Adaptive Hidden Markov Noise Modelling for Speech Enhancement

JIONGJUN BAI

A Thesis submitted in fulfillment of requirements for the degree of
Doctor of Philosophy of Imperial College London

Communication and Signal Processing Group
Department of Electrical and Electronic Engineering
Imperial College London

2012

Copyright Declaration

Declaration of Originality

This thesis consists of the research work conducted in the Department of Electrical and Electronic Engineering at Imperial College London. I declare that the work presented in this thesis is my own, except where acknowledged in the thesis.

Jiongjun Bai

Abstract

A robust and reliable noise estimation algorithm is required in many speech enhancement systems. The aim of this thesis is to propose and evaluate a robust noise estimation algorithm for highly non-stationary noisy environments. In this work, we model the non-stationary noise using a set of discrete states with each state representing a distinct noise power spectrum. In this approach, the state sequence over time is conveniently represented by a Hidden Markov Model (HMM).

In this thesis, we first present an online HMM re-estimation framework that models time-varying noise using a Hidden Markov Model and tracks changes in noise characteristics by a sequential model update procedure that tracks the noise characteristics during the absence of speech. In addition the algorithm will when necessary create new model states to represent novel noise spectra and will merge existing states that have similar characteristics. We then extend our work in robust noise estimation during speech activity by incorporating a speech model into our existing noise model. The noise characteristics within each state are updated based on a speech presence probability which is derived from a modified Minima controlled recursive averaging method.

We have demonstrated the effectiveness of our noise HMM in tracking both stationary and highly non-stationary noise, and shown that it gives improved performance over other conventional noise estimation methods when it is incorporated into a standard speech enhancement algorithm.

Acknowledgments

I consider myself to be very lucky for having the opportunity to know and work under the supervision of Mike Brookes. I am grateful to him for allowing me to undertake this PhD under his supervision and his guidance has always been influential in my work. This thesis would not have been completed without his insightful suggestions and feedback. I have learned a lot from him throughout my PhD, and has been many times inspired by his vast knowledge and creative minds. It is truly an honor for me to have worked with him.

It was a pleasure to be a member of the Speech and Audio Processing Group. The weekly group meetings and discussions are always insightful. The very diverse and interesting members stimulated an enjoyable atmosphere. I wish to express my special thanks to Dr. Patrick Naylor and fellow group members for their comments and suggestions in my research.

Finally, my never ending gratitude goes towards my mum and dad. I dedicate this thesis to both of them, as I am indebted to my parents for their endless love and support during my study.

Contents

List of Figures

List of Acronyms

ANOVA	analysis of variance
AR	auto-regressive
DFT	discrete Fourier transform
EM	expectation–maximization
EVD	eigenvalue decomposition
GMM	Gaussian mixture model
HMM	hidden Markov model
ISTFT	inverse short time Fourier transform
KLT	Karhunen Lòeve transform
LTASS	long-term averaging speech spectrum
PESQ	perceptual evaluation of speech quality
MCRA	Minima controlled recursive averaging
ML	maximum likelihood
MMSE	minimum mean square error
MS	minimum statistics
LPC	linear predictive coding

SGMM	sequential Gaussian mixture model
SNR	signal to noise ratio
gSNR	global signal to noise ratio
sSNR	segmental signal to noise ratio
STOI	short-time objective intelligibility
STFT	short time Fourier transform
SVD	singular value decomposition
UM	unbiased MMSE-based
VAD	voice activity detector

List of Mathematical Notation

a_{ij}	the probability of a transition from state i to j
$\mathbf{A}$	the transition probability matrix
b_i	the probability of an observation for the state i
H_s	number of speech states
H_n	number of noise states
k	frequency bin index
κ	speech level frame rate
L	the number of recent observation frames that are stored for model updating
m	Mel-frequency bin index
M	Mel-spaced filter bank matrix
N_k	noise power in frequency bin k
O_t	the power spectrum of the observed signal at time t
$P_j\left(t_1, t_2\right)$	total probability in state j from frames t_1 to t_2
$Q_j\left(t_1, t_2\right)$	weighted state occupancy count in state j from frames t_1 to t_2
$R_{ij}\left(t_1, t_2\right)$	weighted transition prevalence sum from state i to state j from frames t_1 to t_2

