise at 5dB SNR using speaker-dependent models adapted with varying amounts of noise data

are seen to reduce as the number of noise vectors used for training are increased, mirroring the results displayed in Section 4.5.2.5 which used the KL divergence metric to compare noise distributions using a similar experimental setup. Performance increases until 2.5 seconds of noise data is available after which no further change in performance occurs. Noise signals are all approximately 235 seconds long and so it is clear that the system is effective when only a fraction of the total noise signal is available. To further examine this point Figure 5.7 displays spectrograms of estimated spectral amplitudes using models adapted with varying amounts of noise data. Enhancement is shown to be good in all cases. Marginally more distortion is visible when only one second of noise data is available whilst very little difference is

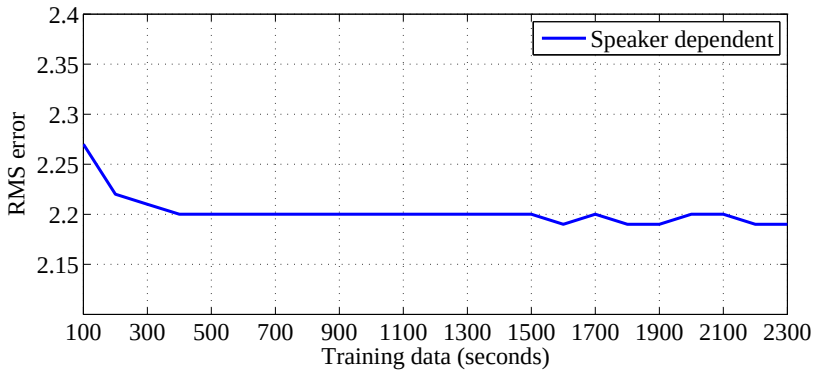


Figure 5.8: Effect of varying the amount of data used to train a speaker-dependent model on estimated filterbank error in white noise at 10dB SNR

observed between the ideal case and those using two and five seconds of noise data.

In terms of the amount of data required to train the overall model of speech, Figure 5.8 examines the effect of varying the amount of training data. Performance was measured in a similar way to the noise model tests. The amount of training data used to train the model was varied between 100 seconds and 2300 seconds in 100 second intervals. Performance is shown to increase with the amount of training data until $\geq$ 500 seconds is used. Using additional training data is shown not to improve performance.

5.4.1.2 Gender-dependent models

Speaker-dependent testing has shown that the system is highly effective at estimating clean spectral envelope using models trained on data from the same speaker in either matched conditions or adapted to the noise. In practical terms it is not useful to have a system that works only on a single speaker and so in this section the use of gender-dependent models is examined. A model is trained for both genders, with 40 speakers used for model training in each case. Each speaker provided just over 18 minutes of data to give models trained on a total of around 12 hours of data each.

First, the amount of data required for gender-dependent model training is tested and compared to the amount of data required for speaker-dependent model training.

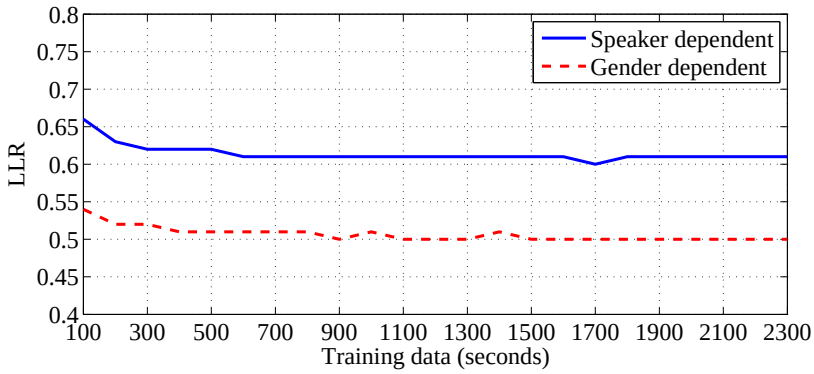


Figure 5.9: Effect of varying the amount of training data used to train speaker and gender dependent enhancement models on spectral envelope distortion of enhanced speech as measured using LLR in white noise at 10dB SNR

Clean spectral envelope was estimated from noisy speech using models trained on varying amounts of training data. Models were trained on stereo data matched to the test environment. Previous results have shown that RMS filterbank error and LLR are highly correlated and so results in this section will be measured using only LLR. The results of this test are presented in Figure 5.9. Performance increases until the amount of training data reaches $\geq$ 600 seconds (10 minutes). After this point the use of additional training data does not improve performance. This is similar to the case of the speaker-dependent system where $\geq$ 500 seconds of training data was required showing that relatively little additional data is required for the gender-dependent case. Gender-dependent results appear to be superior to speaker-dependent results, however the two experiments were run on different datasets and are therefore not directly comparable.

For the adapted model systems, noise adaptation is performed in the same way as with the speaker-dependent system. In addition, a system which also adapts for speaker variability is introduced to account for speaker mismatch. A two stage system as described in Section 4.5.3 is used. Speaker adaptation is first performed on the gender-dependent models of clean speech using MAP adaptation to give speaker-dependent models. These are then adapted for noise using the UT in the normal way. There are therefore three systems to be tested: matched models, noise

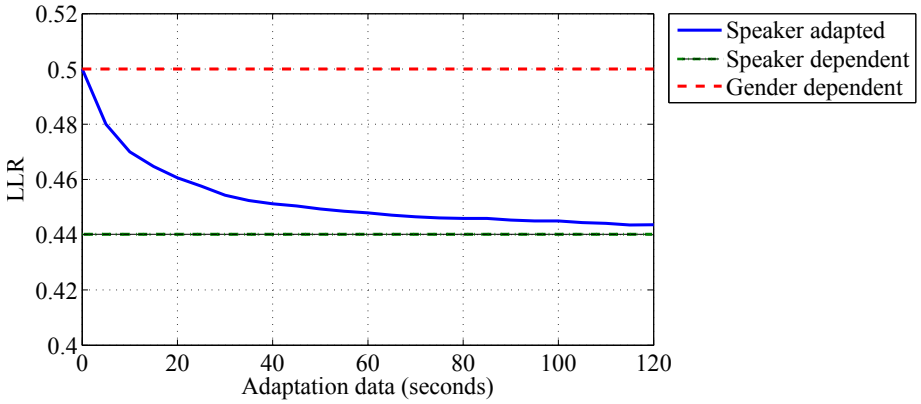
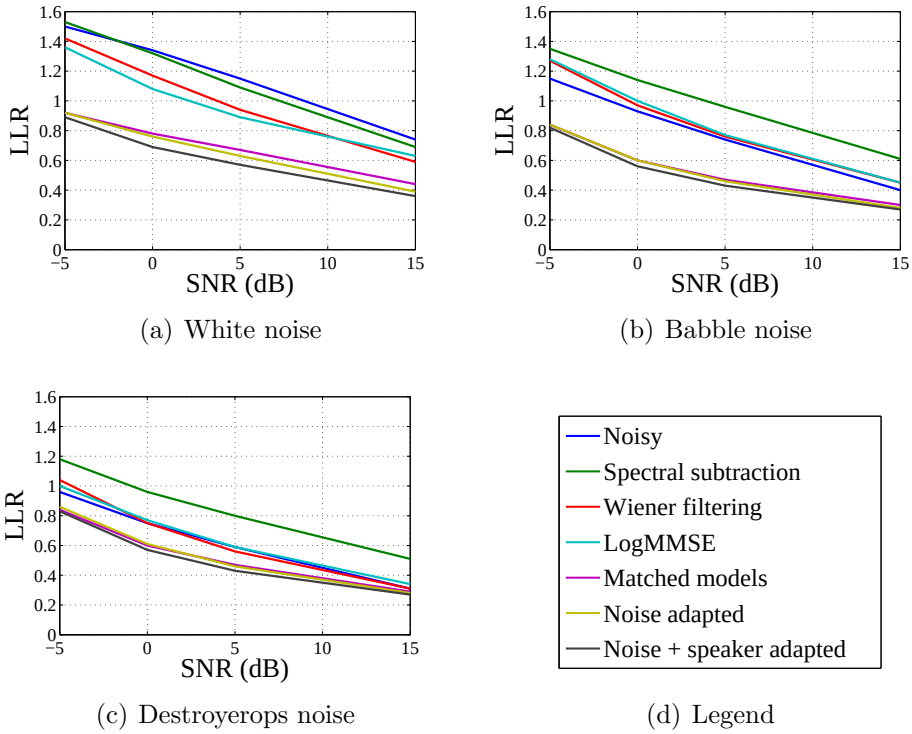


Figure 5.10: Effect of varying the amount of speaker data used for adaptation when enhancing female speech contaminated with 10dB SNR white noise using a.) speaker-adapted model, b.) speaker-dependent model and c.) gender-dependent model

adapted models and models adapted for both speaker and noise.

Speaker adaptation is performed using additional speaker-dependent data and so it is therefore useful to determine how much data is required for adaptation. Figure 5.10 shows the effect of speaker adaptation on enhancement by taking an isolated case and varying the amount of new speaker data used for adaptation. The case of enhancing speech of a single female talker in white noise at an SNR of 10dB was considered. Speaker and gender dependent results were included to determine best and worst case performance. The amount of adaptation data was varied between 0 to 120 seconds, covering the range expected to be available in realistic conditions. An immediate advantage is seen over the gender dependent model when as little as 5 seconds of data is available. Performance continues to increase until 80 seconds of data is available where the level of performance increase is significantly reduced. Speaker adaptation is shown to provide a useful increase in performance with relatively little data and so will be used in later experiments.

Performance is now measured across the three noises previously tested, that is: white, babble and destroyerops. Figure 5.11 shows performance of the female-only system whilst Figure 5.12 shows performance of the male-only system. In all cases performance is significantly better than conventional methods as per the speaker dependent system. Noise adapted models are shown to perform marginally better



(a) White noise (b) Babble noise

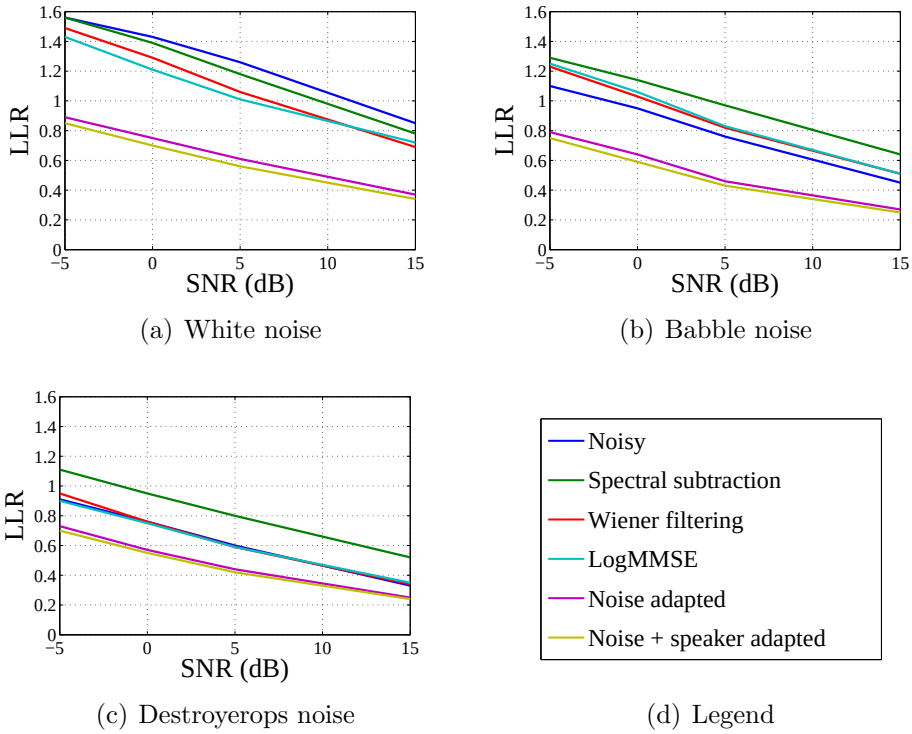
(c) Destroyerops noise (d) Legend

Figure 5.11: Performance of using gender-dependent models for clean spectral envelope estimation from noisy speech spoken by female speakers measured using LLR

than matched models. This is explained by the noise mixing process. Noise was added on a per-speaker basis. As such the absolute level of the noise between speakers will not be uniform due to variations in recording levels. As expected, speaker adaptation gives further reductions in spectral distortion in all cases. Performance was observed to be nearly identical between male and female specific systems.

5.4.1.3 Speaker-independent models

Using gender dependent models was shown to increase spectral distortion, though this was reduced considerably through the use of speaker adaptation. Gender dependent models require a system to determine the gender of the speaker and so to reduce the complexity of the system this section examines the performance of a fully speaker-independent system. Models are trained on the same data as the gender dependent models to give a total of 24 hours of training data taken from 80 speakers



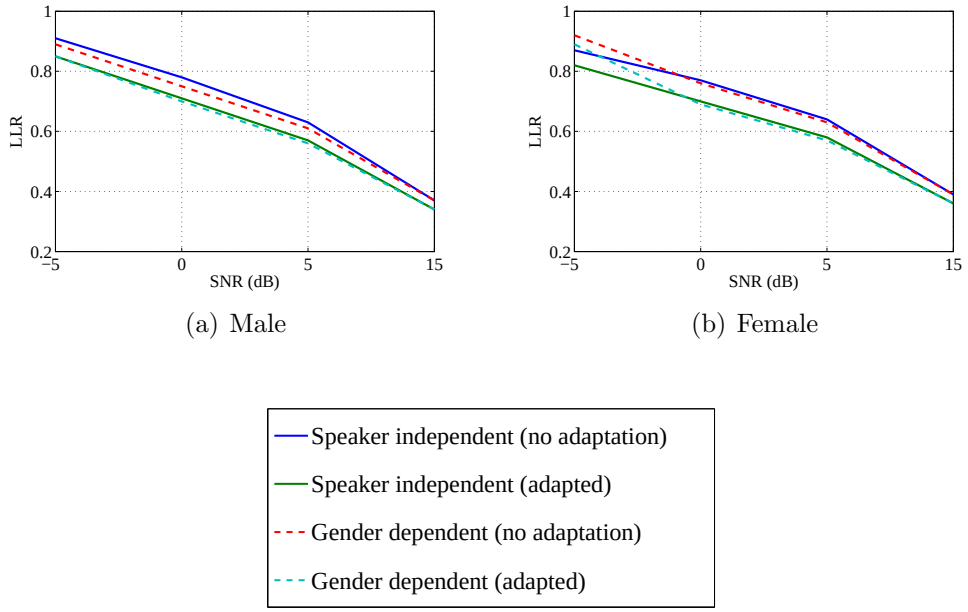
(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 5.12: Performance of using gender-dependent models for clean spectral envelope estimation from noisy speech spoken by male speakers measured using LLR

with an equal male/female split.

Two factors are examined. First, relative performance with gender dependent models is examined in Figure 5.13 before second, performance is compared to conventional methods of enhancement in Figure 5.14. Very little difference in performance is observed when speaker-independent models are used for estimation instead of gender-dependent models. In the case of systems using no speaker adaptation the use of speaker independent models increases spectral distortion when enhancing male speech compared to the male-only case. No significant differences were noted in the case of female speech and when using speaker adaptation demonstrating the effectiveness of speaker adaptation for estimation. Next, speaker independent performance is compared to conventional methods. Performance has been shown to be similar to the gender dependent case and so it is of no surprise that the same relative performance is also observed compared to conventional methods.



(a) Male

(b) Female

(c) Legend

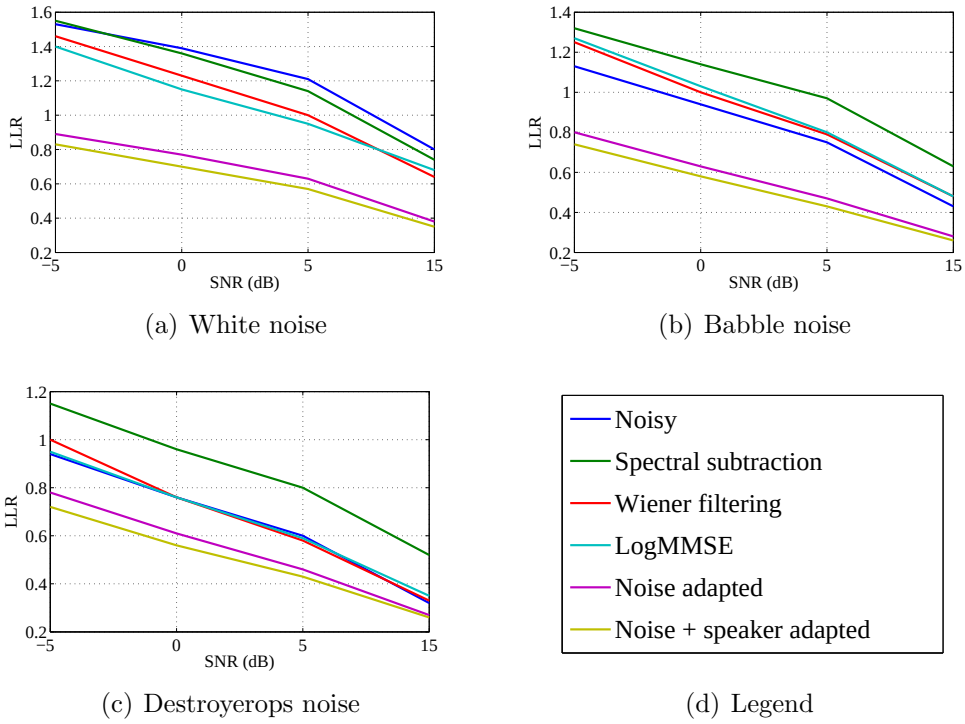
Figure 5.13: Comparison of performance of using speaker-independent models versus gender-dependent models for the purpose of spectral envelope estimation from noisy speech contaminated with white noise

5.4.2 Localised models

This section examines performance of a localised modelling system as described in Section 5.3. Experiments measuring phoneme-level feature correlation in Section 3.5.3.1 suggest that localised models should improve estimation accuracy. Such a system introduces additional complexities, however, with the accurate selection of enhancement models expected to be a key limiting factor. This section begins by examining speaker-dependent systems which are expected to give best performance before moving on to gender-dependent modelling.

5.4.2.1 Speaker dependent

This section is based on previously published work [Harding and Milner, 2012a]. In this case, street noise from the AURORA2 dataset [Hirsch and Pearce, 2000] was used to degrade speech at three SNRs: 0dB, 5dB and 15dB. Speaker-dependent



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 5.14: Performance of using speaker independent models for clean spectral envelope estimation from noisy speech spoken by male and female speakers measured using LLR

Table 5.1: Speaker-dependent class recognition accuracy (%) in street noise

	Clean	15dB	5dB	0dB
Phoneme	74.23	61.92	45.04	33.20
Articulation	77.94	72.04	61.31	52.03

models, trained on conditions matched to the operating environment, are used in the case of both enhancement and recognition.

Estimation model parameters were determined in the same way as per the global system. The optimal number of mixture components per estimation model was varied between 4 and 128 depending on the voicing class and amount of data available; models with very little training data were assigned a relatively low number of mixture components to avoid over-fitting. In terms of recognition model parameters, an HMM-GMM system using a left-right HMM topology with 3 emitting states was used, with 128 mixture components used to model distributions within the recognition system. The recognition system was trained on the same speaker data as the estimation models with conditions matched to the target environment.

Performance of spectral envelope estimation using phoneme and articulation classes is compared to global modelling. The accuracy of the recogniser is measured on the phoneme and articulation classes and the results shown in Table 5.1. This shows articulation class classication to be more robust to noise, having only seven possible class labels compared to the 41 phonemes.

An investigation is now made of the spectral envelope estimation accuracy made by the phoneme class, articulation class and global systems. Figure 5.15 shows mean RMS filterbank error when compared to the original clean features. To show the effect of frame classication accuracy in spectral envelope estimation, the RMS error of the phoneme and articulation class systems are shown first using reference labels (no classification errors) and then using the noisy HMM-based classifications (as shown in Table 5.1). Best performance is given by the phoneme class system using reference labels. This is expected as this method of classification has the most accurate localisation of the feature space. When the HMM-based recogniser provides

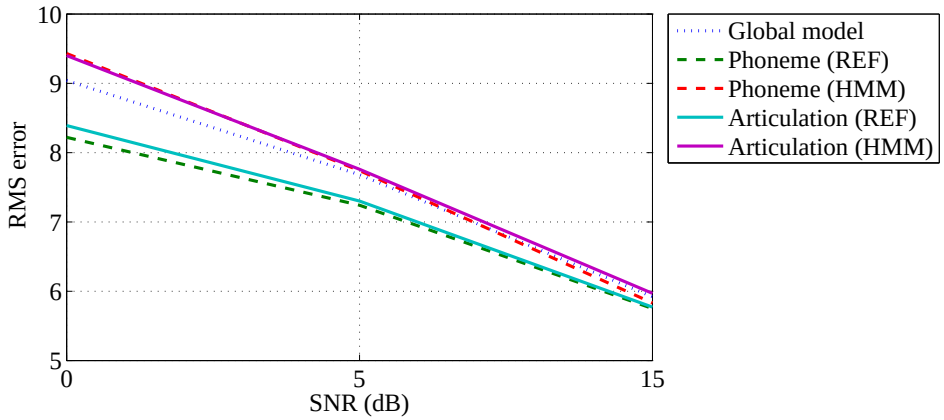


Figure 5.15: Spectral envelope RMS error of female speech enhanced using stereo-trained, speaker-dependent localised models using i.) phoneme classes, ii.) articulation classes in street noise. The use of reference labels (REF) is compared to the obtaining class labels from the HMM-based system (HMM). Performance using a global model is also shown for reference.

class information the errors it introduces cause a deterioration of spectral envelope estimation which increases above both the articulation class and global systems at 0dB SNR. Despite the articulation class labels being more accurate than phoneme labels, the less detailed localisation of the feature space yields performance roughly equal to that of the phoneme-based system at SNRs of 0dB and 5dB.

Speaker dependent results have shown that a system of feature estimation using localised models outperforms the global system when given reference class labels. However, when using a realistic recognition system in highly noisy environments recognition accuracy falls sufficiently low as to cause the overall system errors to increase to above that of the system using a single model of enhancement. Only phoneme models are considered from now on as they were previously shown to offer best performance.

5.4.2.2 Gender dependent

In this section we examine the case of expanding the system to use gender-dependent recognition and enhancement models. Whilst this would normally be expected to reduce performance due to the increased speaker variability, more data is available

for training which may improve robustness. A total of $\approx$ 33 minutes of data was available for training speaker-dependent models, of which 400 seconds was silence. Data was not spread evenly across phonemes, with some phonemes represented by as little as 3-5 seconds of training data. In the case of gender-dependent training significantly more data is available. Models are trained on about 10 hours of data, 2 hours of which is silence. Subsequently, models were trained on a minimum of 31 seconds of data each, with the majority of models trained on at least 8 minutes of data.

Gender dependent models introduce additional speaker variability which is known to degrade the performance of both the recognition and enhancement systems. As per the global enhancement system, MAP adaptation is used to adapt gender-dependent models to speaker dependent models using a small amount of speaker-dependent data. This data may also be used to adapt the recognition models using similar techniques. Previously, recognition models were trained on noisy data matched to the operating environment, however this is not always practical. Instead, for these experiments we consider the case of adapting the recognition system for variations in both speaker and noise.

Previously recognition models were trained on data matched to the target environment. This is not always practical and so instead models are now trained on clean data and adapted for variations in speaker and noise. Noise adaptation for the recognition system can be achieved in two ways. The techniques described in Section 4.5.2 may be used to adapt clean trained models in the case of model adaptation. Alternatively, features may be compensated and used with clean trained models. A two-pass enhancement system may therefore be considered whereby features are enhanced using the global system described in Section 5.2 and used to classify utterances before localised models are used to give the final enhanced feature vectors. This process is illustrated in Figure 5.16.

In terms of speaker adaptation for the recognition system, all methods considered are model-based and consist of the global and class-based MLLR and CMLLR trans-

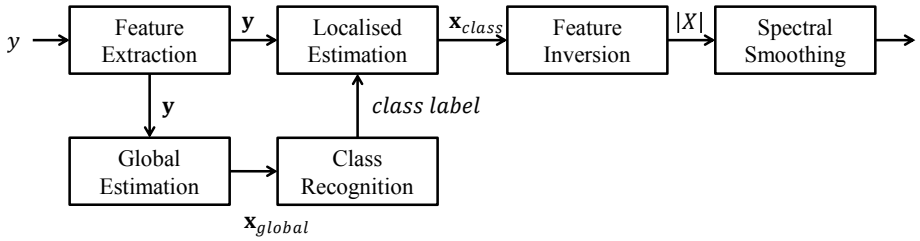


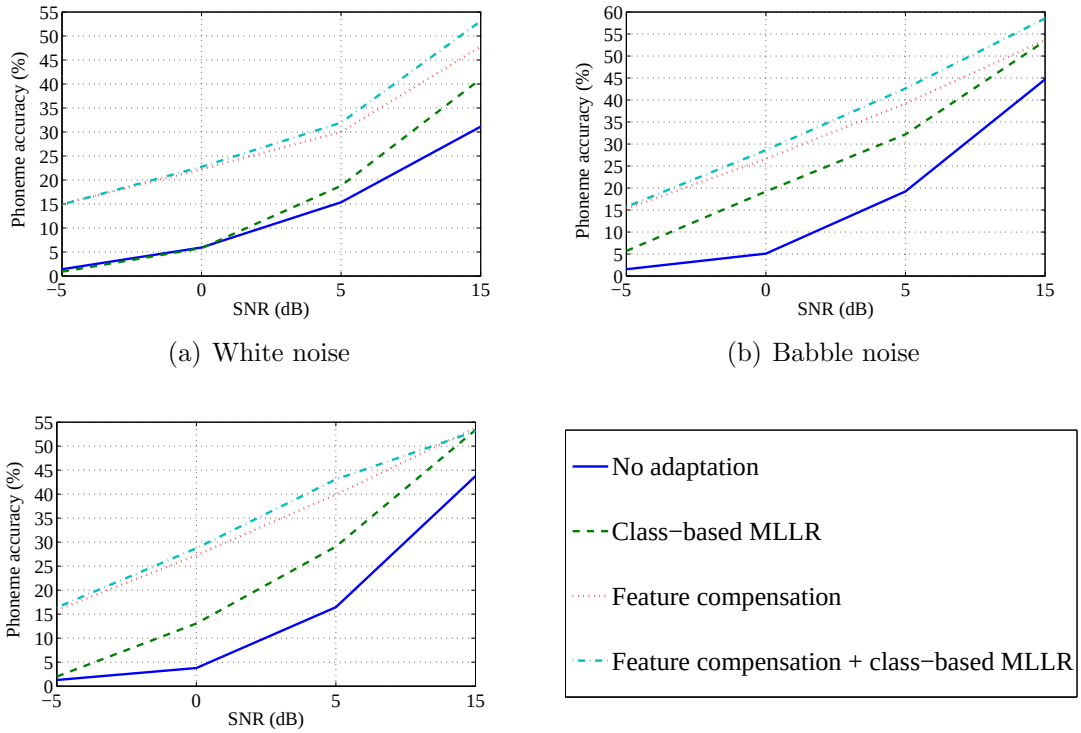
Figure 5.16: Two-pass enhancement enhancement system of i.) enhancement using a global model, before ii.) enhancement using a localised system using global enhanced features as input to the classification system

Table 5.2: Phone recognition performance for a speaker-dependent HMM classification system in clean conditions

	Correct (%)	Accuracy (%)
No adaptation	64.22	61.56
Global MLLR	65.21	62.74
Class-based MLLR	**67.19**	**64.55**
Global CMLLR	65.10	62.38
Class-based CMLLR	66.92	64.44

forms [Gales and Woodland, 1996]. Table 5.2 shows phoneme accuracy comparing the four systems with unadapted models using clean data. 120 seconds of speaker adaptation data with no added noise was used in each case. Class-based MLLR is shown to offer best performance and so is chosen for use in this work. It should be noted that even in clean conditions, in a third of cases the wrong model will be chosen. Even in the best case noise will degrade performance of the recogniser and so the aim is to limit the effect of noise as much as possible.

Next, methods of noise compensation are evaluated. Four systems are considered: no adaptation, class-based MLLR, feature compensation and finally feature compensation with class-based MLLR. In the case of class-based MLLR for noise adaptation speaker-dependent noisy data matched to the operating environment was used, whilst clean data was used in the case of MLLR with compensated features. Figure 5.17 shows the performance of the female-only system. A two-stage system with class-based MLLR speaker adaptation is shown to perform best across all noises for female speech and so is also used for the male-only system. Recognition results for the male system are shown in Figure 5.18 and are shown to be roughly equivalent



(a) White noise

(b) Babble noise

(c) Destroyerops noise

(d) Legend

Figure 5.17: Phoneme accuracy of female-only phoneme recognition systems trained on clean speech and tested using i.) noisy speech features, ii.) model adaptation using MLLR, iii.) features compensated for the noise using the proposed system trained on stereo data, and iv.) compensated features and MLLR

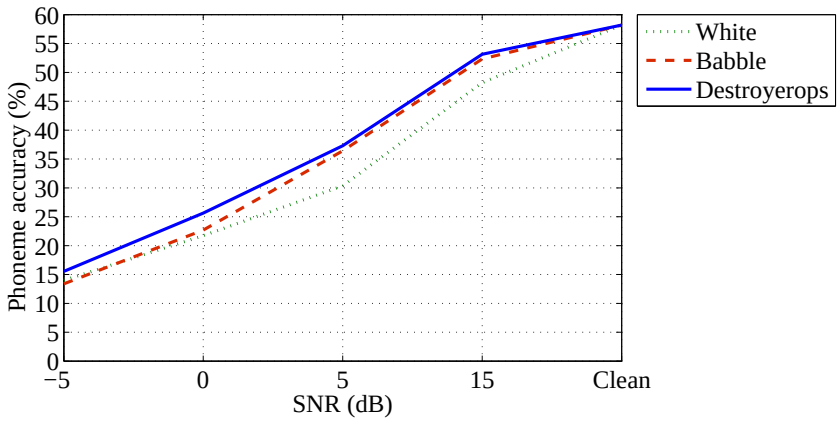
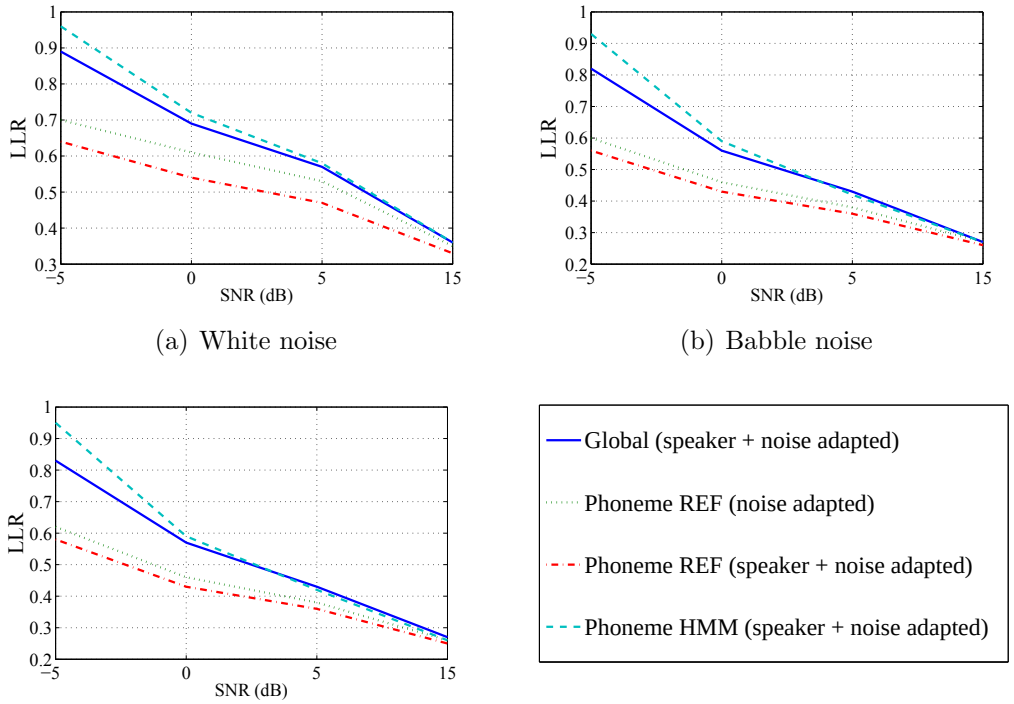


Figure 5.18: Phoneme accuracy of male-only phoneme recognition system using compensated features and class-based MLLR

to those using female speech.

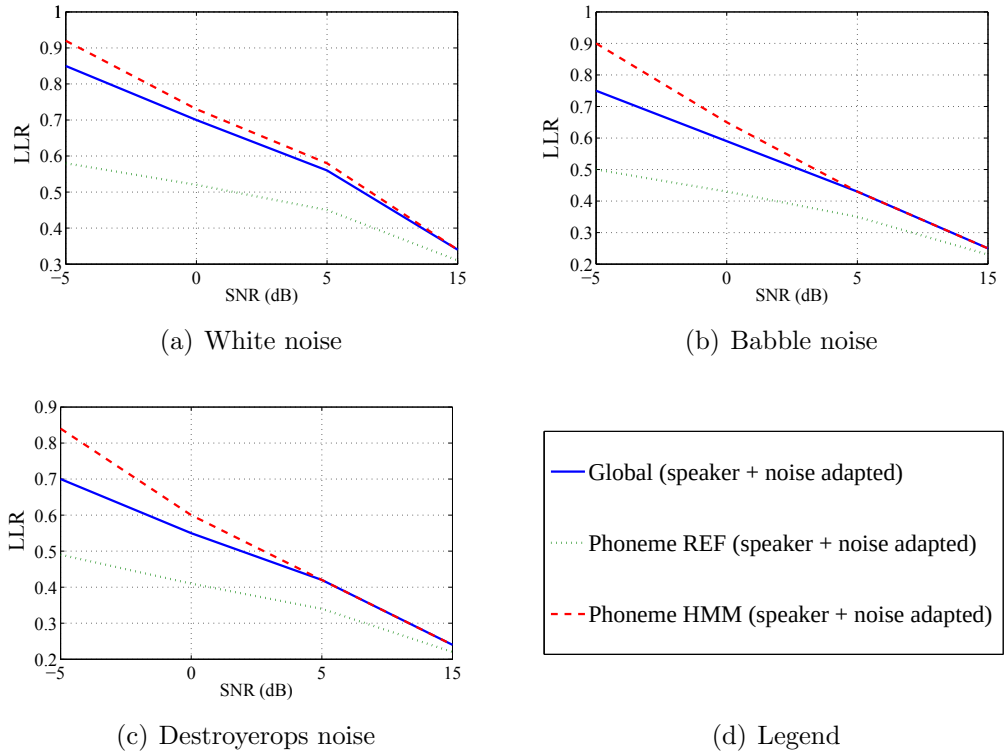
Next, we consider the performance of spectral envelope estimation using localised models selected by the two-pass recognition system. Figure 5.19 shows female-only results whilst Figure 5.20 shows male-only results. Four systems are tested. The global system evaluated in Section 5.4.1 is compared to three localised systems. These are: i.) phoneme models with reference labels with enhancement models adapted for noise, ii.) adapted for speaker and noise and iii.) phoneme models with labels obtained using the two-pass recognition system with speaker and noise adaptation for the enhancement models. The phoneme-based system using reference labels with enhancement models adapted for speaker and noise is clearly shown to offer best performance. A system which adapts enhancement models for noise only, but that is otherwise identical, offers similar performance. In both cases a considerable improvement over the global system is observed. As per the speaker-dependent system this is attributed to the more accurate localisation of the feature space. Whilst these systems have been shown to offer very good performance, they assume prior knowledge of the enhancement model sequence and time alignment. The best phoneme recognition system was therefore used with the best-case enhancement model configuration to determine the best overall realisable system. The recognition system chosen for use is therefore the two-pass recognition model which uses



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 5.19: Mean LLR of female speech enhanced using the proposed phoneme-based two-pass enhancement system comparing the use of reference and realistic class labels to the global system



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 5.20: Mean LLR of male speech enhanced using the proposed phoneme-based two-pass enhancement system comparing the use of reference and realistic class labels to the global system

enhanced features with clean-trained, speaker adapted, recognition models whilst the enhancement system uses models adapted for speaker and noise. At high SNR ($\geq$ 5dB) performance is shown to match that of the global system whilst at lower SNRs the system is shown to offer worse performance. Based on the superior performance of the localised system using reference models and the low recognition performance at low SNRs this reduction in performance is attributed to erroneous model selection; at -5dB SNR the wrong models are selected in $\approx$ 85% of cases compared to $\approx$ 40% at 15dB.

5.4.3 Non-Gaussian noise

The previous tests examined the use of systems using global and localised enhancement models. Common to both cases was the use of the UT to adapt models for noise which assumes the use of a Gaussian distribution to model the noise. Not all noises can be modelled in such a way; noises such as machine gun noise with several distinct spectral modes require the use of multi-modal distributions which may instead be modelled as a mixture of Gaussians using a GMM or states using an HMM. Appropriate strategies for handling such noise were determined in Section 4.5.2.4. Two systems were proposed. In both cases a GMM modelling the noise is assumed to be available. First, an HMM-based method was proposed for PMC style adaptation. This introduces several challenges, primarily how to obtain model parameters. Second, a new method of 'serial model combination' (SMC) was proposed. This method effectively treats each mixture component of the noise GMM as a separate noise model for adaptation and then stacks the resultant noise adapted models to form a single joint density model for enhancement. Preliminary results presented in Section 4.5.2.4 showed the system to be effective and so the use of such a system is further considered in this section. An enhancement system using a single global model is preferred for use in this section based on the results presented in Sections 5.4.1 and 5.4.2.

First, the objective results of enhancing speech corrupted by machine gun noise

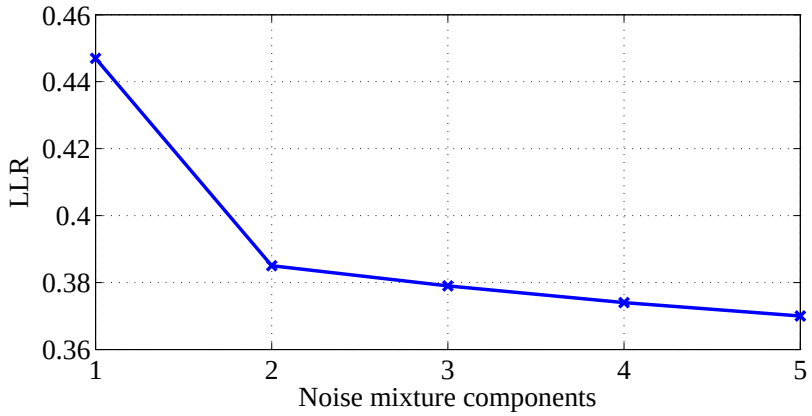
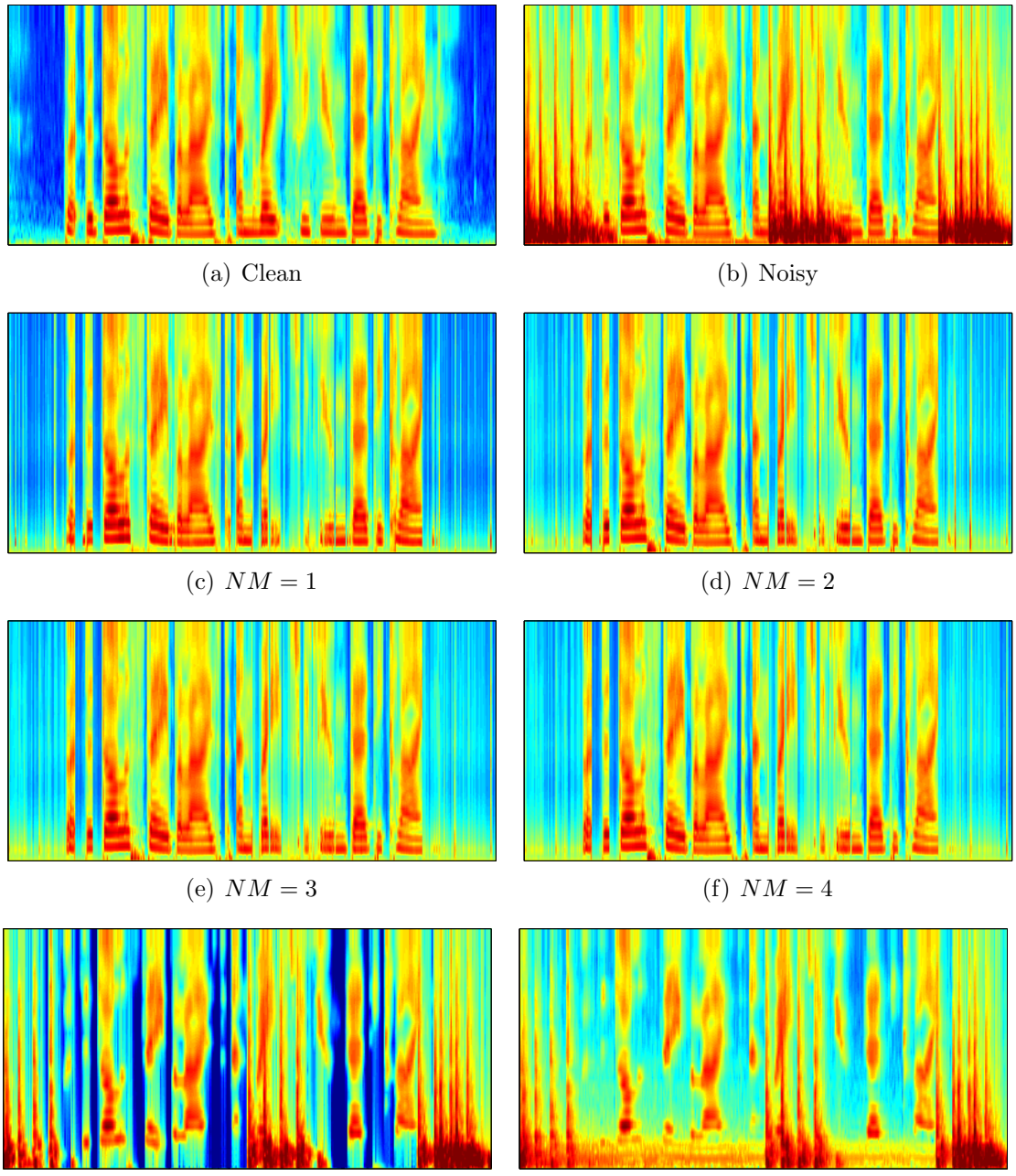


Figure 5.21: Performance of SMC enhancement in terms of LLR using speech corrupted by machine gun noise at -20dB SNR and enhanced using the global system with SMC noise adaptation applied, displayed as a function of the number of mixture components used to model the noise

at -20dB SNR are presented in Figure 5.21. Speaker independent data was used for training and enhancement with MAP adaptation applied for speaker adaptation and SMC with UT used for noise adaptation. The number of mixture components which comprise the noise model were varied between 1 and 5 to give adapted models with between 256 and 1280 mixture components. A significant increase in performance is observed when the number of noise mixture components is increased from one to two. After this point there is a small increase in performance as the number of components is increased. Based on the distribution shown in Figure 4.9(b) it is not surprising that performance increases when more than one noise component is used, however what is surprising is that performance continues to increase beyond three noise mixture components. This increase in performance is therefore attributed to over-fitting of the noise data.

Next, the effect of varying the number of mixture components used by the SMC enhancement scheme is evaluated in terms of spectral amplitudes of a single utterance in Figure 5.22. Two conventional methods are included for comparison purposes, namely Wiener filtering and log MMSE. Very little obvious differences are noticeable between Figures 5.22(c)-(f), where noise mixes are varied between one and four, with all configurations removing the vast majority of noise energy. In all



(a) Clean (b) Noisy

(c) $NM = 1$ (d) $NM = 2$

(e) $NM = 3$ (f) $NM = 4$

(g) Wiener filtering (h) log MMSE

Figure 5.22: Spectrograms of an example utterance comparing enhancement using SMC to conventional methods using speech corrupted by machine gun noise at -20dB SNR

cases this is at the expense of speech distortion. The noise consists of regions of low frequency energy of medium duration (recoil) with very short duration, high energy bursts of wideband energy within each region of noise (gunshot). Where the noise coincides with speech energy the first two to three harmonics are completely masked by the low frequency energy. In the case of isolated gunshots the entire frame is dominated by the noise. Noise may occur in periods of silence or speech and in some cases speech energy is completely missing from the estimated features where silence frames have been mistakenly estimated from frames of very noisy speech. This is most apparent towards the middle of the utterance where a burst of noise has resulted in a short period where speech information has been lost in the estimation process. This is most visible in Figure 5.22(c) where noise was modelled as a Gaussian distribution whilst Figures 5.22(e) and (f) are seen to perform best in this case, though both cases still suffer from information loss. Comparing now the conventional methods to the SMC technique and clear differences are observed. Neither Wiener filtering nor log MMSE are shown to recover any speech energy in regions corrupted by the machine gun noise. The noise is shown to be unaffected by the filtering whilst even periods where no noise previously existed have been distorted. In both cases low frequency harmonics have been completely lost across the entire utterance. This is attributed to the noise estimation algorithms assuming stationary noise and that all frames are affected equally by the noise.

Whilst SMC has been shown to be more effective than straight-forward adaptation for certain types of noise there are concerns relating to the computational efficiency of such a system. Assuming clean speech is modelled by a GMM with $M = 256$ mixture components a system which adapts this model using a noise model with three mixture components would result in an adapted model with $M = 768$ components. The estimated spectral values are formed as a weighted average of estimates from all mixture components and so as the number of noise mixtures is increased, so is the computational complexity at runtime. Figure 5.23 therefore examines the case of using the n-most likely mixture components as determined by the posterior probability, $f(\phi_k|\mathbf{y})$. Once the n-best mixture components have been

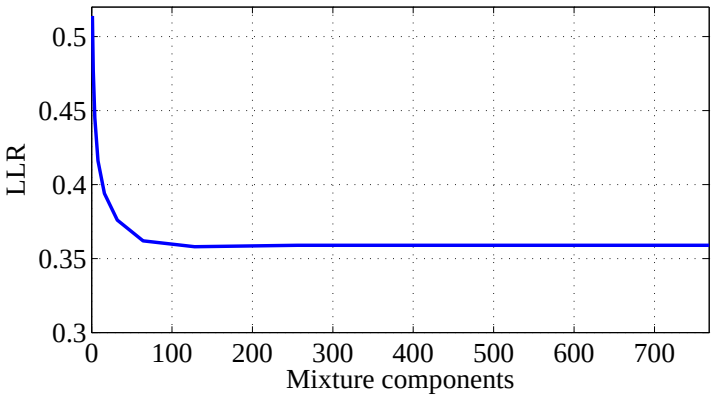


Figure 5.23: Effect of varying the number of mixture components used for estimation in an SMC-adapted system on the LLR of enhanced speech

identified the mixture prior probabilities are equalised as appropriate. Performance is shown to increase until $n = 128$ mixture components are used for estimation. Past this point there is no advantage to using additional estimates. To reduce runtime complexity it is therefore possible to reduce the number of estimates computed and thus increase computational performance.