S_k	clean speech power at frequency k
T	the current time frame
T_L	$T_L = T - L$ the observation frames that must be stored
$U_j\left(t_1, t_2\right)$	weighted state observation sum of the mean in state j from frames t_1 to t_2
$\alpha_i\left(t\right)$	the forward probability of an HMM for the state i at time t
$\beta_i\left(t\right)$	the backward probability of an HMM for the state i at time t
ξ	*a priori* SNR
η_k	speech present probability at frequency k
λ	average forgetting factor
μ_i	mean noise power spectrum for the state i
σ_i^2	noise power variance for the state i
ν_i	mean speech power spectrum for the state i
ς_i^2	speech power variance for the state i
γ	speech Level
$\varpi_i\left(k\right)$	minimum noise power spectrum for the state i
ζ_s	Speech model
ζ_n	Noise model
Γ	speech presence threshold
ϵ	average smoothing parameter
$*^{(T)}$	model parameter based on information available at time T

Chapter 1

Introduction

1.1 Context

Speech enhancement systems aim to improve the quality and intelligibility of speech that has been corrupted in some way, most commonly by additive noise. Improvement in intelligibility has obvious benefits while improvement in quality is highly desirable as it can reduce listener fatigue, particularly in situations in which the listener is exposed to high levels of noise for long periods of time. Many speech enhancement techniques have been developed to reduce or suppress the background noise.

Despite the possible fatigue as mentioned above, human beings are good at adapting to a noisy environment, especially if the noise is persistent or repetitive [55]. For example, if someone just moves to live near a railway line, the noise from trains passing by might be unpleasant and disturbing at first. Over some time, the person will become accustomed to the noise and barely notice that it is still present. However, the person is still aware of other new type of noise, such as a fire alarm, though this too may be ignored in the future if the new noise becomes repetitive. Such selective blocking and adaptation are immensely powerful; we would like to replicate this characteristic in our speech enhancement algorithm.

Almost all speech enhancement algorithms require an estimate of the noise power spectrum or its equivalent [10, 29]. The accuracy of this estimate has a major impact on the overall quality of the speech enhancement: overestimating the noise will lead to distortion of the speech, while underestimating it will lead to unwanted residual noise. The problem of noise identification or suppression is easiest if the noise is stationary at least over intervals of several seconds, so that the noise characteristics remain unchanged during intervals when the presence of speech makes noise estimation difficult. In this case, a common approach is to take a weighted average of the noisy speech power spectrum during speech absence as the noise estimate. Early systems controlled the averaging process by using a voice-activity detector (VAD) [111] to identify noise-dominated frames. To avoid the VAD requirement, Martin estimated the noise spectrum by taking the minimum of the temporally smoothed power spectrum in each frequency bin and then applying a bias compensation factor [90, 92]. This method is effective in estimating both stationary and time-varying noise even when speech is present but, because it relies on temporal averaging, it is unable to follow abrupt changes in the noise spectrum.

In realistic environments, especially when using a mobile device, the noise normally includes multiple components, and can vary rapidly due to relative motion between source and receiver or because the sound sources themselves are intermittent (e.g. ringing phones or door slams). Several authors [107, 133] have recognised that such non-stationary noise environments are better modelled as a set of discrete states than as a single time-varying source. In this approach, each state corresponds to a distinct noise power spectrum and the state sequence over time is conveniently represented by a Hidden Markov Model (HMM). HMMs have been widely used in speech recognition [123, 42], since the range of typical utterances of speech can be pre-trained and included in the speech HMM. Unlike speech, noise arises from a large variety of different environments, and thus it is far less predictable. A robust noise HMM either needs to be trained on an impossibly large number of noise sources or else, like a human listener [8], needs to adapt rapidly to the noise sources present in any situation.

1.2 Research Statement

The aim of this thesis is to propose a robust noise estimation algorithm for single channel speech enhancement in adverse environments where the noise characteristics are highly non-stationary. Noise can be introduced at many points in a recording chain and some forms of noise, such as clipping or CODEC distortion, are normally signal dependent. In this thesis, however, we are concerned with additive acoustic noise which we assume is independent of the speech signal.

In this thesis we address the problems of adaptively tracking the noise characteristics and of efficiently updating the HMM-based noise model. We develop ways of detecting the occurrence of new noise sources and of rapidly incorporating them into the noise model. We also address the issues that arise in tracking the noise characteristics when speech may be present in the signal. The standard HMM training procedure [98] can only work on a fixed length of data. If there is any new arrival of data, we have to retrain the model from scratch. A computationally efficient on-line HMM re-estimation framework for noise estimation is required.

1.3 Thesis Structure

In Chapter 2, we first give an overview of a speech enhancement system, and then provide a literature review of various popular noise estimation methods that have been developed, including minimum statistics, minima controlled recursive averaging and the Hirsch histogram. Finally we discuss various way of assessing noise estimators. Many of the noise estimation methods described in Chapter 2 are based on first-order recursive averaging, which effectively assumes a slowly changing one-state model. Thus they cannot provide a good estimate of fast-changing or intermittent noise. A richer model is needed when dealing with highly non-stationary noise.

This is followed by Chapter 3, where we introduce a multi-state hidden Markov model for noise estimation. In this chapter, we model the noise in the absence of the speech.

We present an online hidden Markov model (HMM) recursive update framework that can track the noise and update the noise model. The noise characteristics within each state of the model are assumed to be slowly changing. A statistics measure (Z-test) is proposed to detect whether there is any abrupt change of the noise; when this occurs, a new state will be created to accommodate such noise. The HMM noise estimator described in this chapter can only work in the absence of speech and so we assume that there is a Voice Activity Detector (VAD) available that identifies when speech is present. For evaluation purposes, we use the initial noise-only segment of each test file to train the noise model and then leave the model fixed during speech presence.

In Chapter 4, in order to detect and update the noise even during the speech activity, we have incorporated a speech model into our existing noise HMM. The inclusion of the speech model improves the identification of novel noise types by ignoring any possible speech-like signals. Furthermore, a modified Minima Controlled Recursive Averaging (MCRA) method is used to update the noise characteristics within each state even when speech is present. An evaluation of this robust noise estimator is presented under different adverse environments.

Finally, in Chapter 5, we summarize the work presented in this thesis and give an outline of possible future work.

1.4 Scope and Original Contribution

The following aspects of this thesis are believed to be original contributions:

1. The online re-estimation HMM framework presented in Chapter 3, which can recursively update the HMM parameters without re-training the model from scratch.
2. The log likelihood measure to detect the presence of a novel noise type and the methods of creating and merging HMM states presented in Chapter 3.
3. The noise estimation method using trained noise HMM and LTASS presented in Chapter 3.

4. The HMM noisy speech model, presented in Chapter 4, which can recursively update the HMM parameters during speech activity.

5. The modified Minima controlled recursive averaging method to update the mean noise power spectrum within each HMM state presented in Chapter 4.

6. The log likelihood measure to detect a novel noise type during speech activity and the noise re-training scheme presented in Chapter 4.

Chapter 2

Literature Review

In this chapter, we will first review different speech enhancement algorithms in terms of their methodologies and their dependence on a good noise estimator. Different noise estimation methods will be discussed in terms of how they exploit distinct noise characteristic from speech and thus separate them, with brief discussion on their possible drawbacks when estimating highly non-stationary noise. Lastly, we review different possible methods of evaluating the performance of a noise estimator.

2.1 Speech Enhancement Algorithms

Speech enhancement algorithms aim to improve the quality and intelligibility of speech degraded by noise. For applications where no time delay is allowed, the most widely used approach is the Kalman filter [50], but if a small delay is permitted, it is more common to perform enhancement in the frequency domain. The advantage is that speech and noise are partially separated in the spectral domain, and their spectral components are somewhat decorrelated. Furthermore, many psycho-acoustic models are spectrally based and can be conveniently applied in this domain. The commonly used spectral domains are: (i) complex spectral amplitudes, (ii) spectral magnitudes, (iii) spectral powers, (iv) log spectral powers, (v) Mel-spaced spectral amplitudes, (vi) Mel-spaced log

spectral powers, (vii) Mel cepstral coefficients. Such spectral domain methods require the sampled input signal to be decomposed into overlapping frames using the Short Time Fourier Transform (STFT) (see Sec. 3.4.1 for more details) in order to estimate the power spectrum as it changes over time. The original signal can be perfectly reconstructed with overlap-addition if no processing is done on the frame spectra, provided that the analysis and synthesis windows are chosen appropriately [4].

Following Loizou [85], we divide speech enhancement algorithms into four categories which we will discuss below. We assume in all cases that the noise N is additive and independent of the wanted speech signal S, such that the complex amplitude of the observed signal O is given as, $O(t,k) = S(t,k) + N(t,k)$, where t and k are time and frequency index respectively. For all speech enhancement algorithms that will be discussed below, the estimated speech can be view as $\hat{S}(t,k) = G(t,k)O(t,k)$, where $G(t,k)$ is the gain function of the proposed speech enhancement algorithm.