5.5 Summary

Two methods of spectral envelope estimation were proposed in this chapter, both based on MAP estimation. First, a system using a global model for estimation was proposed. Next, a system using localised models was proposed to exploit the greater correlation between clean and noisy features that exists in some acoustic classes. The performance of such systems was measured in Section 5.4 using a range of speakers and noises. Speaker dependent performance was shown to be very good, clearly outperforming conventional methods such as Wiener filtering and log MMSE in terms of spectral envelope distortion. Best performance was obtained using models trained on data with the same noise as the test data, though adapting for noise using the UT was shown to offer similar performance even when relatively little information about the noise was available. In the case of gender dependent

and speaker independent systems the introduction of additional speaker variability was shown to reduce performance slightly, though the effect of this was limited with the use of MAP speaker adaptation. Next, a system using localised models of the feature space was tested in Section 5.4.2. The use of localised models was shown to provide significantly better performance in terms of spectral distortion, however this system came with the additional challenge of model selection for enhancement. Despite promising results using prior model sequences the introduction of a realisable recognition system for model selection reduced performance to below that of the global system. In both cases noise was assumed to be modelled as a Gaussian however this is not always optimal and so a method of adapting models using a GMM of the noise was tested in Section 5.4.3. The use of GMMs to model the noise was shown to improve spectral estimation considerably. The optimal method of spectral amplitude estimation is therefore determined to be a system using a global model trained on clean speech and subsequently adapted for speaker variations using MAP adaptation and for noise using either the standard UT approach for the case of Gaussian noises and SMC using UT for non-Gaussian noises.

Chapter 6

Fundamental Frequency Estimation

This chapter describes a method of robust fundamental frequency ($f0$) estimation. A method of estimation similar to the one used for spectral envelope estimation is proposed. MFCCs are extracted from noisy speech and $f0$ estimated using MAP. Two methods of noise robustness are evaluated, namely: model adaptation and feature compensation, whilst speaker variability is compensated using MAP adaptation. The chapter begins by examining existing methods of $f0$ estimation. The proposed method of estimation is then described, the performance of which is evaluated and compared against existing methods.

Contents

6.1 Introduction

This chapter presents a method of robust estimation of the fundamental frequency ($f0$). $f0$ is closely related to, but not synonymous with, the pitch of a speech signal. The fundamental frequency of a signal is the rate at which the vocal folds open and close (Section 3.2) and can be directly measured from the signal. Pitch on the other hand is more subjective, describing the tone of a signal [Talkin, 1995]. Signals with large $f0$ are generally perceived as being 'high pitch' whilst those with lower $f0$ are considered to be 'low pitch' [Fry, 1979]. Despite this link, the perception of pitch may also vary according to the duration and loudness of the sound. In this work we are primarily concerned with accurately measuring the $f0$.

There are many challenges to accurate $f0$ estimation [Rabiner et al., 1976]. Measuring the $f0$ of a perfectly periodic and clean signal is relatively easy, however speech is not usually perfectly periodic nor clean. This is due to variations in the excitation signal as well as variations in spectral detail within each period of the signal caused by vocal tract filtering. In addition, during transitions between voiced and unvoiced speech (or vice-versa) it is difficult to determine the precise cut-off point between the two regions and this can also lead to incorrect measurements of $f0$.

Several methods of estimation are considered and split into two categories. First, 'conventional' methods are described in Section 6.2.1 before model-based methods of estimation are considered in Section 6.2.2. Conventional methods are defined as methods that directly measure some property of the signal in order to determine $f0$. Such methods may operate in the time-domain, frequency-domain or both, with typical measurements being peak and valley measurements, zero-crossing rate and autocorrelation in the time domain and peak-detection in the frequency domain [Hess, 1992]. Frequency domain analysis may be extended to the cepstral domain where the periodic structure of the fundamental and its harmonics are detected as a high quefrency peak in the cepstrum [Rabiner et al., 1976]. Model-based methods are defined as data-driven techniques that use a statistical model to compute estimates

of $f0$. The proposed method of estimation is presented in Section 6.3.

6.2 F0 Estimation Review

This section describes existing methods of $f0$ estimation divided into two categories. First, conventional methods of estimation are considered in Section 6.2.1. Second, newer model-based methods are described in Section 6.2.2.

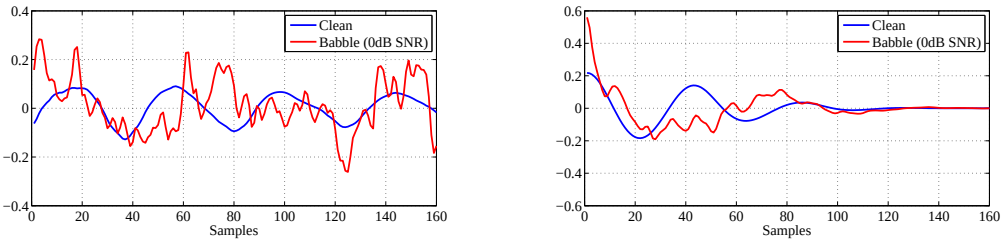
6.2.1 Conventional methods of $f0$ estimation

Conventional methods of $f0$ estimation have been described as methods that measure the periodicity of a signal in either the time-domain, frequency-domain or cepstral-domain. This section describes a number of techniques which include autocorrelation, including the average magnitude difference function (AMDF) variant, and a hybrid method which uses measurements of $f0$ in both the time and frequency domain to form an overall estimate. In all cases $f0$ is estimated on a short-time frame-by-frame basis with frames typically 20ms in duration with a 10ms overlap so that the signal can be assumed stationary.

6.2.1.1 Auto-correlation

An autocorrelation-based $f0$ estimation algorithm is distributed as part of the PRAAT toolbox [Boersma, 2002]. Autocorrelation-based methods such as PRAAT estimate $f0$ values by measuring the position of the largest non-zero lag peak in the autocorrelation analysis of the signal. This peak corresponds to the point at which different segments of the signal are most similar and so should correspond to one period of the signal in frames of voiced speech. The autocorrelation function (ACF) is measured from windowed frames of speech as:

$$R(\tau) = \sum_{m=0}^{N} w(m)x(m)w(m+\tau)x(m+\tau), \tag{6.1}$$



(a) Time-domain waveform (b) Autocorrelation analysis

Figure 6.1: Example of autocorrelation analysis in babble noise at 0dB SNR showing clean and noisy speech in a.) the time domain, b.) autocorrelation domain

where $w(m)$ is the nth sample of the windowing function, $x(m)$ is the mth sample of the time-domain waveform and τ is the autocorrelation lag. A Hann or Hamming window is used in most cases. $f0$ is measured as the largest non-zero lag peak in R. This method of estimation is reliable in the case of perfectly periodic speech in clean conditions however variations in the vocal tract filter and the addition of noise are known to degrade the accuracy of estimations [Rabiner et al., 1976]. These degradations can cause the true value of τ relating to the $f0$ to be masked by another peak.

To illustrate the effect of noise on $f0$ estimation Figure 6.1 gives an example of estimation in the presence of additive noise. Figure 6.1(a) shows a voiced frame of female speech in the time-domain whilst Figure 6.1(b) shows the autocorrelation analysis of the same frame. In both cases speech was sampled at a rate of 8kHz. Clean speech and speech affected by babble noise at 0dB SNR are shown. Examining first the clean time-domain waveform in Figure 6.1(a) shows the peak-to-peak period to be $\approx$ 42 samples. This is visible on the autocorrelation analysis plot in Figure 6.1(b) as a peak at $\tau = 42$. This relates to $\hat{f0} = \frac{8000}{42} = 190.5Hz$. Looking now at the time-domain plot of the noisy signal the waveform is shown to have been considerably distorted. The autocorrelation analysis of the noisy signal shows no peak at the correct position with candidates at $\tau = 11$ and $\tau = 78$, corresponding to $f0$ values of 727.3Hz and 102.6Hz respectively. PRAAT uses a normalised autocorrelation function, compensated for the windowing function, in an attempt to

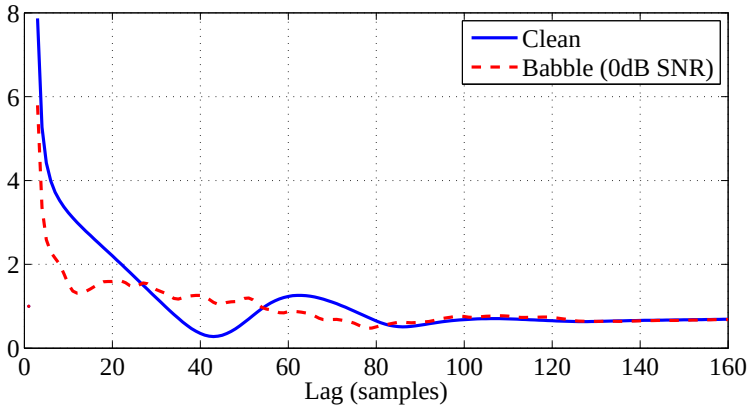


Figure 6.2: Illustration of applying stage 3 of the YIN fundamental frequency estimation algorithm to i.) clean speech and ii.) speech corrupted by babble noise at 0dB SNR

improve the robustness of the method [Boersma, 1993].

6.2.1.2 Average magnitude difference function (AMDF)

YIN is an example of AMDF $f0$ estimation and was proposed by De Cheveigné and Kawahara [2002] as an improvement over autocorrelation methods. Instead of using straight-forward autocorrelation analysis to measure $f0$ the algorithm is based on the average magnitude difference function (AMDF) first proposed by Ross et al. [1974]. A number of post-processing stages are introduced to improve robustness. $f0$ is measured from the AMDF as the lowest value at non-zero lag. Strong resonances in the first formant may cause the formant location to be detected instead of the fundamental and so YIN introduces a cumulative mean normalised difference function to reduce errors. This new function operates by normalising the current lag value, τ, by the average value up until that point. The result of this CMN function is displayed in Figure 6.2 for the same frame of speech as shown in Figure 6.1(b). The smallest value now corresponds to the $f0$ lag. In clean conditions the result is shown to match the autocorrelation method. In the presence of noise this method is also shown to perform poorly with the correct $f0$ overlooked in favour of a larger lag (and so lower $f0$).

In a further attempt at reducing errors YIN also introduces an absolute threshold on dip-detection to avoid erroneously selecting unrelated dips as well as parabolic interpolation to improve $f0$ detection when the fundamental is not an integer multiple of the sampling frequency. Finally, large errors in $f0$ are reduced by limiting the search range of the current frame based on previous results. This prevents values with a difference of $> 25\%$ from the previous frame being measured.

6.2.1.3 Hybrid methods

The final method of $f0$ estimation to be considered is the algorithm used by the ETSI Extended Advanced Front End (XAFE) DSR system for speech reconstruction [Sorin and Ramabadran, 2003]. The ETSI XAFE method combines estimates made in the time and frequency domains to form its final estimate. In the frequency-domain, thresholding is used to generate candidate frequencies by selecting a maximum of twenty peaks which exceed a predefined threshold. The search range is limited to 52-420Hz to cover the range of expected values of $f0$. The frequency resolution is doubled using Dirichlet interpolation to improve accuracy. These candidate peaks are reduced to two based on further processing, including measurement of the difference between candidates and previous values of $f0$.

In the time-domain the speech is low-pass filtered with a cut-off frequency of 800Hz and then downsampled from 8kHz to 2kHz. The ACF is then taken. The two candidate values found by frequency-domain analysis are then compared to the autocorrelation function and the most likely result taken [Medan et al., 1991]. If neither candidate correlates sufficiently well with the ACF, the frame is classified as unvoiced based on the result of a number of other tests.

6.2.2 Model-based $f0$ estimation

A model-based approach to estimation was first proposed by Barnard et al. [1991] using time-domain samples as features with neural networks used to determine $f0$.

Later approaches applied Bayesian estimation techniques to the problem with Rodet and Doval [1992] taking a maximum-likelihood approach to estimation. Several later studies abandoned time-domain features in favour of MFCCs with En-Najjary et al. [2003] using maximum-likelihood estimation to estimate $f0$ from MFCCs extracted from clean speech whilst Shao and Milner [2004] used maximum *a-posterior* estimation for the same task. Later, Milner and Darch [2011] applied the same maximum *a-posteriori* approach to the estimation of $f0$ from MFCCs extracted from noisy speech. Milner and Darch [2011] demonstrated this approach to perform best when the model domain is matched to the feature domain, i.e. models are trained on data recorded in the same conditions as the operating environment. Such data is not always available and so several methods of noise compensation were considered to improve performance. These included feature-based compensation including MVA processing [Chen et al., 2005] and spectral subtraction [Berouti et al., 1979] as well as model-based compensation using the approach proposed by Gales [1995]. The model-adaptation approach was shown to perform best with performance almost matching the best case.

6.3 Proposed Method of $f0$ Estimation

A system of using MAP estimation to estimate $f0$ from MFCCs extracted from noisy speech is proposed in this section. MAP estimation was shown to be effective at estimating clean spectral envelope from noisy MFCCs in Chapter 5 whilst correlation between MFCCs and $f0$ was shown to be high in Section 3.5.3.2. This method of estimation requires model of the joint density of the MFCCs and $f0$ is required *a-priori*. Methods of incorporating noise robustness into model parameters were discussed in Chapter 4 and include the Unscented Transform for Gaussian noises and serial model combination (SMC) for non-Gaussian noises. Alternatively clean-trained models may be used by enhancing features prior to estimation (Chapter 5). In terms of speaker variations, MAP adaptation was described in Chapter 4 and proven effective for spectral envelope estimation in Chapter 5.

Considering first the case of $f0$ estimation from clean speech, estimates are computed as:

$$\hat{f0} = \arg\max_{f0} f(\mathbf{x}|f0), \tag{6.2}$$

where $\mathbf{x}$ is a vector of MFCCs extracted from clean speech. Assuming the use of GMMs to model the joint density this can be rewritten as:

$$\hat{f0} = \sum_{k=1}^{K} h_k(\mathbf{x}) \arg\max_{f0} \left(f(f0|\mathbf{x}, \phi_k) \right), \tag{6.3}$$

where ϕ_k is the kth mixture component of the GMM and $h_k(\mathbf{x})$ is the posterior probability of $\mathbf{x}$ belonging to the kth mixture component, i.e.:

$$h_k(\mathbf{x}) = \frac{\alpha_k f(\mathbf{x}|\phi_k)}{\sum_{k=1}^{K} \alpha_k f(\mathbf{x}|\phi_k)}, \tag{6.4}$$

where α_k is the prior probability of ϕ_k. In practical terms the estimate is computed as:

$$\hat{f0} = \sum_{k=1}^{K} h_k(\mathbf{x}) \left(\boldsymbol{\mu}_k^{f0} - \boldsymbol{\Sigma}_k^{f0\mathbf{x}} (\boldsymbol{\Sigma}_k^{\mathbf{xx}})^{-1} (\mathbf{x} - \boldsymbol{\mu}_k^{\mathbf{x}}) \right). \tag{6.5}$$

Best performance is expected when model statistics are matched to the target speaker however sufficient data is not generally available. Speaker-independent models are relatively easy to obtain using one of the many corpora available (Appendix A) for training data, with the use of conventional $f0$ estimation approaches used to obtain $f0$ values for training. Speaker adaptation methods may then be used to adapt model parameters to the target speaker (Section 4.5.1).

When considering estimation from noisy speech best performance is expected when model parameters are matched to the target environment, i.e.:

$$\hat{f0} = \sum_{k=1}^{K} h_k(\mathbf{y}) \left(\boldsymbol{\mu}_k^{f0} - \boldsymbol{\Sigma}_k^{f0\mathbf{x}} (\boldsymbol{\Sigma}_k^{\mathbf{yy}})^{-1} (\mathbf{y} - \boldsymbol{\mu}_k^{\mathbf{y}}) \right), \tag{6.6}$$

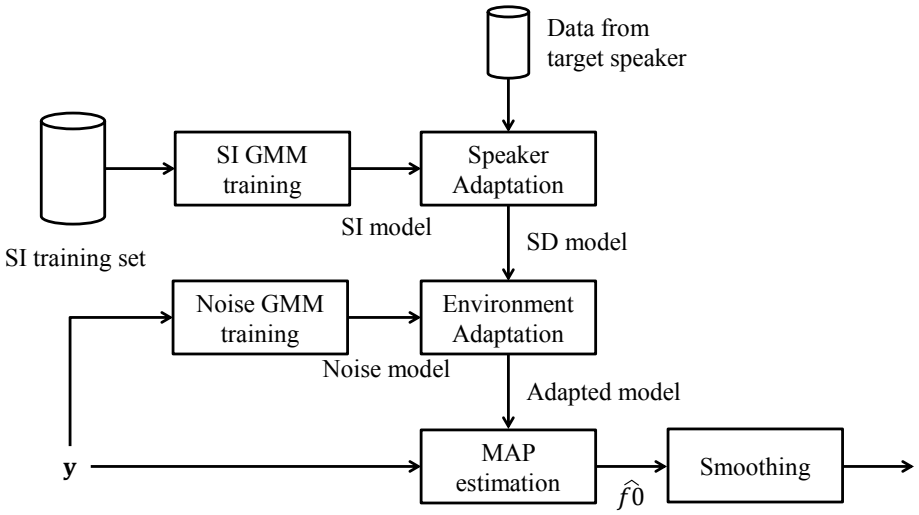


Figure 6.3: Flowchart of proposed system of $f0$ estimation using model-adaptation to compensate for speaker and noise

where $\mathbf{y}$ are MFCCs extracted from noisy speech. Again, this data is unlikely to be available for each specific environment at runtime. There are two approaches to solving this problem. First, model parameters may be adapted to more closely match the feature domain. Second, features may be compensated to more closely match the clean-trained model domain.

Models may be adapted in a number of ways as demonstrated in Section 4.5. For the purposes of $f0$ estimation Milner and Darch [2011] applied the log-normal adaptation approach by Gales [1995] for noise adaptation, however this method was found by Hu and Huo [2006] to perform poorly compared to the Unscented Transform (UT). In terms of feature compensation, a method of feature enhancement for the purpose of spectral envelope estimation using MFCCs was described in Chapter 5. These features may also be used for $f0$ estimation by using the compensated MFCCs as input to the system using clean-trained models. In the case of using compensated MFCCs speaker-independent models are used and so models must still be compensated for speaker. In this work both model-adaptation and feature compensation methods will be considered. Figure 6.3 shows the proposed model-compensation approach whilst Figure 6.4 shows the feature-compensation based approach. In each case no temporal information is incorporated into the

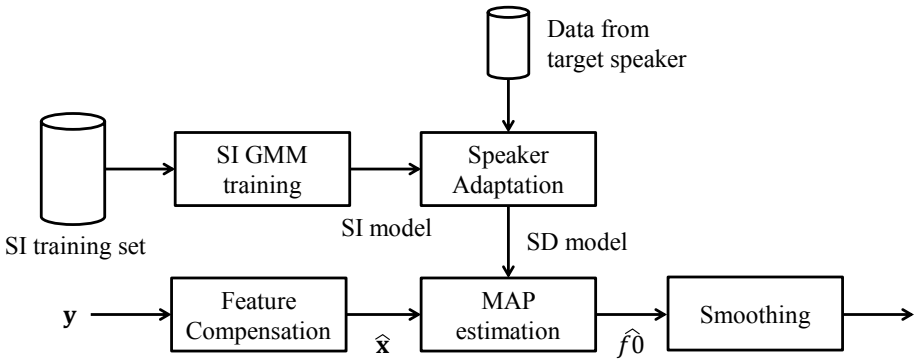


Figure 6.4: Flowchart of proposed system of $f0$ estimation using features compensated for speaker and noise using the global enhancement system described in Chapter 5

estimation process. Median filtering is used to reduce the effect of discontinuities whilst moving average-based filtering is used to smooth the resulting $f0$ contour. This is included in Figures 6.3 and 6.4 as a 'smoothing' process.

When estimating the spectral envelope a system of localised modelling was proposed whereby a series of models representing acoustic classes (i.e. phoneme, articulation class) were used for estimation (Section 5.3). In terms of $f0$ estimation a similar method was shown by Milner and Darch [2011] to perform worse than the global modelling system whilst similar results were observed in informal testing. These results are not too surprising as $f0$ does not correlate highly with phoneme classes. A system of $f0$ estimation using localised modelling is therefore not considered for this work.

This work therefore differs from the work carried out by Milner and Darch [2011] in two ways. First, a method of speaker adaptation (MAP adaptation, Section 4.5.1) is incorporated into the model adaptation process. Second, in the case of model-based noise compensation, the UT is used instead of the log-normal approach whilst a new method of feature compensation is also considered.

6.4 Results

This section presents results of $f0$ estimation in both clean and noisy conditions. Estimation from clean speech is first attempted to determine best-case performance of the proposed system. Next, the system is tested across a range of speaker configurations and noises. Three configurations are tested in terms of speaker variability, namely: speaker-dependent, gender-dependent and speaker-independent. Speaker-dependent data is taken from the NuanceCatherine dataset whilst gender-dependent and speaker-independent data is taken from the WSJCAM0 corpus. For model training the PRAAT $f0$ estimation tool is used.

Performance is measured using two metrics. First, the percentage $f0$ error, E_{f0} is used to measure the difference between reference and estimated values and is computed as:

$$E_{f0} = \frac{1}{N}\sum_{m}^{N}\left(\frac{\left|\hat{f0}(m) - f0(m)\right|}{f0(m)}\right) \times 100\% \qquad \forall f0(m) > 0, \tag{6.7}$$

where N is the total number of frames whilst $\hat{f0}(m)$ is the mth frame of the estimated value and $f0(m)$ is the reference value. The proportion of fine errors, defined as the proportion of frames with $E_{f0} \leq 20\%$, is also of interest. These are important as large errors in $f0$ may cause audible artifacts when used for reconstruction which may affect the perceived quality of reconstructed speech.

Estimation model parameters are optimised in Section 6.4.1. These include the feature dimensionality and number of mixture components which comprise the GMM. $f0$ is then estimated from clean speech in Section 6.4.2 before finally $f0$ estimation from noisy speech is considered in Section 6.4.3.

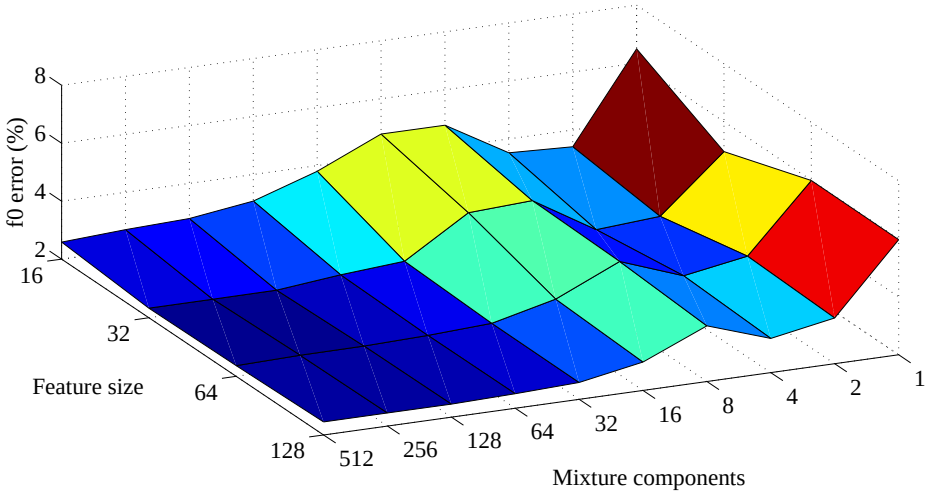


Figure 6.5: Result of varying feature size and number of GMM mixture components for the purpose of $f0$ estimation from MFCCs using MAP estimation in 10dB SNR white noise

6.4.1 Parameter optimisation

This section examines optimising the parameters required for $f0$ estimation using the proposed method of MAP estimation. These are: MFCC feature vector size and number of GMM mixture components. MFCC feature vectors are not truncated and so the feature vector size relates both to the number of filterbank channels and also the number of DCT coefficients. The results of an experiment testing both parameters simultaneously are presented in Figure 6.5. MAP estimation was used to predict $f0$ values from speaker-dependent data taken from the NuanceCatherine dataset. Speech was corrupted by white noise at 10dB SNR. MFCC feature vectors were used, with the feature size varied between 32 and 128. A GMM was used to model the joint distribution of noisy features and $f0$. The number of mixture components which comprise the GMM were varied between 1 and 512. Increasing the feature size is seen to reduce $f0$ error. This is expected as increasing the number of Mel-filterbank channels allows more fine $f0$ and harmonic detail to be represented. Likewise, increasing the number of mixture components also reduces $f0$ error until $M \approx 128$. Results when feature size ≥ 32 and $M \geq 128$ are similar with $M = 256$ modelling MFCCs with 32 filterbank channels achieving best results. This agrees

Table 6.1: Comparison of $f0$ estimation error using proposed system (MAP) and two existing methods of estimation using clean speech

Algorithm	Adaptation	$f0$ error (%)	Fine error %
MAP	–	4.24	96.57
MAP	Speaker	3.81	96.73
ETSI XAFE	–	8.43	91.82
YIN	–	3.55	97.09

with the optimal parameters found for spectral envelope estimation and so these parameters will also be used for $f0$ estimation.

6.4.2 Estimation from clean speech

This section examines the result of estimating $f0$ from MFCCs extracted from clean speech. MFCC features with 32 filterbank channels and 32 DCT coefficients were extracted at 100fps from speech with an 8kHz sample rate based on results in Section 6.4.1. Two MAP-based systems were tested. Speaker-independent data from the WSJCAM0 corpus was used with data taken from 40 speakers with an equal split in terms of genders. 24 hours of data in total was used to train the model. Testing was performed on a further 30 minutes of data taken from a different set of 10 speakers with an even split in terms of gender. First, a standard configuration of the system was tested in which no adaptation takes place. Next, in an attempt at reducing the effect of speaker variability a system using the same model but this time adapted for each speaker was tested. Adaptation was performed using MAP adaptation with 120 seconds of data used from each new speaker. MAP adaptation was applied to adapt model parameters relating to both MFCCs and $f0$. $f0$ values in the adaptation data were determined using the PRAAT pitch estimation tool. Table 6.1 presents results of testing both systems with the ETSI XAFE and YIN pitch detection algorithms included for comparison purposes. Testing was performed using reference voicing classifications. PRAAT was not included in these tests as it was used to determine the initial $f0$ values used for model training. YIN is shown to offer best performance with the lowest $f0$ error and highest proportion of fine errors

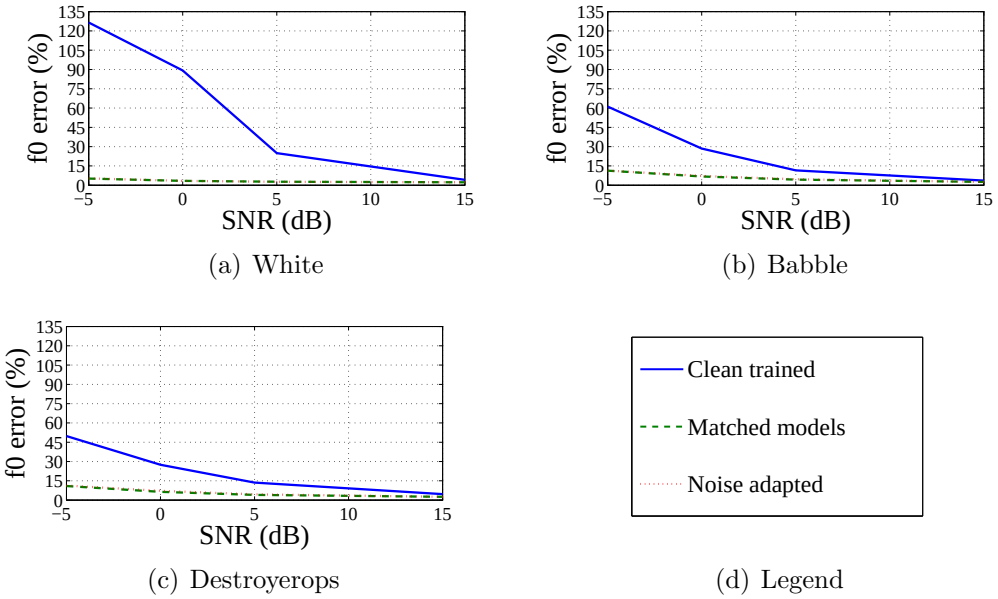
vs. gross errors. The speaker-adapted MAP system is shown to perform surprisingly similarly to YIN. This is promising as it confirms that a significant amount of source information is retained in the feature vector. Performance with non-adapted models is slightly worse than adapted models whilst the ETSI XAFE tool is shown to offer worst performance.

6.4.3 Estimation from noisy speech

The estimation of $f0$ from MFCCs extracted from noisy speech is now considered. This section is split into three parts. First, a speaker-dependent system is considered to determine optimal performance in Section 6.4.3.1. Next, the system is expanded to use separate models for male and female speech in Section 6.4.3.2. The result of using MAP adaptation to reduce the effect of the additional speaker variability is tested in the case of gender-dependent models. Finally, the speaker-independent case is considered to determine overall system performance in Section 6.4.3.3. Two methods of noise compensation are tested. First, a method of model-adaptation is used whereby GMMs are trained using the parameters determined in Section 6.4.1 and adapted for noise using the Unscented Transform (Section 4.5.2.3) for Gaussian noises whilst serial model combination is used in the case of non-Gaussian noises (Section 4.5.2.4). Second, in the case of speaker-independent estimation a system of feature compensation is also evaluated. MFCC features enhanced for the purpose of spectral envelope estimation are used as input to the $f0$ estimation system using clean-trained models. Enhanced features are obtained from a speaker-independent system using the Unscented Transform for noise adaptation and MAP adaptation for speaker adaptation.

6.4.3.1 Speaker dependent

The result of using the MAP estimation approach with speaker-dependent models is examined in this section. Two models were trained. First, models were trained on data with conditions matched to the test environment (matched models). Next, a



(a) White (b) Babble

(c) Destroyerops (d) Legend

Figure 6.6: Fundamental frequency estimation error (%) using speaker-dependent models trained on female speech in a.) white noise, b.) babble noise and c.) destroyerops noise

model was trained on clean speech for later adaptation. In both cases models were trained on ≈ 40 minutes of data and tested on a further ≈ 12 minutes of previously unseen data. Figure 6.6 compares performance of using clean-trained (unadapted) models, matched models and clean-trained models adapted for the noise using the UT in terms of percentage $f0$ error. The use of clean trained models is shown to offer very poor performance with significant errors reported across all noise types when SNR < 15dB. The use of models matched to the test environment are shown to offer best performance with adapted models offering near-identical performance.

In the case of noise-adapted models the reference noise distribution was used. This information is clearly unlikely to be available in realistic scenarios and so Figure 6.7 simulates the use of a VAD-based noise estimation algorithm using Monte-Carlo sampling. $f0$ was estimated using clean trained models and the UT was again used for model adaptation. Noise statistics required for adaptation were obtained from random samples of the noise signal which was assumed to be known *a-priori*. Samples were taken in 50ms bursts with the total amount of data used to obtain

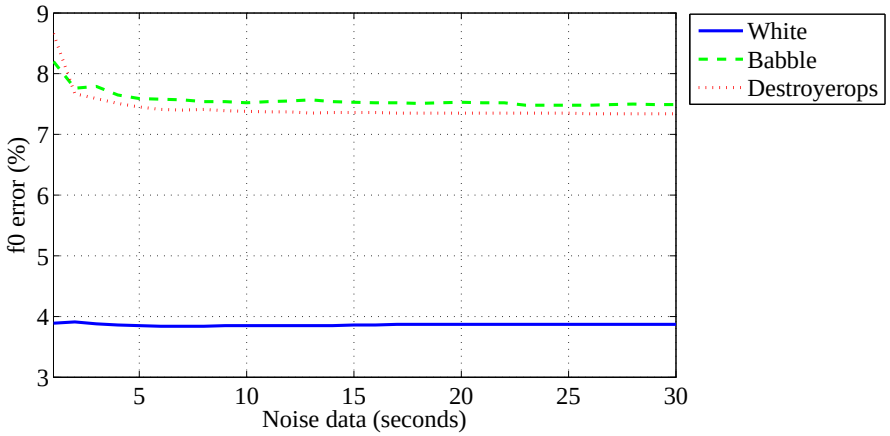


Figure 6.7: Effect of varying the amount of noise data used to adapt speaker-dependent models originally trained on clean speech in terms of f0 error (%) across various noises at 0dB SNR

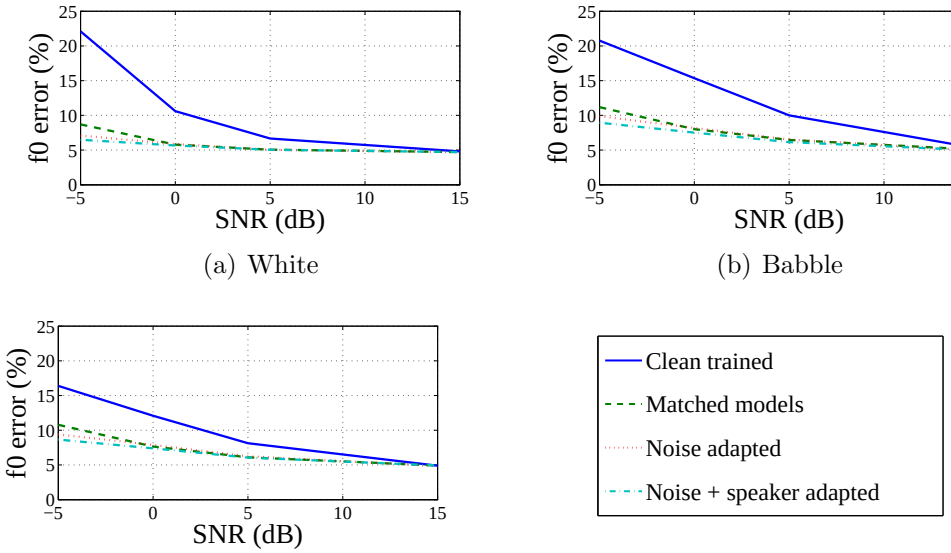
noise statistics varied between 1 and 30 seconds. In white noise no significant difference is noted when the models are adapted using as little as one second of noise data. This amount of noise data is easily obtainable in real-world scenarios. There will typically be a period of ≈ 0.5 seconds at the start and end of utterances which may be used for noise estimation purposes without the need to employ any additional noise estimation strategy. Significant differences are noted for both non-stationary noises, babble and destroyerops, when < 6 seconds of noise data is used to train models. Performance does not increase when more than 6 seconds of noise signal is available. A large difference between the performance of estimation in stationary and non-stationary noises is apparent. This is attributed to the effect of noise on the MFCC feature vectors. In the case of white noise only the first few MFCCs will be affected, relating to energy and spectral slope. Non-stationary noises consist of a number of components which vary in both frequency and amplitude and therefore have a more wide ranging effect on the MFCCs.

6.4.3.2 Gender dependent

Next, the case of using separate models for male and female speech is considered. So far all experiments have assumed that sufficient data and processing resources

are available to build speaker-dependent models. This is rarely the case and so models must be built that perform well across a range of previously unseen speakers. Models trained on speakers of the same gender as the target speaker are therefore trained using 12 hours of data from 20 different speakers. Five speakers of each gender were then selected independently of the speakers used for training. By using models trained on the same gender the mismatch in feature and model statistics is limited, however not negated entirely. This mismatch is known to cause a decrease in performance in applications such as speech recognition and has also been found to reduce performance of spectral envelope estimation (Section 5.4.1). In particular, values of $f0$ vary significantly between genders with female speakers having values of $f0$ roughly double that of male speakers. MAP adaptation is therefore used in an attempt to reduce the effect of such variability. This requires additional speaker data and 120 seconds of clean speech data per speaker was used for adaptation where appropriate. MAP adaptation is used to adapt both MFCC and $f0$ model parameters. As described in Section 4.5.1 clean speech data from the target speaker is used to adapt model parameters. $f0$ is estimated from this data using PRAAT to obtain values from which model parameters are adapted.

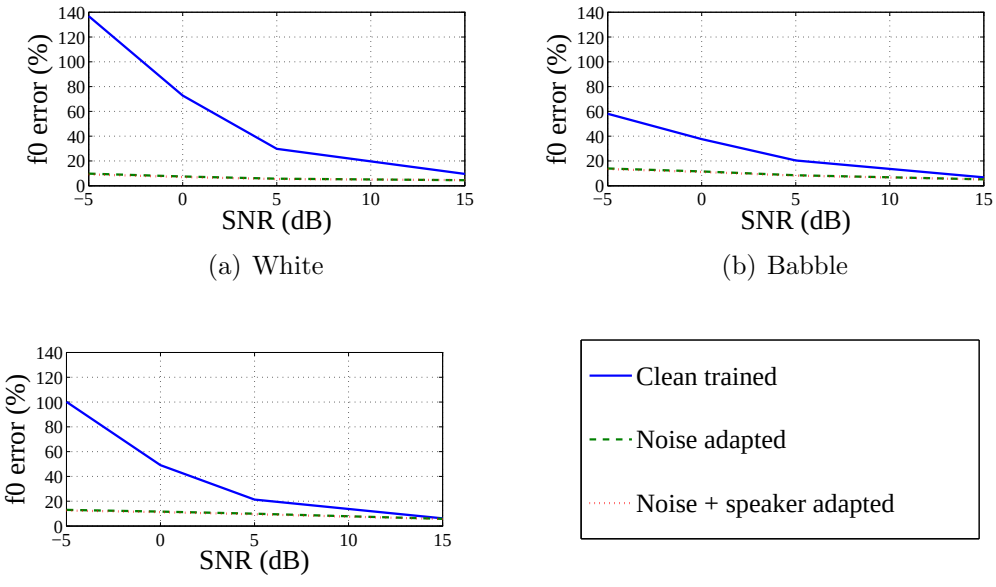
Four systems were tested. First, clean-trained models with no adaptation are tested to determine the worst-case performance. Next, matched models are tested. Finally, two adapted systems are included. Firstly, the noise adapted system as described previously is tested whilst a system adapted for variations in both speaker and noise is also included. Figure 6.8 presents results of female-only testing whilst Figure 6.9 shows male-only results. Female-only results are first considered. As with speaker-dependent testing clean-trained models offer worst performance. This is expected as considerable mismatch between feature and model statistics exists due to the effect of noise in the feature domain. Performance across the three remaining configurations is almost identical at positive SNRs. At -5dB the system adapted for both speaker and noise is shown to perform best. This is also unsurprising as it is the only system to account for variations in speaker and environment. Performance remains relatively stable across SNR in white noise with errors increasing by



(a) White (b) Babble

(c) Destroyerops (d) Legend

Figure 6.8: Fundamental frequency estimation error (%) using gender-dependent models trained on female speech in a.) white noise, b.) babble noise and c.) destroyerops noise



(a) White (b) Babble

(c) Destroyerops (d) Legend

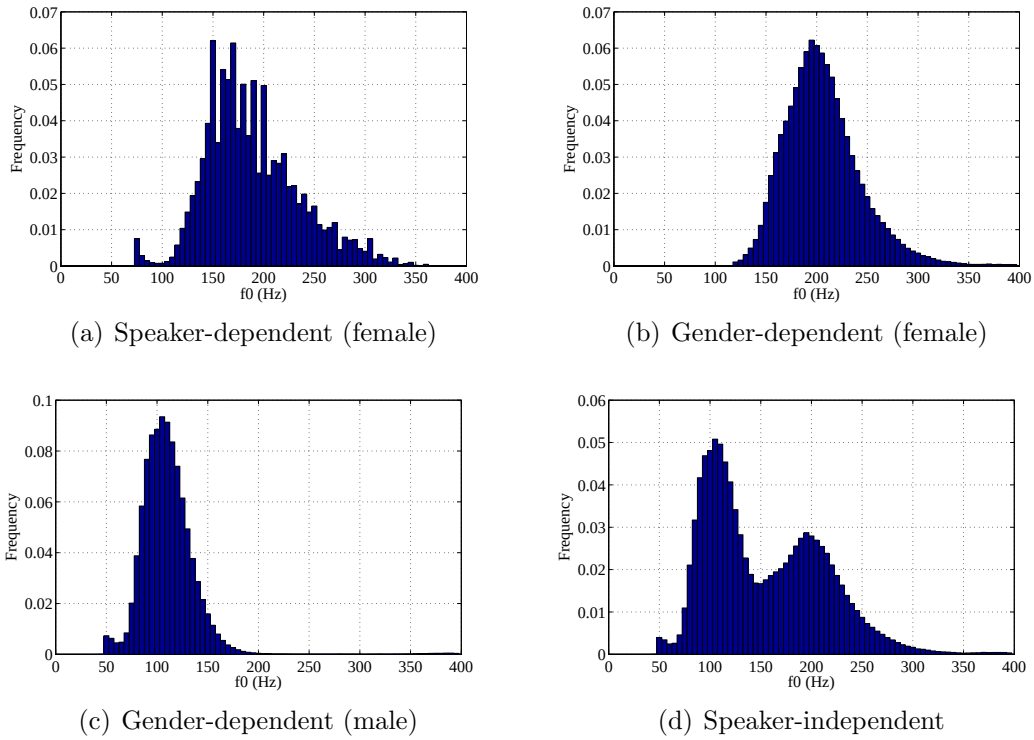
Figure 6.9: Fundamental frequency estimation error (%) using gender-dependent models trained on male speech in a.) white noise, b.) babble noise and c.) destroyerops noise

only 2% in absolute terms between 15 and -5dB SNR. Performance degrades more significantly in non-stationary noises with errors almost doubling between 15 and -5dB SNR. Despite this, errors are still shown to stay below 10% at -5dB SNR. Performance is shown to be similar to the case of speaker-dependent testing.

In the case of male-only testing only three systems were tested. Adapted models have been shown to offer performance at least equal to matched models and so only clean-trained and adapted models are now considered. General trends are shown to be similar to female-only testing. Clean-trained models are shown to offer worst performance with $f0$ error increasing rapidly as the level of noise is increased. The system adapted for both speaker and noise is shown to perform best but speaker adaptation is shown to have relatively little effect on the male-only system. This is attributed to the variance of male values of $f0$. Female values of $f0$ were measured to have a larger variance between speakers and so the effect of adapting the model parameters is more apparent. The distribution of $f0$ values are illustrated later in this chapter in Figure 6.10. Percentage $f0$ error is higher for the male system when compared like-for-like with the female system results. This is due to significant differences in mean fundamental frequency across genders. The mean value of $f0$ across all speakers was measured at 114Hz for male speech and 208Hz for female speech. A 10% error in $f0$ for female speech therefore relates to an absolute error of $\approx 21Hz$ whilst a similar absolute error in male speech would result in a percentage error of 18%. The absolute $f0$ error across speakers is therefore seen to be independent of gender to a large extent.

6.4.3.3 Speaker independent

Gender-dependent testing in Section 6.4.3.2 showed that performance was relatively unaffected by using gender-dependent models over speaker-dependent models, even when no speaker adaptation was used. In this section the use of a completely speaker-independent model is therefore considered. Data was pooled from the training data used for gender-dependent model training to give a model trained using a



(a) Speaker-dependent (female)

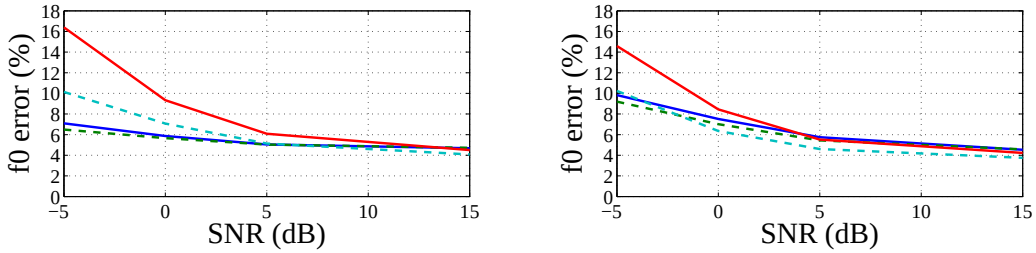
(b) Gender-dependent (female)

(c) Gender-dependent (male)

(d) Speaker-independent

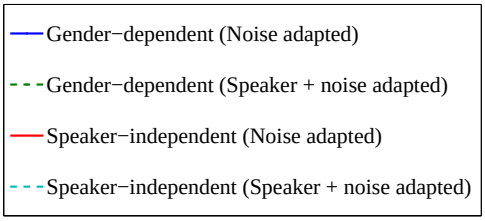
Figure 6.10: Distributions of reference $f0$ values used to train a.) speaker-dependent female model, b.) gender-dependent female model, c.) gender-dependent male model and d.) speaker-independent model

total of 24 hours of training data from 40 different speakers. Mean $f0$ values are significantly different between male and female speech and this poses an additional problem when modelling distributions. Figure 6.10 compares the distribution of $f0$ values for speaker-dependent (female), gender-dependent and speaker-independent data. Speaker-dependent values of $f0$ from a female speaker are shown follow an approximately log-normal distribution whilst this distribution becomes approximately normal in the gender-dependent case for both female and male-specific speech. The speaker-dependent histogram in Figure 6.10(a) is inconsistent owing to insufficient data. Comparing the male and female distributions shows female speech to have larger variance whilst male values are more tightly distributed around the mean. A small number of halving errors are visible in the case of speaker-dependent and male-only distributions and appear as peaks in the histogram at 50-75Hz. Combining these distributions gives the multi-modal distribution shown in Figure 6.10(d).



(a) Female

(b) Male

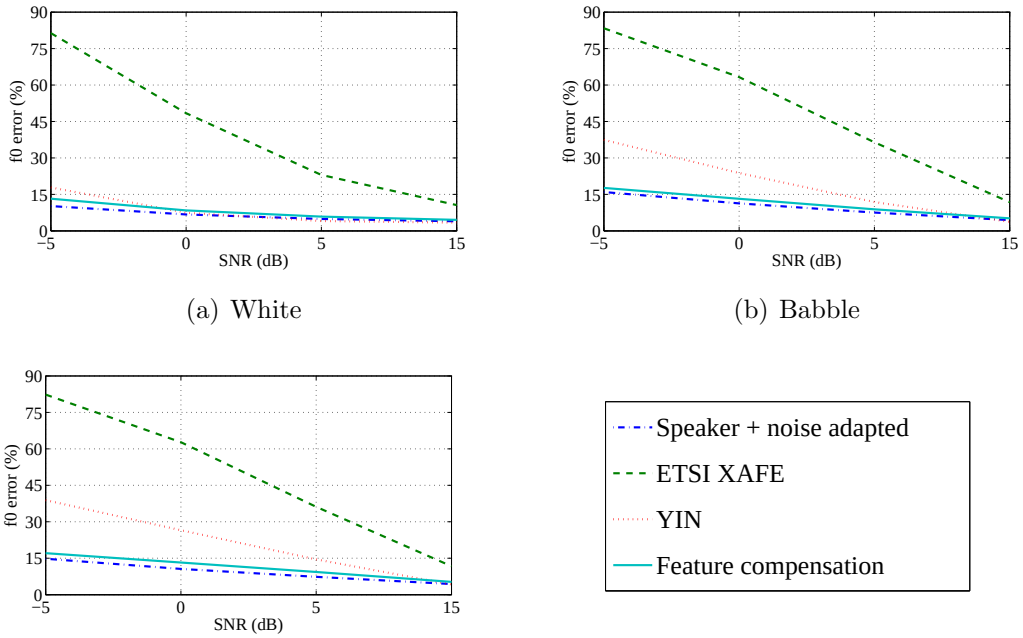


(c) Legend

Figure 6.11: Comparison of performance of gender-dependent system versus speaker-independent system in terms of fundamental frequency error (%) in white noise.