2.1.1 Spectral-subtractive Algorithms

Since the noise is additive, an estimate of the clean signal spectrum can be obtained by subtracting an estimate of the noise spectrum from the noisy speech spectrum in the spectral power, or more commonly, the spectral magnitude domain [10]. The spectral subtraction gain function is given by

$$G_{SS}(t,k) = \max\left(\frac{|O(t,k)| - \left|\hat{N}(t,k)\right|}{|O(t,k)|}, 0\right) \tag{2.1}$$

where $\left|\hat{N}(t,k)\right|$ denotes the estimated noise magnitude. This class of algorithm is usually computationally simple as the enhanced signal can be obtained by computing the inverse discrete Fourier transform of the estimated signal spectrum using the phase of the noisy signal, and therefore only a forward and an inverse Fourier transform are required. The subtraction process typically introduces some speech distortion as well as residual noise artefacts known as musical noise. An over-subtraction factor κ [9, 84] is often used to reduce the residual noise after subtraction especially when the signal to

noise ratio (SNR) is poor. In addition, a spectral floor ϖ is often used to prevent the resultant spectral components from going below a preset minimum value, which is shown as,

$$G_{SS}(t,k) = \max\left(\frac{|O(t,k)| - \kappa\left|\hat{N}(t,k)\right|}{|O(t,k)|}, \varpi\right) \tag{2.2}$$

Many methods have been proposed to alleviate, and in some cases eliminate most of the speech distortion and musical noise introduced by the spectral subtraction process [56, 73, 86].

2.1.2 Wiener Filtering

The Wiener filter [128] reduces the amount of noise present in noisy speech by comparison with an estimate of the desired clean speech. It is a linear estimator of the clean speech spectrum, and it is optimal in the mean square sense. However, the ideal Wiener filter requires knowledge about the statistics of the clean speech power spectrum which is normally unavailable. The Wiener filtering algorithm can be implemented either iteratively or non-iteratively. A model of the clean speech spectrum, such as the AR speech production model [81], can be used iteratively to update the model and estimate the Wiener filter. For non-iterative methods, the Wiener filter can be expressed as a function of the "*a priori* SNR" ξ: the ratio of the clean signal power spectrum to the noise power spectrum. The enhanced speech spectrum can be obtained by multiplying the noisy speech spectrum by the Wiener filter, where noise would be suppressed according to the *a priori* SNR in each frequency bin. The Wiener filter gain is given by

$$G_{WF}(t,k) = \frac{\xi(t,k)}{1+\xi(t,k)} \tag{2.3}$$

A good estimate of the *a priori* SNR is needed for non-iterative Wiener filter methods, since a low-variance estimate of the *a priori* SNR can eliminate musical noise [12]. Ephraim and Malah proposed a decision-directed method of estimating the a priori SNR [29] that is widely used and gives good performance [67]. A hybrid Wiener spectrogram filter [24] exploits the correlation between different time frames to further reduce the

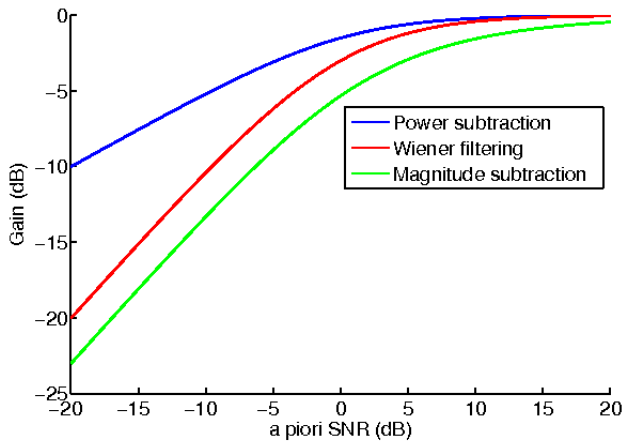


Figure 2.1: Gain of various classical enhancement methods at different SNR [125].

residual noise. Both the Wiener filter and spectral-subtraction can be viewed as estimating a gain that varies over time and frequency as a function of *a priori* SNR [125], which is shown in Fig. 2.1. At high SNRs, the gain converges to 0 dB, little suppression is done and most of the noisy signal will be treated as speech. Conversely, at low SNR, the gain is very low since there is little speech power compared to noise power. From the graph it can be seen that the Wiener filter gain characteristic is similar to that of magnitude subtraction but with an offset of about $+3$ dB in SNR.

2.1.3 Statistical-Model-Based Methods

The statistical-model-based algorithms are based on explicit stochastic models of the speech and noise [74]. Based on the speech and noise models and the observed noisy speech, the enhancement algorithm calculates either the minimum mean squared error (MMSE) or the maximum likelihood (ML) estimate of the clean speech. Given a set of measurements (e.g. noisy speech) that depend on some unknown parameters (e.g. clean speech), we wish to find a nonlinear estimator for these parameters of interest. Under the assumption of a deterministic signal with additive Gaussian noise, the ML estimate of the spectral amplitudes can be determined [94]. Using a Gaussian model for the complex spectral amplitudes of both speech and noise, Ephraim and Malah developed MMSE estimators for the spectral amplitudes [29] and log amplitudes [30] and conclu-

ded that the latter choice gave a much lower residual noise. This widely used approach has been extended by others [91, 13, 33, 5] to encompass other super-Gaussian distributions for the speech spectral coefficients. A basic MMSE estimator is given below,

$$G_{MMSE}(t,k) = \frac{\xi(t,k)}{1+\xi(t,k)} \exp\left(\frac{1}{2}\mathbf{E}_1\{\gamma(t,k)\}\right) \tag{2.4}$$

where ξ and γ are the *a priori* SNR and the *a posteriori* SNR respectively, and $\mathbf{E}_1$ is the exponential integral function [2]. For large values (>20 dB) of the SNR, the MMSE gain function is similar to the Wiener gain, where $G_{MMSE} \simeq \frac{\xi}{\xi+1}$.

These methods were initially developed under the assumption that speech was present at all times. In reality, there are many periods of speech pauses in between words and syllables. Furthermore, speech may not be present at a particular frequency during voiced speech segments. Therefore a better noise suppression rule may be produced if we assume a two-state model for whether speech is present or absent. Thus we could estimate the probability that speech is absent at a particular frequency bin, and incorporate this speech-presence uncertainty in the preceding estimators to reduce the residual noise substantially [16, 14, 57]. Similar to Wiener filtering, all the statistical-model-based algorithms require a good estimation of the SNR of the noisy speech, i.e. a better noise estimation will improve the quality of the enhanced speech signals when using above algorithms.

2.1.4 Subspace Algorithms

The subspace algorithms are based on the principle that the clean signal is generally confined to a subspace of the noisy Euclidean signal space [96]. By decomposing the vector space of the noisy signal into a subspace occupied primarily by the clean signal and a subspace occupied primarily by the noise signal, it is possible to estimate the clean signal by applying either the singular value decomposition (SVD) of a Toeplitz data matrix [21] or eigenvalue decomposition (EVD) onto the noisy signal covariance matrix [32, 103]. By nulling the noise subspace components and retaining the speech

subspace components, an enhanced speech signal is obtained with most of the noise signal is suppressed. The subspace algorithm can be described as $\vec{\mathbf{S}} = G_s\vec{\mathbf{O}}$, where $\tilde{\mathbf{S}}$ and $\vec{\mathbf{O}}$ are vectors of estimated speech and noisy speech respectively, and G_S is the gain matrix defined as,

$$G_S = \Delta\left(\Lambda - |N|^2 I\right)\Lambda^{-1}\triangle^{\#} \tag{2.5}$$

where $()^{\#}$ denotes conjugate transpose, and Δ and Λ is the eigenvalue decomposition of the noisy signal convariance matrix, i.e. $\tilde{\mathbf{O}}\vec{\mathbf{O}}^{\#} = \Delta\Lambda\Delta^{\#}$.

The original subspace algorithms assume the additive noise is white, if this assumption is not true, the quality of the enhanced speech may be severely affected due to incorrect estimation of the covariance matrix. Later algorithms apply pre-whitening of the noisy speech signal to enhance the noise estimation [66, 20]. The implementation of subspace algorithm is often computationally expensive as a SVD or EVD is required per frame.

2.1.5 Evaluating the Enhanced Speech

We may assess the improvement achieved by an enhancement in either intelligibility or quality, or a combination of both. Intelligibility measures assess the fraction of spoken words that can be correctly identified by a listener. They could be evaluated from subjective listening tests by calculating the percentage of words or phonemes that are identified correctly. These listening tests are usually based on the use of nonsense syllabus [34], monosyllabic words [26], rhyming words [126] or sentences [72] as speech material. Alternatively, intelligibility can also be estimated indirectly from the word accuracy of a speech recognizer on a standardized task such as the Aurora test database [63]. Recently a number of objective metrics such as the Short Time Objective Intelligibility (STOI) measure [119] have been developed that correlate strongly with subjective intelligibility tests. When comparing different speech enhancement algorithms, two common statistical tests, the t-test and analysis of variance (ANOVA) test , can be used to assess significant differences between algorithms in terms of intelligibility scores .