In the case of speaker-dependent and gender-dependent estimation one key benefit was the reduced frequency of gross errors compared to conventional methods of estimation. This was due to the explicit knowledge of the distribution of $f0$ values which prevented erroneously high or low values being predicted. By expanding the training data to cover male and female speech this advantage is reduced. For example, the probability of female speech taking a value of 100Hz was almost zero in the gender-dependent case, but in the speaker-independent case this probability is increased significantly. The proportion of gross errors is therefore expected to increase in the case of speaker-independent estimation, with speaker adaptation expected to be more effective than in the gender-dependent case.

Figure 6.11 compares the result of estimating $f0$ from speech affected by white noise using speaker-independent models to those using gender-dependent models. As expected, the speaker-independent system adapted for noise performed worst for both male and female $f0$ estimation. Including speaker adaptation improved results significantly in both cases with male $f0$ estimation results improving to offer best



(a) White (b) Babble

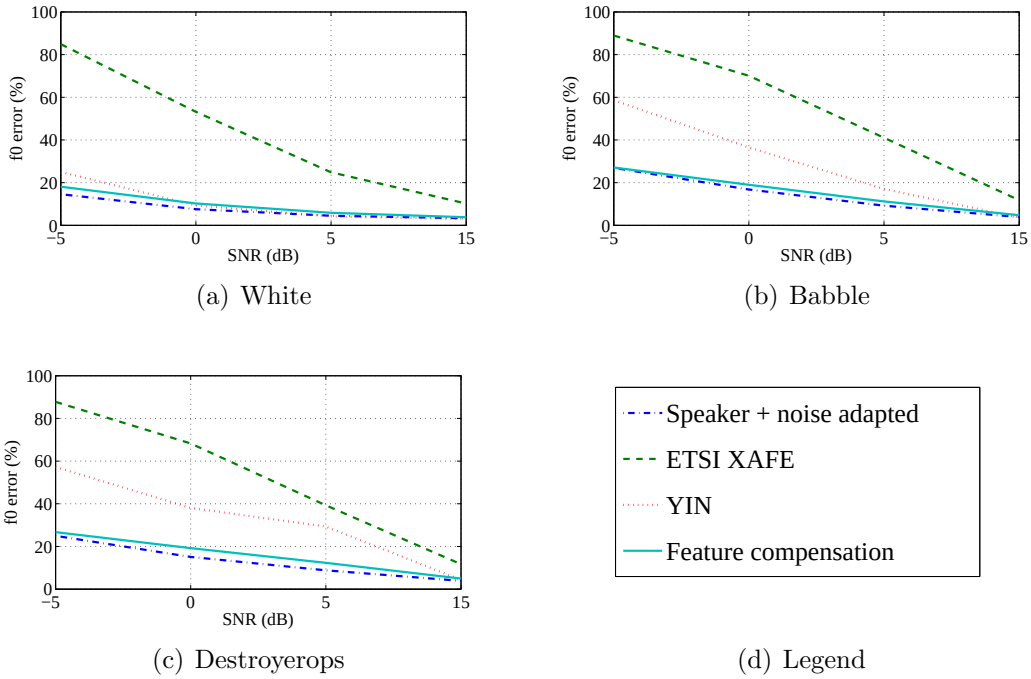
(c) Destroyerops (d) Legend

Figure 6.12: Fundamental frequency estimation error (%) using speaker-independent models in a.) white noise, b.) babble noise and c.) destroyerops noise

performance at positive SNR and equivalent performance to the gender-dependent system at -5dB SNR. Speaker adaptation did not improve performance as much in the case of estimating $f0$ from female speech, with gender-dependent systems still offering best performance.

Next, performance of the MAP-based estimation system is compared to conventional methods in Figure 6.12. Two MAP-based systems are compared to two conventional approaches. In the case of MAP-based systems the system using speaker and noise adaptation is tested. In addition, a system using speaker-independent models with features enhanced using the estimation method described in Section 5.2 is also tested. In this case models are adapted for speaker only. The two conventional methods tested are the ETSI XAFE [Sorin and Ramabadran, 2003] and YIN estimation methods [De Cheveigné and Kawahara, 2002].

The ETSI XAFE $f0$ estimation is the least effective method of estimation with performance similar to that of MAP estimation using unadapted clean-trained mod-



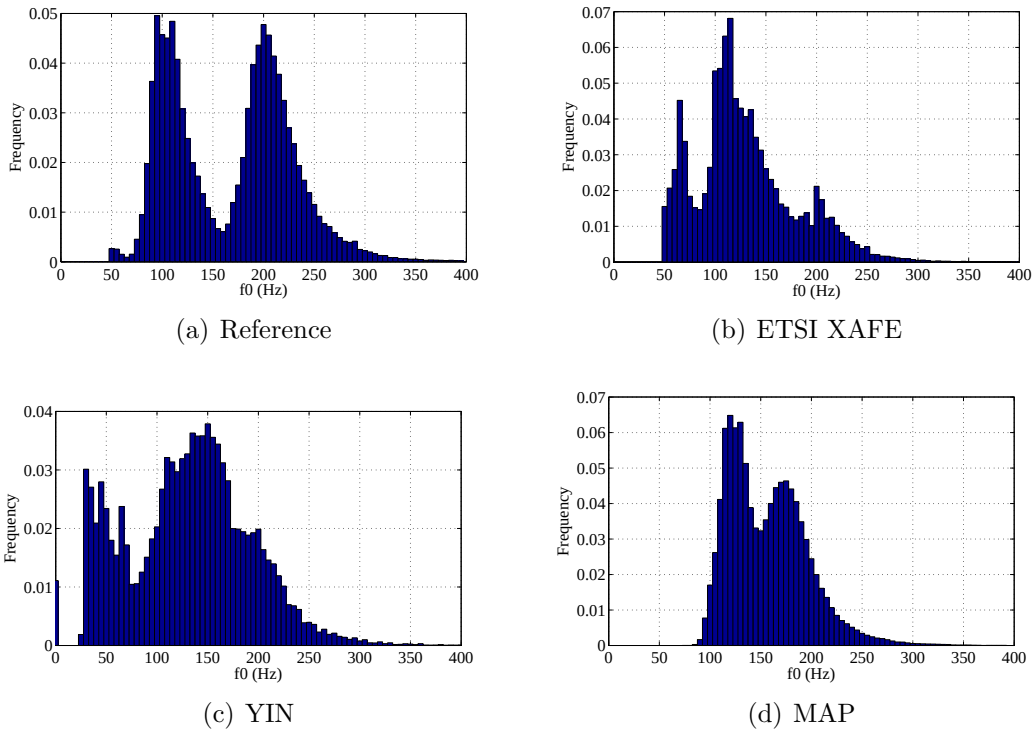
(a) White (b) Babble

(c) Destroyerops (d) Legend

Figure 6.13: Gross fundamental frequency estimation error (%) using speaker-independent models in a.) white noise, b.) babble noise and c.) destroyerops noise

els. The two MAP-based systems considered in these results offer best performance in all cases with the model-adapted system performing best overall. The use of compensated features is almost as effective as adapting model parameters and so offers a way of reducing the overall complexity of the speech enhancement system as such features will already be available from spectral envelope estimation. In white noise YIN performs almost as well as the MAP-based methods at low SNR whilst at 15dB SNR YIN actually offers best performance across all noises. This is in line with results of estimating $f0$ from clean speech presented in Table 6.1 also showing YIN to perform well. This performance advantage disappears as the level of noise increases with error rates at -5dB SNR double those achieved by MAP-based methods in both babble and destroyerops noise.

So far results have focused on measuring percentage $f0$ error. The proportion of gross errors is also of interest and so Figure 6.13 now compares performance of $f0$ estimation methods in terms of gross error proportions. The trend of results



(a) Reference

(b) ETSI XAFE

(c) YIN

(d) MAP

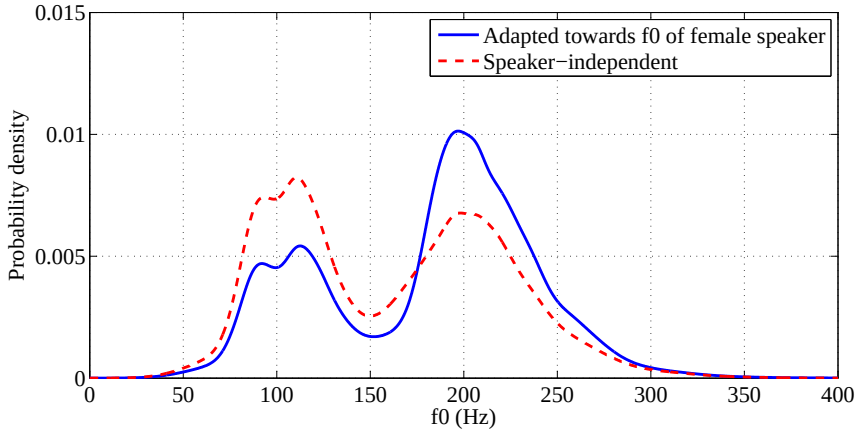
Figure 6.14: Comparison of distributions of reference and estimated $f0$ values from speech mixed with babble noise at 0dB SNR where: a.) reference values, b.) estimated using ETSI XAFE system, c.) estimated using YIN and d.) estimated using proposed system using speaker and noise adaptation

are shown to mirror percentage $f0$ error in most cases. The MAP-based system using model adaptation is again shown to perform best. At -5dB SNR although the method is still shown to perform best, more than 20% of voiced frames are shown to have an $f0$ error of greater than 20%.

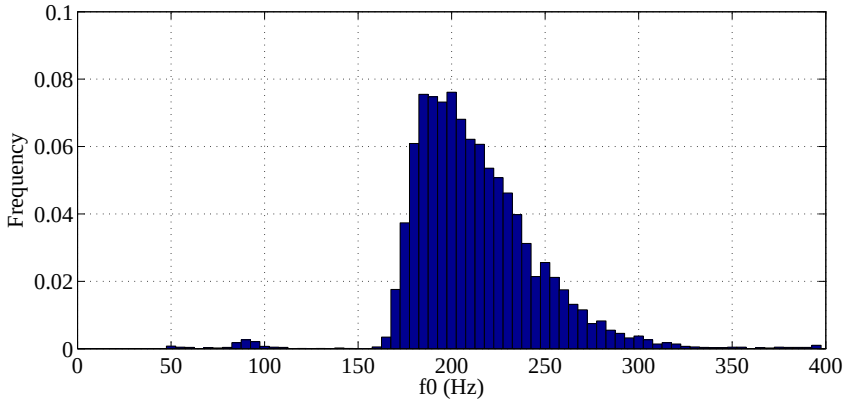
Figure 6.10 examined the distributions of the training data. To examine the effect of estimation, Figure 6.14 now compares the distributions of the test set in terms of reference and estimated values of $f0$. The distribution of reference $f0$ values from the test set are shown in Figure 6.14(a) whilst the distribution of those estimated from speech affected by babble noise at 0dB SNR using the model-adapted MAP estimation method is shown in Figure 6.14(d). The distribution of results using the two conventional methods are shown in Figures 6.14(b) and 6.14(c). Figure 6.14(a) shows a clear separation between male and female speech with male speech $f0$ values

centered around 108Hz and female values centred around 211Hz. In the case of the ETSI XAFE estimation method in Figure 6.14(b) a significant number of frames are shown to have been under-estimated with an additional peak between 50-90Hz suggesting this is due to halving errors. A similar effect is shown when using YIN for estimation. Figure 6.14(c) shows male speech to be significantly affected with a large proportion of male speech frames having halving or doubling errors. Neither of the conventional methods appear to be affected significantly by doubling errors in frames of female speech. Lastly, the distribution of values estimated by the MAP-based approach is considered in Figure 6.14(d). No halving or doubling errors are shown to have occurred, though the distribution is still shown to have been distorted. Whilst two clear peaks are visible, male $f0$ values are now centred around 123Hz whilst female values are centred around 175Hz. This is a significant shift from the 108 and 211Hz centres of the reference values giving errors of +15Hz and -36Hz for male and female speech respectively, with values appearing to have been pulled towards the global mean of the distribution (159Hz). This is attributed to the effect of speaker adaptation. Figure 6.15 illustrates the effect of speaker adaptation on the distribution of $f0$ values in the estimation model when adapting for a female speaker. Figure 6.15(a) shows the effect of speaker adaptation on the probability density function (pdf) of $f0$ values in the estimation model whilst Figure 6.15(b) shows the distribution of $f0$ values of the target speaker. Whilst the emphasis on low values of $f0$ is reduced and the probability of higher values of $f0$ is increased there is still a significant non-zero probability of $f0$ values occurring outside of the range of values illustrated in Figure 6.15(b). $f0$ is estimated as a weighted average across all mixture components and so the effect of these 'out of range' values will be to reduce the estimated value of $f0$. In the case of male speech the opposite is observed estimated values of $f0$ increasing in value.

Finally, Figure 6.16 shows examples of $f0$ contours estimated using the two conventional methods and the best MAP-based approach for the utterance *"Look out of the window and see if it's raining"*. The estimated contour is compared to reference values. $f0$ values were estimated from speech corrupted by babble noise at

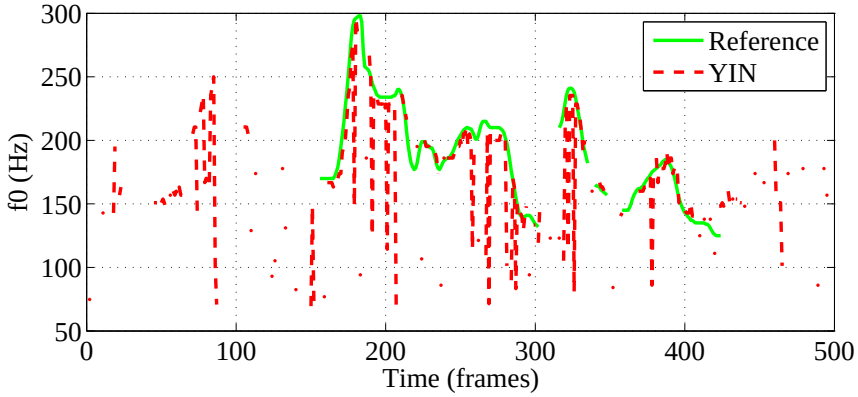


(a) pdfs of $f0$ values from estimation models

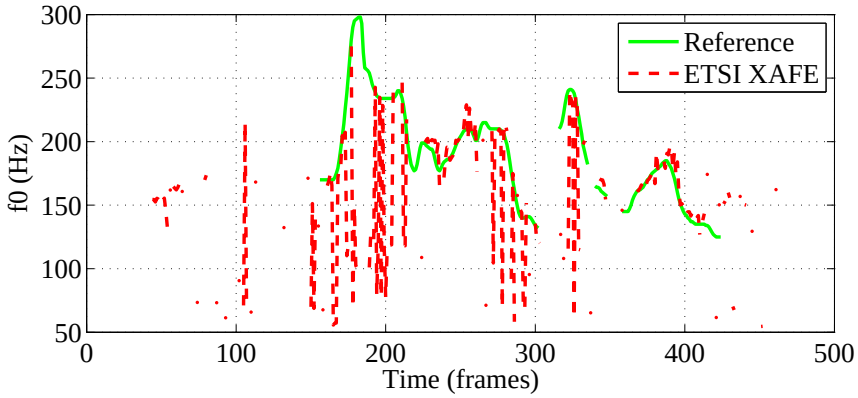


(b) Distribution of $f0$ values of target speaker (female)

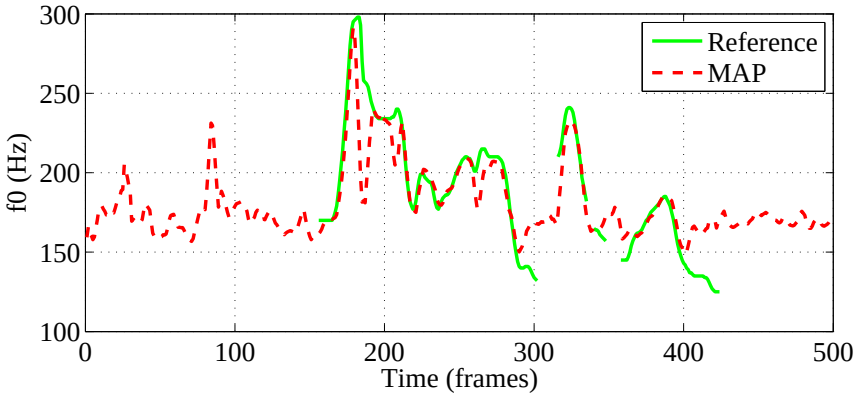
Figure 6.15: Effect of speaker adaptation on distribution of $f0$ modelled by the joint density model where a.) compares the unadapted distribution of $f0$ versus the adapted distribution and b.) shows the distribution of actual $f0$ values from the target speaker



(a) YIN



(b) ETSI XAFE



(c) MAP (noise adapted)

Figure 6.16: Comparison of $f0$ tracks estimated from a single utterance mixed with babble noise at an SNR of 0dB using a.) YIN, b.) ETSI XAFE and c.) proposed system using noise adaptation

0dB SNR. The contours estimated using the conventional methods both contain a significant number of halving errors. Surprisingly, no doubling errors occur in ei-

ther case. There are also a number of cases where $f0$ values have been estimated in periods of non-speech. Overall, neither method offers the accurate $f0$ contour required for clean speech reconstruction. Looking now at the estimate obtained using the MAP-based method and a significantly smoother contour is observed. $f0$ estimation is forced for all frames and so each frame has a value associated with it, even in non-speech and unvoiced regions, and so this method relies also on a voicing classification being made (Chapter 7). In regions of voiced speech activity $f0$ is shown to have been accurately estimated with very few errors. No halving or doubling errors are observed.

6.4.3.4 Non-Gaussian noise

Previous experiments have shown that a MAP-based approach to $f0$ estimation using MFCCs is effective in noisy environments. So far all noises have been modelled as Gaussian distributions. Not all noises can be modelled as such and so this section considers the use of the serial model combination (SMC) noise adaptation scheme described in Section 4.5.2.4 for $f0$ estimation. Machine gun noise is used as an example of non-Gaussian noise.

Machine gun noise is a particularly challenging noise when considering the task of accurate $f0$ estimation. Relatively low energy, low frequency noise is interspersed with very high energy bursts of wide-band impulsive noise. An example of machine gun noise is shown in Figure 6.18. Frames affected by such bursts are expected to offer no useful information relating to $f0$. Despite this, MAP estimation is still expected to perform relatively well for this task: even in frames where relatively little $f0$ information is available from the noisy speech the estimated value of $f0$ will not deviate significantly from the mean of the model distribution. The inclusion of median filtering and smoothing should also lessen the effect of such noise.

SMC was shown to be effective for spectal envelope estimation in Section 5.4.3 with noise models consisting of three mixture components offering best performance. Performance of such a system for $f0$ estimation is now measured. Speaker-

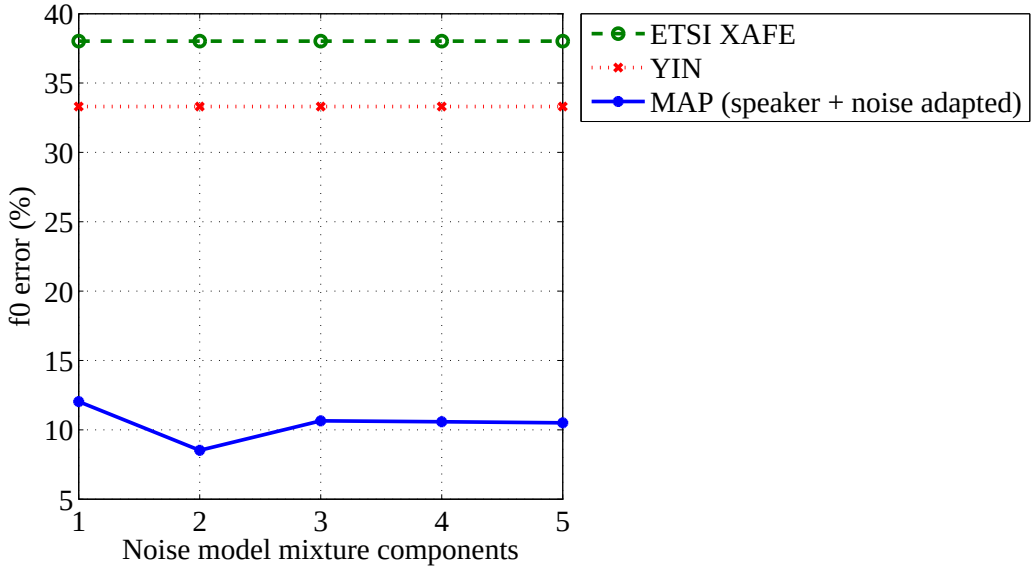
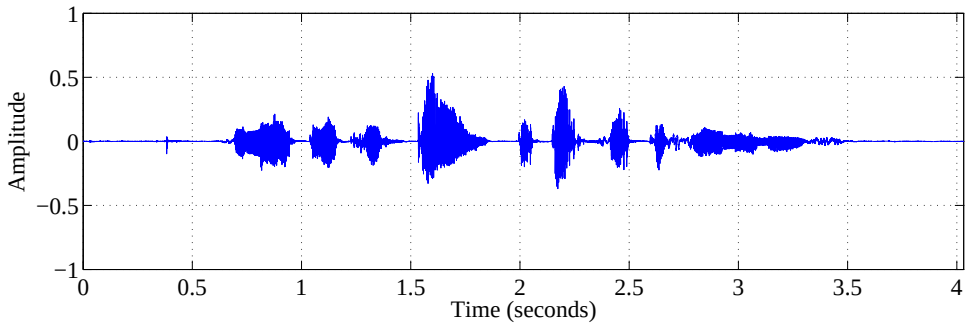


Figure 6.17: Effect of varying number of mixture components in noise model using SMC on $f0$ error in machine gun noise (-20dB SNR)

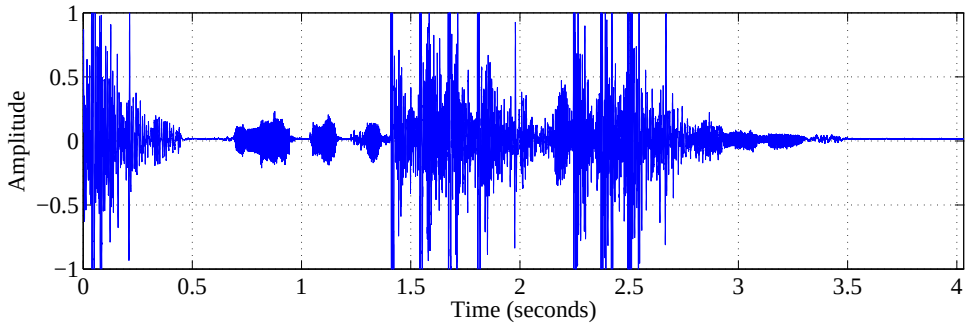
independent models were trained using the same training data as previous experiments, that is 24 hours of data from 40 speakers with an equal male/female split. The system was tested on a further 30 minutes of speech data spoken by a total of 5 male and 5 female speakers. Models were trained in clean conditions and then adapted for speaker using MAP adaptation. SMC was applied for noise adaptation using GMMs of the noise, which was assumed known *a-priori* for the purpose of these experiments. These results are presented in Figure 6.17 where results of $f0$ estimation using the ETSI XAFE and YIN algorithms have also been included for comparison purposes.

The MAP-based system clearly offers best performance with performance of both conventional methods degraded significantly by the noise. In terms of the number of mixture components used to model the noise, two mixture components is demonstrated to perform best.

Figure 6.18 shows the effect of machine gun noise in the time domain on the utterance *"He might be a tough guy but that's what this union is"* spoken by a female speaker. Figure 6.19 now shows the result of estimating $f0$ from the same

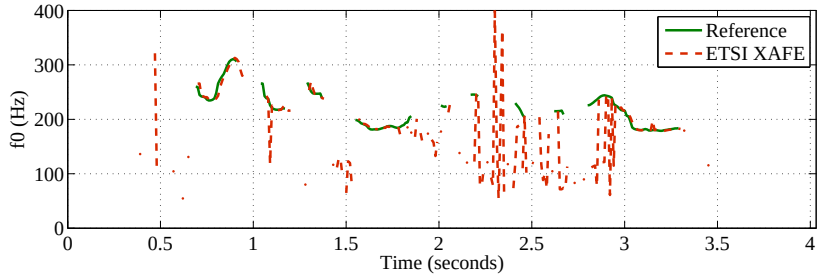


(a) Clean speech

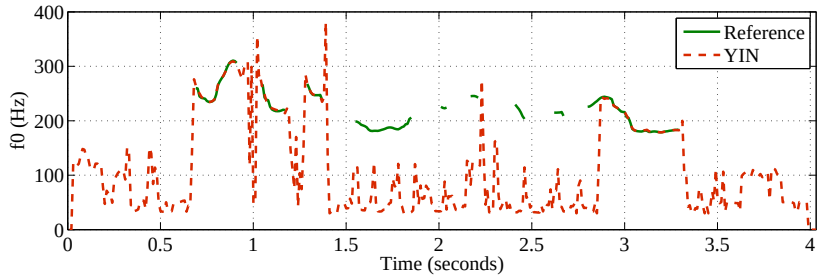


(b) Noisy speech (machine gun noise at -20dB SNR)

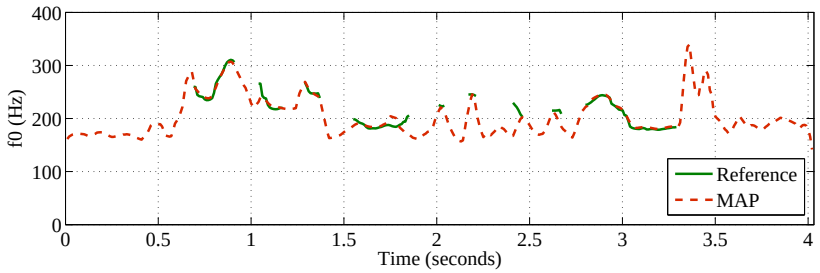
Figure 6.18: Time domain example of speech corrupted by machine gun noise at -20dB SNR



(a) ETSI XAFE



(b) YIN



(c) MAP (noise adapted)

Figure 6.19: Comparison of $f0$ tracks estimated from a single utterance mixed with machine gun noise at an SNR of -20dB using a.) YIN, b.) ETSI XAFE and c.) proposed system using noise adaptation

utterance. In each case the $f0$ contour estimated from noisy speech is compared to a hand corrected $f0$ contour estimated from clean speech using PRAAT. The systems compared include the conventional ETSI XAFE and YIN methods as well as the MAP-based method using speaker adaptation and SMC with noise modelled by a GMM with two mixture components. By comparing Figures 6.18 and 6.19 it is clear that the two conventional methods offer reasonably accurate estimates in periods unaffected by the machine gun noise. Despite offering worst overall performance, in the case of this utterance the ETSI XAFE algorithm is shown to perform best out

of the conventional methods. YIN is shown to be severely affected by the machine gun noise. Based on the results presented in Figure 6.17 the MAP-based method is expected to perform best. This is clearly the case in Figure 6.19 where a highly accurate estimate of $f0$ has been achieved.

6.5 Summary

A system of estimation using MAP and MFCCs was proposed in this chapter. The proposed system was similar to the one used for spectral envelope estimation. One of the key challenges in using such a system is how to obtain the required joint density models. The use of methods of speaker and noise adaptation to compensate clean-trained model parameters and the use of features compensated for noise were therefore also considered in the case of $f0$ estimation.

The proposed system was compared to two conventional methods of estimation: YIN and ETSI XAFE. The MAP estimation method using model-based compensation schemes was found to perform best. The use of speaker-independent models was found to degrade performance compared to speaker and gender-dependent models, with speaker adaptation essential to achieve good performance. The speaker-independent system adapted for speaker and noise was also shown to outperform both conventional methods when estimating $f0$ from speech affected by high levels of noise. The use of serial model combination (SMC) for noise adaptation was demonstrated to be effective in very high levels of machine gun noise when using the SMC adaptation method proposed in Section 4.5.2.4.

Chapter 7

Voicing Classification

In this chapter a method of data-driven voicing classification is described. Conventional methods of voicing classification typically use features such as zero-crossing rate and energy to classify frames of speech however these features are not robust to noise. In this work a range of machine learning methods are tested for the purpose of robust voicing classification. A broad range of classifiers are considered and include parametric, probabilistic and non-probabilistic, artificial neural networks and regression. A system using GMMs trained on speaker-independent clean speech and subsequently adapted for speaker and noise was found to perform best for this work and so will be used for this method of speech enhancement.

Contents

7.1 Introduction

This chapter describes a system of robust voicing classification required for this method of speech enhancement. Voicing classification is a three way classification problem and is closely linked to voice activity detection (VAD). The purpose of a voicing classifier is to classify regions of speech based on their voicing class, i.e. $c \in \{nonspeech, voiced, unvoiced\}$. VAD can be seen as a subset of this problem whereby utterances are classified as $c \in \{nonspeech, speech\}$ where $speech = \{voiced, unvoiced\}$. Both classifiers have numerous applications such as noise estimation, speech recognition and speech coding [Kim and Chang, 2000; Ramírez et al., 2004; Sangwan et al., 2002]. Many different methods have been proposed to solve these classification problems and they operate typically by measuring properties of signal such as zero crossing rate, spectral energy and spectral distortion [Benyassine et al., 1997]. These features are not robust to noise and so these conventional methods either adapt threshold values according to noise levels or require the signal to be cleaned using speech enhancement methods [Sorin and Ramabadran, 2003]. More recently, machine learning (ML) techniques have been applied to these classification problems whereby features are extracted from the audio and either Gaussian mixture models (GMMs) [Darch et al., 2006] or support vector machines (SVMs) [Enqing et al., 2002; Ramírez et al., 2006] applied to classify the audio as non-speech, voiced speech or unvoiced speech.

In this work we are interested in robust voicing classification and so this chapter begins with an investigation into the use of machine learning (ML) methods for the problems of voice activity detection (VAD) and voicing classification (VC) in Section 7.2. The review focuses primarily on VAD and VC in noisy environments to determine if there is any advantage in using ML methods for robust classification versus conventional methods. Next, based on this investigation, a system for robust VC for use in this work is proposed in Section 7.3. Two methods of achieving robustness to the noise are considered; systems using enhanced features are compared to a GMM-based system where variations in speaker and environment are compen-

sated by model-adaptation. Results are presented in Section 7.4 where the proposed systems are tested for their performance across different levels of speaker variability, noises and noise levels. Finally, Section 7.5 concludes the chapter with a discussion of the proposed system and how it fits in with the overall speech enhancement system.

7.2 Data-Driven Voicing Classification

This section describes a preliminary study on the use of machine learning classifiers for the purpose of voicing classification and voice activity detection[1]. A range of ML classifiers are applied to the two tasks and their performance is compared against conventional techniques. A broad range of classifiers are considered and include parametric, probabilistic and non-probabilistic, artificial neural networks and regression. Some of these classifiers have previously been applied to speech processing applications while other classifiers chosen have not. Importantly, for the tasks of VAD and VC, the tests use speech that has been contaminated by noise as would be encountered in real situations. As computing processing power and storage increases it is useful to consider whether such machine learning classifiers have application to VAD and VC. Such methods also have advantages in that they learn classification boundaries from training data rather than requiring thresholds or constants to be determined as with conventional methods.

To apply machine learning techniques to VAD and VC a speech feature vector must be decided upon. MFCCs were found to be appropriate for this task in the review of features in Section 3.4 where correlation between MFCCs and voicing class was measured. Previous investigations into VC, such as those by Darch et al. [2006], also found MFCCs to be effective and so for this work the MFCC vector will be used as the basic feature.

Static MFCC vectors, $\mathbf{x}$, (comprising coefficients $C1$ to $C12$) are extracted from

[1]This review was published at Interspeech [Harding and Milner, 2012b]

speech sampled at 8kHz using 23 overlapping mel-spaced triangular filterbank channels at a rate of 100 frames per second in accordance with the ETSI XAFE standard [Sorin and Ramabadran, 2003]. In addition to the basic MFCC vector, the zero'th coefficient can also be included as this gives a measure of energy which is useful in classification. Temporal derivatives can also be augmented to the static features and give additional information regarding rates of change of the vectors. Investigation into the effects of these combinations is presented in Section 7.2.2.1.

The classifiers are described in Section 7.2.1 along with a justification for their inclusion in this study. Experimental results for VAD and VC are presented and analysed in Section 7.2.2.

7.2.1 Classifiers

This subsection gives a brief description of each of the classifiers used in the comparison, with the aim of describing the algorithm at a high level so that basic principles and difference and similarities with other classifiers are highlighted. Implementations from the WEKA API [Hall et al., 2009] were used with the exception of the GMM classifier which used an in-house implementation. It is not possible to evaluate every possible classifier but the criteria decided upon for inclusion was reasonably broad ranging so as to make a useful comparison of different methods.

7.2.1.1 Gaussian mixture model (GMM)

The GMM is a parametric probabilistic classifier that models the distribution of multivariate input data using a mixture of K Gaussian distributions. GMMs have been shown to be effective at modelling the distribution of MFCC vectors in many applications such as speech recognition and synthesis and notably voice activity detection [Darch et al., 2006]. This makes them a good baseline for comparing performance against other ML classifiers.

During a training stage, feature vectors, $\mathbf{x}$, are pooled according to their class, c,

to give class-specific vector pools, where $c \in \{nonspeech, voiced, unvoiced\}$ in the case of voicing classification. Expectation maximisation (EM) clustering is applied to each vector pool to create a GMM for each class, ϕ^c [Darch et al., 2006]. An unseen vector, $\mathbf{x}$, is classified according to the GMM with highest probability. The discrete cosine transform (DCT) used in the MFCC feature extraction process removes correlation in the log filterbank domain. However, whilst this is true for static features, augmenting the feature vector with temporal derivatives does reintroduce correlation and so a full covariance matrix is retained to take into account these cross-correlations within the feature vector space.

7.2.1.2 Support vector machine (SVM)

SVMs are a non-probabilistic binary linear classifier [Mitchell, 1997]. SVMs require feature vectors, $\mathbf{x}$, to be linearly separable based upon their class. Where feature vectors are not linearly separable a kernel function is selected that best maps the vectors into a new feature space where the vectors are linearly separable. The dividing hyperplane that provides the largest possible margin between the transformed data is then calculated and stored as a set of support vectors. New vectors are classified by calculating which side of the dividing hyperplane they fall. SVMs clearly rely on the appropriate selection of kernel function and this work considers the standard polynomial kernel in the WEKA API [Hall et al., 2009].

7.2.1.3 Multilayer perceptron (MLP)

MLPs are an extension of the linear perceptron and a form of artificial neural network [Mitchell, 1997]. They can be viewed as being related to SVMs, differing mainly in the method of class separation [Collobert and Bengio, 2004]. Like SVMs, MLPs aim to find the maximum margin between vectors based on the class. Instead of using a kernel function to transform the feature space, MLPs use multiple linear perceptrons to separate non-linearly separable vectors on their class in the existing feature space. Unseen vectors are classified by calculating the decision region in

which they fall based on the arrangement and weighting of the linear perceptrons. From visual inspection, feature vectors from the datasets described in Appendix A were found not to be linearly separable. This suggests that MLPs should perform well due to their ability to form complex decision regions.

7.2.1.4 C4.5 decision trees

The C4.5 algorithm builds decision trees and can be considered a form of statistical classifier [Mitchell, 1997]. The decision tree is built by calculating the information gain that results from splitting a training data set on the class for each coefficient in the MFCC vector. The individual MFCC that gives the largest information gain is chosen as the split for the current node. The MFCC chosen to split at each subsequent node is determined in the same way until all vectors in the subset are labelled with the same class. Unseen vectors are classified by the decision rules determined by the split on each tree node (MFCC) until a class is determined. For VAD in clean conditions the most discriminative MFCC is likely to be $C0$ due to large differences in energy between speech and non-speech. $C1$ (spectral slope) is also likely to be effective in determining speech from non-speech.

7.2.1.5 Tree ensembles (Rotation Forest)

Ensemble classifiers are multiple classifier systems that use a number of models to obtain better performance. This work uses the Rotation Forest classifier which comprises a number of decision trees [Rodriguez et al., 2006]. Each decision tree is trained on a random subset of training data with principal component analysis (PCA) applied to each subset. All principal components are retained to preserve the information in the variance of the data whilst decorrelating the feature space. Although static coefficients are already decorrelated by the DCT the correlation introduced into the feature vector by the temporal derivatives should be removed by PCA.

7.2.1.6 Naïve Bayesian networks

Bayesian networks are probabilistic classifiers that model the joint distribution of feature vectors with a set of conditional probabilities using a directed acyclic graph [Mitchell, 1997]. A fully connected graph suggests that all coefficients in the vector are dependent on one another. An alternative to the fully connected Bayesian network is naïve Bayes which makes the assumption that each individual MFCC is dependent only on the class, vastly simplifying the complexity of the model. This assumption allows each MFCC to be modelled using an independent Gaussian distribution. The method is reasonably similar to the GMM classifier (Section 7.2.1.1) when K =1 and assuming diagonal covariance. In informal testing, performance of naïve Bayes was found to be equivalent to that of a full Bayesian network and so the naïve Bayes classifier is used in this work.

7.2.1.7 Logistic regression

Logistic regression is a form of binomial regression analysis with no assumption made as to the distribution of the data [Mitchell, 1997]. The log outcomes of the class are modelled as a linear combination of the feature vectors, with the best fit calculated using maximum likelihood estimation. New vectors are classified by calculating the log odds of the vector belonging to a particular class using the regression model built in training.

7.2.2 Results

This subsection presents results and analysis of voice activity detection and voicing classification across the set of classifiers and also compares accuracy against conventional systems. Results are first presented to determine the optimal feature vector.

The speech used was taken from the WSJCAM0 dataset [Robinson et al., 1995] and downsampled to 8kHz. For testing in stationary noise, white noise was added,

Table 7.1: Effect of MFCC feature type on voice activity detection accuracy at an SNR of 10dB in white noise

	SVM	MLP	LogRes
$C(1-12)$	0.81	0.88	0.79
$C(0-12)$	0.85	0.89	0.86
$C+\Delta C$	0.85	0.96	0.86
$C+\Delta C+\Delta\Delta C$	0.88	0.97	0.88

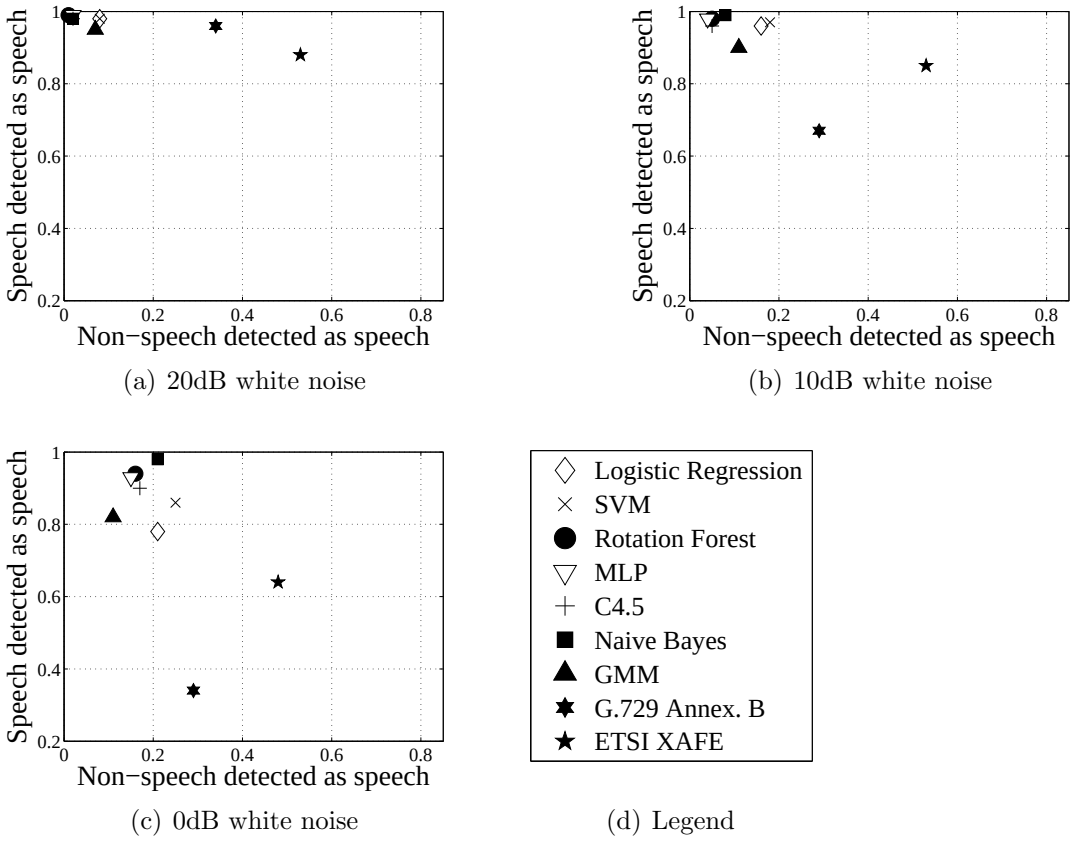
while for non-stationary noise, street noise from the NOIZEUS dataset was added at 0dB, 10dB and 20dB SNR [Hu and Loizou, 2006]. The SNR of the noise was computed based on the active speech level [P.56, 1993] to reduce the effect of long periods of silence on the SNR calculation. The modified Intermediate Reference System (IRS) filter used in ITU-T P.862 was then applied to simulate the frequency response of a telephone handset before MFCC features were extracted.

Classifiers were trained on 20 male and 20 female talkers and tested on 5 male and 5 female talkers that were previously unseen. Ten utterances were selected per talker to give a total of 400 utterances (≈1200 sec) for training and a further 100 (≈300 sec) for testing. The test set comprised 36% silence, 22% unvoiced and 42% voiced speech. Reference VAD data was obtained using an energy threshold applied to noise-free speech. A pitch track was then calculated using PRAAT [Boersma, 2002] and combined with the energy thresholding to give labels of non-speech, unvoiced and voiced. The test set was subsequently hand corrected where necessary.

7.2.2.1 Feature selection

To determine the optimal MFCC vector a preliminary voice activity detection test was performed. Results are presented in Table 7.1 using speech that has been contaminated with white noise at an SNR of 10dB. For comparison three different classifiers are used – SVM, MLP and logistic regression.

Across the three different classifiers, results show that adding $C0$ and including velocity (ΔC) and acceleration ($\Delta\Delta$C) temporal derivatives all increase performance. The gain made by each addition varies across the classifiers but the overall



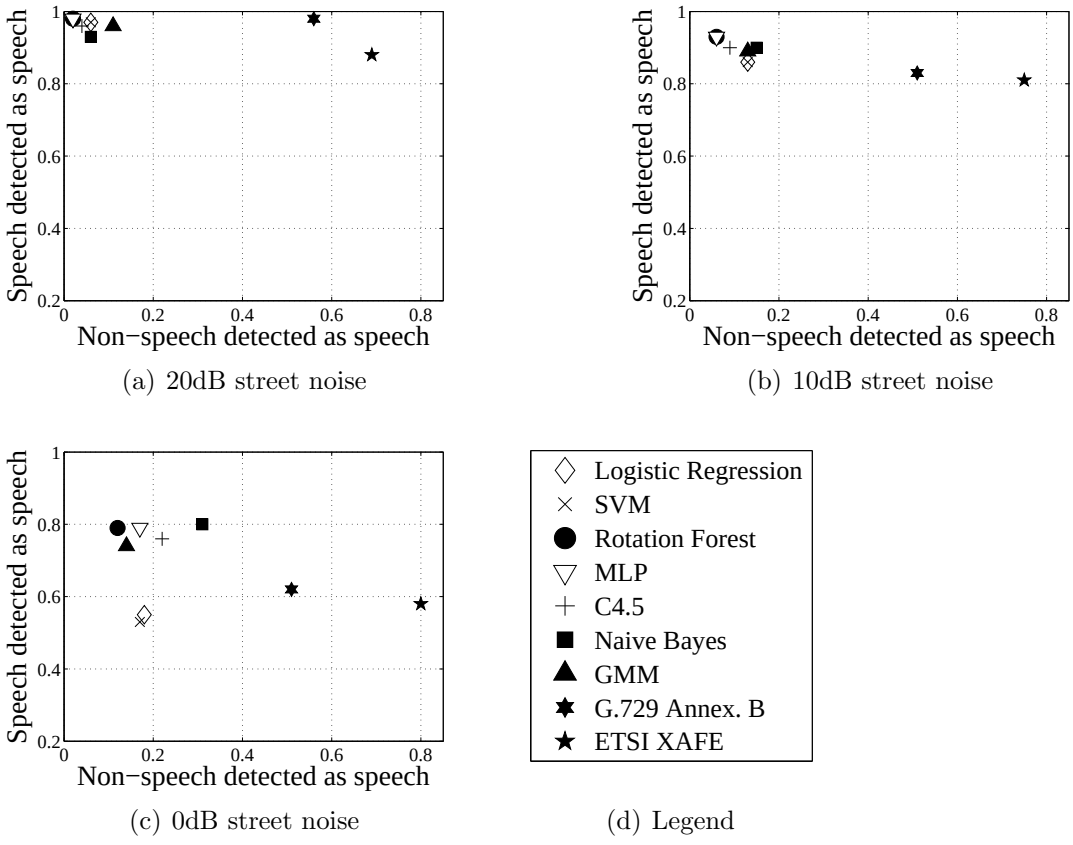
(a) 20dB white noise

(b) 10dB white noise

(c) 0dB white noise

(d) Legend

Figure 7.1: Performance of voice activity detection in white noise at SNRs of 20dB, 10dB and 0dB in ROC space



(a) 20dB street noise

(b) 10dB street noise

(c) 0dB street noise

(d) Legend

Figure 7.2: Performance of voice activity detection in street noise at SNRs of 20dB, 10dB and 0dB in ROC space

result is an absolute increase in accuracy of between 7% and 9%. Therefore, for the remainder of experiments the MFCC vector comprises $C0$ to $C12$ with velocity and acceleration augmented.

7.2.2.2 Voice activity detection

This subsection examines classifier performance on voice activity detection. Each classifier is evaluated in receiver operating characteristic (ROC) space in terms of the true positive rate (speech detected correctly as speech) and false positive rate (non-speech detected as speech). Figures 7.1 and 7.2 shows performance of the classifiers in white noise and street noise at SNRs of 20dB, 10dB and 0dB. To provide baseline results, the performance of two industry standard VADs, namely the G.729 and ETSI XAFE, are also included [Benyassine et al., 1997; Sorin and Ramabadran, 2003].

Comparing performance between the two noise types shows accuracy to be worse in the non-stationary street noise where sounds from cars, sirens, etc. introduce misclassifications. In terms of SNR, at higher SNRs classifier accuracies are reasonably close in the ROC space but as SNRs fall the classifier performances disperse with a shift towards the higher error region (bottom-right) of the ROC space. For both noise types the most accurate classifiers are Rotation Forest, GMM and the MLP and the worst performing are SVM and logistic regression. However, all of these classifiers outperform the two baseline VADs which are seen to perform poorly even in the relatively high SNR of 20dB. As SNRs fall their performance degrades rapidly, showing their sensitivity to noise. The poor performance of the conventional methods in low levels of noise is attributed to the applications for which they are designed. In both cases the two conventional methods are designed for speech transmission and so a low false negative rate is desirable. This subsequently has a negative effect on the false positive rate as demonstrated by these results.

The three best performing classifiers all share the ability to model cross-covariances, which distinguishes them from the other classifiers. Even though the DCT employed

Table 7.2: Voicing classification accuracy in white noise at SNRs of 20dB, 10dB and 0dB

Classifier	20dB				10dB				0dB			
	NS	UV	V	OVL	NS	UV	V	OVL	NS	UV	V	OVL
SVM	**0.99**	0.67	0.95	0.91	0.98	0.37	0.94	0.84	**0.97**	0.05	0.88	0.74
MLP	**0.99**	**0.88**	0.95	**0.95**	0.98	0.79	0.94	**0.93**	0.96	0.39	**0.92**	**0.82**
Rotation Forest	**0.99**	0.84	**0.96**	**0.95**	0.98	0.72	**0.95**	0.92	0.95	0.38	0.91	**0.82**
C4.5	0.98	0.83	0.92	0.92	0.96	0.71	0.92	0.89	0.86	0.41	0.86	0.77
Naïve Bayes	0.97	0.77	0.90	0.90	**0.99**	0.67	0.84	0.86	0.94	0.41	0.80	0.77
GMM	0.88	0.80	0.86	0.85	0.89	0.56	0.90	0.83	0.88	0.42	0.82	0.76
Logistic Regression	**0.99**	0.67	0.94	0.90	0.97	0.39	0.94	0.84	0.94	0.10	0.89	0.75
ETSI XAFE	0.47	0.83	0.91	0.73	0.47	**0.88**	0.83	0.71	0.52	**0.90**	0.51	0.59

in MFCC feature extraction should remove the correlation within the log filterbank, augmenting the vector with velocity and acceleration derivatives reintroduces some correlation. It is postulated that this may cause varying levels of difficulty for many of the classifiers tested, with the exception of Rotation Forest, GMM and MLP which gives rise to their superior performance. This suggests that applying a further transform to the feature vector, for example PCA, would decorrelate the coefficients of the entire feature vector and potentially improve the performance of other classifiers.

7.2.2.3 Voicing classification

Voicing classification extends VAD into the three class problem of determining between non-speech, unvoiced speech and voiced speech. For some classifiers only a binary decision is possible – for example SVM and logistic regression. In these cases two instances of the classifier were used, with the first being a VAD and the second applied to speech frames to classify between voiced and unvoiced speech, therefore allowing a three class output. The results of voicing classification in white noise are displayed in Table 7.2 and in street noise in Table 7.3. The tables show the classification accuracy for non-speech (NS), unvoiced (UV) and voiced (V) frames. A measure of the overall accuracy (OVL) is also shown and is computed from the total number of frames correctly classified. Results are presented at SNRs of 20dB, 10dB and 0dB. To serve as a baseline, the voicing classification accuracy from the ETSI XAFE standard is included [Sorin and Ramabadran, 2003].

Table 7.3: Voicing classification accuracy in street noise at SNRs of 20dB, 10dB and 0dB

Classifier	20dB				10dB				0dB			
	NS	UV	V	OVL	NS	UV	V	OVL	NS	UV	V	OVL
SVM	**0.98**	0.77	0.92	0.91	0.89	0.63	0.82	0.81	0.69	0.32	0.68	0.61
MLP	**0.98**	**0.87**	**0.94**	**0.94**	0.94	0.76	**0.89**	**0.88**	0.85	0.46	0.74	0.72
Rotation Forest	**0.98**	0.81	**0.94**	0.93	**0.95**	0.72	**0.89**	**0.88**	**0.86**	0.44	**0.77**	**0.74**
C4.5	0.96	0.79	0.89	0.89	0.89	0.69	0.80	0.81	0.73	0.46	0.63	0.63
Naïve Bayes	0.93	0.73	0.90	0.87	0.90	0.62	0.81	0.80	0.83	0.33	0.62	0.64
GMM	0.88	0.81	0.85	0.85	0.86	0.70	0.76	0.79	0.84	0.49	0.58	0.66
Logistic Regression	0.97	0.77	0.92	0.91	0.89	0.63	0.82	0.81	0.69	0.35	0.66	0.61
ETSI XAFE	0.31	0.83	0.90	0.88	0.25	**0.88**	0.78	0.81	0.20	**0.92**	0.42	0.44

The results follow a similar pattern to the VAD results, with voicing classification accuracy worse in the non-stationary street noise and reducing as SNRs fall. In terms of the accuracy of individual classifiers, as was observed for VAD, the Rotation Forest and MLP have highest overall classification performance. This is likely to be for the reasons discussed for VAD and related to the ability of these classifiers to deal with correlated data. Overall voicing classification accuracy for the ETSI XAFE baseline tends to be poor and gives lowest performance for the majority of test conditions.

Considering now the accuracy of identifying the individual voicing categories, unvoiced speech is clearly the most difficult to identify correctly. As SNRs fall, the machine learning methods rapidly become ineffective at identifying unvoiced speech as there are relatively few distinguishing features between noise and unvoiced speech. This leads to the majority of unvoiced frames being incorrectly classified as non-speech. Unvoiced classification is further affected by the 4kHz bandwidth of the speech and the application of the IRS filter to simulate the telephony channel. Both of these reduce high frequency energy which is an important cue for unvoiced speech. Conversely, the ETSI XAFE method is seen to retain a high score for unvoiced classification. However this is at the expense of correctly identifying non-speech frames and is explained by the increasing noise levels causing the ETSI XAFE method to classify non-speech as unvoiced speech. Unvoiced classification performance may be improved by introducing a bias into the classifiers which favours the unvoiced class. Whilst this would improve the unvoiced classification accuracy this could reduce performance in other classes.

Machine learning classification accuracy for voiced and non-speech frames, although deteriorating as SNRs reduce, is, however, more robust to noise than for unvoiced classification. This classification problem is more simple as voiced frames tend to be of higher energy than unvoiced frames and will therefore have a higher local SNR. Voiced frame energy is focused in lower frequency regions which are retained during feature extraction thereby providing useful discriminative information in the feature vectors.

7.3 Proposed Method of Voicing Classification

This section describes systems of robust voicing classification for this method of speech enhancement. For speech reconstruction using the HNM a simplified voicing classification is required consisting of only two classes, i.e. a problem of classifying $c \in \{notvoiced, voiced\}$ where $notvoiced = \{nonspeech, unvoiced\}$. Section 7.2 reviewed a wide range of ML classification methods alongside conventional methods. MLPs and Rotation Forest were found to provide the best performance when the training environment was matched to the test environment. This configuration is unrealistic for real-world scenarios where the environment may not be well represented in the training data. Domain adaptation is therefore required to obtain good performance using ML methods. The choice of method is thus reduced to those which may be adapted to account for such variations.

As described in Chapter 6, there are two approaches to this problem: feature compensation and model adaptation. In the former, features are extracted from noisy speech and 'cleaned' for use in a clean-trained classification model [Deng et al., 2000]. In the latter, the model is adapted to the domain of the noisy features [Vair et al., 2006; Gales, 2011].

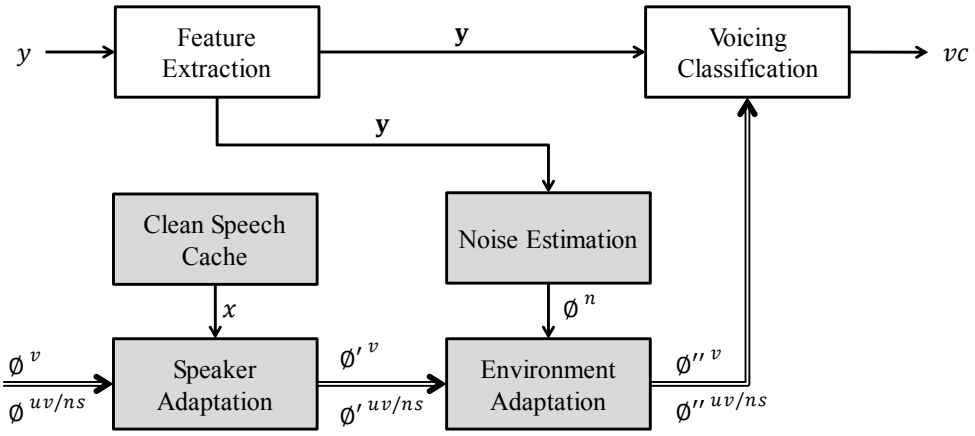


Figure 7.3: Proposed VC system using model-based speaker and noise compensation

7.3.1 Model adaptation

Starting with the case of model adaptation, given information about the target speaker and environment, models trained on 'universal' data can be adapted to the domain of the target. A range of speaker and noise adaptation methods for MLPs and GMMs already exist, developed primarily for use in speech recognition systems. Of these, GMMs have had the most focus due to their use in current state-of-the-art HMM-GMM based recognition systems [Gales, 2011] and so this work will focus on adapting a GMM-based system. These adaptation methods are described in Chapter 4 where it is shown that easily obtainable models of the environment can be used to adapt GMMs to specific environments whilst small amounts of speaker-specific data can also be used to form speaker-dependent models. A GMM-based approach using model-adaptation is therefore considered.

Figure 7.3 illustrates the proposed GMM-based system. First, universal background models (UBMs) are built from vectorpools of clean, speaker independent, speech to give our initial GMMs: ϕ^v, modelling voiced speech, and $\phi^{uv/ns}$ which models both unvoiced and non-speech. The next stage is to adapt these speaker independent models to the current speaker to give speaker dependent models. Small amounts of additional data from the target speaker is used to adapt the models using MAP adaptation as described in Section 4.5.1. This results in speaker-dependent

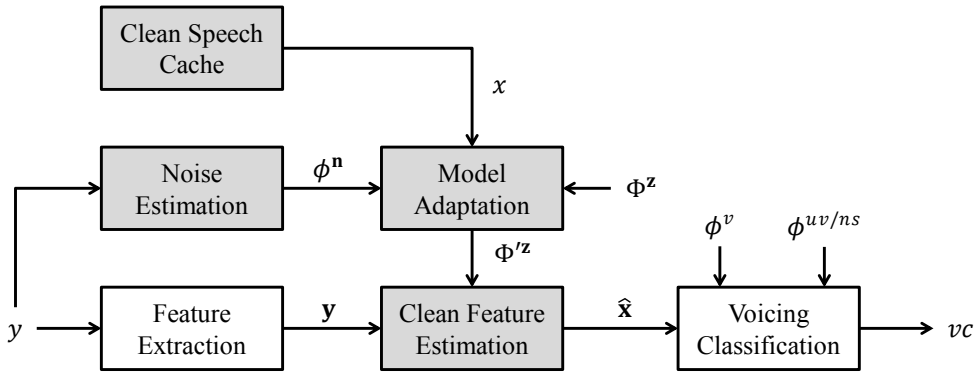


Figure 7.4: Proposed VC system using compensated features

models ϕ'^{v} and $\phi'^{uv/ns}$.

The final stage is to adapt the speaker-dependent GMMs, which presently model clean speech, to the current environment. First, a GMM of the noise, ϕ^{n}, is trained using noise data extracted from the noisy speech, y. This is combined with the clean speech GMMs using the unscented transform (UT) as described in Section 4.5.2 to give our final speaker-dependent, environment-dependent GMMs ϕ''^{v} and $\phi''^{uv/ns}$. These models may then be used to classify the frames of noisy speech, $\mathbf{y}$, as per the process described in Section 7.2.1.1.

7.3.2 Feature compensation

For the case of methods using feature compensation, models are trained on clean, speaker independent data. Features extracted from the noisy speech and then used to form an estimate of clean features. In this work a system for feature compensation has been developed in Chapter 5 for the purpose of spectral envelope estimation. The use of compensated features is beneficial as any classification method may now be used. The two best systems as determined in the earlier review of machine learning methods in Section 7.2, MLP and Rotation Forest, are therefore considered in this case along with the GMM-based system to enable direct comparison to the model-adapted system. Figure 7.4 illustrates the process of voicing classification with compensated features.

Starting with the models used for feature compensation, $\phi^{\mathbf{n}}$ and $\Phi^{\mathbf{z}}$ represent models of the noise and joint density of clean and noisy speech respectively, where $\mathbf{z} = [\mathbf{y}, \mathbf{x}]^T$. $\Phi'^{\mathbf{z}}$ is therefore the adapted joint density model. Finally, ϕ^{v} and $\phi^{uv/ns}$ represent clean-trained classification models for voiced and unvoiced/non-speech as per the model-adapted system. Two classification models are shown on Figure 7.4 however in the case of other classifiers only one model may be required.

7.4 Results

This section presents results of experiments that compare the proposed GMM-based model-adaptation system with the Rotation Forest, MLP and GMM classifiers using compensated features. In addition, these methods are compared against the conventional ETSI XAFE voicing classifier. The section begins by describing a set of experiments which are used to optimise the parameters of the proposed systems in Section 7.4.1 before overall results using these parameters are presented in Section 7.4.2. Overall results are summarised in Section 7.4.3.

7.4.1 Parameter optimisation

This section presents the results of experiments used to determine the optimal parameters of the model-adaptation system. Parameters are optimised on the GMM-based system and where parameters are shared between systems these assumed to also be appropriate for the MLP and Rotation Forest systems. There are two parameters to optimise in the case of the GMM-based system: the feature size and number of mixture components in the GMMs. Systems using feature compensation will use the parameters already determined in the review of ML methods in Section 7.2.

Section 3.5.3.3 measured the correlation between MFCCs and voicing class. The base-configuration has been fixed as the optimal MFCC feature vector for spectral envelope enhancement which comprises 32 filterbank channels transformed using a 32 point DCT and comprises only of static features. Most useful information was

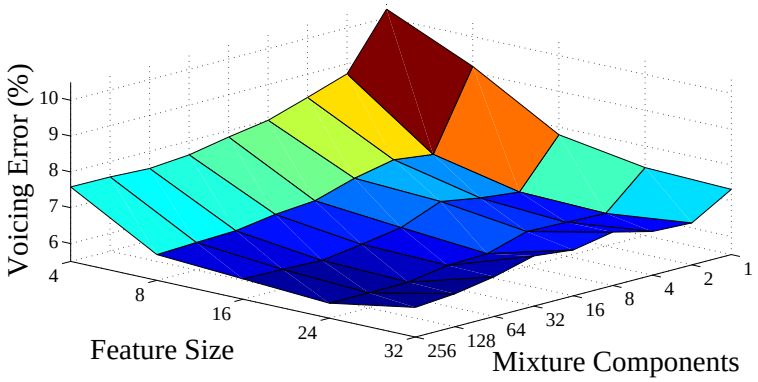


Figure 7.5: Effect of varying feature and model sizes on voicing classification error using models trained and tested on clean speech

shown to be contained in the first 15 of the 32 MFCC coefficients and so the first parameter to be optimised is therefore the number of DCT coefficients retained for use in the final system.

Next, we consider the number of mixture components required by the GMMs to effectively model the feature distributions. Darch et al. [2006] has shown that no significant advantage is achieved by using different numbers of mixture components to model voiced and non-voiced speech. This work therefore considers only the case where the number of mixture components used for the voiced model is equal to the number used in the non-voiced model.

Voicing classification error was calculated for a range of parameters. The speaker-dependent data from the NuanceCatherine dataset was used for testing, with white noise mixed with the speech at an SNR of 10dB. Figure 7.5 shows the effect of varying the feature size simultaneously with the number of mixture components.

The feature dimensionality is seen to have the largest effect on classification accuracy with larger feature sizes preferred. A feature size of 24 was found to be optimal in this case. Results in Section 3.5.3.3 showed there to be minimal information relating to voicing class between $c(15)$ and $c(24)$ and so it is surprising that optimal performance is found at 24 coefficients rather than 16. Focusing now on the modelling parameter, a larger number of mixture components is preferred

over smaller models with a minimum error found at 128 mixture components.

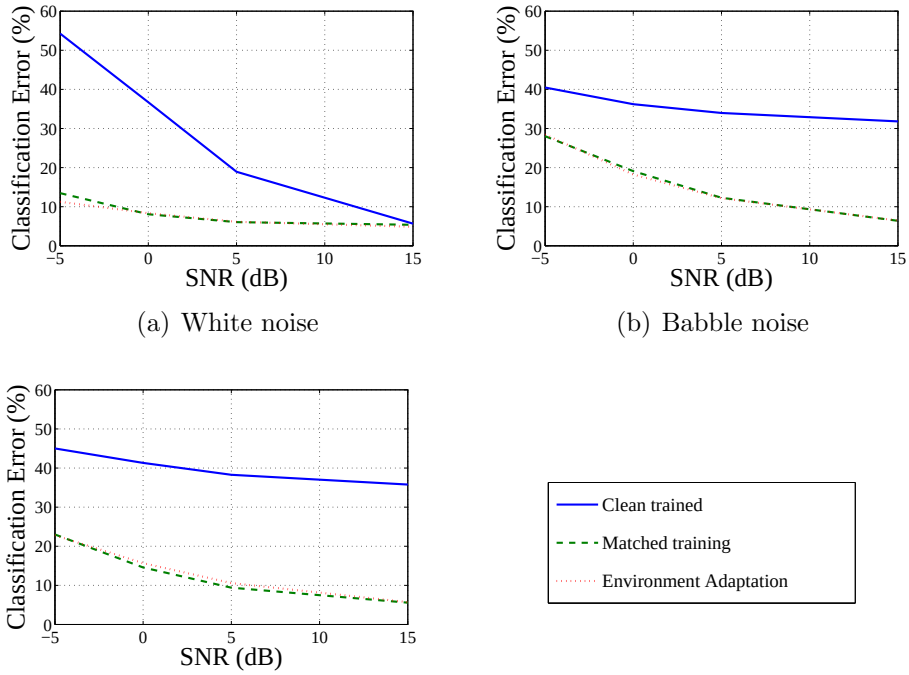
These parameters differ from those used for spectral envelope and fundamental frequency estimation, highlighting the differences in requirements for classification versus estimation. A smoother spectral envelope is sufficient for voicing classification whilst the number of mixture components is also reduced to give a less detailed model of the feature distributions.

7.4.2 Voicing classification results

This section presents results of a range of experiments used to determine the most suitable method of robust voicing classification for use in this speech enhancement system. A range of methods are considered, including the GMM-based classifier using model-adaptation and other ML classifiers using enhanced features.

Results are split into three parts. First, results of experiments testing the proposed GMM model-adaptation system in a number of configurations are presented in Section 7.4.2.1. Second, the result of using compensated features with clean-trained ML models is presented in Section 7.4.2.2. Three classifiers are evaluated in this section: Rotation Forest, Multilayer Perceptron (MLP) and GMM. Rotation Forest and MLP were found to offer best performance in the review of ML methods in Section 7.2 whilst the GMM classifier will allow direct comparisons to the model-adapted system. Finally, the model-adapted and feature-compensated methods are compared against the conventional voicing classifier from the ETSI Aurora XAFE standard in Section 7.4.3 where the most suitable system is selected.

In each case systems are tested on speaker-dependent, gender dependent and speaker-independent data. Three noises are tested: white noise, babble noise and destroyerops at -5dB, 0dB, 5dB and 15dB SNR.



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 7.6: Performance of proposed GMM voicing classification system trained on speaker-dependent data using: i.) clean speech, ii.) noisy speech matched to the testing environment and iii.) model adaptation

7.4.2.1 GMM-adaptation method

This section presents results of using the model-adaptation system developed in Section 7.3. Starting with the case of speaker-dependent data, the system was trained and tested on different segmentations of the NuanceCatherine dataset using three types of GMMs: uncompensated (trained in clean conditions), matched (trained on noisy data matched to the test environment) and adapted to the test environment (trained in clean conditions and adapted to the environment). Results are displayed in Figure 7.6.

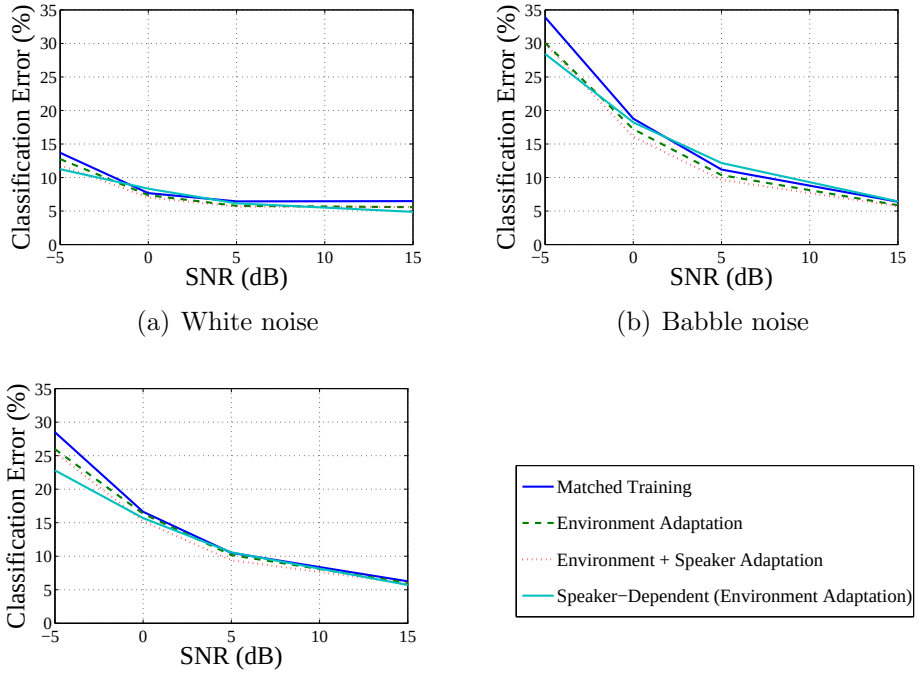
In very low levels of white noise (15dB SNR) there is little benefit to using compensated models, however performance degrades substantially at lower SNRs or when tested in non-stationary noises such as babble or destroyerops noise. In these cases, the environment adaptation and matched train/test systems clearly out-perform the clean-trained models. In most cases the matched train/test system

marginally outperforms the environment adapted GMMs, though in some cases the adapted GMMs offer slightly better performance. This is encouraging as it shows there is little difference between the optimal system and our proposed adapted system.

Next, the case of a gender-dependent system is considered. Data from the WSJ-CAM0 dataset was used with 20 speakers used to train each system with five different speakers used for testing. Results of the female-only system are in most cases comparable to those found in the speaker-dependent system. The largest differences are found at -5dB with a significant increase in error found in the gender-dependent system. Interestingly, the environment-adapted system outperforms the matched train/test system in almost all cases. This is attributed to the noise mixing process. Noise was added on a per-speaker basis. Environment adaptation was also performed on a per-speaker basis whilst the matched train/test system was trained across all speakers resulting in slight discontinuities in absolute noise level between speakers. Another notable result from this experiment is the relatively minor effect that speaker adaptation has on results when compared to spectral envelope and fundamental frequency estimation results. Approximately 160 seconds of speaker adaptation data was used per speaker, which reduced error rates by as much as 7% relative, though this relates to a decrease in absolute error of between only 0.1 and 1.1%.

Speaker adaptation was found to have a negligible effect on classifying male speech with no significant improvements found over the environment-only adaptation. Comparing the results of male and female-dependent systems, there is a significant increase of between 2 and 8% absolute error rate for male speech compared to female classification which relates to a 14-71% relative increase in error for comparable systems.

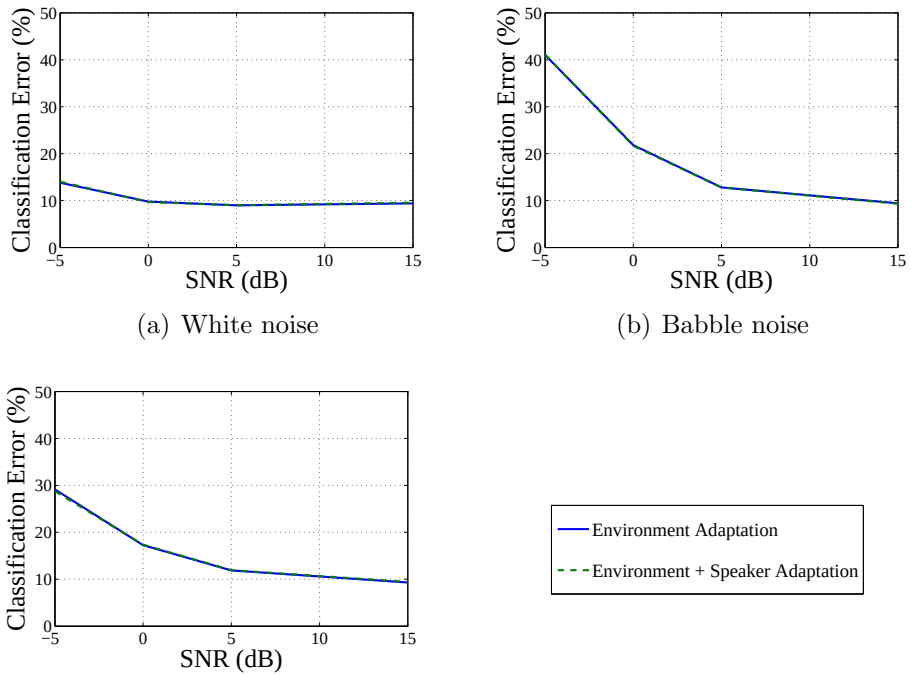
Finally, the case of a fully speaker independent system is tested. For this system the same speakers used in the gender dependent test were used in combination to train the models to give a total of 40 speakers. Uncompensated and matched models



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

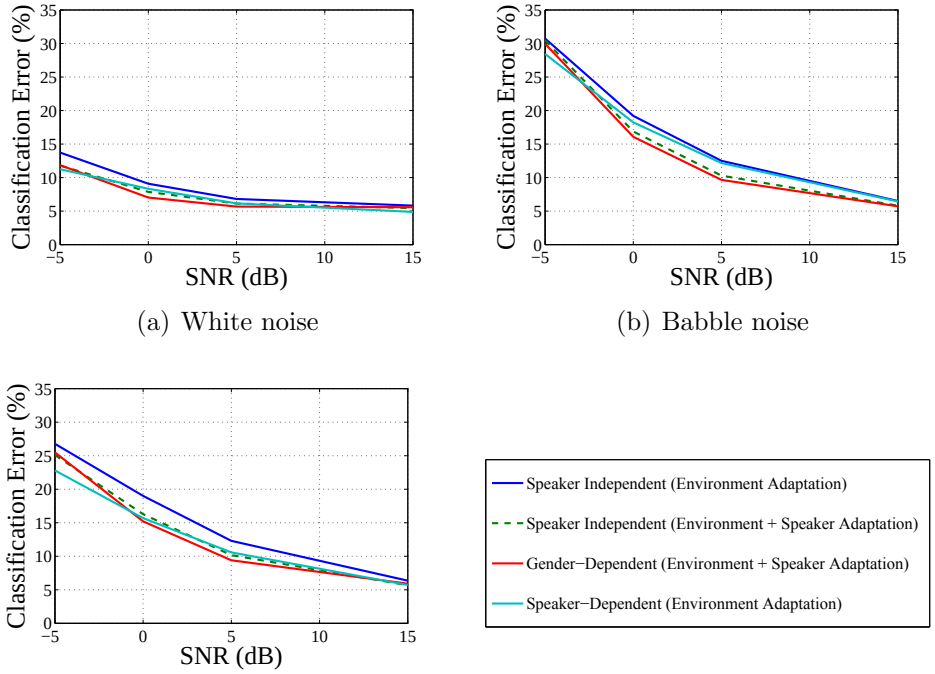
Figure 7.7: Performance of proposed GMM voicing classification system trained on female-only data using: i.) noisy speech matched to the testing environment, ii.) model adaptation for noise, iii.) model adaptation for speaker and noise and iv.) speaker-dependent system using noise adaptation



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 7.8: Performance of proposed GMM voicing classification system trained on male-only data using: i.) model adaptation for noise and ii.) model adaptation for speaker and noise



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 7.9: Performance of proposed GMM voicing classification system tested on female-only data and trained on: i.) gender-independent data using noise adaptation, ii.) gender-independent data using speaker and noise adaptation, iii.) gender-dependent data using environment and noise adaptation and iv.) speaker-dependent data using environment adaptation

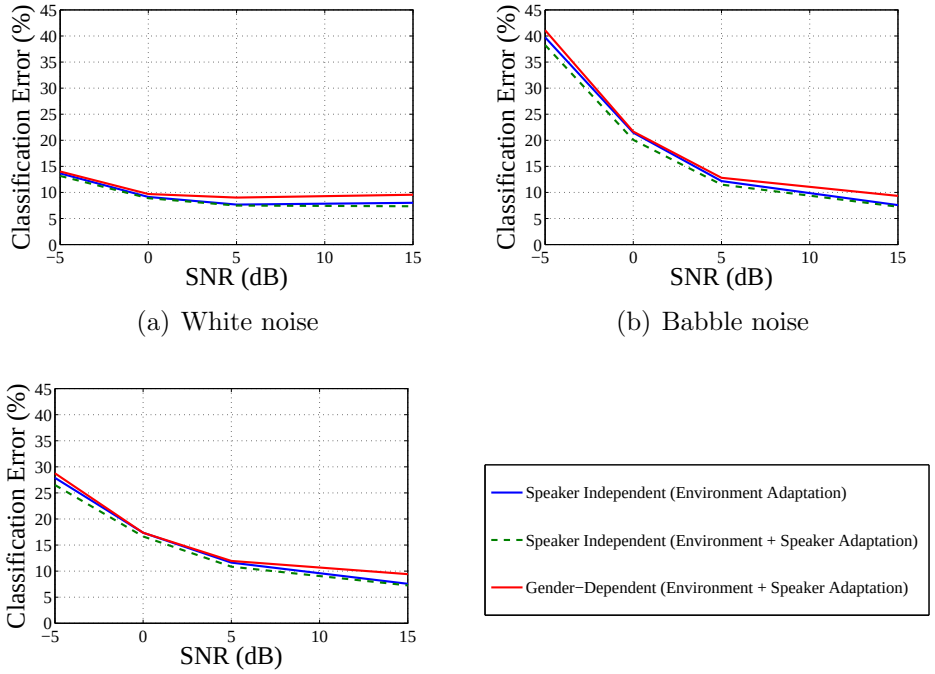
were not tested for speaker-independent data as the adapted system has been shown to offer the best performance in realisable conditions. Results of testing female speech, presented in Figure 7.9, are shown to be within 2% of the gender-dependent system, even without speaker adaptation, which shows the robustness of the method to both gender and speaker. Interestingly the speaker-independent system is shown to perform better than the speaker-dependent system in most cases. This is believed to be due to the increased amount of training data available allowing more accurate models to be trained alongside the effectiveness of the speaker-adaptation method.

Next, male speech was tested using the speaker-independent model and compared to gender-dependent results in Figure 7.10. Again, the speaker-independent system is seen to outperform the gender-dependent system. As well as being attributed to the increased amount of training data, these results show that voicing classification is neither speaker nor gender-dependent to any large extent meaning a fully speaker-independent system is possible.

Overall, the GMM-adaptation method is shown to offer good results which scale well with variability in both speaker and environment. The system uses the same features extracted for spectral envelope and fundamental frequency estimation but relies on adapting each voicing class GMM, adding to the computational complexity of the system. In the next section feature compensation methods are examined to determine if the enhanced features available from spectral envelope estimation may be used with similar effect for voicing classification.

7.4.2.2 Classifiers using compensated features

This section presents the result of using features extracted from noisy speech and subsequently cleaned using the process described in Chapter 5 as input to clean-trained classifiers. Three methods of classification are considered: GMM, MLP and Rotation Forest. Temporal derivatives were not used in the model-adaptation approach as it is computationally expensive to adapt dynamic parameters [Gales, 1995]. No such restrictions exist in this case and so temporal derivatives are calcu-



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

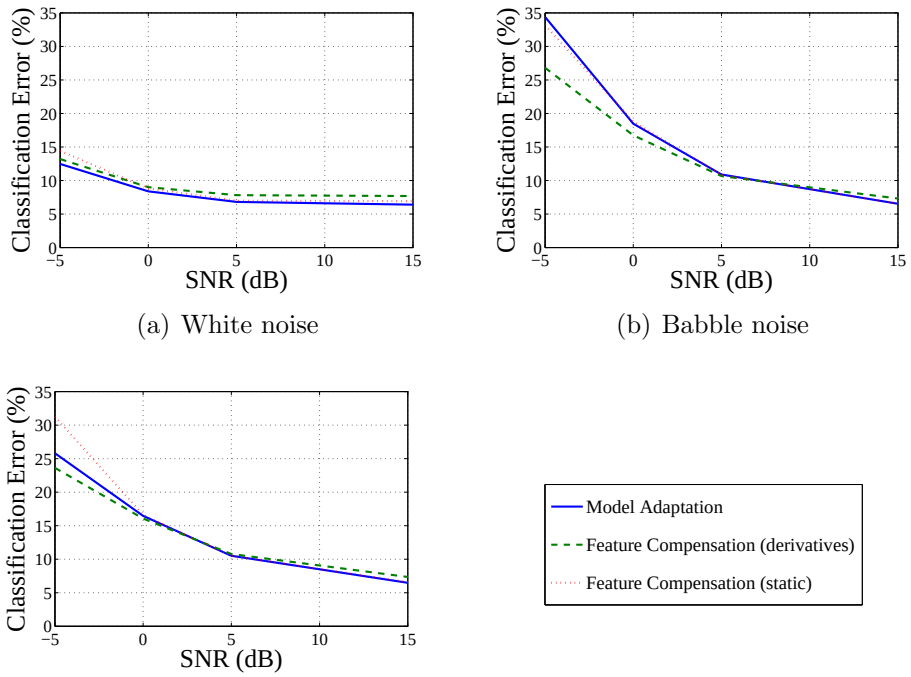
Figure 7.10: Performance of proposed GMM voicing classification system tested on male-only data and trained on: i.) gender-independent data using noise adaptation, ii.) gender-independent data using speaker and noise adaptation and iii.) gender-dependent data using environment and noise adaptation

lated from cleaned features.

All models were trained and tested on the same dataset as the GMM-adaptation method, that is 20 male and 20 female speakers for training and a combined total of 10 different male and female speakers for testing. All experiments in this section are performed using speaker-independent models.

We start with the comparison between the best model-adaptation system from Section 7.4.2.1 and the GMM classifier using enhanced features. Both systems use MAP-adaptation for speaker adaptation and the Unscented Transform for noise adaptation, the difference between the systems being the stage at which these transforms operate. The model-adaptation approach uses these techniques to adapt the models to the noisy feature domain whilst in the feature compensation method the features are adapted to the clean feature domain for use with clean-trained models. In addition, we also test the feature-compensation approach with temporal derivatives. These systems are compared in Figure 7.11 in stationary and non-stationary noises.

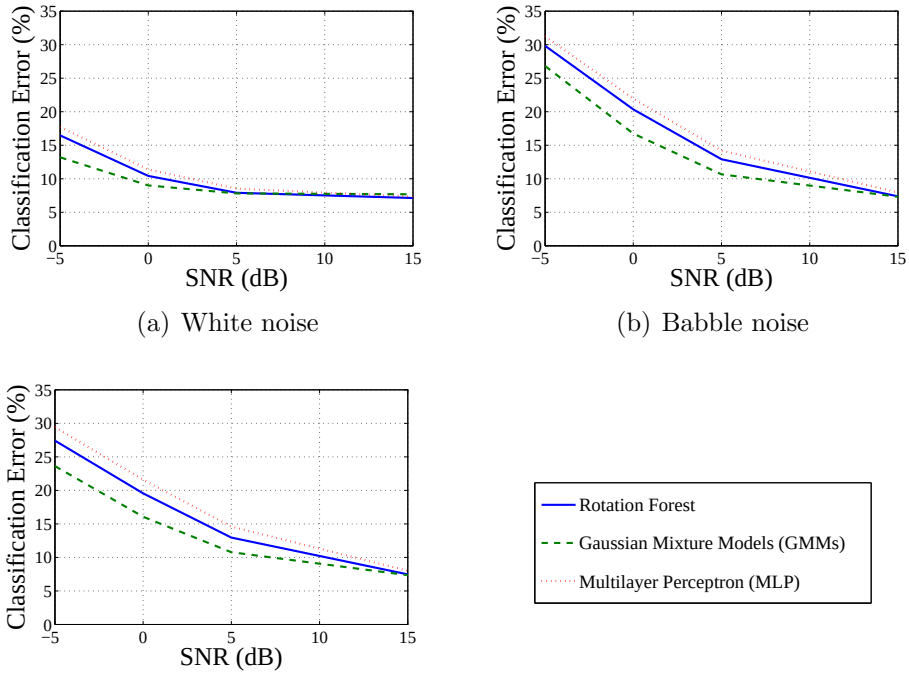
Examining results across all three noises shows that the type of noise affects the overall performance of the methods. In stationary noise best performance is obtained using the model adaptation approach with the feature compensation approaches varying in preference across SNR with features including derivatives offering best performance at -5dB SNR and static features offering better performance at higher SNRs. Examining now the case of non-stationary noises (babble and destroyerops), the feature compensation approaches are seen to perform much more strongly. In both cases feature compensation with temporal derivatives outperform both other methods, which are shown to be roughly equivalent except in the case of destroyerops noise at -5dB SNR where the model adaptation method performs better. The superior performance of the feature compensation method with temporal derivatives is attributed to the quality of feature estimation. As shown in Chapter 5, the RMS error of features estimated from signals affected by white noise is lower than features estimated from sources contaminated with non-stationary noises such



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 7.11: Performance of proposed GMM voicing classification system trained and tested on gender-independent data and compensated for noise using i.) model adaptation, ii.) enhanced features including temporal derivatives and iii.) enhanced features using static coefficients



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 7.12: Comparison of voicing classification error of best Machine Learning classifiers trained on clean speech and tested on features extracted from noisy speech and compensated for the noise using the system described in Chapter 5

as babble or destroyerops noise. Introducing temporal information which spans several frames therefore introduces a set of coefficients which are robust to within-frame errors introduced by the estimation process.

Next, we consider the case of using the best ML approaches, namely Rotation Forest and MLP classifiers, for classification. These methods use enhanced features including temporal derivatives and are also compared to the GMM feature-compensation system using the same features. These systems are compared in Figure 7.12. Rotation Forest is shown to perform better than the MLP as would be expected from the results of the review earlier in this chapter (Section 7.2). Surprisingly, however, GMMs are shown to perform significantly better than both competing methods in all but the case of white noise at 15dB SNR. This is in contradiction to the results shown in Section 7.2 where they were one of the worst performing ML methods. There are two factors which could affect this result. Firstly, previous

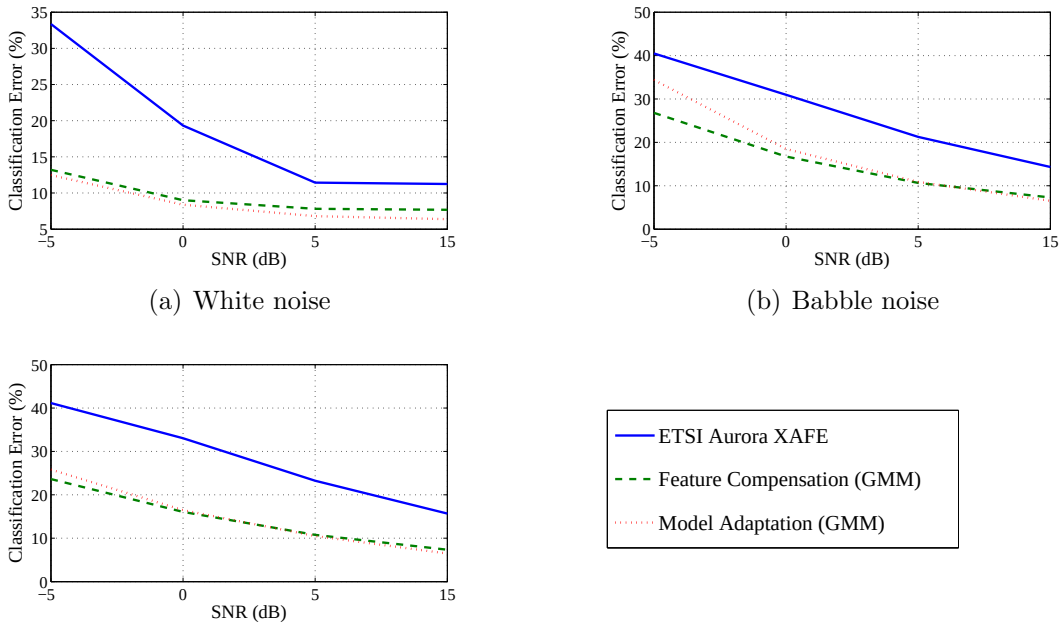
results were of a three-class voicing classification task whilst this task is a more simple two-class voiced vs. non-voiced classification. Secondly, previous results only considered the case of testing in conditions matched to the training environment. This task considers enhanced features with clean-trained models which won't provide an exact match due to estimation errors. It is therefore postulated that the GMM method is more robust to such estimation errors.

Based on the results of experiments presented in this section the best method for use with compensated features is the GMM classifier using features with temporal derivatives calculated from the enhanced features. This classifier will be compared to the best model-based approach and also the conventional ETSI Aurora XAFE voicing classifier in Section 7.4.3.

7.4.3 Overall results

In the previous sections various methods of robust voicing classification, including methods using model adaptation and feature compensation have been evaluated. This section aims to compare the best configurations of both approaches and compare them to the conventional voicing classifier from the ETSI Aurora XAFE standard to determine the best method for this application.

Figure 7.13 compares the three methods: the conventional ETSI Aurora XAFE standard, the proposed GMM model-adaptation system and the GMM classifier using compensated features with temporal derivatives. Both GMM-based systems are proven to be significantly more robust than the ETSI Aurora XAFE method across all noises and SNRs. Little difference between the GMM methods is visible, with the only significant difference noticeable in babble noise at an SNR of -5dB where the feature compensation method is shown to be more robust than the model-adaptation system. Overall, the feature-compensation based system is therefore the most suitable for use in this work. Not only does it offer classification robust to variations in speaker and noise, but it uses feature vectors which are already available as part of the spectral envelope estimation process and thus reduces the



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Legend

Figure 7.13: Comparison of voicing classification error of the ETSI Aurora XAFE system and the proposed GMM classification system using i.) enhanced features and ii.) model adaptation

complexity of the overall speech enhancement system.

7.5 Summary

In this chapter a wide range of systems for voice activity detection and voicing classification have been reviewed (Section 7.2), with machine learning methods such as Rotation Forests and Multilayer Perceptrons offering superior performance over conventional methods such as the G729 Annex.B VAD and the ETSI Aurora XAFE VAD and VC when the training environment was matched to the testing environment. Two methods of channel compensation were therefore also proposed: model-adaptation and feature compensation. In the case of model-adaptation, a novel system of adapting GMMs for both noise and speaker was proposed in Section 7.3 with results in Section 7.4.2.1 showing the adapted GMM-based system to be robust to variations in gender, speaker and noise. In the case of feature compensation,

Rotation Forest, MLP and GMM classifiers were tested with 'cleaned' features estimated from features extracted from noisy speech. Overall, the GMM-based classifier using feature compensation was found to offer best performance in terms of both overall classification accuracy and computational complexity.

Chapter 8

Phase Estimation

This chapter examines a range of phase models for use in this method of speech enhancement. Noise is known to affect both the magnitude and phase spectra of speech signals however most existing methods of speech enhancement make no attempt to enhance the phase spectrum. This work therefore examines a range of phase models to determine the best method of phase estimation. These include: noisy signal phase, zero-phase, minimum-phase and random phase, whilst the phase of clean speech is also included as a measure of optimal performance. The quality of speech reconstructed using each phase model is measured objectively using PESQ whilst a listening test was also performed to determine the preferred system.

Contents

8.1 Introduction

In this chapter a range of phase models are investigated to determine the best method of phase estimation for this method of speech enhancement. Most methods of speech enhancement retain the phase of the noisy speech and make no attempt at estimating the phase of the clean speech [Loizou, 2007]. This is because it is widely agreed that the ear is insensitive to shifts in phase [Paliwal, 2003]. However, shifts in *relative* phase between frequency components are less well understood, with Weiss et al. [1975] suggesting that rapid fluctuations in relative phase can cause perceptual artifacts in reconstructed signals. Earlier studies, such as those by Wang and Lim [1982], claimed that the effect of noise on phase in conditions where speech remains intelligible are relatively minimal. Despite this there is also evidence that the use of the phase of clean speech is preferable over using the phase of noisy speech in more recent studies [Moon et al., 2010]. Finally, listening tests have shown that phase is important to the perceptual quality and intelligibility of speech, with additive noise distorting the phase spectra to a perceivable extent [Paliwal and Alsteris, 2005]. It is therefore important to understand the extent of the perceivable distortions that will be caused by shifts in the phase caused by addition of noise.

Loizou [2007] demonstrated that the MMSE estimate of the clean speech phase is the phase of the noisy speech and gave a threshold of about 8dB SNR, below which noise distorts the phase to such an extent as to cause perceivable artifacts. We therefore examine various other methods of phase estimation in an attempt to improve on the perceptual quality of reconstructed speech with SNRs below this threshold. The methods considered include: noisy signal phase, zero-phase, minimum-phase and random phase. The phase of the original clean speech is not available for enhancement but is also included in experiments to determine optimal performance. These methods are described in detail in Section 8.2 and their robustness to noise is considered.

Most speech enhancement methods, including the one described in this work, operate on the concept of analysis followed by enhancement and then resynthesis.