Quality measures assess various aspects of speech besides intelligibility, such as the speech being more natural, pleasant, and acceptable. Subjective quality assessments generally fall into one of two categories: those in which listeners are asked to express a preference amongst two or more stimuli and those in which they assign a numerical value to the quality of a single stimulus (i.e. quality rating). In some forms of the latter type, listeners assign separate values to different aspects of the quality such as speech distortion or background noise level. The relative preference task methods are usually subjective in nature, requiring listener's opinion on their preferences over paired-comparison tests, and might not be reliable. Objective quality measures estimate quality algorithmically by measuring aspects of the noisy signal and, in the case of intrusive measures, comparing the noisy and clean signals. Intrusive measures include the segmental SNR [122], spectral distance measures based on LPC (e.g. Itakura-Saito measures [53]), perceptually motivated measures such as the Bark distortion measures [127], and the perceptual evaluation of speech quality (PESQ) [70]. A well-presented evaluation of several objective measures is given in [65] where it was found that PESQ showed the strongest correlation to the listening test [1]. The PESQ score given in (2.6) is computed as a linear combination of the average disturbance value d_{sym} and the average asymmetrical disturbance value d_{asym},

$$PESQ = 4.5 - 0.1d_{sym} - 0.0309d_{asym} \tag{2.6}$$

where the range of the PESQ score is 0.5 to 4.5. High correlations ($\rho > 0.92$) with subjective listening tests were reported in [105] for telecommunication applications. A number of non-intrusive measures have recently been proposed that do not require access to the clean speech signal but that nevertheless correlate well with PESQ [51].

2.1.6 Summary

Although the speech enhancement algorithms above exploit different distinct features of speech to distinguish it from unwanted noise, they share the common objective of reducing the noise while introducing minimum speech distortion. Loizou [85] compared

different speech enhancement algorithms selected from each of the classes described above through subjective listening tests [1]. The MMSE based methods [29, 30] consistently performed the best with the highest quality and least speech distortion, while the subspace algorithms [64] yielded the lowest quality but were good at preserving the speech intelligibility. A few algorithms improved the quality of the enhanced speech significantly in some conditions, but none of them provided significant improvement when babble noise, recorded in a restaurant, was added to the speech.

2.2 Noise Estimation Algorithms

All the speech enhancement algorithms described above require an estimate of the noise power spectrum, or equivalently the signal to noise ratio, in each time frame.. For instance, the estimated noise spectrum is used directly in the spectral subtractive algorithms [10]. Alternatively it is used to evaluate the Wiener filter in the Wiener algorithm [128], to estimate the *a priori* SNR in the (log) MMSE algorithms [30] and to estimate the noise covariance matrix in subspace algorithms [32]. The accuracy of this estimate has a major impact on the overall quality of the speech enhancement: overestimating the noise will lead to distortion of the speech, while underestimating it will lead to unwanted residual noise. In this section, we will discuss various types of noise estimation algorithms and the ways in which they distinguish between the speech and noise of a signal.

2.2.1 Voice Activity Detection

The simplest approach for noise estimation is to use a voice activity detector (VAD) to identify when speech is absent and to estimate the noise power spectrum by averaging that of the input signal during these speech-free intervals. Speech absences occur not only at the beginning and end of a sentence but also in the middle of a sentence, primarily at the closures of the stop consonants. The appropriate averaging time-constant depends on the assumed stationary of the noise. In general, VAD algorithms output a

binary decision and declare a segment of speech to be voice active, i.e. VAD=1, if some measured quantity exceeds a predefined threshold. The decision rule of a VAD is based on one or more measurable quantities whose values differs between noise and speech [99, 35, 111, 121], where the decision threshold is often determined empirically. Most of the conventional VADs [111, 39, 109] assume that the noise statistics are stationary over a longer period of time than those of speech, which makes it possible to estimate the time-varying noise statistics in spite of the occasional presence of speech.

Generally, the VAD method does not work well if the noise signal is highly non-stationary or the signal to noise ratio (SNR) is low [27]. Even if the VAD algorithm is accurate, it still might not be sufficient in speech enhancement applications, especially in a highly non-stationary noise environment, where the noise characteristics might change frequently during the intervals when speech is present. Hence the noise spectrum should be ideally continuously estimated and updated from the noisy speech even during speech activity. In order to achieve this, it is necessary to make use of prior knowledge about differences between the characteristics of noise and speech. Common assumptions are that:

1. the short-time power spectrum of noise is more stationary than that of speech

2. the power of the noisy speech signal in individual frequency bands often decays to the power level of the noise

3. the frequency of periodic noise sources changes very slowly with time; this is in contrast to voiced speech whose pitch changes more rapidly.