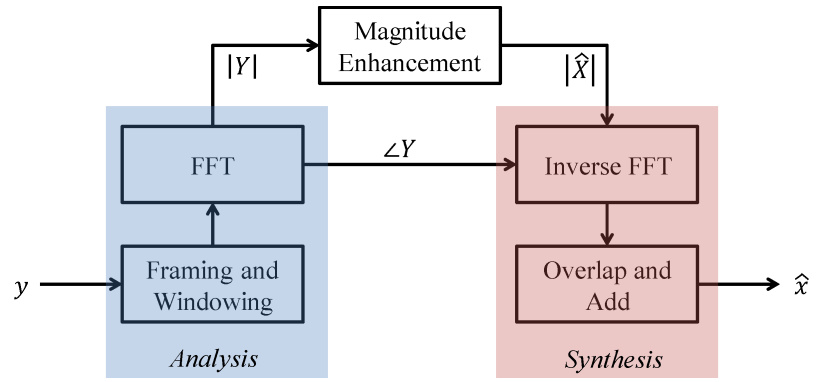


Figure 8.1: Diagram of typical analysis/synthesis based speech enhancement system

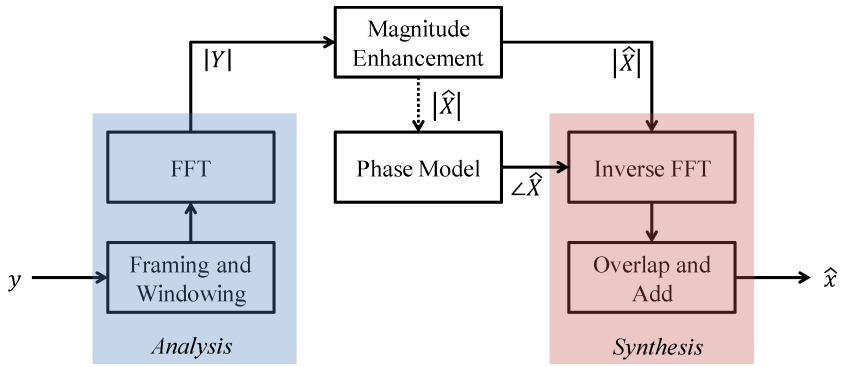


Figure 8.2: Diagram of phase model in the standard analysis/synthesis framework

This process is illustrated in Figure 8.1. At the analysis stage the phase spectrum is extracted from the input signal. Typically, only the magnitude spectrum is enhanced and then recombined with the previously extracted phase to give a modified complex spectrum which is transformed back to the time domain at the synthesis stage [Loizou, 2007].

In the case of the alternative phase models considered in this work a slightly different system is required. Figure 8.2 shows how the standard framework of analysis, enhancement and then synthesis is altered to allow the inclusion of the phase models. The phase is no longer extracted from the original signal but is instead estimated independently of the original signal phase. Optionally, the magnitude spectra may be used in the estimation of the new signal phase, i.e. when using models such as the minimum-phase model.

Objective and subjective experiments are carried out to determine the overall quality of reconstructed speech using each of the phase models and the subjective preferences of the systems in Sections 8.3.1 and 8.3.2 respectively. Results are summarised in Section 8.4.

8.2 Phase Models

This section describes the phase models considered for use in this method of speech enhancement. The methods considered are: noisy signal phase, zero-phase, minimum-phase and random phase. The phase of the original, clean, speech phase is not available for enhancement but is also included in this investigation to allow 'oracle' experiments used to determine optimal performance. The process of extracting the original signal phase is described in Section 8.2.1. The two 'naïve' models, zero-phase and random phase, are discussed in Section 8.2.2. They are described as 'naïve' as they make several assumptions which are not necessarily linked to the physical properties of the original signal phase. Finally, the minimum-phase model is described in Section 8.2.3.

8.2.1 Original signal phase

The most widely used source of signal phase in speech enhancement applications is the phase of the original signal [Loizou, 2007]. The phase is computed from the complex spectrum which is obtained through the use of an FFT of a windowed frame of the time-domain signal as per Equation 8.1 where $\theta(k)$ is the kth bin of the phase spectrum, $Y(k)$ is the kth bin of the complex spectrum and $\Re$ and $\Im$ denote real and imaginary components respectively.

$$\theta(k) = \angle Y(k) = \arctan\left(\frac{\Im(Y(k))}{\Re(Y(k))}\right). \tag{8.1}$$

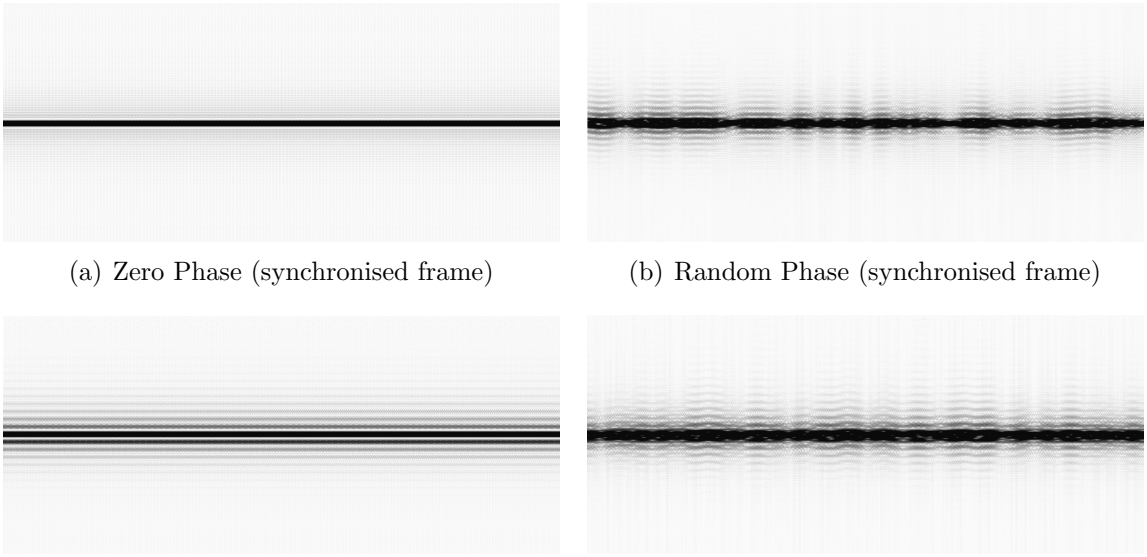
For phase extraction the frame length and window are normally selected to match those used for calculating the magnitude spectrum. When calculating the magnitude spectrum a frame length of 10-20ms is typically used with a Hamming or Hann window. Results presented in Shannon and Paliwal [2006] and Loweimi et al. [2011] show that the length of the analysis frame and window type are important factors in the quality of reconstructed speech. For this work a 20ms frame length with a Hamming window is used to match the configuration used for spectral feature extraction.

8.2.2 Zero and random phase models

A naïve model of the phase is to assume that the phase is unimportant and unrelated to the original signal. This vastly simplifies the system but introduces a number of assumptions which may not be valid in all cases. The two methods of naïve phase estimation evaluated in this work are the zero-phase model and the random-phase model. In the case of the zero-phase model all points in the phase spectra are set to zero whilst in the case of the random-phase model each bin is assigned a random value.

These models make two main assumptions. First, it is assumed that the phase is unrelated to the original signal and second, that it is not a function of time or frequency and so it is assumed that no phase interactions exist between frequency components. Weiss et al. [1975] suggests that zero-phase model may be suitable as the relative phase of the sinusoids will be constant and so should not degrade the quality of speech. To ensure continuity in phase values between frames, phase discontinuities must either be compensated for by computing the phase offset between frames or by synchronising frames to the fundamental frequency.

The effect of phase discontinuities is illustrated in Figure 8.3 where narrowband spectrogram plots of sinusoids synthesised using the zero and random phase models are displayed. A single sinusoid was synthesised as $x(m) = \sin(2\pi fm + \phi)$ where the frequency was given a value $f = 2000Hz$ and $\phi = 0$ in the case of the zero phase



(a) Zero Phase (synchronised frame)

(b) Random Phase (synchronised frame)

(c) Zero Phase (unsynchronised frame)

(d) Random Phase (unsynchronised frame)

Figure 8.3: Narrowband spectrograms of sinusoids synthesised using zero and random phase models using frame widths synchronised and unsynchronised with pitch period

model and $\phi = \text{rand}()$ in the case of random phase. The signal was resynthesised using a frame-based approach with the frame length set to 20ms at a sample rate of 8kHz. Frames were combined with no overlap.

The zero-phase model is shown to be effective when the frame width is matched to the period of the sinusoid which ensures the phase returns to zero at the end of each frame. When frames become out of sync with the fundamental period the phase is reset to zero at frame boundaries causing discontinuities and therefore artifacts in the resulting signal, with tones at erroneous frequencies appearing around the true sinusoid frequency. The random phase model, shown in Figures 8.3(b) and 8.3(d) is shown to be unaffected by frame synchronisation, but is affected by other artifacts related to discontinuities between frames caused by the phase being reset to random values between frames. The resulting signal appears to have been modulated in frequency and amplitude, the effect of both being audible.

Despite the potential of the zero phase model when frames are synchronised, for this application the complexity of altering the system to be pitch-synchronous is

considered to be too high and therefore the standard fixed frame system is used for the zero and random phase models.

8.2.3 Minimum-phase model

The minimum-phase model synthesises phase values based on the spectral envelope and fundamental frequency of a signal [Quatieri and McAulay, 2002]. The overall estimate of the phase, $\hat{\theta}(k)$, comprises two components:

$$\hat{\theta}(f) = \hat{\phi}(f) + \hat{\Phi}(f). \tag{8.2}$$

The first component, $\hat{\phi}(f)$, relates to the phase offset from the excitation and is defined as:

$$\hat{\phi}(f) = 2\pi f m, \tag{8.3}$$

where f corresponds to the frequency of the current component and m is the sample number of the reconstructed time domain waveform. The HNM reconstruction model synthesises voiced speech as a sum of harmonic sinusoids. This means $\hat{\theta}(f)$ and need only be computed at the harmonic frequencies and so $\hat{\phi}(f)$ becomes:

$$\hat{\phi}(lf_0) = 2\pi l f_0, \tag{8.4}$$

where l is the harmonic index and f_0 is the fundamental frequency. Phase offsets are tracked between frames to avoid inconsistencies by incorporating an additional term $p(j)$:

$$\hat{\phi}(lf_0) = 2\pi l f_0 p(j-1), \tag{8.5}$$

where $p(j-1)$ is the value of the recursive function p at frame index $j-1$ and $p(j)$ is defined as:

$$p(j) = T - \frac{T}{2} - p(j-1) \pmod{\frac{1}{f_0}}, \tag{8.6}$$

where $p(0) = 0$ and $T = \frac{N}{Fs}$ is the frame period given a frame length of N samples and a sample rate of Fs. A frame overlap of 50% is compensated for through the use of the term $\frac{T}{2}$.

The second component, $\hat{\Phi}(f)$, is estimated from the vocal tract filter assuming a minimum phase system. The minimum phase delay is computed in a two-stage process using a Hilbert transform [Oppenheim et al., 1989]. The first stage is to extract cepstral coefficients from the spectral envelope:

$$c(n) = \frac{2}{N_{fft}} \sum_{k=1}^{N_{fft}/2} \log\left(|\hat{X}(k)| \cos(2\pi nk)\right) \quad \text{for} \quad 1 \leq n \leq D, \tag{8.7}$$

where N_{fft} is the length of the DFT and D is the number of cepstral coefficients, with $D \geq 44$ sufficient for good performance [Quatieri and McAulay, 2002]. The Hilbert transform of these cepstral coefficients is then taken to give $\hat{\Phi}(f)$ as:

$$\hat{\Phi}(f) = -2 \sum_{n=1}^{D} c(n) \sin(2\pi nf). \tag{8.8}$$

As per $\hat{\phi}$ this component is only sampled at harmonic frequencies. The overall phase model is therefore defined as:

$$\hat{\theta}(lf_0) = \hat{\phi}(lf_0) + \hat{\Phi}(lf_0). \tag{8.9}$$

In the case of unvoiced frames $\hat{\theta}(k) = R$ where R is a random number and $0 \leq R \leq 2\pi$.

Table 8.1: Minimum-phase test configurations

Name	Amplitude	F0
MIN_REF_REF	REF	REF
MIN_REF_MAP	REF	MAP
MIN_MAP_REF	MAP	REF
MIN_MAP_MAP	MAP	MAP

8.3 Results

This section presents results of a number of experiments carried out to determine the optimal method of estimating the sinusoid phase values for use in this method of speech enhancement. All of the phase models previously described in this section are evaluated, namely: original signal phase (from clean and noisy speech), minimum-phase model, zero-phase model and random-phase model. This section presents both objective and subjective quality results of using these models to reconstruct speech from parameters estimated from the clean and noisy speech. All experiments in this section use speech from the WSJCAM0 corpus and, where applicable, destroyerops noise from the NOISEX dataset. Speech from five male and five female speakers was used for testing. Speech was sampled at rate of 8kHz with each speaker contributing 50 utterances to give a total of 500 utterances with an average duration of about 4 seconds to give a total of 30 minutes of test data.

Unlike the other phase estimation models, the minimum-phase model depends on the sinusoid amplitudes and frequencies to form an estimate of the phase as described in Section 8.2.3. As such, a number of additional experiments are carried out to determine the extent on which this model relies on accurate estimation of correct amplitude and frequency values. Table 8.1 displays the range of configurations considered. In the case of amplitude and $f0$ 'REF' relates to parameters obtained from clean speech whilst 'MAP' denotes that parameters have been estimated from noisy speech using the speaker independent MAP estimation techniques described in Chapter 5 and 6. The zero and random phase models are not functions of the original speech and so the output of these models is constant across each of the

configurations listed in Table 8.1.

This section begins by presenting results of objective tests measuring the overall quality of speech reconstructed using each of the phase models in Section 8.3.1. Next, the result of a listening test performed to determine the subjective preference of the systems is presented.

8.3.1 Objective results

This section presents the results of a set of experiments carried out to determine the objective quality of reconstructed speech using each of the phase models. It is therefore laid out as follows: Section 8.3.1.1 begins by presenting results of an experiment used to determine the relative performance of the phase models by reconstructing speech given reference amplitude and fundamental frequency values with phase values estimated using each of the models (i.e. the `MIN_REF_REF` configuration for the minimum-phase model). The sections which then follow relate to the other configurations of the minimum-phase model displayed in Table 8.1. Section 8.3.1.2 evaluates the reliance of the minimum phase model on accurate spectral amplitudes by comparing the `MIN_REF_REF` and `MIN_MAP_REF` configurations whilst the `MIN_REF_MAP` and `MIN_MAP_MAP` configurations are introduced in Section 8.3.1.3 to determine the effect of $f0$ estimation.

8.3.1.1 Effect of phase models on speech reconstruction from reference parameters

The experiments presented in this section examine the effect of phase estimation on the reconstruction of clean speech. As such, magnitude spectra, $f0$ and voicing were all extracted from clean speech with phase values obtained from each of the phase estimation models. The HNM reconstruction model was used to reconstruct speech as described in Section 3.3.3.

Five different sources of phase were considered. Phase values were extracted from

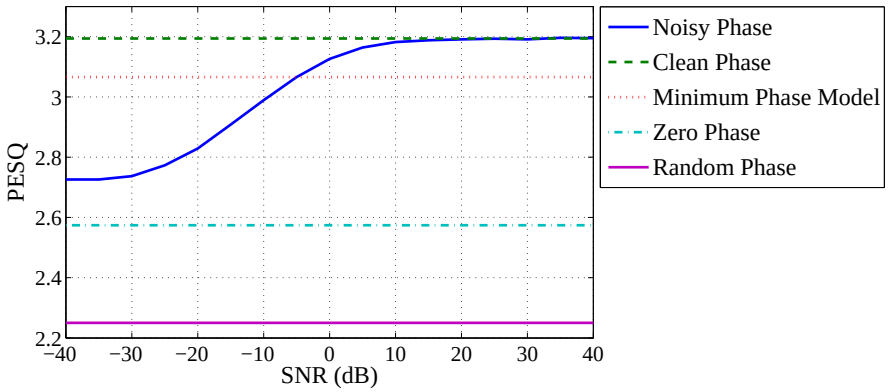
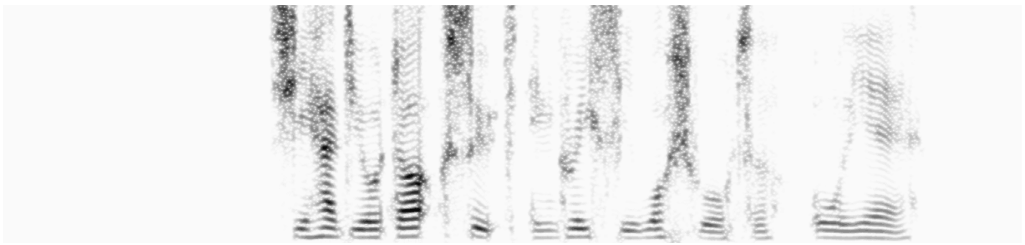


Figure 8.4: Comparison of the overall quality of speech reconstructed using a number of phase models as measured objectively using PESQ

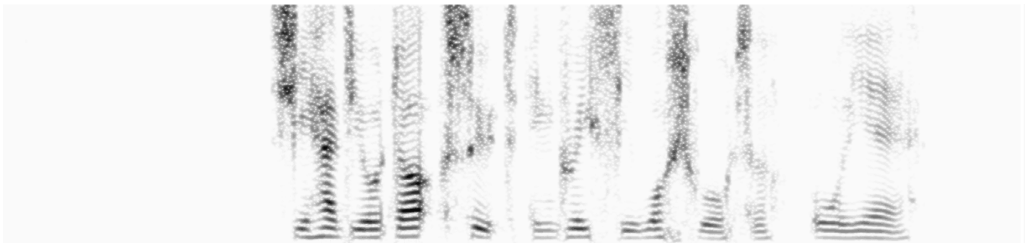
clean and noisy speech. In the case of noisy speech, destroyerops noise was added to clean speech at SNRs of between -40dB and +40dB before phase extraction. Three methods of artificial phase were also evaluated and include: the minimum phase model, zero phase and random phase. Figure 8.4 shows the objective quality of speech reconstructed using these methods, as measured using PESQ.

Comparing first the 'clean' phase with the phase extracted from noisy speech reveals that the two methods are equal when the SNR is $\geq$10dB. At lower SNRs the noisy phase is seen to reduce the quality of reconstructed speech at a rate consistent with the increase in noise level. At SNRs of $\leq$-30dB no further reduction in quality occurs. In conditions with an SNR of $\geq$10dB the local SNR of the harmonics are sufficiently high that any phase distortion is not noticeable. At lower SNRs harmonic phases are distorted due to large errors in the complex spectra introduced by the noise. This causes a more noise-like quality to the signal. In the case of -20dB SNR some of the original noise is audible in the reconstructed signal when using the noisy phase.

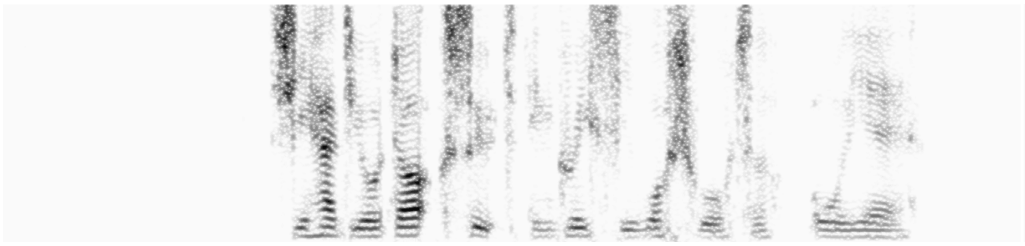
Figure 8.5 examines the effect of using 'noisy' phase values further by comparing reconstructions of the utterance *"She had your dark suit in greasy wash water all year"* using the clean phase and noisy phase at +20dB, 0dB and -20dB SNR. Speech reconstructed using noisy phase at 20dB SNR (Figure 8.5(b)) can be seen to be very



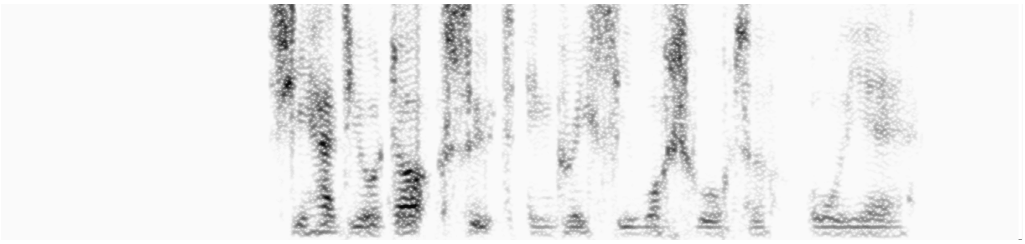
(a) Reconstruction from reference parameters



(b) Reconstructed using phase from signal with destroyerops noise added at +20dB SNR



(c) Reconstructed using phase from signal with destroyerops noise added at 0dB SNR

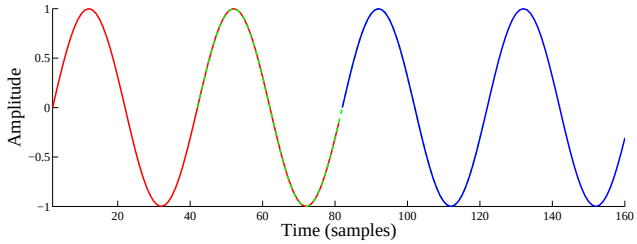


(d) Reconstructed using phase from signal with destroyerops noise added at -20dB SNR

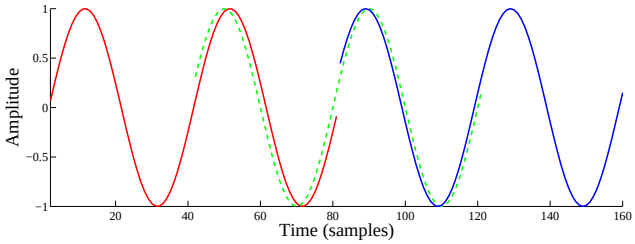
Figure 8.5: Spectorgrams showing the effect of noisy phase on speech by reconstructing clean speech using the HNM using phase extracted from the same utterance corrupted by destroyerops noise at SNRs of 20dB, 0dB and -20dB

similar to the clean reconstruction (Figure 8.5(a)). At 0dB SNR a slight amount of noise can be seen around harmonics whilst at -20dB SNR the harmonic structure has been significantly distorted (Figure 8.5(d)).

This phenomenon is now examined in more detail for the case of a single sinusoid. A sinusoid with constant amplitude and frequency was generated using a frame-



(a) $e_{ph} = 0$



(b) $e_{ph} = 4$

Figure 8.6: Time-domain plot of sinusoid frames showing no phase error (a) and an error of $e_{ph} = \frac{\pi}{4}$ (b) for a sinusoid with constant amplitude and $f = 200Hz$

based approach as per the HNM. The amplitude and frequency components were kept constant between frames with phase offsets tracked between frames. Errors were introduced as random additive component to the phase in the range $-\frac{\pi}{\gamma} \leq 0 \leq \frac{\pi}{\gamma}$ where:

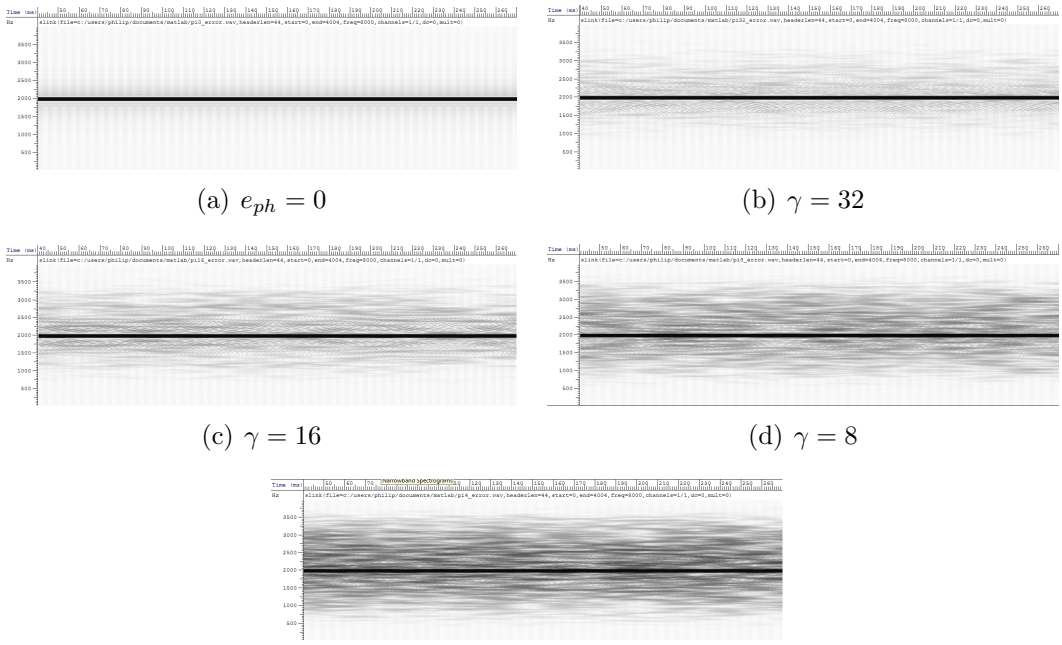
$$e_{ph} = \begin{cases} 0 & \text{if } \gamma = 0 \\ -\frac{\pi}{\gamma} \leq 0 \leq \frac{\pi}{\gamma} & \text{else} \end{cases}, \tag{8.10}$$

and is applied to the reconstruction model as:

$$x(m) = \sin\left(2\pi f m + \phi + e_{ph}\right), \tag{8.11}$$

where $x(m)$ is the mth sample of the output signal, f is the sinusoid frequency and ϕ is the original phase.

Figure 8.6 shows overlapping frames in the time-domain before overlap and add for the case of $e_{ph} = 0$ and $\gamma = 4$. Significant time-offsets are observed between frames in the case of $\gamma = 4$. Whilst overlap and add will average out the effect



(a) $e_{ph} = 0$ (b) $\gamma = 32$

(c) $\gamma = 16$ (d) $\gamma = 8$

(e) $\gamma = 4$

Figure 8.7: Narrowband spectrograms of reconstructed sinusoid signal showing the effect of phase errors in the frequency domain for a sinusoid with constant amplitude and $f = 2kHz$

of these discontinuities to some extent, artifacts are likely to remain in frequency domain analysis of the signal and are visible as additional frequency components similar to those in Figure 8.3. Next, we examine the narrowband spectrograms of reconstructed signals with varying phase errors. Figure 8.7 shows the effect of phase errors in the frequency domain. When $e_{ph} = 0$ no phase errors are introduced. As the error is increased noise begins to become visible around the sinusoid frequency. Whilst there are visible artifacts in all but the reference case, no artifacts are easily audible until $\gamma = 8$, supporting the claim made in Loizou [2007] that phase errors only begin to become perceivable when the error reaches a threshold of between $\frac{\pi}{8}$ and $\frac{\pi}{4}$.

The degradation in objective quality displayed in Figure 8.5 can therefore be attributed to phase distortions caused by noise introducing uncertainty as to the exact time-position of the sinusoids causing a noise-like distortion around harmonics.

Considering now the three artificial phase models the minimum phase model is

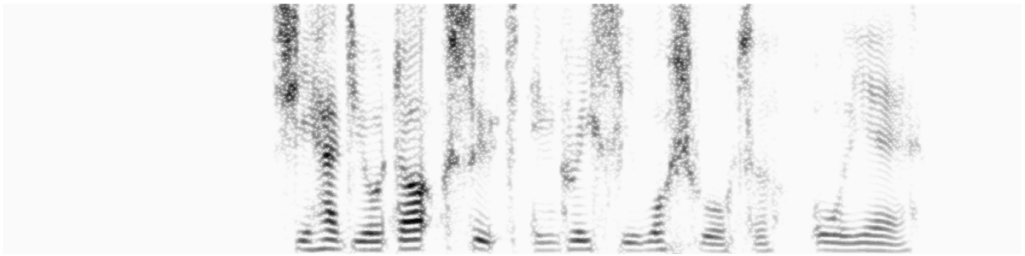
shown to be the most realistic alternative to the noisy phase in the results shown in Figure 8.4. Whilst the minimum phase model reduces the quality of reconstructed speech compared to using the clean phase, at SNRs of $\leq -5dB$ the minimum phase model provides speech of higher quality than when reconstructed using the phase of the noisy speech. Both the zero and random-phase models are shown to degrade overall quality significantly, below that even of the noisy phase at -40dB SNR.

Figure 8.8 now compares spectrograms of speech reconstructed using each of the three artificial phase models to speech reconstructed using the original (clean speech) phase. Speech reconstructed using the minimum-phase model is shown to be remarkably similar to that reconstructed using the original phase. No inter-harmonic noise present in the minimum-phase reconstruction. Whilst this may seem to be an appealing quality, this effect is actually found to reduce the naturalness of the speech by introducing a 'buzziness' to the signal. Moving on to the zero-phase model, the harmonic structure is shown to be have been significantly degraded, especially at high frequencies. Speech reconstructed using the random-phase model is shown to have no remaining harmonic structure with all frames essentially reconstructed as unvoiced.

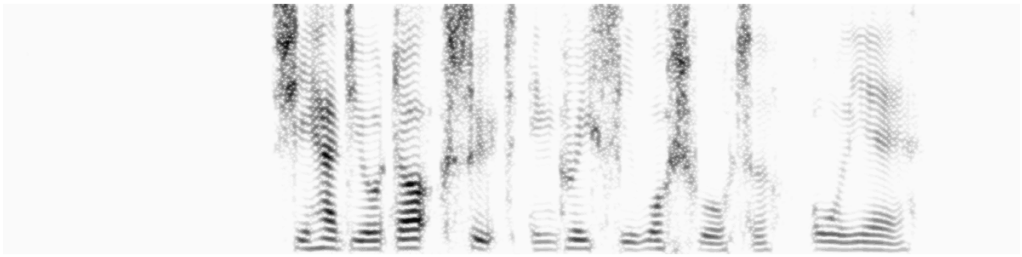
Although the minimum phase model has been shown to offer a credible alternative to the noisy phase, all of the results presented in this section estimate the phase from the spectral envelope of clean speech. Only an estimate of the clean spectral envelope will be available in the final system and so the effect of using spectral envelope estimated from noisy speech for phase estimation using the minimum-phase model is examined in Section 8.3.1.2.

8.3.1.2 Effect of spectral envelope estimation on the minimum phase model

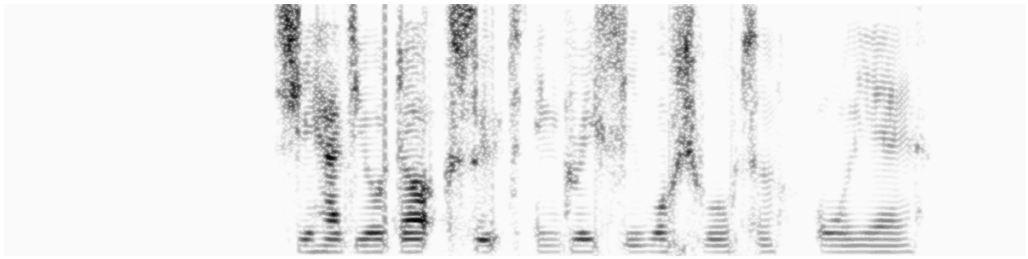
This section examines the effect of using the minimum phase model with estimated spectral amplitudes. F0 and voicing are estimated from the clean speech while the spectral amplitudes are sampled from the speaker-independent, speaker-adapted,



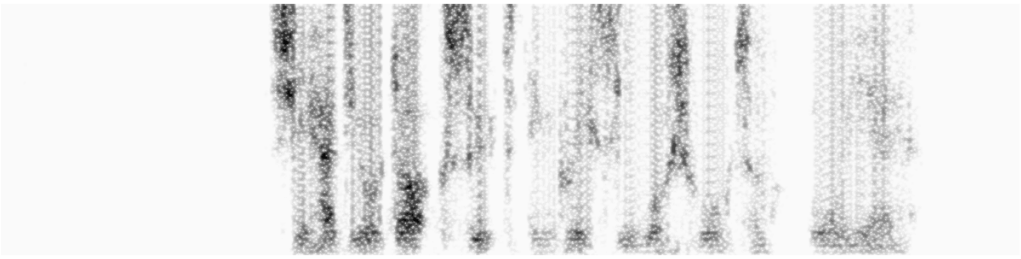
(a) Reconstruction from reference parameters



(b) Reconstructed using phase from minimum-phase model



(c) Reconstructed using phase from zero-phase model



(d) Reconstructed using phase from random-phase model

Figure 8.8: Comparison of narrowband spectrograms of utterance reconstructed using artificial phase models

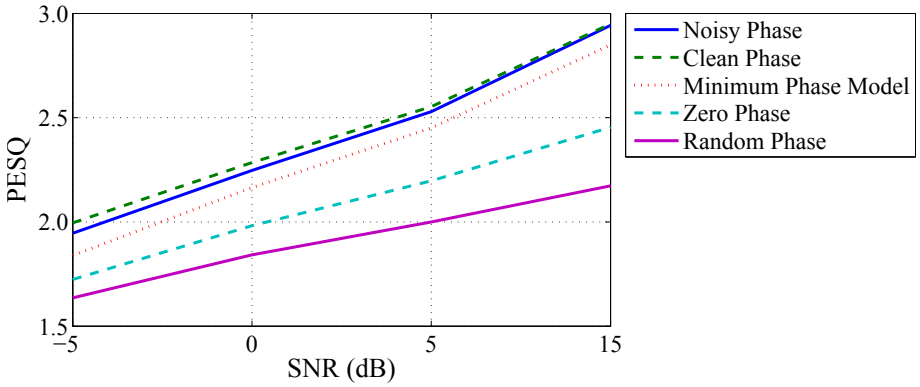


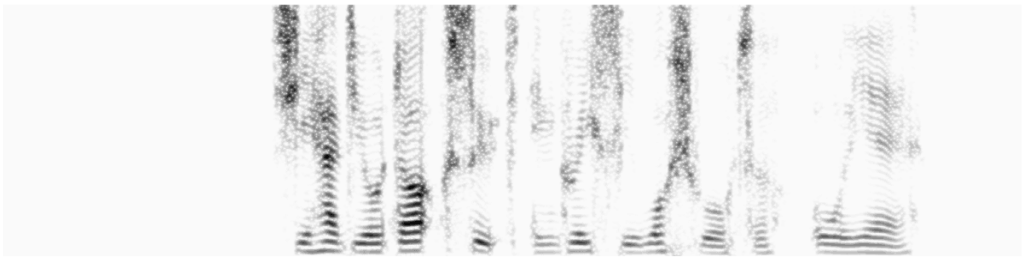
Figure 8.9: Objective quality of speech reconstructed using spectral envelope estimated from noisy speech, reference $f0$ and voicing and a range of phase models

MAP estimate of the spectral envelope described in Chapter 5. This relates to the `MIN_MAP_REF` configuration from Table 8.1. Figure 8.9 shows objective quality, measured with PESQ, using this configuration with a range of phase models. Objective quality has been significantly reduced for all methods when compared to results using spectral amplitudes from clean speech in Figure 8.4. SNRs in this test range from -5dB to +15dB SNR, reflecting the realistic operating range of the final system.

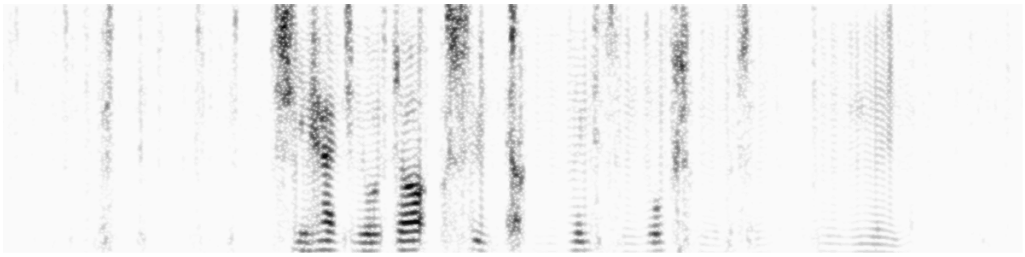
The zero and random phase models are still shown to offer the worst performance out of the models, with the clean phase providing benchmark performance. Despite performance of the minimum phase model matching that of the noisy phase at -5dB SNR in the previous section, when using spectral features estimated from noisy speech the performance of the minimum-phase model degrades significantly. This leaves noisy phase as the best realisable method of phase estimation for these conditions, closely tracking the performance of the clean phase.

Figure 8.10 compares spectrograms of reconstructions of the utterance *"She had your dark suit in greasy wash water all year"*, spoken by a male speaker, using the phase estimated from clean speech, noisy speech and the minimum-phase model at 0dB SNR. All utterances used estimated spectral amplitudes and reference $f0$ and voicing as per the rest of the experiments in this section.

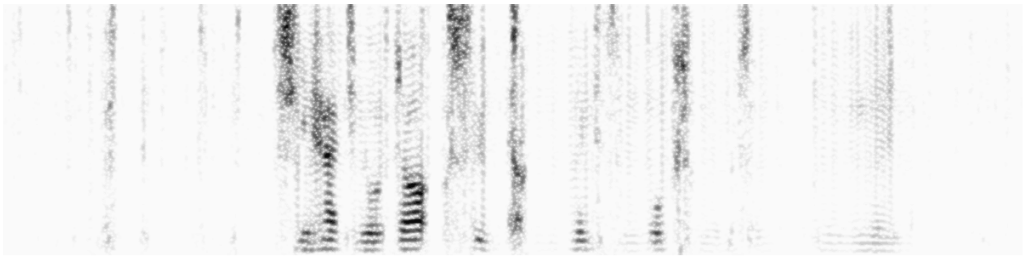
Distortions introduced by the spectral envelope estimation are easily visible across all examples when compared to the reference reconstruction (Figure 8.10(a)). In



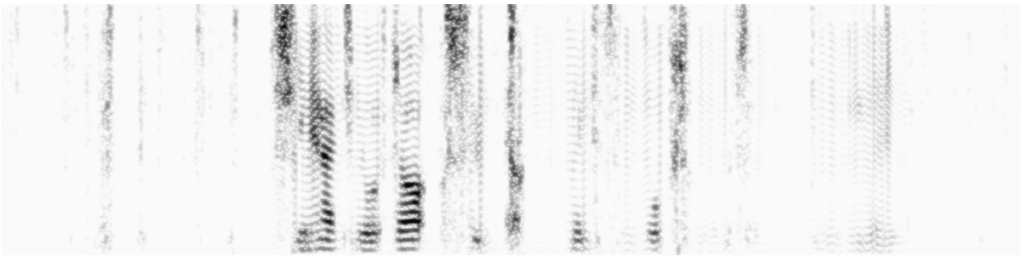
(a) Clean phase and reference spectral amplitudes



(b) Clean phase and estimated spectral amplitudes



(c) Noisy phase and estimated spectral amplitudes



(d) Minimum phase and estimated spectral amplitudes

Figure 8.10: Narrowband spectrograms comparing the effect of using the minimum phase model with spectral amplitudes estimated from speech at 0dB SNR and reference f0/voicing

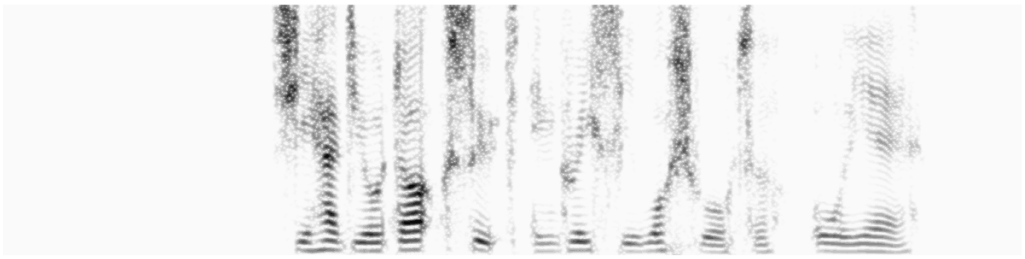
terms of variation between the examples reconstructed from estimated spectral amplitudes very few differences are immediately visible with the only significant difference located during the first segment of voiced speech where, as with the previous example (Figure 8.9, Section 8.3.1.2), the inter-harmonic noise has been reduced at the expense of reducing the naturalness of the speech.

We may therefore conclude in this section that the distortions introduced by the spectral amplitude estimation stage also degrade the quality of the phase estimate produced by the minimum-phase model to below that of the noisy phase. The next section continues this investigation by examining the effect errors in $f0$ have on the phase of reconstructed speech.

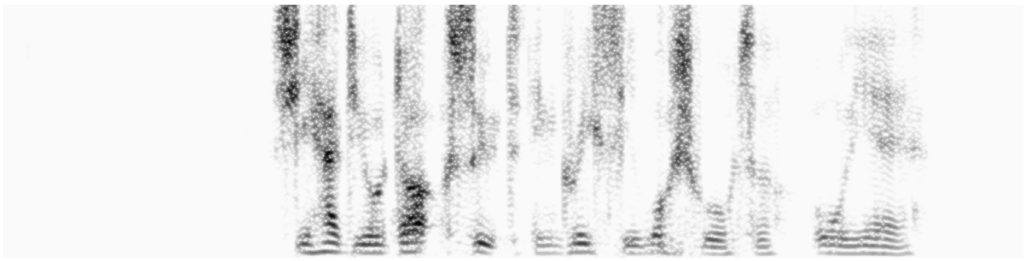
8.3.1.3 Effect of fundamental frequency estimation on the minimum phase model

This section examines whether errors in $f0$ caused by the estimation process, as described Chapter 6, affect the phase of reconstructed speech using a range of phase models. The results of two experiments are presented; those using reference spectral envelope, estimated $f0$ and reference voicing (`MIN_REF_MAP`) as well as those using using estimated spectral envelope and $f0$ and reference voicing (`MIN_MAP_MAP`). Voicing classification is kept at reference values in all cases as we are interested in the effect of $f0$ errors rather than the combined effect of $f0$ and voicing errors.

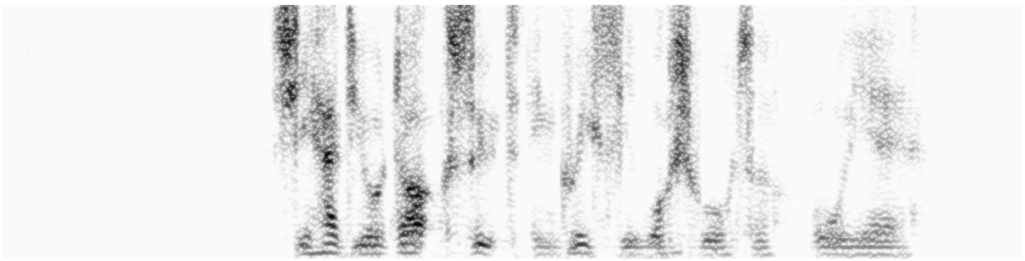
First, we consider the case of using estimated $f0$ and voicing with clean spectral envelope to see the effect $f0$ has on phase. Given a sufficiently high error in fundamental frequency it may be beneficial to use a model that better tracks the new harmonic trajectories. Figure 8.11 compares spectrograms of the utterance *"She had your dark suit in greasy wash water all year"* reconstructed using $f0$ estimated from noisy speech at 0dB SNR of destroyerops noise and phase from clean speech, noisy speech and the minimum-phase model. The relative $f0$ error of this utterance is 9.95% with a mean absolute error of 13.83Hz. Figure 8.11(b) shows the result of using clean phase spectra. When compared to the reference reconstruction (Fig-



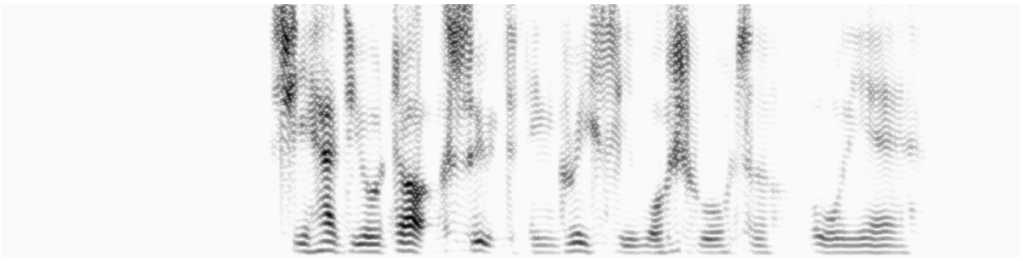
(a) Reconstructed using reference acoustic features



(b) Reconstructed using $f0$ estimated from noisy speech at 0dB SNR and clean phase



(c) Reconstructed using $f0$ and phase estimated from noisy speech 0dB



(d) Reconstructed using $f0$ estimated from noisy speech at 0dB SNR and minimum phase

Figure 8.11: Narrowband spectograms illustrating the effect of noisy phase on speech reconstruction using estimated $f0$ and clean spectral envelope

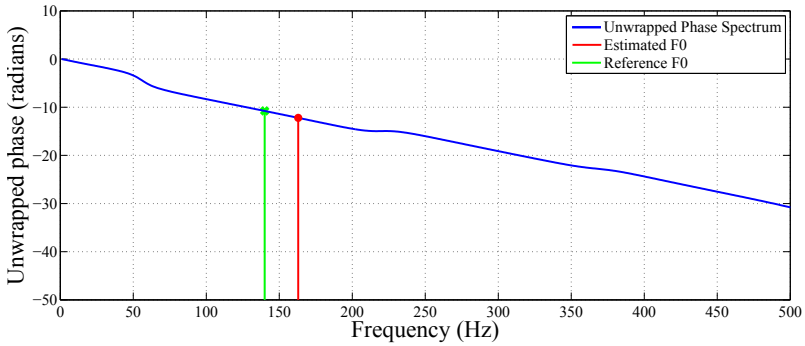


Figure 8.12: Example of incorrectly sampling phase value

ure 8.11(a)), significant distortion is observed around harmonics, especially those in the mid to high frequency regions. A similar effect is observed in Figure 8.11(c) when using the noisy phase, though more artifacts are visible due to the noise already introducing phase errors. In the case of both the clean and noisy phases these distortions occur when errors in fundamental frequency cause the incorrect phase value to be sampled from the phase spectrum. Examining now the speech reconstructed using the minimum-phase model in Figure 8.11(d), a considerable improvement in harmonic tracking is noticeable, though as with the example in Figure 8.10(d), this model introduces a 'buzziness' to the reconstructed signal which reduces the naturalness of the speech.

The relationship between errors in fundamental frequency estimation and phase is now examined. Figure 8.12 shows an example of an error in $f0$ causing the wrong phase value to be sampled. In this case, $f0 = 140Hz$ but a $f0$ error of $23Hz$ has caused the phase value at $163Hz$ to be sampled instead. Comparing the two sampled values shows a phase error of 1.43 radians to have occurred, or $\approx \frac{\pi}{2}$. Figure 8.13 illustrates the relationship between phase and $f0$ errors across a large number of frames (> 60000). As expected, there is a strong linear relationship shown between the two errors. Annotations have been included on the graph to show the points relating to the examples in Figure 8.7. The range between $e_{ph} = \frac{\pi}{8}$ and $e_{ph} = \frac{\pi}{4}$ is the range at which errors due to the phase are assumed to become audible [Loizou, 2007]. This gives a range of between $4 - 14Hz$ for which phase sampling errors caused by

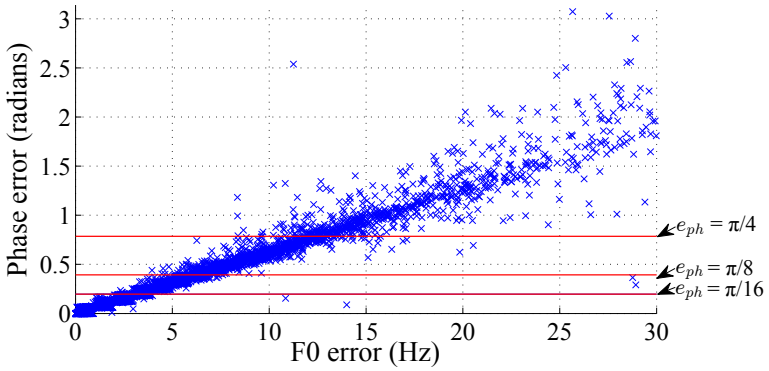


Figure 8.13: Demonstration of the relationship between $f0$ error and phase error

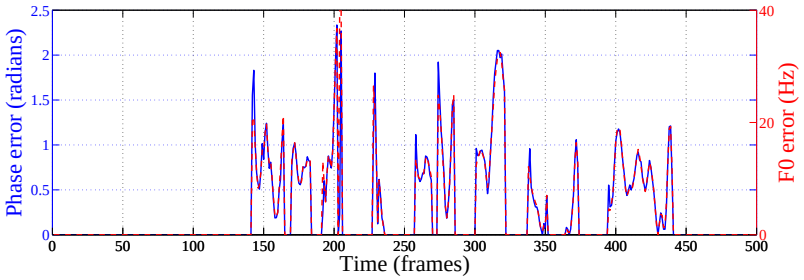


Figure 8.14: Relationship between $f0$ and (clean) phase errors across a single utterance at 0dB SNR destroyerops

$f0$ will begin to be perceivable to human listeners. Figure 8.14 now demonstrates the relationship between $f0$ error and the corresponding phase error for the first harmonic across a single utterance (the same utterance as in Figure 8.11). Here, the linear relationship is clear with an exact mapping between the magnitude of $f0$ error and phase error highlighting the importance of accurate $f0$ estimation on phase error.

The effect of phase sampling errors caused by fundamental frequency estimation is displayed in Figure 8.15 as a function of the SNR from which the fundamental frequency was estimated. Fundamental frequency was estimated from the noisy speech using the speaker independent, speaker adapted system from Chapter 6. This should give an idea of the real-world consequences of the relationship shown in Figures 8.13 and 8.14. The phase error was measured for the first harmonic across a range of utterances from five male and five female speakers in three types of noise:

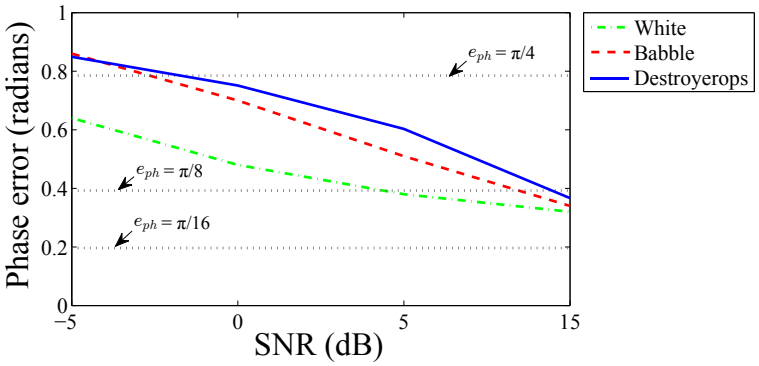


Figure 8.15: Effect of SNR on average phase error for 1st harmonic of voiced frames

white, babble and destroyerops.

At 15dB SNR no perceptually relevant phase errors occur, whilst at -5dB SNR phase errors are likely to be noticeable in destroyerops and babble noises. All other points fall between $\frac{\pi}{8}$ and $\frac{\pi}{4}$ and so are in the range at which they may start to become noticable [Loizou, 2007]. This suggests that, perceptually, the impact of phase errors for the first harmonic should be fairly limited in most cases.

Figure 8.16 now examines the effect of phase errors caused by errors in $f0$ on the 6th harmonic. In this case, all errors are significantly above $\frac{\pi}{4}$ suggesting that harmonics within the frequency ranges covered by this test (550Hz to 750Hz) will be significantly distorted. Examining these regions in Figures 8.11(b) and 8.11(c) shows this to be the case. Whilst phase errors are higher the energy of the 6th harmonic is typically lower than that of the first harmonic and so these errors may be less perceivable to human listeners.

We now move on to the results of objectively measuring the quality of this configuration, displayed in Figure 8.17. Despite the minimum phase model tracking the modified harmonic trajectories more effectively, the objective quality is measured to be worse than both the clean and noisy phases. This can be attributed to the 'buzzy' timbre of the reconstructed signal reducing the naturalness of the speech.

Finally, we replace the spectral envelope with the estimated spectral envelope giving the results displayed in Figure 8.18. This confirms the results in Section 8.3.1.2

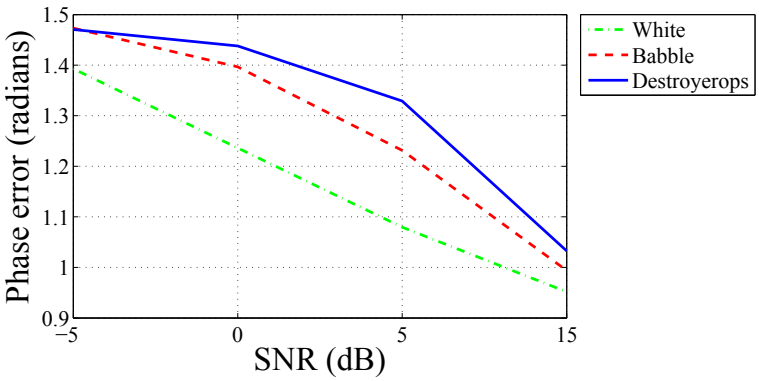


Figure 8.16: Effect of SNR on average phase error for 6th harmonic of voiced frames

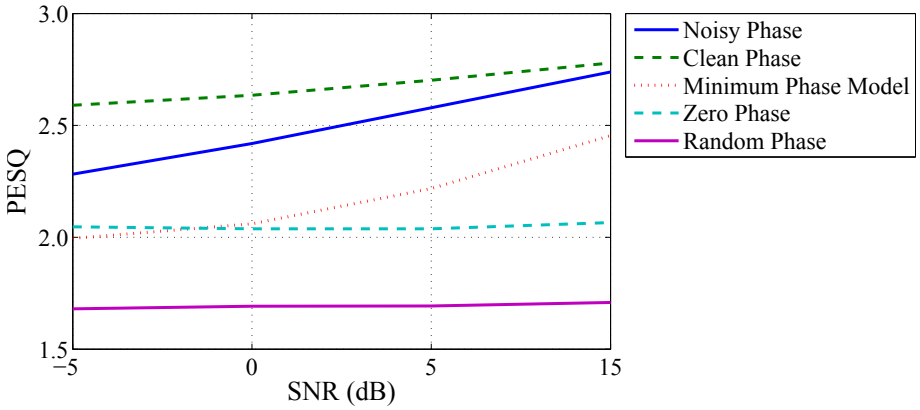


Figure 8.17: Objective quality of speech reconstructed using clean spectral envelope, estimated $f0$ and voicing and a range of phase models

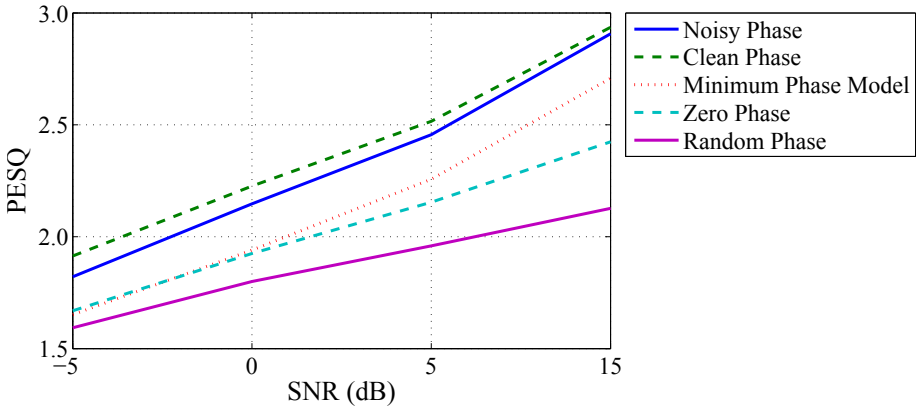


Figure 8.18: Objective quality of speech reconstructed using spectral envelope, $f0$ and voicing estimated from noisy speech and a range of phase models

which show that the minimum phase model is degraded further by the estimated spectral envelope.

Whilst PESQ shows using the minimum-phase model results in speech of worse quality than obtained using the noisy phase further testing is required to determine the final preferred system. PESQ has been shown to correlate strongly with subjective quality, however no study is known to have taken place as to its sensitivity to significant differences in phase. For this reason a comparative mean opinion score (CMOS) test is carried out in the following section as a means of determining subjectively the preferred system.

8.3.2 Subjective results

This section presents results of a set of comparative mean opinion score (CMOS) listening tests carried out to determine the subjective performance of the phase models considered for use in this work. Building on Section 8.3.1, which presented results of experiments measuring the objective quality of speech when reconstructed using each of the phase models across a range of configurations, the results presented in this section serve to determine the preferred system for use in this work.

Results in Section 8.3.1 showed the best two realisable systems to be the minimum-phase model and the phase of the noisy speech. In this section we therefore only

consider these two models alongside the phase of clean speech to provide benchmark of optimal performance.

The CMOS test was decided over the traditional MOS as we are most interested in the subjective preference between the systems rather than the overall quality of each system. 20 listeners participated in the listening tests. Results were obtained in accordance with Annex E of the ITU-T Recommendation P.800 [ITU-T, 1996]. For each configuration the listeners were presented with a 'reference' utterance which they were asked to compare to the 'assessed' utterance which was then played. They were then asked to rate the quality of the assessed utterance using the reference utterance as the baseline using a seven-point comparison category rating (CCR) as described in Section 2.6.1.1.

For this work there are four scenarios which we are interested in. These are shown in Table 8.1, though in this section we will not be considering the case of reference spectral envelope and $f0$ (`MIN_REF_REF`). For each scenario there are then three further configurations which will be examined: clean phase vs. noisy phase, clean phase vs. minimum-phase and minimum-phase vs. noisy phase. This section is split into subsections evaluating each scenario. Section 8.3.2.1 examines the effect of using estimated spectral envelope with reference $f0$ (`MIN_MAP_REF`). Next, the effect of using reference spectral envelope with estimated $f0$ is shown in Section 8.3.2.2 (`MIN_REF_MAP`). Finally, results of testing a system using spectral envelope and $f0$ estimated from noisy speech are presented in Section 8.3.2.3 (`MIN_MAP_MAP`).

8.3.2.1 Effect of using estimated spectral envelope

This section presents results of comparing speech using each phase model where speech has been reconstructed using spectral envelope estimated from noisy speech and $f0$ from clean speech. Objective quality results presented in Section 8.3.1.2 showed that performance of the minimum-phase model degraded with the use of estimated spectral envelope to below that of the noisy phase. Subjective results are presented in Figure 8.19. Due to the small sample size (20 listeners) some

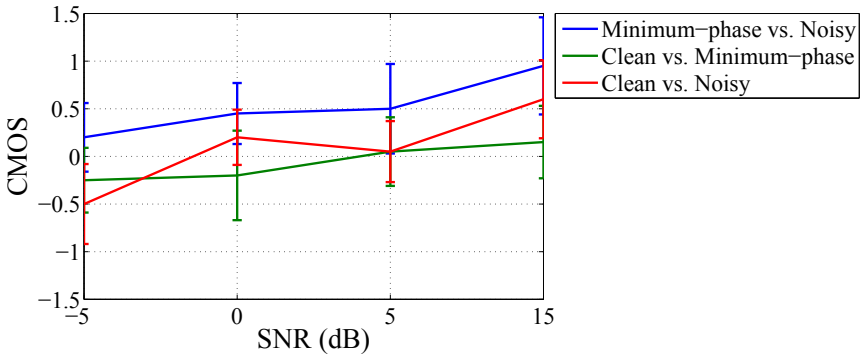


Figure 8.19: CMOS results of using estimated spectral envelope and reference $f0$ (`MIN_MAP_REF`). Error bars show confidence intervals at a significance level of $p = 0.05$. Negative values indicate a preference to the 'reference' configuration (first listed).

experimental error is expected in the results and is reflected by the size of the error bars which represent the 95% confidence interval. Despite this, general trends are visible within the results. Comparing first the use of clean phase to the noisy phase, results are consistently within half a category of 'About the Same', though surprisingly there is a slight tendancy to prefer the noisy phase at -5dB SNR which is attributed to experimental error resulting from the small sample size. Comparing the clean phase to the minimum-phase model shows a very slight preference towards the clean phase at all but 15dB SNR. This is also the case when comparing the noisy phase to the minimum-phase model with the noisy phase showing a slight preference.

These results reflect the objective results presented in Section 8.3.1.2, though the results are still relatively close (within one category in all cases).

8.3.2.2 Effect of using estimated $f0$

In this section we examine the effect of using $f0$ estimated from noisy speech with spectral envelope from clean speech. Objective results in Section 8.3.1.3 showed the noisy phase to again outperform the minimum-phase model across all SNRs. Subjective results are presented in Figure 8.20 and would appear to mirror the objective results. There is a strong preference towards the clean and noisy phases when compared to the minimum-phase model. There is also a preference towards

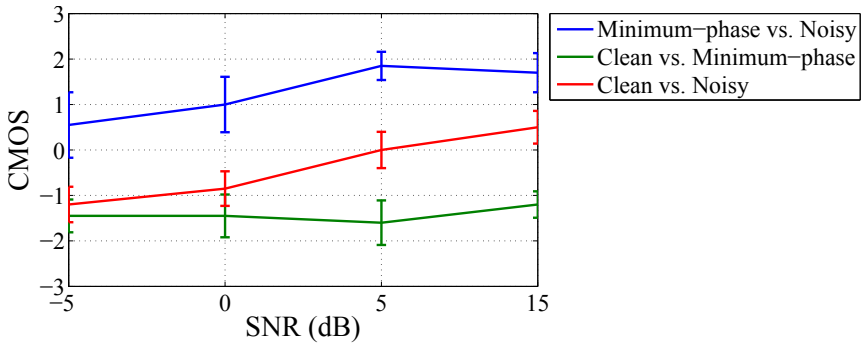


Figure 8.20: CMOS results of using reference spectral envelope and estimated $f0$ (`MIN_REF_MAP`). Error bars show confidence intervals at a significance level of $p = 0.05$ Negative values indicate a preference to the 'reference' configuration (first listed).

the clean phase when compared to the noisy phase, with the degree of preference correlating strongly with SNR.

8.3.2.3 Realisable system

This section now evaluates the phase models with realisable parameters of both spectral envelope and $f0$, with both parameters estimated from noisy speech. As with the experiments examining spectral envelope and $f0$ separately, objective results have shown the noisy phase to be preferable over the minimum-phase model. Results of subjective testing are displayed in Figure 8.21. As with the results in Section 8.3.2.2 examining $f0$, the clean phase is preferred over the noisy phase with the degree of preference linked to the SNR. Across all SNR the noisy and clean phases are preferred to the minimum-phase model, however at low SNR it is interesting to note that in both cases the difference between the 'original' signal phases and the minimum-phase model actually reduces to within half a category rating suggesting that other degradations in the signal are more prominent, masking the effect of the phase models to the listeners. At -5dB and 0dB SNR the differences between the noisy phase and the minimum phase model are between 'About the Same' and 'Slightly Better' in favour of the noisy phase. At higher SNR a larger difference is observed with results falling between 'Slightly Better' and 'Better'.

The noisy phase is shown to be the preferred system and so the noisy phase will

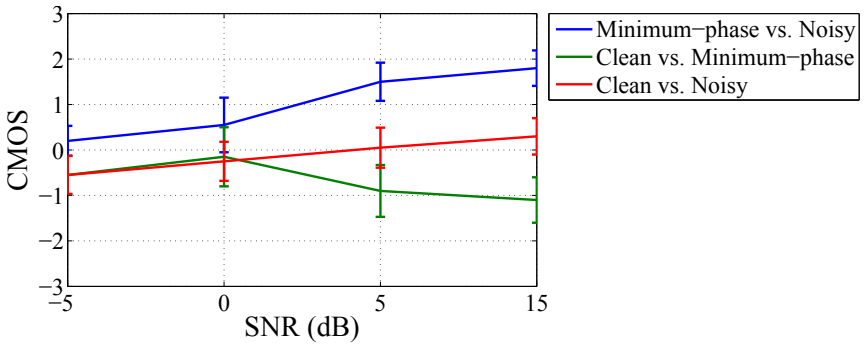


Figure 8.21: CMOS results of using estimated spectral envelope and $f0$ (`MIN_MAP_MAP`). Error bars show confidence intervals at a significance level of $p = 0.05$. Negative values indicate a preference to the 'reference' configuration (first listed).

be used in this method of speech enhancement.

8.4 Summary

This chapter has reviewed a range of phase models for use in a speech enhancement system. Through the use of objective and subjective tests, the phase of the noisy speech was found to be the optimal estimate of the clean-speech phase and so has been selected for use in this system.

Chapter 9

Speech Enhancement System

This chapter presents results of enhancement using the proposed method of speech enhancement. The system is driven by a set of acoustic features which are estimated from noisy speech using the methods of estimation previously described in this thesis. Two existing methods of enhancement which use the same acoustic features are also described for comparison purposes. First, a method of direct feature inversion and second, a method of model-based Wiener filtering. Performance is also compared to three methods of conventional speech enhancement, namely: spectral subtraction, Wiener filtering and log MMSE. Performance is evaluated in terms of subjective and objective quality.

Contents

9.1 Introduction

A method of speech enhancement by reconstruction has been proposed in this thesis. The aim of this chapter is to describe the implementation details of the proposed method and to subsequently measure the performance of the system against existing methods of enhancement.

The chapter begins with a description of the proposed method in Section 9.2. Chapter 2 described a number of existing methods of speech enhancement. Three of the described conventional methods of enhancement are tested, namely: spectral subtraction [Berouti et al., 1979], Wiener filtering (*a-priori* SNR) [Scalart et al., 1996] and log MMSE [Ephraim and Malah, 1985]. Performance of the proposed method is also compared against two more recent methods of enhancement: i.) a model-based Wiener filter as proposed by Hadir et al. [2011] and ii.) a method of MFCC feature inversion as proposed by Boucheron and Leon [2012]. MATLAB implementations of the three conventional methods of enhancement written by Loizou [2007] were used whilst in-house implementations of the model-based Wiener filter and direct feature inversion method were used.

Overall speech quality is measured objectively using PESQ whilst a listening test is also performed to give subjective results in terms of signal quality, background noise intrusiveness and overall quality.

9.2 Speech Enhancement System

This section describes implementation details of the tested methods of speech enhancement. The proposed method of speech enhancement is described in Section 9.2.1 whilst the two competing methods, the model-based Wiener filter and direct MFCC inversion method are described in Sections 9.2.3 and 9.2.2.

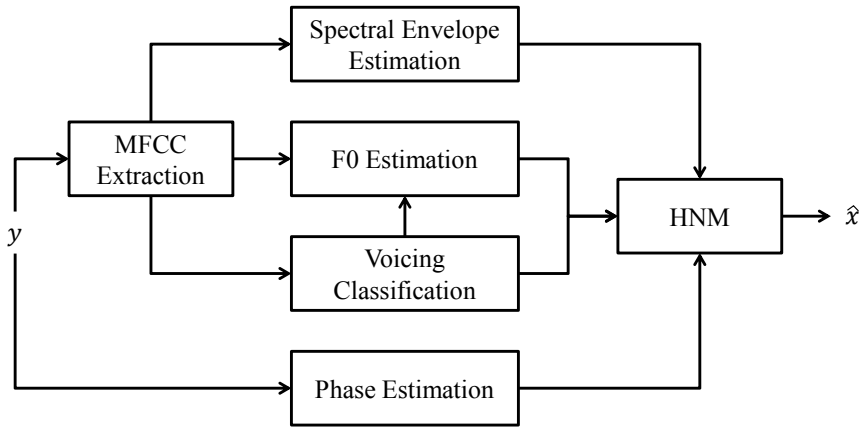


Figure 9.1: Diagram of proposed speech enhancement by reconstruction system

9.2.1 Proposed method of enhancement

The proposed method of speech enhancement by reconstruction is described in this section. The HNM is used to reconstruct cleaned speech and is driven by a set of four acoustic features. These are: spectral envelope, fundamental frequency, voicing classification and phase. Robust estimates of these acoustic features are made from the noisy speech using the methods of estimation described in this thesis to give the system illustrated in Figure 9.1. MFCC features are first extracted from the noisy speech. Estimates of the clean spectral envelope and fundamental frequency are made from these MFCC features using MAP estimation. Voicing classifications are made from the same MFCC features using a GMM-based system whilst phase is extracted directly from the noisy speech. These acoustic features are then used to drive the HNM reconstruction model as:

$$\hat{x}(m) = \sum_{l=1}^{L} |\hat{X}(l\hat{f}_0)| \cos(2\pi l \hat{f}_0 m + \theta_y(l\hat{f}_0)) + \hat{n}(m), \tag{9.1}$$

where $|\hat{\mathbf{X}}|$ is the estimated spectral envelope, $\hat{f}_0$ is an estimate of the fundamental frequency and $\theta_y(l\hat{f}_0)$ is the phase of the noisy speech sampled at the lth harmonic where L is the total number of harmonics in the frame, computed as $L = \lfloor \frac{8000/2}{\hat{f}_0} \rfloor$. $\hat{n}(m)$ represents filtered noise, derived from the estimated spectral envelope as de-

scribed in Section 3.3.3. Reconstructed frames of speech are combined using overlap and add.

9.2.2 Direct inversion

An approach of speech enhancement by MFCC feature inversion is now described [Boucheron and Leon, 2012]. Clean spectral amplitudes are computed as the pseudo-inverse of cleaned MFCC feature vectors (Section 3.4.2.2). By assuming sufficient source information is retained in the MFCC features the estimated spectral envelope may be used to directly reconstruct speech [Boucheron and De Leon, 2008]. A significant amount of source information has been shown to be present in these inverted features and so the harmonic structure of voiced frames is expected to be retained. The enhanced complex spectrum is therefore computed as:

$$\hat{X}(k) = |\hat{X}_M(k)|e^{j\angle Y(k)}, \tag{9.2}$$

where $|\hat{X}_M(k)|$ is the pseudo-inverse of the cleaned MFCC features and $\angle Y(k)$ is the phase of the noisy speech. This approach assumes the noisy phase is the optimal estimate of the clean phase (Wang and Lim [1982]; Loizou [2007]; Chapter 8). The cleaned complex spectrum, $\hat{X}(k)$, is then transformed to a time-domain waveform using an inverse DFT and combined with other frames using overlap and add.

9.2.3 Model-based Wiener filter

A model-based Wiener filtering approach to speech enhancement proposed by Hadir et al. [2011] is described in this section. As described in Section 2.2.2, noise is filtered from speech in the frequency domain using a Wiener filter as:

$$|\hat{X}(k)| = H(k)|Y(k)|, \tag{9.3}$$

where $|Y(k)|^2$ is the kth spectral bin of the power spectrum of the noisy speech and $H(k)$ is the Wiener filter, computed as:

$$H(k) = \frac{|X(k)|^2}{|Y(k)|^2}, \tag{9.4}$$

where $|X(k)|^2$ is the kth spectral bin of the power spectrum of the clean speech. Given an estimate of the clean power spectral envelope the model-based Wiener filter is computed as:

$$H(k) = \frac{|\hat{X}_M(k)|^2}{|Y_M(k)|^2}. \tag{9.5}$$

The estimate of the clean spectral envelope, $|\hat{X}_M(k)|^2$, is the pseudo-inverse of the cleaned MFCC vectors computed using the estimation system described in Chapter 5 whilst $|Y_M(k)|^2$ denotes the pseudo-inverse of MFCCs extracted from the noisy speech. The pseudo-inverse of the noisy power spectral envelope was used to preserve the fine spectral detail of the original signal as per Hadir et al. [2011]. The filtered magnitude spectrum, $|\hat{X}(k)|$, is then combined with the phase of the noisy signal to give the enhanced complex spectrum:

$$\hat{X}(k) = |\hat{X}(k)|e^{j\angle Y(k)}. \tag{9.6}$$

The final stage is to combine frames using overlap and add (Section 3.3.3.4).

9.3 Results

This section presents results of a series of experiments performed to determine the quality of speech produced by the proposed methods of speech enhancement. Performance is measured using subjective as well as objective testing. In all cases speaker-independent data from the WSJCAM0 corpus is used. The four acoustic features required for reconstruction were estimated as follows:

Spectral envelope A speaker-independent GMM was trained on clean speech and adapted for speaker variations using MAP adaptation. The Unscented Transform was then used to adapt for noise to give a model of the joint density of clean and noisy MFCC features (Chapter 5).

Fundamental frequency A GMM was trained on a joint feature of MFCCs extracted from voiced frames of clean speech and the corresponding fundamental frequency. Speaker independent data was used for training and so MAP adaptation was used for speaker adaptation. Speaker adaptation data consisted of the same format of joint feature as used for model training with fundamental frequency for adaptation obtained using PRAAT [Boersma, 2002]. Noise adaptation was once again achieved using the Unscented Transform (Chapter 6).

Voicing classification A GMM-based system using model adaptation to adapt for mismatches in speaker and noise was used for voicing classification. MAP adaptation was used for speaker adaptation whilst the Unscented Transform was used for noise adaptation (Chapter 7).

Phase The noisy phase was found to be best for reconstruction (Chapter 8).

In terms of noise adaptation data the statistics of the noise are assumed to be known in full *a-priori* whilst 120 seconds of clean speech from the target speaker is used for speaker adaptation. Whilst this level of information about the noise will rarely be available, the purpose of these experiments is to determine the optimal performance of the overall method of speech enhancement. The effect of using realistic estimates of the noise statistics is reported in the relevant chapters of this thesis.

Models were trained on a total of 24 hours of training data from 20 male and 20 female speakers. Data from ten additional speakers was used for testing with 50 utterances spoken by each speaker to give a total of 25 minutes of test data. Four different noise types are tested. White noise, babble noise and destroyerops noise

are assumed to be Gaussian and are tested at SNRs of -5, 0, 5 and 15dB. Machine gun noise, a highly non-stationary, non-Gaussian noise is tested at -20dB SNR to determine the performance of the SMC method of adaptation. Noise was artificially added to clean speech at the required SNRs. A sampling rate of 8kHz was used with frames of 20ms duration extracted at a rate of 100fps to give a 50% overlap.

A total of six methods of enhancement are tested in this section. Thee conventional methods are tested, namely: spectral subtraction [Berouti et al., 1979], Wiener filtering (*a-priori* SNR) [Scalart et al., 1996] and log MMSE [Ephraim and Malah, 1985]. The proposed system is also tested alongside two other state of the art methods and these are labelled as:

HNM (MAP) corresponds to the proposed reconstruction-based method of speech enhancement described in Section 9.2.1.

Wiener (MAP) corresponds to the model-based Wiener filter described in Section 9.2.3.

Direct (MAP) corresponds to the method of direct feature inversion described in Section 9.2.2.

Results begin with a measurement of objective quality in Section 9.3.1 before results of listening tests measuring subjective quality are then presented in Section 9.3.2. The effect of errors in fundamental frequency estimation on the quality of reconstructed speech is then examined in Section 9.3.3.

9.3.1 Objective quality measurement

Performance is first measured in terms of speech quality as measured objectively using PESQ. In the case of Gaussian noises the standard implementation of the UT was used for noise adaptation whilst the SMC variant of the UT was used for non Gaussian noises. This section is split into two further sections: first, performance

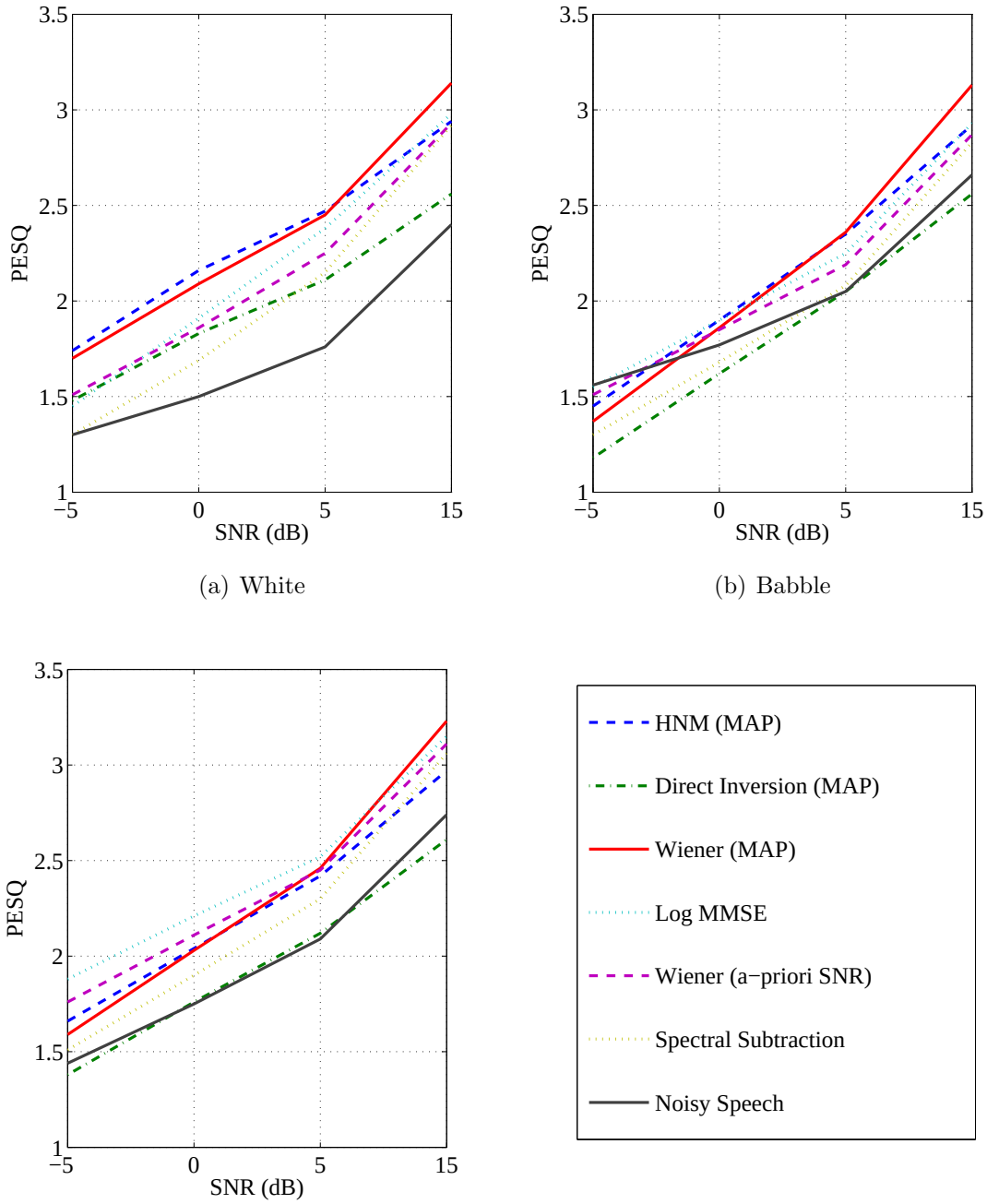
in Gaussian noises is reported in Section 9.3.1.1. Second, performance is measured in non-Gaussian noise in Section 9.3.1.2.

9.3.1.1 Performance in Gaussian noises

The result of objectively measuring the quality of enhancement of speech corrupted by Gaussian noises is presented in Figure 9.2. In the case of white noise all methods of speech enhancement are demonstrated to improve the quality of speech over the noisy speech. The HNM and model-based Wiener filter (Wiener (MAP)) are shown to perform best across all SNRs whilst the method of direct feature inversion using MAP estimated features is shown to perform relatively poorly with performance matching roughly that of the conventional methods of speech enhancement. Of the conventional methods of enhancement, log MMSE is shown to perform closest to the HNM and model-based Wiener systems.

Across all noises it is interesting to compare the relative performance of the Wiener (MAP) and HNM (MAP) methods of enhancement. At high SNR the Wiener (MAP) system performs best as the fine detail of the speech is preserved. As the SNR falls the relative performance of the HNM method improves to give best performance at $\leq$ 5dB SNR. This is attributed to the response of the Wiener filter which is relatively smooth across frequency due to the MFCC inversion process and across time due to the smoothing included in the spectral envelope estimation process. This results in inter-harmonic noise remaining in the signal whilst in some cases some of the speech signal has been removed due to over-smoothing. At the same SNRs the HNM model only reconstructs components thought to be related to the speech signal and so performs best.

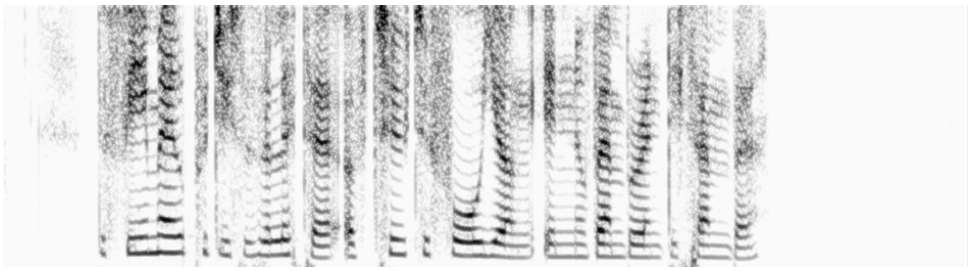
Figure 9.3 illustrates this effect using narrowband spectrograms of the utterance *"The female produces a litter of two to three young in November and December"* spoken by a female speaker and enhanced using the proposed method of enhancement (Figure 9.3(c)), the model-based Wiener filter (Figure 9.3(d)) and the method of direct feature inversion (Figure 9.3(e)). The Wiener (MAP) system clearly removes



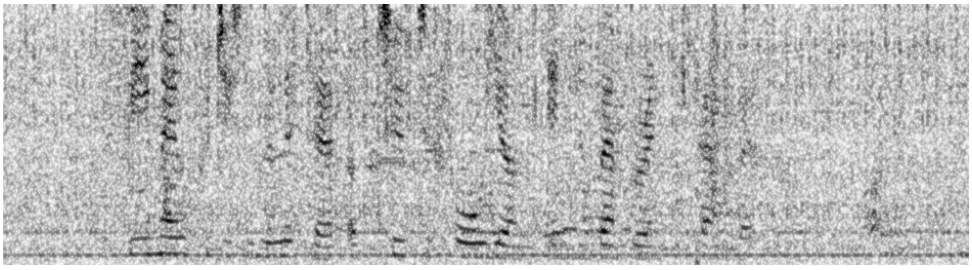
(a) White (b) Babble

(c) Destroyerops (d) Legend

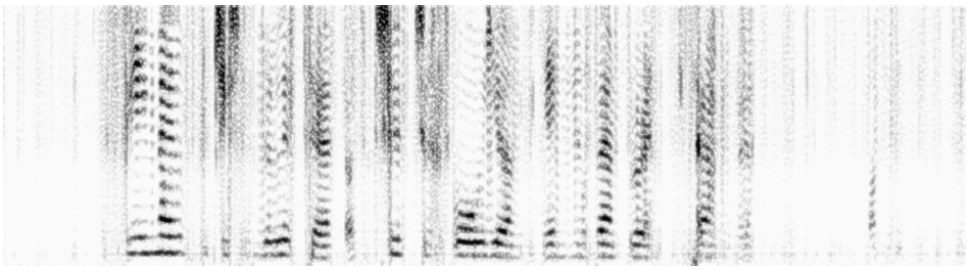
Figure 9.2: Objective quality of speech enhancement systems in three noises as measured using PESQ



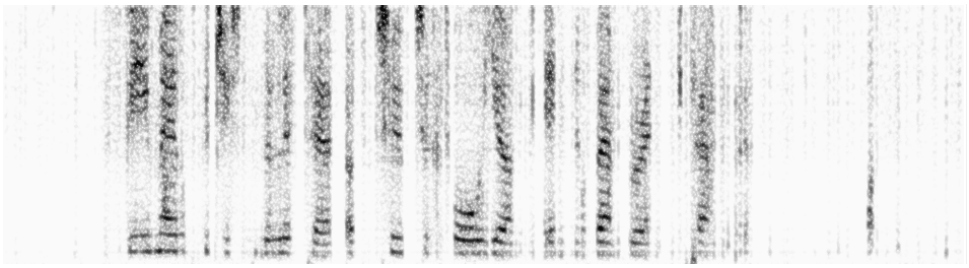
(a) Clean



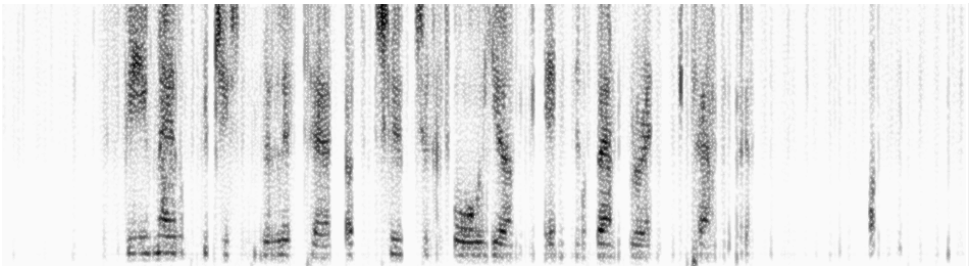
(b) Noisy (Destroyerops noise at 0dB SNR)



(c) HNM (MAP)



(d) Wiener (MAP)



(e) Direct (MAP)

Figure 9.3: Comparison of enhancement using HNM (MAP), Wiener (MAP) and Direct (MAP) systems in destroyerops noise at 0dB SNR

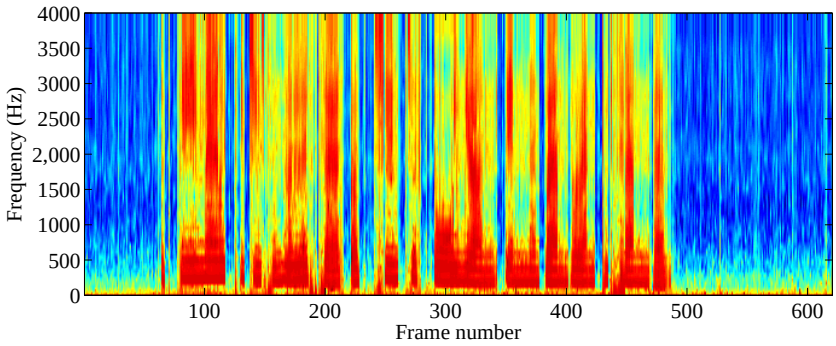


Figure 9.4: Log-spectral frequency response of Wiener (MAP) filter

a significant amount of the noise however some fine spectral detail has also been corrupted. In particular, low amplitude harmonics have been completely removed whilst almost all inter-harmonic noise remains. Figure 9.4 further illustrates this effect by showing the filter response of this system in the log-spectral domain. Relatively little attenuation occurs during periods of high speech energy (and therefore high local SNR). The filter is shown to be relatively smooth across frequency and time. More spectral detail relating to the original, clean, utterance is visible in the case of the HNM (MAP) system, with no inter-harmonic noise in the reconstructed signal.

The effect of errors in $f0$ estimation and voicing classification are shown to have very little effect on the reconstructed speech whilst a small amount of residual noise is present owing to inaccuracies in spectral envelope estimation. Whilst some source information is shown to have been retained in the Direct (MAP) system the resynthesised speech is heavily corrupted with significant amounts of inter-harmonic noise. This is attributed to smoothing in the spectral domain caused by extraction and subsequent pseudo-inversion of the MFCC features.

When enhancing speech corrupted by non-stationary noises, namely the babble and destroyerops noises, performance is shown to be worse than in the case of white noise. At high SNR the Wiener (MAP) and HNM (MAP) systems are shown to perform well, however as the level of noise increases the conventional methods are shown to offer best performance in some cases, with log MMSE performing best in

Table 9.1: Objective quality of enhancement systems in the presence of machine gun noise at -20dB SNR

System	Noise mixture components	PESQ
HNM (MAP)	1	1.85
	2	2.09
	3	2.12
Direct inversion (MAP)	1	1.62
	2	1.83
	3	1.97
Wiener (MAP)	1	1.97
	2	2.18
	3	2.21
Spectral subtraction	-	1.26
Wiener (*a-priori* SNR)	-	1.14
Log MMSE	-	1.20
Unprocessed	-	1.38

destroyerops noise. The reduced performance of the HNM (MAP) system in these conditions is attributed to $f0$ and voicing classification errors. In destroyerops noise the $f0$ error increases from 10.61% at 0dB SNR to 14.79% at -5dB SNR whilst voicing classification errors increase from 16.46% at 0dB to 25.80% at -5dB SNR.

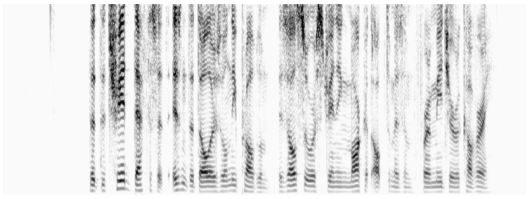
9.3.1.2 Performance in non-Gaussian noise

The case of non-Gaussian noise is now considered. Machine gun noise was added to clean speech at an SNR of -20dB. Speaker-independent models were adapted for speaker using MAP adaptation as per previous experiments whilst noise adaptation was this time performed using serial model combination (SMC). GMMs were trained from the known noise signal and used for adaptation as this configuration was previously found to perform best. The number of mixture components of the noise model were varied between 1 and 3, with three mixture components found to offer best performance in Chapter 5.

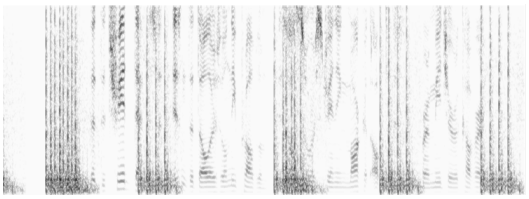
The result of objectively measuring the resulting speech quality after enhancement is presented in Table 9.1. Performance of the conventional methods of speech enhancement is shown to be very poor with all three methods reducing the overall

quality of speech. Of the three MAP-estimation based methods of enhancement the Wiener (MAP) is measured to offer best performance whilst the approach of direct feature inversion performing worst. The superior performance of the Wiener (MAP) system versus the HNM (MAP) system is attributed to the bursty nature of the machine gun noise. In periods of no-noise the frequency response of the Wiener filter will approach unity and so no attenuation or alteration of the original signal will take place. In the case of the HNM-based method of reconstruction a small reduction in quality is suffered due to the modelling error of the reconstruction process in clean conditions and this is thought to be responsible for the difference in performance between the two methods. In terms of signal quality, at times of high noise (gun shot) casual listening reveals better noise suppression in the case of the HNM (MAP) system. Not all of the noise is removed by filtering owing to spectral envelope estimation errors in the case of the Wiener (MAP) system whilst the HNM reconstruction model is unable to reconstruct the noise resulting in a 'cleaner' signal.

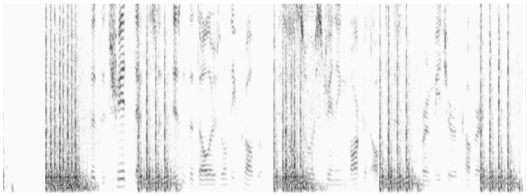
Figure 9.5 illustrates the result of enhancement using the three conventional and three estimation model-based methods for the utterance *"That the trade deficit isn't the dollar's only problem, it's also restraining market optimism for a major recovery"* spoken by a male speaker. Five bursts of machine gun fire are visible in the spectrogram of the noisy speech, each consisting of four shots (Figure 9.5(b)). In the case of the conventional methods of enhancement (Figures 9.5(c)-9.5(e)) no machine gun noise appears to have been suppressed. This is despite the *a-priori* SNR Wiener filter and log MMSE methods introducing distorting the signal; in both cases the first three harmonics have been completely removed. This is attributed to the noise estimation processes assuming the noise is constant across the utterance and filtering out the speech instead of the noise due to the relatively low energy of the speech versus the noise. Focusing now on the proposed methods of enhancement in Figures 9.5(f)-9.5(h), all three methods are shown to have completely removed the noise. Some distortion is apparent in the case of the Direct (MAP) system due to the feature inversion process whilst the Wiener (MAP) and HNM (MAP) systems are shown to provide a good reproduction of the clean signal. Whilst no machine



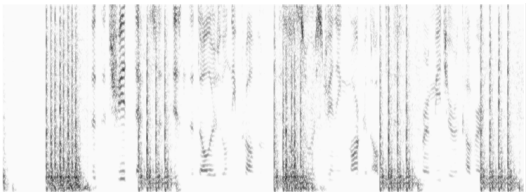
(a) Clean



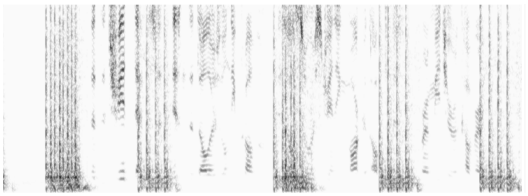
(b) Noisy (Machine gun noise at -20dB SNR)



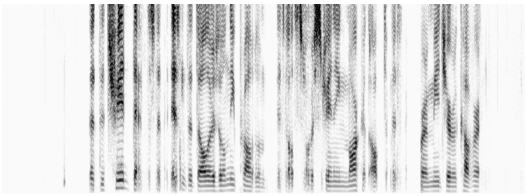
(c) Spectral Subtraction



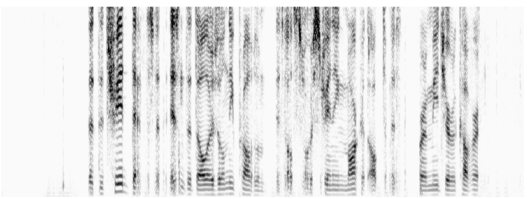
(d) Wiener (*a-priori* SNR)



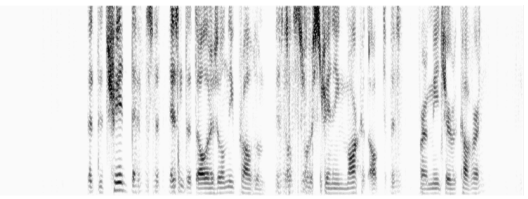
(e) Log MMSE



(f) Direct (MAP)



(g) Wiener (MAP)



(h) HNM (MAP)

Figure 9.5: Comparison of performance of speech enhancement methods in machine gun noise at -20dB SNR

gun noise is visible in either case some distortion is noticeable in periods where the noise has been removed. In a small number of frames the speech energy appears to have been removed completely which is thought to have occurred due to poor quality estimates of the spectral envelope in very high levels of noise.