These assumptions have led to noise estimation algorithms that will be reviewed in the rest of this section. To utilize the first assumption, most noise estimation algorithms analyse the noise signal using short-time spectra, i.e. STFT frame-based processing. The analysis segment should be chosen to be long enough to contain speech pauses and low-energy signal segments, but short enough to track fast changes in the noise level.

2.2.2 Minimum-Tracking Algorithms

The minimum-tracking algorithms assume that even during speech activity, the power of a noisy speech signal in each individual frequency bands will frequently decay to the power level of the noise, i.e. the energy will be dominated by the noise. Hence it is possible to estimate the noise using Minimum Statistics [93] by tracking the minimum of the noisy signal power in each frequency band regardless whether speech is present or not. The noise power is assumed to be the minimum power that has arisen within a past window length of T (typically 0.5 to 1.5 seconds). This window length must be chosen to be large enough to bridge any broad peak of the speech signal. The minimum statistics algorithm is illustrated in Fig. 2.2 in which the upper trace shows the power in one sub-band (centred at 250 Hz) of a noisy speech signal. The lower trace in Fig. 2.2 shows the output of a minimum filter with $T = 0.8$ s. The output of the minimum filter will inevitably underestimate the true mean noise power and it is necessary to compensate for this bias. A fixed compensation factor was used in [93], and the estimated noise power spectrum $\hat{N}(t,k)$ as given in (2.7), where t and k are the time and frequency index respectively.

$$\hat{N}(t,k) = BP_{min}(t,k) \tag{2.7}$$

where the value of the fixed bias B depends on the minimum search window length L, and it was set to $B = 1.5$ for $L = 100$ [93]. $P_{min}(t,k)$ is the minimum within the past L frames of the search frames of the smoothed power spectrum $P(t,k)$, which is given in (2.8).

$$P(t,k) = \lambda P(t-1,k) + (1-\lambda) O(t,k) \tag{2.8}$$

where λ is the smoothing constant typically set to between 0.9 to 0.95.

The algorithm was extended later to include a bias factor that varied with time and frequency [90], and the minimum was found by searching the first-order recursively smoothed periodogram. A better approximation of the compensation factor was proposed by considering the statistics of the smoothed periodogram [92].

The minimum statistics algorithm [90] was reported to perform well in non-stationary

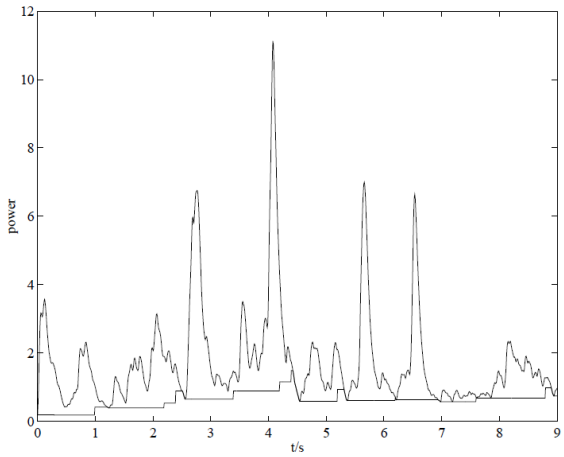


Figure 2.2: Power in the 250Hz sub-band of a noisy speech signal and the output of a minimum filter with $T = 0.8$ s (reproduced from [93]).

noise, but the adaptation time can exceed T when the noise level is increasing [101]. A similar approach was used in [25] but, instead of taking the minimum over T, the noise speech spectrum is smoothed using two different time constants; a short time constant is used when the energy in a frequency bin is decreasing to ensure rapid adaptation to a new minimum while a long time constant is used when the power increases to prevent adaptation to the speech energy. The approach is computationally efficient but is considered to perform less well than the minimum statistics approach because selecting the long time constant is a compromise between the response to sudden increases in noise and preventing the speech power from modulating the estimated noise power. The adaptation time of the minimum statistics approach can be improved by using a Bayesian approach to obtain a more robust estimate of the speech presence probability [95, 38].

The use of the minimum makes the technique sensitive to outliers, and other quantiles might be used instead in order to improve robustness. This was later extended to a two-pass approach [7] in which the quantile used in the second pass depended on the estimated SNR from the first pass. It was found in [115] that the median gave the best results when evaluated using a speech recogniser. Few people, however, appear to have followed up this work and Manohar [89] demonstrated that this approach performs poorly on non-stationary noise.