9.3.2 Subjective quality measurement

Performance of the proposed method of speech enhancement is now measured subjectively using a series of three-way MOS tests. Twenty listeners participated in each test, with each listener hearing one example of each method at each noise and SNR. The listening tests were performed in a sound-proof room with utterances played through headphones. Tests were performed in accordance with the [ITU-T, 2003] recommendations and so a short familiarisation test was added to the start of each session.

Two listening tests were performed. The objective of the first was to determine the performance of the proposed reconstruction-based method of speech enhancement using speaker-dependent data whilst in the second speaker independent data was used. The mean of the MOS scores across listeners are presented with error bars denoting the 95% confidence level.

The results of the first test are now presented. In this test a single female speaker from the NuanceCatherine dataset was used for training and testing. 40 minutes of data were used for model training with a further 20 minutes used for testing. Utterances were randomly selected from the test set for the listening test. Car noise was added to speech at 20dB, 10dB and 5dB SNR. Speech with no added noise was also included in testing. The HNM was used to reconstruct speech using four acoustic feature configurations. Table 9.2 details the configuration of each system, where $|\mathbf{X}|$ denotes the clean spectral envelope and $|\hat{\boldsymbol{Y}}|$ denotes the spectral envelope of the original speech which may be either clean or noisy depending on the SNR. In the case that clean spectral envelope was estimated from noisy speech, MAP was used for estimation with speaker-dependent models trained in the same conditions

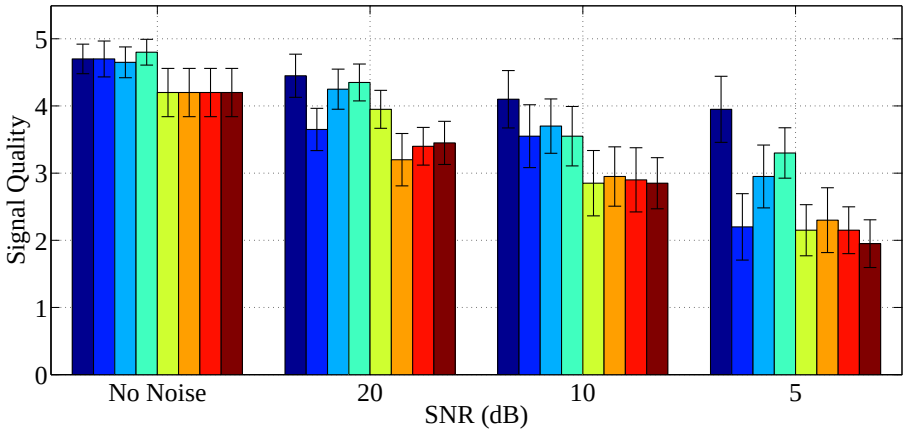
Table 9.2: System configurations for first listening test

Method	Spec. Env.	F0	Voicing	Phase
HNM($f0$, $\|\mathbf{Y}\|$)	Noisy	XAFE (Clean)	XAFE (Clean)	Noisy
HNM($\hat{f}0$, $\|\mathbf{Y}\|$)	Noisy	XAFE (Noisy)	XAFE (Noisy)	Noisy
HNM($f0$, $\|\hat{\boldsymbol{X}}\|$)	MAP	XAFE (Clean)	XAFE (Clean)	Noisy
HNM($\hat{f}0$, $\|\hat{\boldsymbol{X}}\|$)	MAP	XAFE (Noisy)	XAFE (Noisy)	Noisy

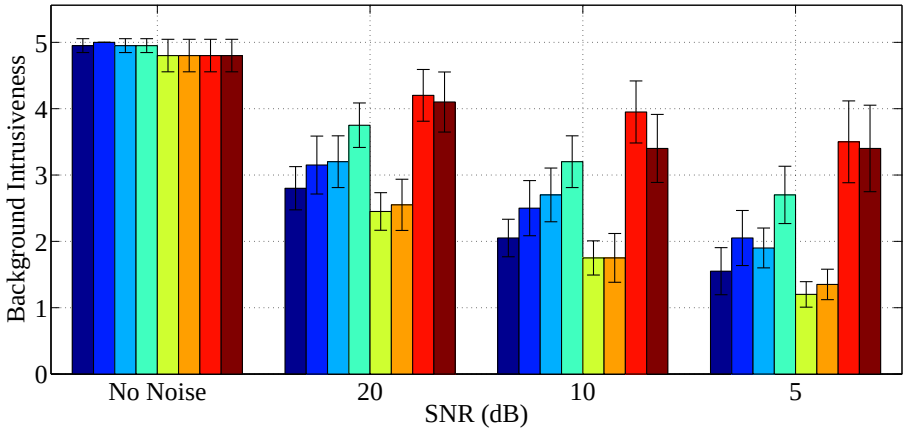
as the test environment (i.e. matched models).

Listening test results for these configurations are presented in Figure 9.6. Starting with signal quality, speech reconstructed using the HNM is shown to be slightly lower quality than the original speech. Quality deteriorates with the addition of noise but remains equivalent to spectral subtraction. In terms of background noise intrusiveness the two methods using estimated spectral envelope are shown to perform significantly better than other methods, including log MMSE. Finally, in terms of overall quality the two reconstruction methods using estimated spectral envelope are shown to be comparable to the conventional method of Wiener filtering. The use of fundamental frequency and voicing estimated from noisy speech as opposed to clean speech is shown to reduce performance in all three categories. Errors in fundamental frequency and voicing affect the excitation of the reconstructed speech. Misclassifications of voicing will cause voiced frames to be reconstructed as unvoiced frames causing a noise-like artifact in the reconstructed speech whilst unvoiced frames reconstructed as voiced frames may cause more tonal artifacts. In terms of background noise and overall quality spectral envelope estimation is shown to perform well, though no significant improvement is noted in terms of signal quality.

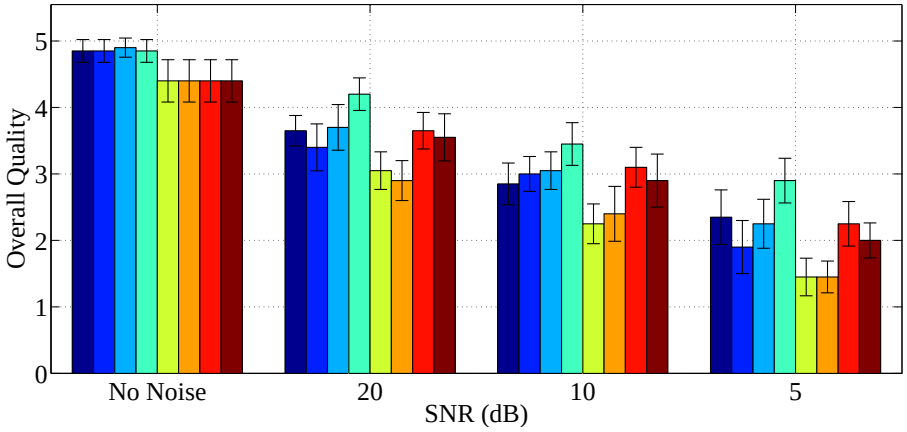
The second listening test is now described. Speaker-independent data was used to test the overall system of speech enhancement. Previously, conventional methods of fundamental frequency and voicing were used to test the performance of the proposed method. In this test the HNM is driven by spectral envelope, fundamental frequency and voicing estimated using the configurations described earlier in Section 9.3. The spectral envelope used by HNM (MAP) system for reconstruction was also used for enhancement using both the Wiener (MAP) and Direct (MAP) systems.



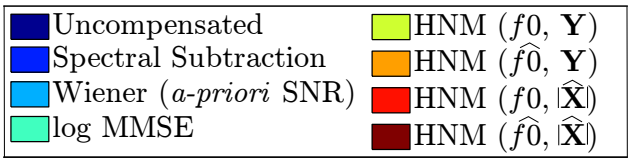
(a) Signal Quality



(b) Background Intrusiveness



(c) Overall Quality



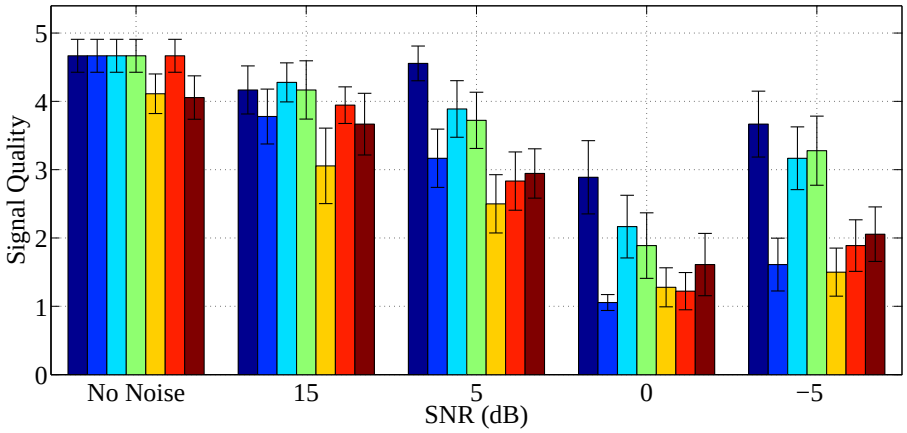
(d) Legend

Figure 9.6: Result of 3-way MOS test measuring signal quality, background noise intrusiveness and overall quality of speech enhancement methods in car noise. Error bars show confidence intervals at a significance level of $p = 0.05$.

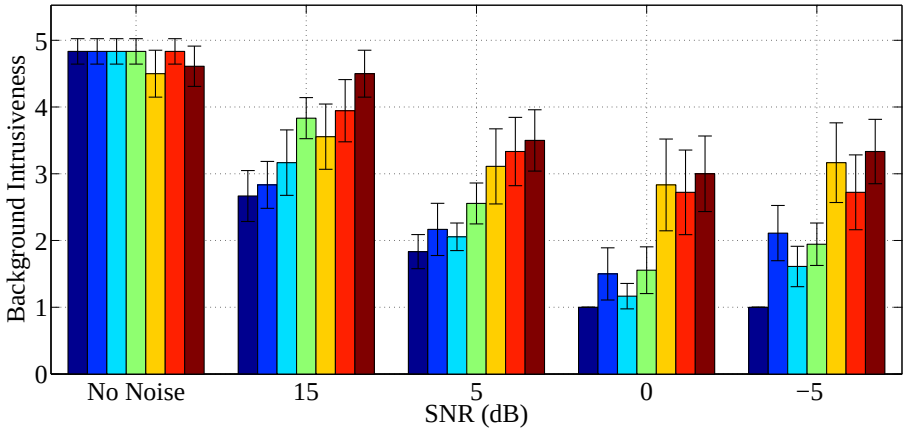
To reduce the duration of this listening test only white noise, babble noise and machine gun noise were included. Results obtained using objective tests with destroyerops noise were similar to those using babble noise and so only babble noise is included. White and babble noises were tested at -5, 0, 5 and 15dB SNRs whilst machine gun noise was tested at -20dB SNR only. Speech with no added noise was also included to measure the level of distortion caused by the process of speech reconstruction and direct feature inversion. All six methods of enhancement were tested alongside the unprocessed speech. This gives a total of 66 test cases resulting in an average test length of 16.5 minutes assuming an average utterance length of 5 seconds.

The result of testing the systems in white noise are presented in Figure 9.7. Conventional methods of enhancement are shown to offer best speech quality. As per the single-speaker results presented in Figure 9.6, the reconstructed speech is shown to perform relatively poorly in this respect and this is attributed to estimation errors in terms of fundamental frequency and voicing. Despite this, the HNM (MAP) system is still shown to outperform the Direct (MAP) and Wiener (MAP) systems at low SNR. The poor performance of the Direct (MAP) system is attributed to the lack of information regarding the excitation in voiced frames whilst the low performance of the Wiener (MAP) system at low SNR in terms of signal quality compared to the HNM (MAP) system is attributed to the HNM (MAP) system reconstructing signal components only related to the original speech. In all cases signal quality is reported to be higher at -5dB SNR than at 0dB SNR. Listeners reported that at these SNR levels it was often difficult to focus on the speech signal due to the very high level of noise and so results at -5dB SNR can be considered unreliable.

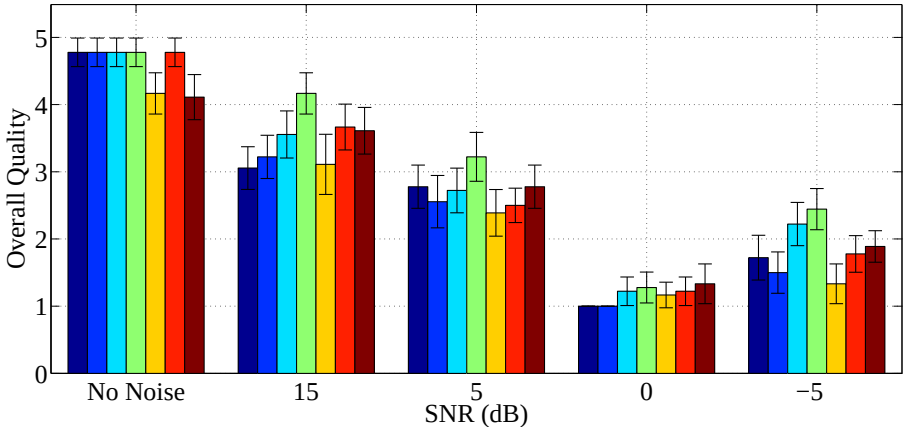
In terms of background noise the three MAP-based systems are shown again to outperform conventional methods of speech enhancement by a large margin. In fact, even at SNRs of 0 and -5dB performance of the three MAP-based methods is shown to be equivalent to that of the Wiener (*a-priori* SNR) method at 15dB SNR.



(a) Signal Quality



(b) Background Intrusiveness



(c) Overall Quality

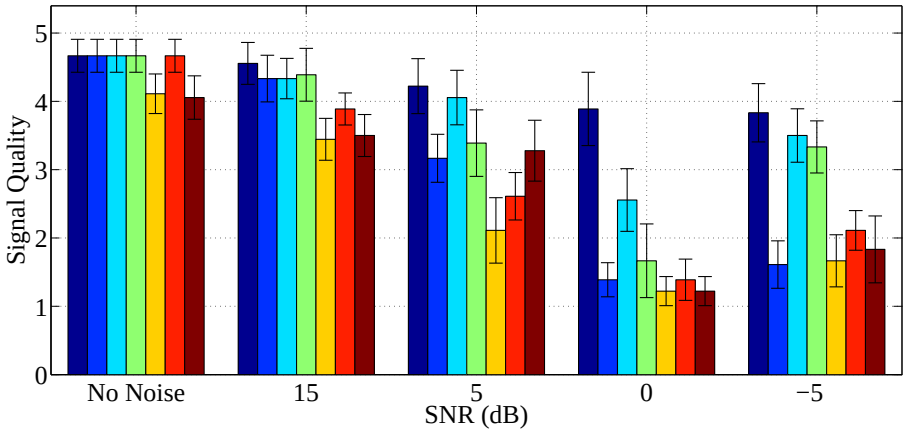
(d) Legend

Figure 9.7: Result of 3-way MOS test measuring signal quality, background noise intrusiveness and overall quality of speech enhancement methods in white noise. Error bars show confidence intervals at a significance level of $p = 0.05$.

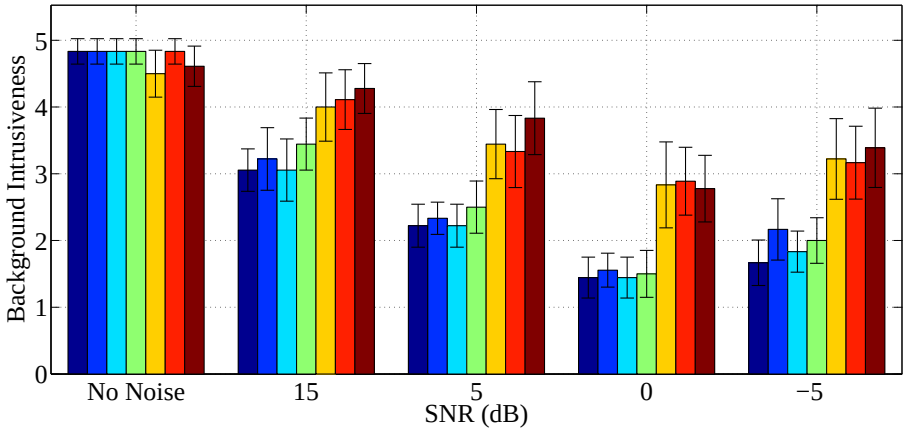
This high performance is clearly related to the accuracy of the estimated spectral envelope as this is the only common factor between the three systems. Out of the three MAP-based systems, the HNM (MAP) system is shown to perform best as no residual noise is reconstructed. Overall, performance is shown to be equivalent to the conventional Wiener filter. Despite large gains in terms of background noise, distortions in signal quality mean overall quality is reduced.

Performance is now discussed in terms of babble noise and these results are presented in Figure 9.8. Comparing the performance in babble noise to white noise (Figure 9.7) shows very little differences. The largest difference in performance between the two noises comes in terms of signal quality. Signal quality is reported to be slightly higher in the case of babble noise and this is reflected by slightly better results in terms of overall quality. This is a promising result as it shows there to be very little difference in terms of background noise removal between stationary and non-stationary noises.

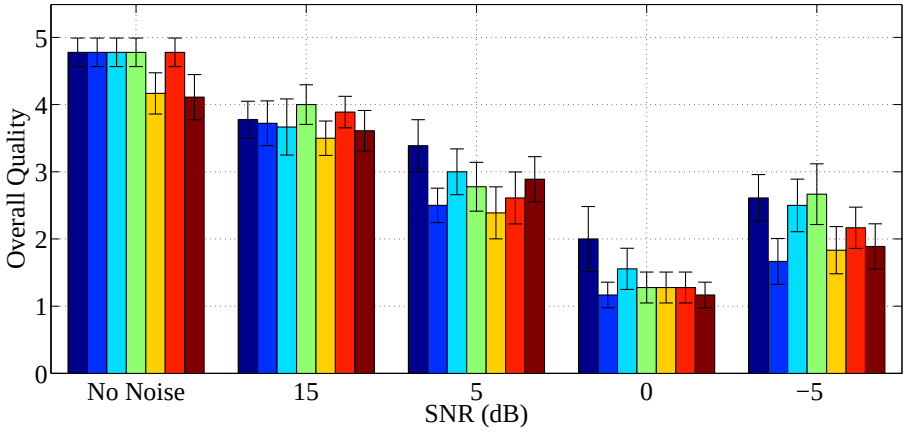
Finally, performance is measured in machine gun noise. Spectral envelope, fundamental frequency and voicing were once again estimated using speaker-independent models, but in the case of machine gun noise serial model combination (SMC, Section 4.5.2.4) was used to adapt models using a GMM of the noise with three mixture components. The results of this test are presented in Figure 9.9. As with the previous results in white and babble noise, signal quality is shown to have been reduced slightly by the reconstruction model. In the case of machine gun noise this is particularly prominent and can be attributed to the nature of the noise. During periods of machine gun fire the speech is completely masked and so no judgement of its quality can be performed. In the case of the MAP-based systems the noise is suppressed to the point where the gun shots are no longer easily audible and so signal distortions are heard at time points where the gun shots previous existed, resulting in lower measured speech quality. Between gun shots there is no noise and so the signal is reconstructed with only a minor degradation in quality caused by inherent modelling errors.



(a) Signal Quality



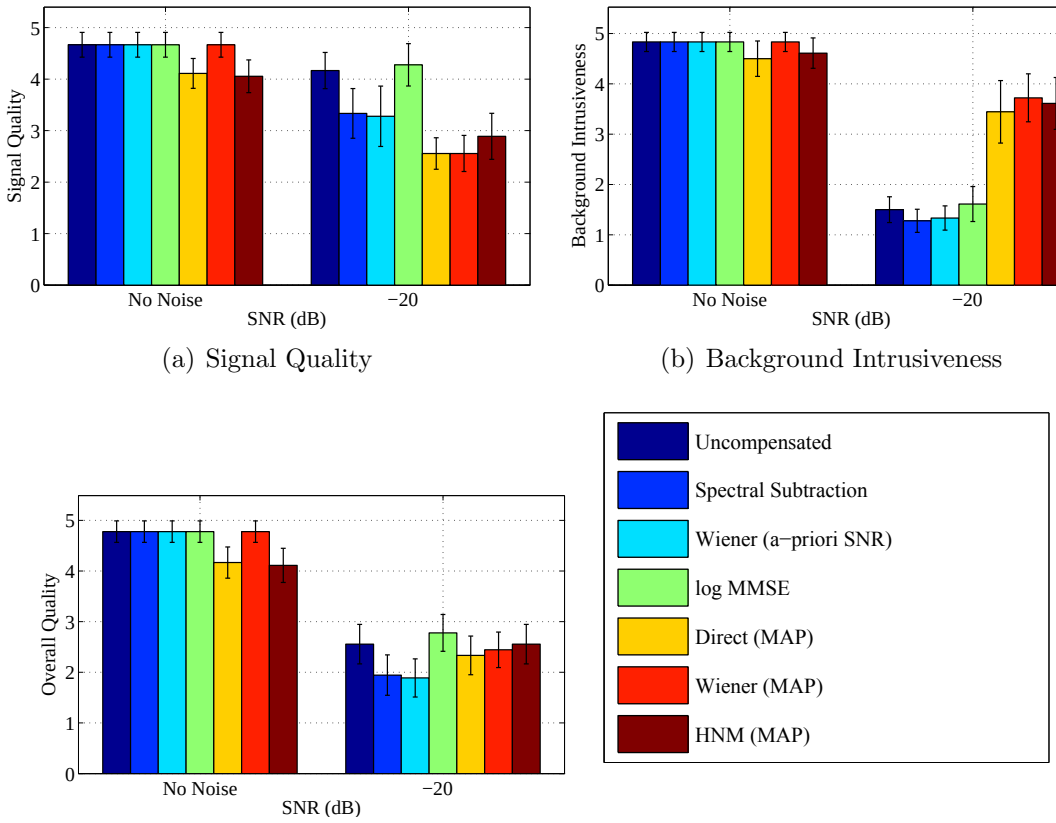
(b) Background Intrusiveness



(c) Overall Quality

(d) Legend

Figure 9.8: Result of 3-way MOS test measuring signal quality, background noise intrusiveness and overall quality of speech enhancement methods in babble noise. Error bars show confidence intervals at a significance level of $p = 0.05$.



(a) Signal Quality (b) Background Intrusiveness

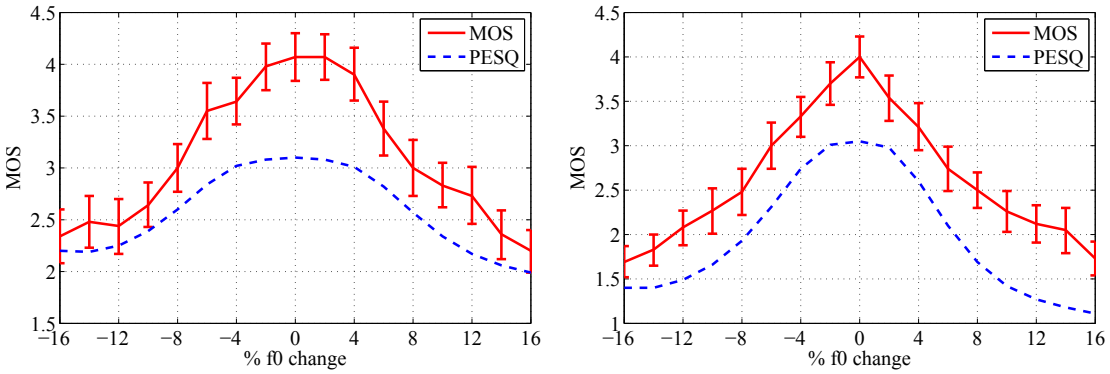
(c) Overall Quality (d) Legend

Figure 9.9: Result of 3-way MOS test measuring signal quality, background noise intrusiveness and overall quality of speech enhancement methods in machine gun noise. Error bars show confidence intervals at a significance level of $p = 0.05$.

Conventional methods of spectral envelope estimation have previously been shown to be ineffective in machine gun noise (Section 5.4.3) and this is reflected by the poor performance of these methods in terms of background noise. All three MAP estimation-based methods offer better performance with little difference in performance between the Direct (MAP), Wiener (MAP) and HNM (MAP) systems. Overall, however, performance is measured to be lower than that of the conventional log MMSE approach of enhancement. This is again attributed to distortions in the speech signal itself rather than residual noise or other artifacts.

9.3.3 Effect of errors in F0 on reconstructed speech quality

The effect of high errors in $f0$ on the quality of reconstructed speech is now examined. Speech was reconstructed using the HNM using acoustic features extracted from clean speech. The $f0$ was then modified by a factor of between -16% and +16% in 2% intervals, i.e: $\hat{f}_0 = s \cdot f_0$ where $0.84 \geq s \geq 1.16$ and f_0 is obtained from the clean speech. The same test utterances were used as per previous experiments, that is a combined total of 25 minutes of male and female speech spoken by 10 speakers. Performance was measured using both PESQ and a MOS listening test to give a measure of objective and subjective performance, respectively. The results of these experiments are split in terms of gender and displayed in Figure 9.10. The effect of $f0$ modification is much more apparent for female speech where $f0$ values are higher and so percentage changes result in larger absolute differences. In terms of objective results, PESQ shows a range of +/- 2% where $f0$ errors are unlikely to affect the quality of reconstructed female speech whilst subjective results show a far greater performance drop off at the same level. This relates to an absolute difference of +/- 4.19% on average. Beyond this range the quality of speech is shown to be degraded to a perceptually noticable level with a relatively steep gradient in terms of percentage change versus MOS. Male speech is shown to be more robust to changes in $f0$ in terms of percentage change however due to lower average values of $f0$ compared to female speech this is unsurprising. A 2% relative change in $f0$ for male speech is



(a) Male (b) Female

Figure 9.10: Effect of modifying $f0$ in the quality of reconstructed speech as measured using subjective MOS tests and objective PESQ evaluation. Error bars show confidence intervals at a significance level of $p = 0.05$

equivalent to a 2.46Hz difference in absolute terms. A range of +/- 4% change in relative $f0$ is deemed to be acceptable when reconstructing male speech, equivalent to a range of +/- 4.92Hz in absolute terms. This is very similar to the case of female speech and suggests that absolute differences in $f0$ are perceived similarly between male and female speech.

In the previous example of destroyerops noise at 0dB SNR the $f0$ error was measured to have increased from 10.61% at 0dB SNR to 14.79% at -5dB SNR. Assuming a constant $f0$ error this relates to a difference in MOS of up to 0.5 based on the results presented in Figure 9.10. This suggests that errors in $f0$ contribute a significant amount to overall quality. This result goes some way to explaining why the quality of the HNM-based system degrades in high levels of noise.

9.4 Summary

The overall method of speech enhancement was evaluated in this chapter and compared against three conventional methods of speech enhancement as well as two further methods based on feature estimation. Through the use of listening tests the proposed method of speech enhancement by reconstruction was shown to perform

strongly in terms of background noise removal, however degradations in speech signal quality, attributed to the accuracy of fundamental frequency estimates, resulted in overall performance equivalent to that of the conventional Wiener filter.

Chapter 10

Conclusions and Further Work

The aim of this thesis was to develop a novel method of single-channel speech enhancement able to compensate for additive noise to produce cleaned speech free of artifacts such as musical noise commonly associated with conventional methods of enhancement. This concluding chapter reviews the work presented in this thesis, identifies key findings and finally presents suggestions of further work. The chapter is therefore split into three sections. First, the work presented in this thesis is reviewed in Section 10.1. Second, conclusions of the project are presented in Section 10.2 before finally further work is suggested in Section 10.3.

Contents

10.1 Review

This section reviews the work presented in this thesis. Chapter 1 introduced the problem of single-channel speech enhancement before a range of current methods of enhancement were described in Chapter 2. Typical methods of speech enhancement were found to operate by filtering the input noisy signal. Speech enhanced using these methods often contained artifacts, known as musical noise, in cases where the noise was not precisely estimated. In an attempt to overcome these artifacts this work proposed a method of speech enhancement by reconstruction whereby clean speech is reconstructed using a reconstruction model driven by a set of acoustic features estimated directly from the noisy speech. Given a suitable model of reconstruction the output speech should not contain any residual noise or artifacts as they will not be reconstructed. Chapter 3 therefore examined a range of speech reconstruction models. A number of models were considered, with the harmonic plus noise model (HNM) deemed to be the most suitable model due the high quality of speech reproduced using the model as well as its low complexity. The HNM was shown to require four acoustic features: i.) spectral envelope, ii.) fundamental frequency, iii.) voicing and iv.) phase. Correlation between parameterisations of the noisy speech and these acoustic features was then measured to predict the success of future estimation. MFCCs were found to offer the most practical parameterisation of the noisy speech and so were chosen as an intermediate feature on which to base acoustic feature estimation. The next five chapters therefore focused on estimation of the required acoustic features.

Chapter 4 described a framework of acoustic feature estimation using MAP estimation. This required a model of the joint density of feature vectors extracted from the noisy speech and the target acoustic feature. Gaussian mixture models (GMMs) were used to model the joint density and so in this case the MAP estimate is equivalent to the minimum mean square error (MMSE) estimator. The model of the joint distribution can be obtained in several ways. Stereo training data may be used to directly obtain the models as per the SPLICE method of feature estimation

proposed by Deng et al. [2000]. This method is not practical in cases where the speaker or noise are not known in advance. Instead, a method based on model-adaptation was proposed. First, a GMM of MFCCs extracted from a large corpus of clean speech was built to give a speaker-independent model of the clean speech. Next, speaker adaptation may optionally be applied given clean speech from the target speaker using MAP adaptation to give a speaker-dependent model of clean speech. Finally, an estimate of the noise can be used to adapt the model for noisy speech using the Unscented Transform (UT). The UT is a data-driven method of parallel model combination. A phase-average mismatch function was used to mix the clean speech model parameters with the noise model to give an estimate of the noisy speech model parameters. Noise is typically assumed to be modelled by a Gaussian distribution, however not all noise was shown to fit this distribution. Additionally, the UT was therefore modified to give a novel method of noise adaptation that can adapt model parameters using a GMM of the noise.

Chapters 5 to 8 then described how the four acoustic features were estimated from the noisy speech. Spectral envelope (Chapter 5) and fundamental frequency (Chapter 6) were estimated using the MAP-based methods of estimation described in Chapter 4. Performance was evaluated in terms of speaker-dependent, gender-dependent and speaker-independent models and in each case compared against conventional methods of estimation. In the case of spectral envelope estimation the MAP-based system was compared against spectral subtraction, Wiener filtering and log MMSE, whilst in the case of fundamental frequency estimation the proposed system was compared against YIN and the ETSI XAFE estimator. In all cases the proposed methods of estimation performed better than the conventional methods of estimation showing the effectiveness of the data-driven approach.

Chapter 7 examined the problem of voicing classification. A data-driven approach was also taken to this problem and so a range of machine learning methods were evaluated. These included: support vector machines (SVM), Rotation Forests, multilayer perceptrons (MLP), naïve Bayes and a GMM-based method of classifica-

tion. Classifiers that were able to model correlation in the feature space performed best, namely: MLPs, Rotation Forest and GMM. Of these, the GMM classifier was determined to be the most suitable method due to the ability to adapt speaker-independent models of clean speech to specific speakers and noises using the transforms described in Chapter 4. Performance of the GMM classifier was compared against the voicing classifier included in the ETSI XAFE. In all cases the proposed method was found to perform significantly better, with the GMM classifier offering relative performance improvements of up to 62% in white noise.

Estimation of the final acoustic feature, phase, was considered in Chapter 8. Four methods of phase estimation were evaluated and included the phase of the original noisy signal and the zero, random and minimum-phase models. Performance was evaluated in terms of the quality of reconstructed speech using each phase model in a range of conditions and was measured using objective tests as well as subjective listening tests. In each case performance was compared against speech reconstructed using the phase of the original, clean, signal. The phase of the noisy speech was determined to be the best estimate of the clean phase. This is in-line with previous studies on phase estimation such as those by Loizou [2007] and Paliwal and Alsteris [2005].

Finally, the proposed method of speech enhancement was evaluated in Chapter 9. The best speaker-independent methods of acoustic feature estimation described in the preceding chapters were used to drive the HNM speech reconstruction model. This gave a speaker-independent system of speech enhancement requiring the following as input:

1.) Noisy speech signal MFCC feature vectors are extracted from the noisy signal. These MFCCs are used as a basis of spectral envelope and $f0$ estimation as well as voicing classification. Phase values are also extracted from the noisy speech signal.

2.) Speaker adaptation data Previously collected clean speech from the target speaker may be used to adapt estimation models to improve performance.

3.) Noise estimate An estimate of the noise is required to adapt clean-trained models to the target environment.

4.) Spectral envelope estimation model A GMM of MFCCs extracted from clean speech is required for spectral envelope estimation. GMMs are trained from a large amount (approx. 20 hours) of speech from a range of speakers of both genders to give speaker-independent models.

5.) $f0$ estimation model A GMM of the joint density of clean MFCCs and $f0$ is required for $f0$ estimation. A conventional method of $f0$ estimation may be used to acquire $f0$ values for training. In this work the same data is used for $f0$ as for spectral envelope estimation, with the autocorrelation-based PRAAT estimator used to estimate $f0$ from the clean speech.

6.) Voicing class models The training data is split into vectors of voiced speech or not voiced speech based on the $f0$ estimated as part of the $f0$ model training process. Two GMMs are then built, one of voiced speech and the other of all other data (unvoiced speech/silence).

In addition to the proposed HNM-based method of speech enhancement the estimated spectral envelope was used to build a Wiener filter to give a model-based Wiener filter as first suggested by Hadir et al. [2011] as well as a method of direct inversion as proposed by Boucheron and Leon [2012]. These methods were also compared to three conventional methods of speech enhancement, namely: spectral subtraction, Wiener filtering (*a-priori* SNR) and log MMSE. Performance was measured in terms of speech quality as measured objectively using PESQ as well as subjectively using 3-way listening tests measuring speech quality, background noise intrusiveness and overall quality.

10.2 Conclusions

Conclusions drawn from this work are presented in this section. The HNM was chosen to reconstruct speech and this model was shown to offer good performance in listening tests presented in Chapter 9. A 3-way listening test was performed in which the HNM was shown to reconstruct speech of between 'Good' and 'Excellent' quality on the 5-point MOS scale (Figure 2.2). This was approximately 0.5 MOS points below that of the unprocessed clean speech and this slight reduction in quality was deemed to be an acceptable modelling error.

In terms of acoustic feature estimation the systems developed for spectral envelope and $f0$ estimation and voicing classification were demonstrated to perform better than conventional methods whilst experiments testing phase models showed the phase of the noisy speech to be the optimal estimate of the clean speech phase.

The estimated acoustic features were used to reconstruct speech using the HNM and performance was measured objectively using PESQ and subjectively with a 3-way listening test. PESQ results showed the HNM and model-based Wiener filter to perform best out of all competing systems in white, babble and machine gun noise with the conventional log MMSE method of enhancement offering slightly better performance in high levels of destroyerops noise. In terms of subjective evaluation, all systems were tested in white, babble and machine gun noise. Speech quality was shown to have been reduced in the case of the proposed HNM-based method of speech enhancement with quality comparable to that of spectral subtraction. Across all noises the level of background noise was shown to be significantly lower than the conventional methods of speech enhancement with ratings of the HNM-based method at -5dB SNR exceeding performance of the conventional methods at 15dB SNR in white and babble noise. Overall performance of the HNM-based method was shown to be roughly equivalent to that of the conventional Wiener filter and in most cases superior to the model-based Wiener filter.

The strong performance of the HNM-based method of enhancement in terms of

background noise is attributed to the absence of musical noise. Enhanced utterances were shown to contain no musical noise across all test SNRs and noises, with very little residual background noise remaining in the signal even at very low SNRs (-5dB SNR). It would therefore be expected that this significant advantage in terms of background noise would be reflected by superior overall quality when compared with conventional methods of enhancement, however this was not always the case. It was found that overall performance was reduced by degradations in speech signal quality; these degradations were attributed primarily to the estimated fundamental frequency. A listening test was performed to measure the quality of speech reconstructed by the HNM driven by acoustic features extracted from clean speech where the $f0$ was warped by between -16 to 16%. The results of this test showed that relatively minor errors in $f0$ caused large degradations in speech quality. A relative error of 4% in $f0$ was shown to reduce perceived quality by up to 0.5 MOS points with errors of 8% degrading speech quality by up to 1.5 MOS points.

10.3 Further Work

The aim of this section is to identify further work which may be undertaken to improve the quality of the proposed method of speech enhancement. The section is divided into two sections. First, suggestions relating to the speech reconstruction model are made before the following section identifies methods of improving estimation of each of the four acoustic features.

10.3.1 Reconstruction model

One of the limiting factors of this method of speech enhancement is the reconstruction model on which it is based. In clean conditions a degradation of approximately 0.5 MOS points was measured in terms of reconstructed speech versus unprocessed speech. This was attributed to modelling errors introduced by the reconstruction model. In preliminary testing STRAIGHT was shown to offer slightly better speech

quality, but at the expense of much higher demands in terms of acoustic features; features of a significantly higher dimension resolution were found to be required for reconstruction. Given more time it would be interesting to determine how difficult it would be to accurately estimate these very high resolution acoustic features to determine whether speech quality could be increased by this model.

10.3.2 Acoustic feature estimation

This section identifies potential methods of improving estimation of the four acoustic features required for reconstruction.

10.3.2.1 Spectral envelope estimation

In Chapter 5 a method of localised estimation was proposed that required frames to be classified based on either their articulation or phoneme class. Using reference classifications performance was shown to be increased considerably with improvements of up to 10% observed at low SNR (0dB SNR). An HMM-GMM based recognition system using context-independent models and an unconstrained grammar was built using HTK [Young et al., 2002] and used to classify frames. When this system was used for enhancement overall performance degraded considerably. This was attributed to the very low recognition accuracy in noisy conditions, with phoneme accuracy as low as 33% at 0dB SNR. Future work in terms of spectral envelope estimation should therefore focus in increasing the accuracy of the frame classification system. In terms of a conventional HMM-GMM based system, performance may be improved with the use of context-dependent models [Lee, 1990] as well as the use of either a more constrained grammar or language model [Odell et al., 1994]. Whilst this would increase the amount of data required for training models and place additional constraints on the method it is expected that these additions could increase the performance of the spectral envelope estimation system considerably.

10.3.2.2 Fundamental frequency estimation

Fundamental frequency accuracy was shown to be an important factor for overall speech quality in listening test results presented in Chapter 9. Considerable improvements over the conventional YIN and ETSI XAFE methods of $f0$ estimation were achieved in non-stationary noises using a method of MAP estimation, however errors at -5dB SNR were still approximately 15% in the case of a speaker-independent system. Based on the listening test results presented in Figure 9.10 show that this relates in a drop in MOS of up to 1.5 points in the case of male speech and up to 2 points in the case of female speech. These are considerable deteriorations in performance and suggest that improving the quality of $f0$ estimates could lead to large gains in overall quality of the proposed method of speech enhancement. Speaker-dependent and gender-dependent systems were shown to offer performance up to 50% better than the speaker-independent system and so large gains in performance could be obtained by using separate models for male and female speakers and developing a gender classification system [Wu and Childers, 1991]. Alternatively, the process of speaker-adaptation performance could be targeted for improvement. Figure 6.15 showed that whilst the proposed method of speaker adaptation offered good performance, performance is not yet optimal.

10.3.2.3 Voicing classification

A method of data-driven voicing classification was developed in Chapter 7. Several machine learning classifiers were tested and several were found to offer good performance. One approach of improving voicing classification accuracy could be to develop an ensemble method of classification. Ensemble methods combine the output of multiple classifiers to improve classification accuracy [Rodriguez et al., 2006]. The output of the ensemble classifier consists of a weighted average of the included classifiers, where the weightings are learnt from the training data. Alternatively, voicing classification can be seen as a problem of time-series classification and so methods developed in this field could also be applied [Bagnall et al., 2012].

10.3.2.4 Phase estimation

The phase of the noisy speech was empirically determined to be the best estimate of the clean signal phase in Chapter 8. At low SNR (< -5dB SNR) the minimum-phase model was shown to outperform the noisy signal phase when clean spectral amplitudes were used. When estimated spectral amplitudes were used to compute the minimum-phase performance deteriorated to below that of the noisy signal phase. If the accuracy of the spectral envelope estimator was improved gains in terms of phase accuracy could also be achieved.

Appendix A

Dataset Descriptions

This appendix describes the datasets used in this work. Several datasets were used and aim of this appendix is to describe each in more detail than given in the rest of this thesis. In terms of speaker data, the NuanceCatherine and WSJCAM0 datasets were used. All speech was recorded in noise-free environments and artificially mixed with noise at the required SNRs. Noise signals were taken from the NOISEX'92 dataset. This Appendix is therefore split into three sections. First, the NuanceCatherine is described in Section A.1 whilst second, the WSJCAM0 dataset is described in Section A.2. Finally, the NOISEX'92 dataset is described in Section A.3.

Contents

Table A.1: Voicing class distribution of the NuanceCatherine dataset (Presented in terms of number of 10ms feature vectors)

Voicing class	Train	Test	Total
Voiced	156357 (65.6%)	82578 (63.4%)	238935
Unvoiced	45456 (19.1%)	28968 (22.3%)	74424
Silence	36525 (15.3%)	18606 (14.3%)	55131
Total	238338	130152	368490

A.1 NuanceCatherine

The NuanceCatherine dataset consists of speech recordings from a single female speaker recorded in a noise-free environment and was annotated at the University of East Anglia (UEA) for Nuance Communications Ltd. A laryngograph was used at the time of recording to monitor vocal tract activity from which fundamental frequency and voicing class were estimated and subsequently hand corrected for model training and testing. Speech was originally recorded at a sampling rate of 16kHz and downsampled using a polyphase downsampling filter to 8kHz for this work (MATLAB `resample` function distributed by MATLAB [2010]). Utterances consisted of phonetically balanced sentences. Voicing class distribution of this dataset is displayed in Table A.1. A total of 1 hour of data was recorded, of which approximately 40 minutes was used for model training and the remaining 20 minutes used for testing.

A.2 WSJCAM0

The WSJCAM0 dataset consists of speech recordings from a large number of male and female speakers recorded in a noise-free environment. The WSJCAM0 dataset was recorded by Fransen et al. [1994]. A total of 140 speakers participated in the recording sessions. 92 speakers spoke 90 utterances of continuous speech read from extracts of the Wall Street Journal newspaper containing words from a vocabulary of 64000 words. The remaining 48 speakers read 40 sentences of continuous prose with a reduced vocabulary of 5000 words. Speech was recorded at a sampling rate

Table A.2: Voicing class distribution of all male speakers in the WSJCAM0 dataset (Presented in terms of number of 10ms feature vectors)

Voicing class	Train	Test	Total
Voiced	1729793 (38.6%)	64281 (39.2%)	1794074
Unvoiced	1817408 (40.6%)	62042 (37.9%)	1879450
Silence	934051 (20.8%)	37548 (22.9%)	971599
Total	4481252	163871	4645123

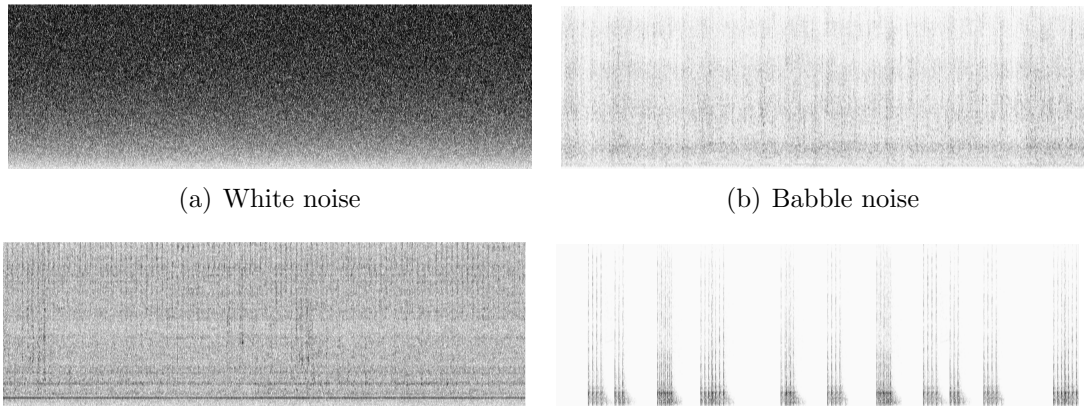
Table A.3: Voicing class distribution of all female speakers in the WSJCAM0 dataset (Presented in terms of number of 10ms feature vectors)

Voicing class	Train	Test	Total
Voiced	1502710 (43.1%)	75818 (46.7%)	1578528
Unvoiced	1319150 (37.8%)	54898 (33.8%)	1374048
Silence	666836 (19.1%)	31787 (19.6%)	698623
Total	3488696	162503	3651199

of 16kHz and downsampled to 8kHz using the same polyphase downsampling filter as was used for the NuanceCatherine dataset. A subset of the corpus was used, with 48 female speakers and 63 female speakers used for model training purposes to give a total of 10 hours of female speech and 12 hours of male speech. A further 5 male and 5 female speakers used for testing with each speaker contributing 6 minutes of audio to give a total of 1 hour test data. The distribution of voicing classes is displayed in Table A.2 for male speech whilst Table A.3 shows the distribution of voicing for female speech. Fundamental frequency was not measured at the time of recording and so PRAAT was used to extract $f0$ for model training and testing purposes.

A.3 NOISEX'92

The NOISEX'92 dataset consists of a number of noise signals recorded as part of a NATO Research Study Group on Speech Processing [Varga and Steeneken, 1993]. Noises were recorded at a sampling rate of 20kHz and subsequently downsampled to 8kHz for this work. Noise was mixed in according with the ITU P.56 standard [P.56, 1993]. A MATLAB implementation by Loizou [2007] was used in this work (function



(a) White noise (b) Babble noise

(c) Destroyerops noise (d) Machine gun noise

Figure A.1: Narrowband spectrogram of noises

`addnoise_asl`). Four noises were used for this work: white, babble, destroyerops and machine gun, and these described in the remainder of this section.

White noise White noise was used as an example of stationary noise. Samples were taken from a normal distribution to give the noise signal displayed in Figure A.1(a). This noise has a flat frequency response and is time invariant.

Babble noise 100 people were recorded speaking in a canteen for a duration of 235 seconds. In some cases individual voices are audible. Figure A.1(b) shows the narrowband spectrogram of this noise.

Destroyerops noise This noise was recorded in the operations room of a destroyer class warship. Figure A.1(c) shows the noise signal to contain both stationary and non-stationary sources of noise. In terms of stationary noise several constant low pitch tones with periodic wide band noise thought to originate from some sort of machinery also present. Some babble noise is audible in the background.

Machine gun noise Figure A.1(d) shows the narrowband spectrogram of machine gun noise. A .50 calibre machine gun was fired in burst mode with periods of silence between gun shots.

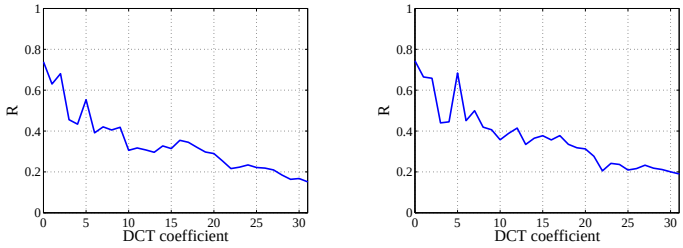
Appendix B

Phoneme Correlation

This appendix contains phoneme-class correlation as described in Section 3.5.3.1. In each case correlation is measured between clean and noisy MFCC feature vectors where noisy MFCCs are extracted from speech contaminated with white noise at an SNR of 0dB. The figures are arranged in articulation classes as displayed in Table B.1.

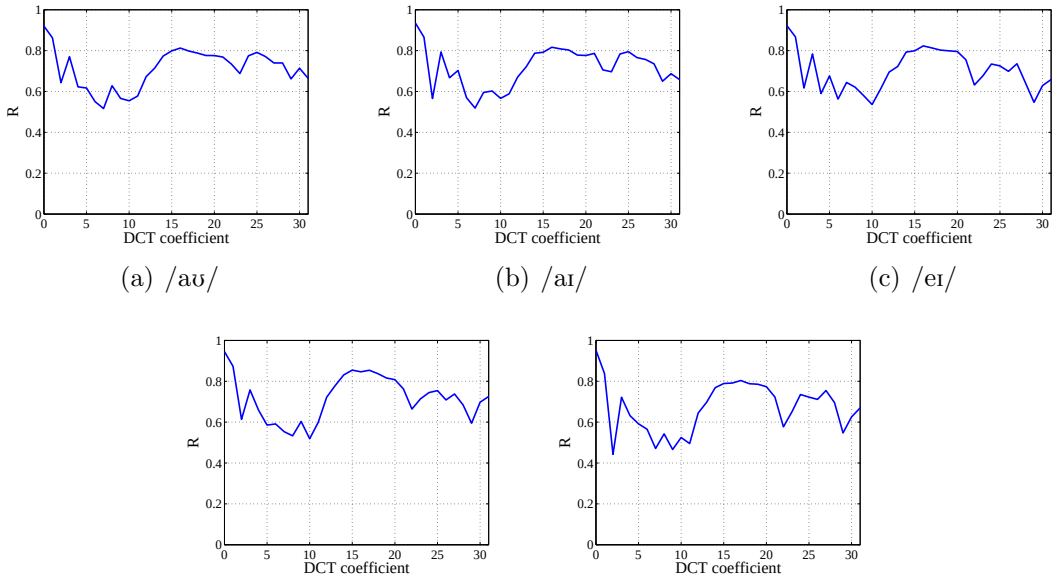
Table B.1: Articulation classes

Articulation Class	Figure
Affricates	B.1
Diphthongs	B.2
Fricatives	B.3
Liquids	B.4
Monophthongs	B.5
Nasals	B.6
R-coloured Vowels	B.7
Semi-vowels	B.8
Stops	B.9
Silence	B.10



(a) /tʃ/ (b) /dʒ/

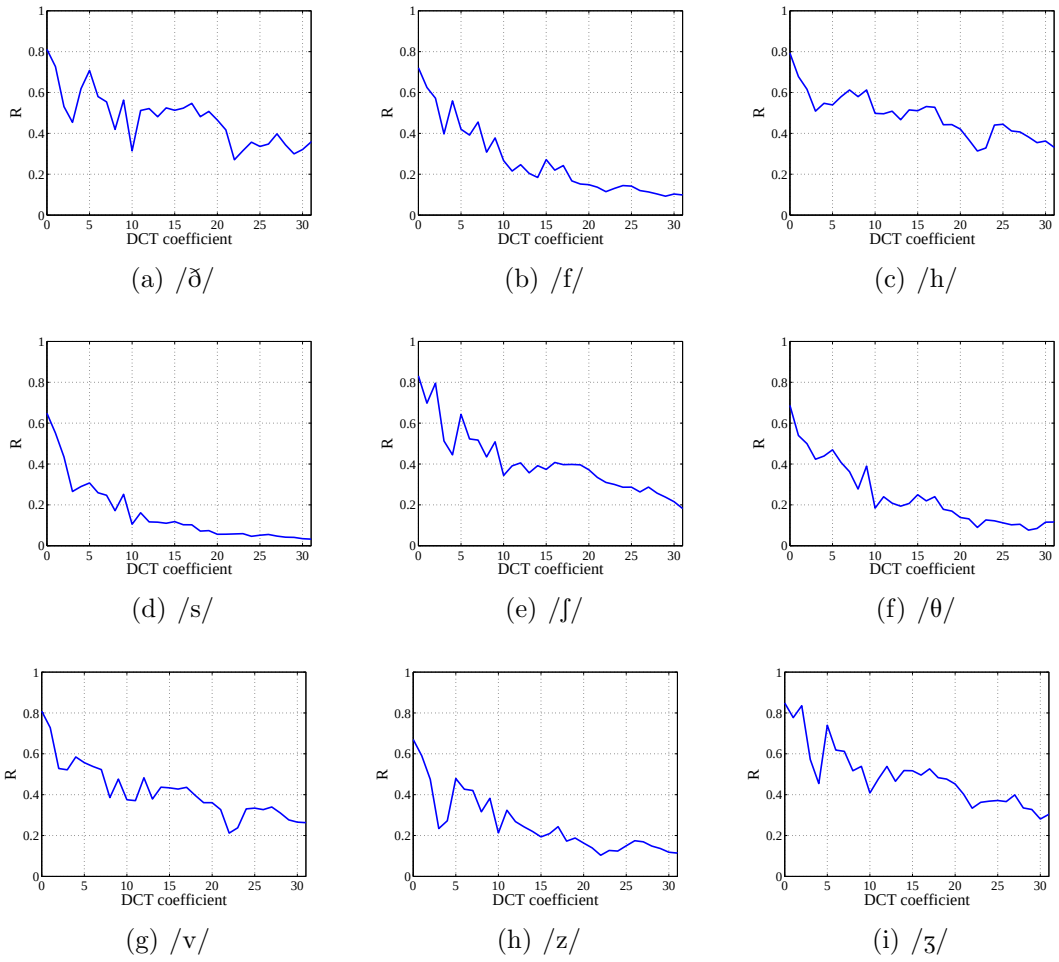
Figure B.1: Phoneme coefficient feature correlation (affricates)



(a) /aʊ/ (b) /aɪ/ (c) /eɪ/

(d) /oʊ/ (e) /ɔɪ/

Figure B.2: Phoneme coefficient feature correlation (diphthongs)

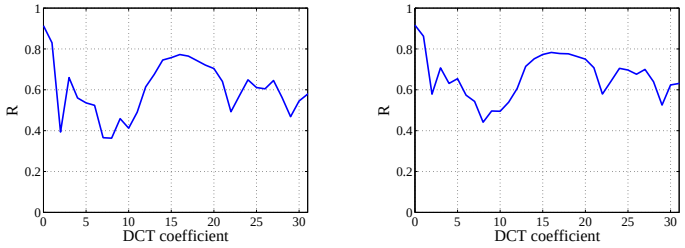


(a) /ð/ (b) /f/ (c) /h/

(d) /s/ (e) /ʃ/ (f) /θ/

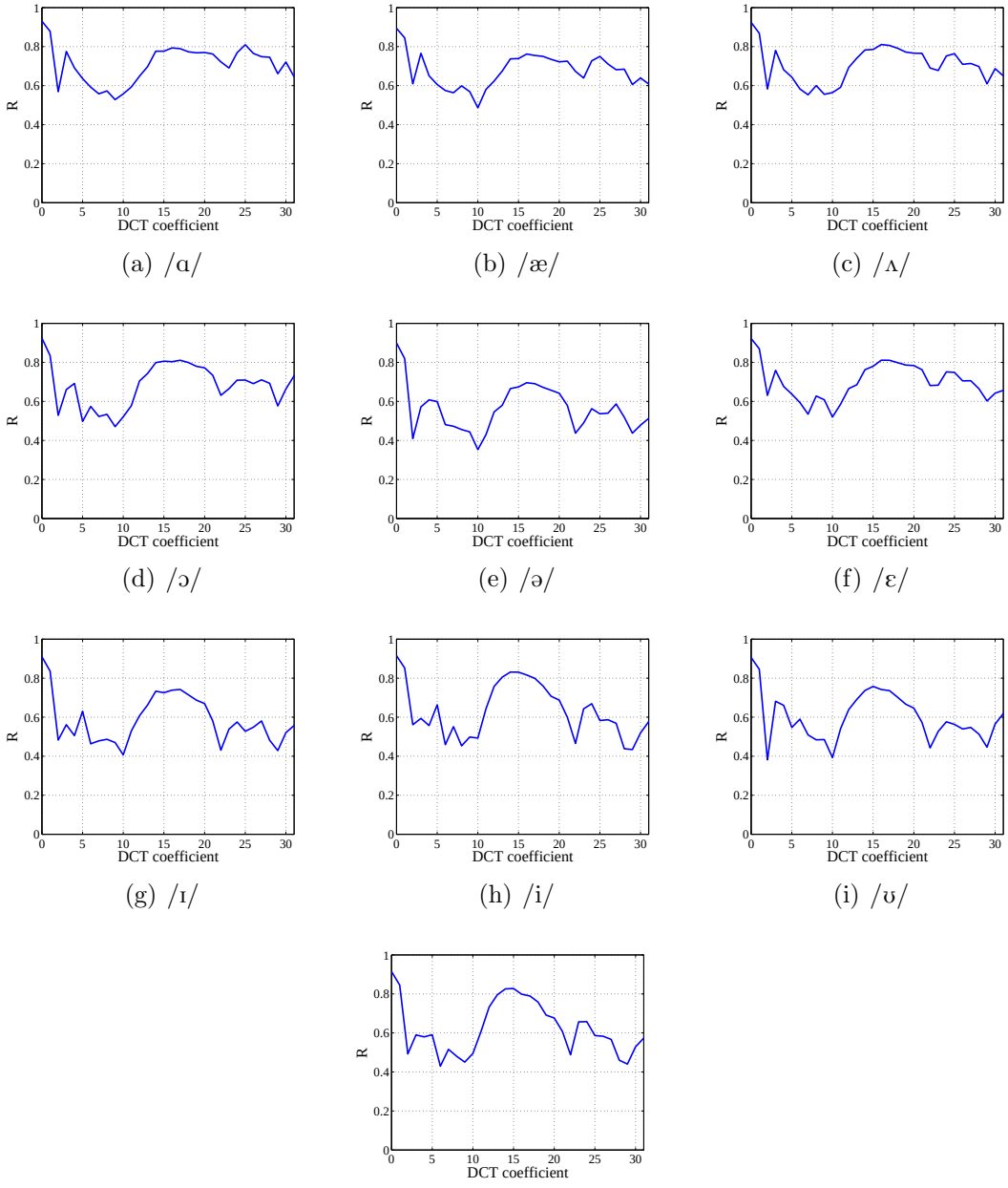
(g) /v/ (h) /z/ (i) /ʒ/

Figure B.3: Phoneme coefficient feature correlation (fricatives)



(a) /ɸ/ (b) /r/

Figure B.4: Phoneme coefficient feature correlation (liquids)



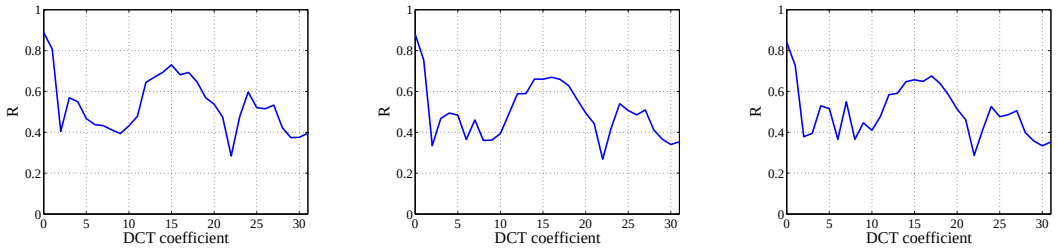
(a) /ɑ/ (b) /æ/ (c) /ʌ/

(d) /ɔ/ (e) /ə/ (f) /ɛ/

(g) /ɪ/ (h) /i/ (i) /ʊ/

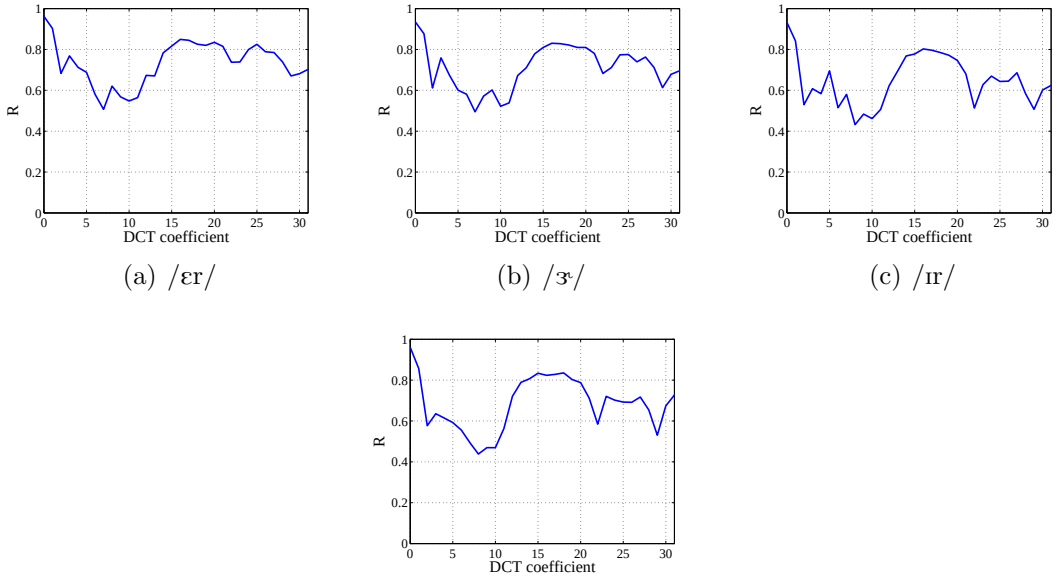
(j) /u/

Figure B.5: Phoneme coefficient feature correlation (monophthongs)



(a) /m/ (b) /n/ (c) /ŋ/

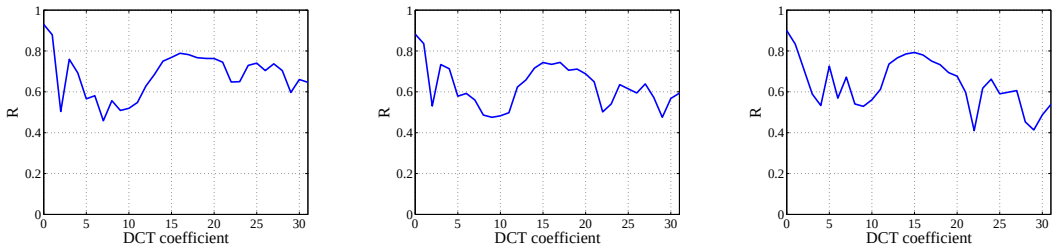
Figure B.6: Phoneme coefficient feature correlation (nasals)



(a) /ɛr/ (b) /ɜˠ/ (c) /ɪr/

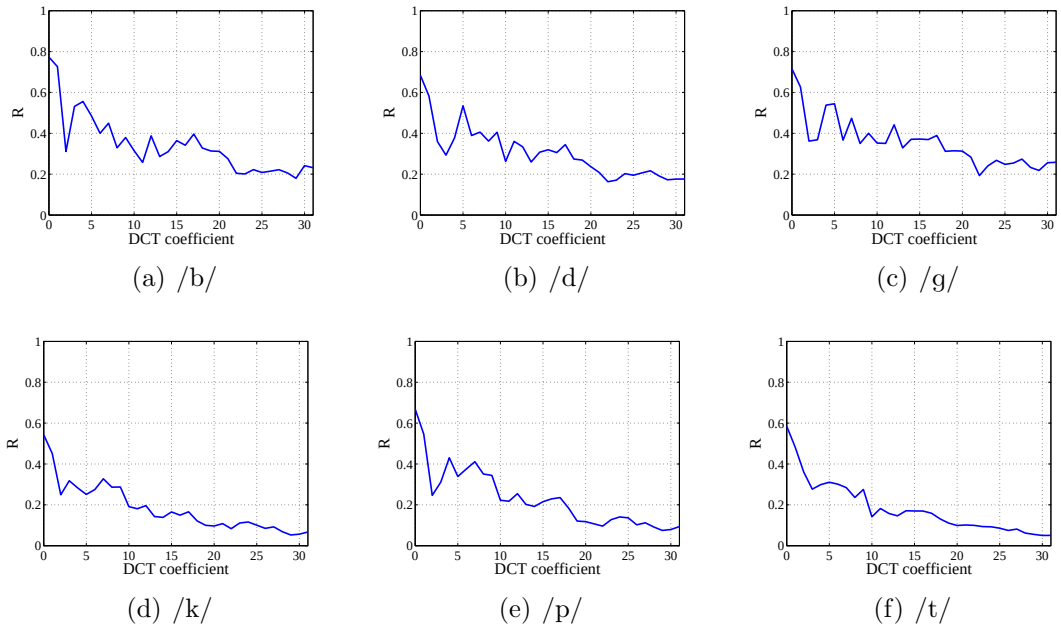
(d) /ɔr/

Figure B.7: Phoneme coefficient feature correlation (R-coloured vowels)



(a) /ʔ/ (b) /w/ (c) /j/

Figure B.8: Phoneme coefficient feature correlation (semi-vowels)



(a) /b/ (b) /d/ (c) /g/

(d) /k/ (e) /p/ (f) /t/

Figure B.9: Phoneme coefficient feature correlation (stops)

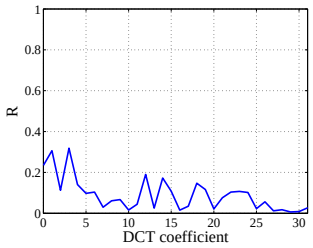


Figure B.10: Phoneme coefficient feature correlation (silence)

Bibliography

Acero, A., Deng, L., Kristjansson, T., and Zhang, J. (2000). HMM Adaptation using Vector Taylor Series for Noisy Speech Recognition. In *International Conference on Spoken Language Processing*, volume 3, pages 869–872.

Acero, A. and Stern, R. (1990). Environmental Robustness in Automatic Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 849–852. IEEE.

Afify, M., Cui, X., and Gao, Y. (2007). Stereo-Based Stochastic Mapping for Robust Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Almajai, I. and Milner, B. (2009). Enhancing Audio Speech using Visual Speech Features. In *Tenth Annual Conference of the International Speech Communication Association*.

ANSI, A. N. S. I. (1997). *American National Standard: Methods for Calculation of the Speech Intelligibility Index*. Acoustical Society of America.

Bagnall, A., Davis, L., Hills, J., and Lines, J. (2012). Transformation Based Ensembles for Time Series Classification. In *SIAM International Conference on Data Mining (SDM)*.

Bai, J., B. M. (2011). Adaptive Hidden Markov Models for Noise Modelling. In *Proceedings of the Nineteenth European Signal Processing Conference*.

Barnard, E., Cole, R., Vea, M., and Alleva, F. (1991). Pitch Detection with a Neural-net Classifier. *IEEE Transactions on Signal Processing*, 39(2):298–307.

Barnett, P. and Knight, R. (1996). The Common Intelligibility Scale. *Proceedings of the Institute of Acoustics*, 17:201–206.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics*, pages 164–171.

Benyassine, A., Shlomot, E., Su, H.-Y., Massaloux, D., Lamblin, C., and Petit, J.-P. (1997). A Silence Compression Scheme for use with G.729 Optimized for

V.70 Digital Simultaneous Voice and Data Applications (Recommendation G.729 Annex B). *IEEE Communications Magazine*, 35(9):64–73.

Berouti, M., Schwartz, R., and Makhoul, J. (1979). Enhancement of speech corrupted by acoustic noise. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 208–211. IEEE.

Boersma, P. (1993). Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound. In *Proceedings of the Institute of Phonetic Sciences*, volume 17, pages 97–110. Amsterdam.

Boersma, P. (2002). PRAAT, A System for Doing Phonetics by Computer. *Glot international*, 5(9/10):341–345.

Boucheron, L. and De Leon, P. (2008). On the Inversion of Mel-frequency Cepstral Coefficients for Speech Enhancement Applications. In *International Conference of Signals and Electronic Systems (ICSES)*, pages 485–488. IEEE.

Boucheron, L. and Leon, P. D. (2012). Low-SNR, Speaker-Dependent Speech Enhancement using GMMs and MFCCs. In *Proceedings of Interspeech*.

Cappé, O. (1994). Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor. *IEEE Transactions on Speech and Audio Processing*, 2(2):345–349.

Chatterjee, S. and Hadi, A. S. (1986). Influential Observations, High Leverage points, and Outliers in Linear Regression. *Statistical Science*, 1(3):379–393.

Chazan, D., Hoory, R., Cohen, G., and Zibulski, M. (2000). Speech Reconstruction from Mel Frequency Cepstral Coefficients and Pitch Frequency. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 1299–1302. IEEE.

Chen, C., Bilmes, J., and Ellis, D. (2005). Speech Feature Smoothing for Robust ASR. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 525–528. IEEE.

Chen, J. and Gersho, A. (1987). Real-time Vector APC Speech Coding at 4800 bps with Adaptive Postfiltering. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 12, pages 2185–2188. IEEE.

Chen, R., Chan, C., and So, H. (2012). Model-Based Speech Enhancement With Improved Spectral Envelope Estimation via Dynamics Tracking. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1324–1336.

Chin, K., Xu, H., Gales, M., Breslin, C., and Knill, K. (2011). Rapid Joint Speaker and Noise Compensation for Robust Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5500–5503. IEEE.

Cohen, I. (2003). Noise Spectrum Estimation in Adverse Environments: Improved Minima Controlled Recursive Averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5):466–475.

Collobert, R. and Bengio, S. (2004). Links Between Perceptrons, MLPs and SVMs. In *International Conference on Machine Learning*, pages 23–30. ACM.

Cui, X., Afify, M., and Gao, Y. (2008). MMSE-based Stereo Feature Stochastic Mapping for Noise Robust Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume , pages 4077 –4080. IEEE.

Darch, J. (2008). *Robust Acoustic Speech Feature Prediction from Mel Frequency Cepstral Coefficients*. PhD thesis, University of East Anglia.

Darch, J., Milner, B., Almajai, I., and Vaseghi, S. (2007). An Investigation into the Correlation and Prediction of Acoustic Speech Features from MFCC Vectors. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Darch, J., Milner, B., and Vaseghi, S. (2006). MAP Prediction of Formant Frequencies and Voicing Class from MFCC Vectors in Noise. *Speech Communication*, 48(11):1556–1572.

De Cheveigné, A. and Kawahara, H. (2002). YIN, A Fundamental Frequency Estimator for Speech and Music. *The Journal of the Acoustical Society of America*, 111:1917.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38.

Deng, L., Acero, A., Jiang, L., Droppo, J., and Huang, X. (2001). High-Performance Robust Speech Recognition using Stereo Training Data. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 301–304. IEEE.

Deng, L., Acero, A., Plumpe, M., and Huang, X. (2000). Large-Vocabulary Speech Recognition under Adverse Acoustic Environments. In *Proceedings of Interspeech*, pages 806–809.

Deng, L., Droppo, J., and Acero, A. (2004). Enhancement of Log Mel Power Spectra of Speech using a Phase-Sensitive Model of the Acoustic Environment and Sequential Estimation of the Corrupting Noise. *IEEE Transactions on Speech and Audio Processing*, 12(2):133–143.

Digalakis, V., Rtischev, D., and Neumeyer, L. (1995). Speaker Adaptation using Constrained Estimation of Gaussian Mixtures. *IEEE Transactions on Speech and Audio Processing*, 3(5):357–366.

Ditech Networks (2007). Limitations of PESQ for Measuring Voice Quality in Mobile and VoIP Networks. Technical report, Ditech Networks.

Doblinger, G. (1995). Computationally Efficient Speech Enhancement by Spectral Minima Tracking in Subbands. In *Proceedings of Eurospeech*, volume 1, page 2.

Droppo, J. and Acero, A. (1998). Maximum a-posteriori Pitch Tracking. In *International Conference on Spoken Language Processing*.

Droppo, J., Acero, A., and Deng, L. (2002). A Nonlinear Observation Model for Removing Noise from Corrupted Speech Log Mel-Spectral Energies. In *International Conference on Spoken Language Processing*, pages 182–185.

Egan, J. (1948). Articulation Testing Methods. *The Laryngoscope*, 58(9):955–991.

En-Najjary, T., Rosec, O., and Chonavel, T. (2003). A New Method for Pitch Prediction from Spectral Envelope and its Application in Voice Conversion. In *Proceedings of Eurospeech*.

Enqing, D., Guizhong, L., Yatong, Z., and Xiaodi, Z. (2002). Applying support vector machines to voice activity detection. In *Sixth International Conference on Signal Processing*, volume 2, pages 1124–1127. IEEE.

Ephraim, Y. and Malah, D. (1984). Speech enhancement using a Minimum-mean Square Error Short-time Spectral Amplitude Estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(6):1109–1121.

Ephraim, Y. and Malah, D. (1985). Speech Enhancement using a Minimum Mean-square Error Log-spectral Amplitude Estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2):443–445.

Ephraim, Y., Malah, D., and Juang, B. (1989). On the Application of Hidden Markov Models for Enhancing Noisy Speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):1846–1856.

Fairbanks, G. (1958). Test of Phonemic Differentiation: The Rhyme Test. *The Journal of the Acoustical Society of America*, 30(7):596–600.

Faubel, F. and Klakow, D. (2010). Estimating Noise from Noisy Speech Features with a Monte Carlo Variant of the Expectation Maximization Algorithm. *Proceedings of Interspeech*.

Faubel, F., McDonough, J., and Klakow, D. (2008). A Phase-Averaged Model for the Relationship Between Noisy Speech, Clean Speech and Noise in the Log-Mel Domain. In *Proceedings of Interspeech*, pages 553–556.

Fletcher, H. and Steinberg, J. (1930). Articulation Testing Methods. *The Journal of the Acoustical Society of America*, 1(2B):17–21.

Fransen, J., Pye, D., Robinson, T., Woodland, P., and Young, S. (1994). WSJCAM0 Corpus and Recording Description. Technical Report CUED/F-INFENG/TR.192, Cambridge University Engineering Department.

French, N. and Steinberg, J. (1947). Factors Governing the Intelligibility of Speech Sounds. *The Journal of the Acoustical Society of America*, 19(1):90–119.

Fry, D. (1979). *The Physics of Speech*. Cambridge University Press.

Fujimoto, M., Watanabe, S., and Nakatani, T. (2012). Noise Suppression with Unsupervised Joint Speaker Adaptation and Noise Mixture Model Estimation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4713–4716. IEEE.

Gales, M. (1995). *Model-Based Techniques for Noise Robust Speech Recognition*. PhD thesis, Cambridge University Engineering Department (CUED).

Gales, M. (2011). Model-Based Approaches to Handling Uncertainty. *Robust Speech Recognition of Uncertain or Missing Data-Theory and Applications*, pages 101–125.

Gales, M. and Woodland, P. (1996). Mean and Variance Adaptation within the MLLR Framework. *Computer Speech and Language*, 10(4):249–264.

Gales, M. and Young, S. (1993). Cepstral Parameter Compensation for HMM Recognition in Noise. *Speech Communication*, 12(3):231–239.

Gales, M. and Young, S. (1995). A Fast and Flexible Implementation of Parallel Model Combination . In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 133 –136 vol.1.

Gales, M. J. F. (1998). Maximum Likelihood Linear Transformations for HMM-based Speech Recognition. *Computer Speech and Language*, 12(2).

Gallager, R. (2008). *Principles of Digital Communication*, volume 1. Cambridge University Press.

Garofolo, J. S. (1993). *TIMIT: Acoustic-phonetic Continuous Speech Corpus*. Linguistic Data Consortium.

Gauvain, J. and Lee, C. (1994). Maximum a-posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298.

Guilmin, G., Bouquin-Jeannes, R., and Gournay, P. (1999). Study of the Influence of Noise Preprocessing on the Performance of a Low Bit Rate Parametric Speech Coder. In *Proceedings of Eurospeech*, volume 3, pages 2367–2370.

Hadir, N., Faubel, F., and Klakow, D. (2011). A Model-Based Spectral Envelope Wiener Filter for Perceptually Motivated Speech Enhancement. In *Proceedings of Interspeech*.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18.

Hanson, B., Wong, D., and Juang, B. (1983). Speech Enhancement with Harmonic Synthesis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 8, pages 1122–1125. IEEE.

Harding, P. and Milner, B. (2011). Speech Enhancement by Reconstruction from Cleaned Acoustic Features. In *Proceedings of Interspeech*, pages 1189–1192.

Harding, P. and Milner, B. (2012a). Enhancing Speech by Reconstruction from Robust Acoustic Features. In *Proceedings of Interspeech*.

Harding, P. and Milner, B. (2012b). On the use of Machine Learning Methods for Speech and Voicing Classification. In *Proceedings of Interspeech*.

Hermus, K. and Wambacq, P. (2006). A Review of Signal Subspace Speech Enhancement and its Application to Noise Robust Speech Recognition. *EURASIP Journal on Advances in Signal Processing*, 2007(1):045821.

Hess, W. J. (1992). Pitch and Voicing Determination. *Advances in Speech Signal Processing*, pages 3–48.

Hirsch, H. and Ehrlicher, C. (1995). Noise Estimation Techniques for Robust Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 153–156. IEEE.

Hirsch, H.-G. and Pearce, D. (2000). The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*.

Hu, Y. and Huo, Q. (2006). An HMM Compensation Approach using Unscented Transformation for Noisy Speech Recognition. *Chinese Spoken Language Processing*, pages 346–357.

Hu, Y. and Loizou, P. (2006). Subjective Comparison of Speech Enhancement Algorithms. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 153–156. IEEE.

Hu, Y. and Loizou, P. C. (2003). A Generalized Subspace Approach for Enhancing Speech Corrupted by Colored Noise. *IEEE Transactions on Speech and Audio Processing*, 11(4):334–341.

Hu, Y. and Loizou, P. C. (2008). Techniques for Estimating the Ideal Binary Mask. In *Proceedings of the Eleventh International Workshop on Acoustic Echo Noise Control*.

Huang, X., Acero, A., and Hon, H. (2001). *Spoken Language Processing*, volume 15. Prentice Hall.

Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals. *The Journal of the Acoustical Society of America*, 57.

ITU (1996). P. 830: Subjective Performance Assessment of Digital Telephone-band and Wideband Digital Codecs. *International Telecommunication Union*.

ITU-T (1996). Methods for Subjective Determination of Transmission Quality (Recommendation P.800). *Geneva, Switzerland*.

ITU-T (2003). Subjective Test Methodology for Evaluating Speech Communication Systems that Include Noise Suppression Algorithm (Recommendation P.835). Technical report, ITU-T.

Jensen, J. and Hansen, J. (2001). Speech Enhancement using a Constrained Iterative Sinusoidal Model. *IEEE Transactions on Speech and Audio Processing*, 9(7):731–740.

Jensen, J. and Hendriks, R. C. (2011). Spectral Magnitude Minimum Mean-square Error Binary Masks for DFT-based Speech Enhancement. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4736–4739. IEEE.

Julier, S. and Uhlmann, J. (2004). Unscented Filtering and Nonlinear Estimation. *Proceedings of the IEEE*, 92(3):401–422.

Kalikow, D., Stevens, K., and Elliott, L. (1977). Development of a Test of Speech Intelligibility in Noise using Sentence Materials with Controlled Word Predictability. *The Journal of the Acoustical Society of America*, 61(5):1337–1351.

Kang, G. and Fransen, L. (1989). Quality Improvement of LPC-processed Noisy Speech by using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(6):939–942.

Kates, J. and Arehart, K. (2005). Coherence and the Speech Intelligibility Index. *The Journal of the Acoustical Society of America*, 117:2224.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring Speech Representations using a Pitch-Adaptive Time-Frequency Smoothing and an Instantaneous-Frequency-Based F0 Extraction: Possible Role of a Repetitive Structure in Sounds. *Speech Communication*, 27(3):187–207.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (2009). An Algorithm that Improves Speech Intelligibility in Noise for Normal-hearing Listeners. *The Journal of the Acoustical Society of America*, 126(3):1486.

Kim, N. S. and Chang, J.-H. (2000). Spectral Enhancement Based on Global Soft Decision. *IEEE Signal Processing Letters*, 7(5):108–110.

Kleijn, W. and Paliwal, K. (1995). *Speech Coding and Synthesis*. Elsevier Science Inc.

Kondoz, A. (2004). *Digital Speech: Coding for Low Bit Rate Communication Systems*. John Wiley & Sons.

Krini, M. and Schmidt, G. (2009). Model-Based Speech Enhancement for Automotive Applications. In *Proceedings of the 6th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 632–637. IEEE.

Kryter, K. (1962). Methods for the Calculation and use of the Articulation Index. *The Journal of the Acoustical Society of America*, 34:1689.

Kullback, S. and Leibler, R. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lee, K.-F. (1990). Context-Dependent Phonetic Hidden Markov Models for Speaker-independent Continuous Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(4):599–609.

Leggetter, C. and Woodland, P. (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, 9(2):171.

Li, J., Seltzer, M., and Gong, Y. (2012). Improvements to VTS Feature Enhancement. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4677–4680. IEEE.

Li, J., Yu, D., Gong, Y., and Deng, L. (2010). Unscented Transform with Online Distortion Estimation for HMM Adaptation. In *Proceedings of Interspeech*, pages 1660–1663.

Liesenborgs, J. (2000). *Voice over IP in networked virtual environments*. PhD thesis, School for Knowledge Technology, Hasselt, Belgium, Thesis.

Lim, J. (1978). Estimation of LPC Coefficients from Speech Waveforms Degraded by Additive Random Noise. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 599–601. IEEE.

Lim, J. and Oppenheim, A. (1978). All-pole Modeling of Degraded Speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(3):197–210.

Lin, L., Holmes, W., and Ambikairajah, E. (2003). Adaptive Noise Estimation Algorithm for Speech Enhancement. *Electronics Letters*, 39(9):754–755.

Linde, Y., Buzo, A., and Gray, R. (1980). An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*, 28(1):84–95.

Loizou, P. (2007). *Speech Enhancement: Theory and Practice*. CRC.

Loizou, P. and Kim, G. (2011). Reasons Why Current Speech-enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):47–56.

Lotter, T. and Vary, P. (2005). Speech Enhancement by MAP Spectral Amplitude Estimation using a Super-Gaussian Speech Model. *EURASIP Journal on Applied Signal Processing*, 2005:1110–1126.

Loweimi, E., Ahadi, S., and Loveymi, S. (2011). On the Importance of Phase and Magnitude Spectra in Speech Enhancement. In *Nineteenth Iranian Conference on Electrical Engineering (ICEE)*, volume , page 1.

Macon, M. and Clements, M. (1996). Speech Concatenation and Synthesis using an Overlap-add Sinusoidal Model. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 361–364. IEEE.

Martin, R. (1994). Spectral Subtraction Based on Minimum Statistics. *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 6:8.

MATLAB (2010). *version 7.10.0 (R2010a)*. The MathWorks Inc., Natick, Massachusetts.

McAulay, R. and Malpass, M. (1980). Speech Enhancement using a Soft-decision Noise Suppression Filter. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(2):137–145.

McAulay, R. and Quatieri, T. (1986). Speech Analysis/Synthesis Based on a Sinusoidal Representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754.

Medan, Y., Yair, E., and Chazan, D. (1991). Super Resolution Pitch Determination of Speech Signals. *IEEE Transactions on Signal Processing*, 39(1):40–48.

Meyer, J., Simmer, K., Kammeyer, K., et al. (1997). Comparison of One-and Two-channel Noise-estimation Techniques. In *Fifth International Workshop on Acoustic Echo and Noise Control (IWAENC)*, pages 11–12.

Meyer, J. and Simmer, K. U. (1997). Multi-channel Speech Enhancement in a Car Environment using Wiener Filtering and Spectral Subtraction. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 1167–1170. IEEE.

Miller, G. and Nicely, P. (1955). An Analysis of Perceptual Confusions Among some English Consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Milner, B. and Darch, J. (2011). Robust Acoustic Speech Feature Prediction From Noisy Mel-Frequency Cepstral Coefficients. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):338 –347.

Milner, B., Darch, J., and Vaseghi, S. (2008). Applying Noise Compensation Methods to Robustly Predict Acoustic Speech Features from MFCC Vectors in Noise. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume , pages 3945 –3948. IEEE.

Milner, B. and Shao, X. (2006). Clean Speech Reconstruction from MFCC Vectors and Fundamental Frequency using an Integrated Front-end. *Speech Communication*, 48(6):697–715.

Milner, B., Shao, X., and Darch, J. (2005). Fundamental Frequency and Voicing Prediction from MFCCs for Speech Reconstruction from Unconstrained Speech. In *Ninth European Conference on Speech Communication and Technology*.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, 1 edition.

Moharir, G., Patwardhan, P., and Rao, P. (2002). Spectral Enhancement Preprocessing for the HNM Coding of Noisy Speech. In *Seventh International Conference on Spoken Language Processing*.

Moon, S.-H., Kim, B., and Lee, I.-S. (2010). Importance of Phase Information in Speech Enhancement. In *International Conference on Complex, Intelligent and Software Intensive Systems (CISIS)*, volume , pages 770 –773.

Moreno, P., Raj, B., and Stern, R. (1996). A Vector Taylor Series Approach for Environment-Independent Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 733–736. IEEE.

Munson, W. and Karlin, J. (1962). Isopreference Method for Evaluating Speech-Transmission Circuits. *The Journal of the Acoustical Society of America*, 34(6):762–774.

Nagarajan, S. and Sankar, R. (1998). Efficient Implementation of Linear Predictive Coding Algorithms. In *Proceedings of the IEEE*, pages 69–72. IEEE.

Nehorai, A. and Porat, B. (1986). Adaptive Comb Filtering for Harmonic Signal Enhancement. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(5):1124–1138.

Neumeyer, L. and Weintraub, M. (1994). Probabilistic Optimum Filtering for Robust Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1. IEEE.

Nilsson, M., Soli, S., and Sullivan, J. (1994). Development of the Hearing in Noise Test for the Measurement of Speech Reception Thresholds in Quiet and in Noise. *The Journal of the Acoustical Society of America*, 95:1085.

Odell, J., Valtchev, V., Woodland, P., and Young, S. (1994). A One-pass Decoder Design for Large Vocabulary Recognition. In *Proceedings of the Workshop on Human Language Technology*, pages 405–410. Association for Computational Linguistics.

Oppenheim, A., Schafer, R., and Buck, J. (1989). *Discrete-Time Signal Processing*, volume 2. Prentice Hall Englewood Cliffs, NJ:.

O'Shaughnessy, D. (1987). *Speech Communications: Human And Machine (IEEE)*. Universities press.

P.56, I.-T. (1993). Objective Measurement of Active Speech Level (Recommendation P.56). Technical report, ITU-T.

Paliwal, K. (2003). Usefulness of Phase in Speech Processing. In *IPSJ Spoken Language Processing Workshop*, pages 1–6.

Paliwal, K. and Alsteris, L. (2005). On the Usefulness of STFT Phase Spectrum in Human Listening Tests. *Speech Communication*, 45(2):153–170.

Quatieri, T. and McAulay, R. (2002). Audio Signal Processing based on Sinusoidal Analysis/Synthesis. *Applications of Digital Signal Processing to Audio and Acoustics*, pages 343–416.

Rabiner, L., Cheng, M., Rosenberg, A., and McGonegal, C. (1976). A Comparative Performance Study of Several Pitch Detection Algorithms. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(5):399–418.

Ramírez, J., Segura, J. C., Benítez, C., de la Torre, Á., and Rubio, A. J. (2004). A New Kullback-Leibler VAD for Speech Recognition in Noise. *IEEE Signal Processing Letters*, 11(2):266–269.

Ramírez, J., Yélamos, P., Górriz, J., and Segura, J. (2006). Svm-based speech endpoint detection using contextual speech features. *Electronics letters*, 42(7):426–428.

Reynolds, D., Quatieri, T., and Dunn, R. (2000). Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1):19–41.

Rhebergen, K. and Versfeld, N. (2005). A Speech Intelligibility Index-based Approach to Predict the Speech Reception Threshold for Sentences in Fluctuating Noise for Normal-hearing Listeners. *The Journal of the Acoustical Society of America*, 117:2181.

Robinson, T., Fransen, J., Pye, D., Foote, J., and Renals, S. (1995). WSJCAM0: A British English Speech Corpus For Large Vocabulary Continuous Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 81–84. IEEE.

Rodet, X. and Doval, B. (1992). Maximum-likelihood Harmonic Matching for Fundamental Frequency Estimation. *The Journal of the Acoustical Society of America*, 92:2428.

Rodriguez, J., Kuncheva, L., and Alonso, C. (2006). Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:1619–1630.

Ross, M., Shaffer, H., Cohen, A., Freudberg, R., and Manley, H. (1974). Average Magnitude Difference Function Pitch Extractor. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 22(5):353–362.

Sambur, M. and Jayant, N. (1976). LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(6):488–494.

Sangwan, A., Chiranth, M., Jamadagni, H., Sah, R., Venkatesha Prasad, R., and Gaurav, V. (2002). VAD Techniques for Real-time Speech Transmission on the Internet. In *Fifth IEEE International Conference on High Speed Networks and Multimedia Communications*, pages 46–50. IEEE.

Scalart, P. et al. (1996). Speech Enhancement Based on a-priori Signal to Noise Estimation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 629–632. IEEE.

Shannon, B. J. and Paliwal, K. K. (2006). Role of Phase Estimation in Speech Enhancement. In *Proceedings of Interspeech*, pages 1423–1426.

Shao, X. (2005). *Robust Algorithms For Speech Reconstruction On Mobile Devices*. PhD thesis, University of East Anglia.

Shao, X. and Milner, B. (2004). Pitch Prediction from MFCC Vectors for Speech Reconstruction. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages I–97. IEEE.

Shinohara, Y. and Akamine, M. (2009). Bayesian Feature Enhancement using a Mixture of Unscented Transformation for Uncertainty Decoding of Noisy Speech. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume , pages 4569 –4572. IEEE.

Sim, B., Tong, Y., Chang, J., and Tan, C. (1998). A Parametric Formulation of the Generalized Spectral Subtraction Method. *IEEE Transactions on Speech and Audio Processing*, 6(4):328–337.

Sorin, A. and Ramabadran, T. (2003). Extended Advanced Front End (XAFE) Algorithm Description, Version 1.1. Technical report, ETSI STQ-Aurora DSR Working Group.

Spanias, A. (1994). Speech Coding: A Tutorial Review. *Proceedings of the IEEE*, 82(10):1541–1582.

Steeneken, H. and Houtgast, T. (1980). A Physical Method for Measuring Speech-transmission Quality. *The Journal of the Acoustical Society of America*, 67:318.

Stouten, V., Demuynck, K., and Wambacq, P. (2003). Robust Speech Recognition using Model-Based Feature Enhancement. In *Eighth European Conference on Speech Communication and Technology (Eurospeech)*, pages 17–20.

Stylianou, Y. (2001). Applying the Harmonic plus Noise Model in Concatenative Speech Synthesis. *IEEE Transactions on Speech and Audio Processing*, 9(1):21–29.

Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (2010). A Short-time Objective Intelligibility Measure for Time-frequency Weighted Noisy Speech. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume , pages 4214 –4217. IEEE.

Tabrikian, J., Dubnov, S., and Dickalov, Y. (2004). Maximum a-posteriori Probability Pitch Tracking in Noisy Environments using Harmonic Model. *IEEE Transactions on Speech and Audio Processing*, 12(1):76–87.

Talkin, D. (1995). A Robust Algorithm for Pitch Tracking (RAPT). *Speech Coding and Synthesis*, 495:518.

Tierney, J. (1980). A Study of LPC Analysis of Speech in Additive Noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):389–397.

Vair, C., Colibro, D., Castaldo, F., Dalmasso, E., and Laface, P. (2006). Channel Factors Compensation in Model and Feature Domain for Speaker Recognition. In *IEEE Speaker and Language Recognition Workshop*, volume , pages 1 –6.

Varga, A. and Moore, R. (1990). Hidden Markov Model Decomposition of Speech and Noise. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 845–848. IEEE.

Varga, A. and Steeneken, H. J. (1993). Assessment for Automatic Speech Recognition: II. NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems. *Speech Communication*, 12(3):247–251.

Vary, P. (1985). Noise Suppression by Spectral Magnitude Estimation–Mechanism and Theoretical Limits. *Signal Processing*, 8(4):387–400.

Vaseghi, S. (2008). *Advanced Digital Signal Processing and Noise Reduction*. Wiley.

Vaseghi, S. and Milner, B. (1995). Speech Recognition in Impulsive Noise. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 437–440 vol.1. IEEE.

Vaseghi, S. and Milner, B. (1997). Noise Compensation Methods for Hidden Markov Model Speech Recognition in Adverse Environments. *IEEE Transactions on Speech and Audio Processing*, 5(1):11–21.

Viikki, O. and Laurila, K. (1998). Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication*, 25(13):133 – 147.

Voiers, W. (1977). Diagnostic Acceptability Measure for Speech Communication Systems. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 204–207. IEEE.

Voiers, W. (1983). Evaluating Processed Speech using the Diagnostic Rhyme Test. *Speech Technology*, 1(4):30–39.

Wang, D. and Lim, J. (1982). The Unimportance of Phase in Speech Enhancement. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 30(4): 679 – 681.

Weiss, M., Aschkenasy, E., and Parsons, T. (1975). Study and development of the INTEL technique for improving speech intelligibility. Technical report, DTIC Document.

Woodland, P. (2001). Speaker Adaptation for Continuous Density HMMs: A Review. In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*.

Wu, K. and Childers, D. G. (1991). Gender Recognition from Speech. Part I: Coarse Analysis. *The Journal of the Acoustical Society of America*, 90:1828.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., and Woodland, P. (2002). The HTK book. *Cambridge University Engineering Department*, 3.

Yu, D., Deng, L., Droppo, J., Wu, J., Gong, Y., and Acero, A. (2008). A Minimum-Mean-Square-Error Noise Reduction Algorithm on Mel-Frequency Cepstra for Robust Speech Recognition. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4041–4044. IEEE.

Zavarehei, E., Vaseghi, S., and Yan, Q. (2007). Noisy Speech Enhancement using Harmonic-Noise Model and Codebook-based Post-Processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1194–1203.

Zhao, D., Kleijn, W., Ypma, A., and De Vries, B. (2008). Online Noise Estimation Using Stochastic-Gain HMM for Speech Enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(4):835 –846.