2.2.3 Time-recursive Algorithms

The time-recursive averaging algorithms exploit the fact that, even when speech is present, the spectral power in some frequency bins will be dominated by the noise. Thus individual frequency bins of the estimated noise spectrum can be updated whenever the speech-presence probability in that particular frequency band is low. Consequently, the noise spectrum can be estimated as a weighted average of past noise estimates and the present noisy speech spectrum [61]. The weight, or smoothing factor, is adaptive, and usually takes large values when speech is absent and small values when speech is present in each frequency bin [6]. The time-recursive averaging algorithms can be regarded as a soft-decision VAD, where the estimated noise spectrum can be updated all the time but with a time-constant that varies according to the probability of speech presence. Sohn [112] proposed a soft-decision VAD that was based on the likelihood ratio that is equivalent to the Itakura-Saito distortion measure or cross entropy between background noise and observed signal [110, 54]. A similar approach is used in Malah [88] where the estimated SNR averaged across all frequencies is used to control adaptation together with an additional frequency-dependent factor that depends on the estimated speech presence in each frequency bin. Lin [82, 83] presented a simpler version, where the noise is updated adaptively from noisy speech with a time constant which is based on a sigmoid function of the noisy speech to noise ratio. Gerkmann [46] proposed a soft speech presence probability estimation without any minima searching methods, which shows a good performance when incorporated with the MMSE-based noise estimation [58].

In minima-controlled recursive averaging algorithms (MCRA) [17, 16], the speech presence probability in each frequency bin is estimated by taking the ratio between the power in the current frame and its minimum over a searching window of length T, which is then used to control which sections of the noisy speech are averaged to estimate the noise power. It gives an estimate with less bias and reduced variance than the original approach [93]. The MCRA algorithm estimates the noise as,

$$\hat{N}(t,k) = \hat{\lambda}(t,k)\,\hat{N}(t-1,k) + \left(1 - \hat{\lambda}(t,k)\right) O(t,k) \tag{2.9}$$

where $\hat{\lambda}(t,k)$ is the time-frequency dependent smoothing parameter,

$$\hat{\lambda}(t,k) = \lambda + (1-\lambda)p(t,k)$$

where λ is the fixed smoothing constant, and $p(t,k)$ is the smoothed speech presence probability depending on whether the ratio $\frac{P_{min(t,k)}}{P(t,k)}$ exceeds a fixed threshold (the threshold is set to 5 in [17]). $P_{min(t,k)}$ is the minimal tracking on the smoothed noisy power spectrum $P_{(t,k)}$ defined in (2.8).

This noise estimator was developed slightly in [108] and an improved version of the noise estimator was given in [15] which uses a two-iteration procedure that refines an initial speech presence detector. This approach requires a local minimum over a fixed length of window, thus fails to adapt quickly to any abrupt changes. Rangachari [101] extends this approach in two ways by using a different way of calculating the minimum spectrum without any fixed windows [25], which resulted in a lower latency (0.5 s instead of 1.5 s) and a frequency-dependent threshold on the ratio of noisy speech spectrum to minimum spectrum which is used to estimate the speech presence probability and thence to control the adaptation rate. The algorithm is improved slightly in [100] by smoothing the speech presence probability over time which implicitly accounts for its correlation between successive frames. The smoothing approach is extended further in [78, 79] where the threshold used when deciding speech presence probability varies with the likelihood of speech presence in the previous frame.

Hendriks [59] calculates a high resolution DFT with four times the required resolution and compares this with the current low resolution noise spectrum estimate to give a VAD decision for each high resolution frequency bin. The low resolution noise spectrum is then updated using any bins (between 0 and 4) that are classified as noise-only together with an empirical bias compensation factor that is a function of the number of noise-only bins. They found that this approach performed comparably to a subspace approach [60] but with much lower computational cost. They found it was better than [90] in almost all cases, especially for rapidly changing noises.

2.2.4 Histogram-based Methods

The histogram-based algorithms exploit the fact that at moderate SNRs the histogram of noisy-speech energy values will include a distinct peak corresponding to the noise. The length of the window used for constructing the histogram will affect the performance of such algorithms. Noise overestimation occurs when the window is not long enough to encompass broad speech peaks, especially in the low-frequency bands, where the speech energy is often high. McAulay [94] proposed an energy histogram algorithm that determined an adaptive energy threshold to decide on the presence of speech, along with fixed upper and lower threshold limits which take priority. The adaptive threshold is chosen to lie at the 80th centile of the histogram of energies that are below the upper fixed threshold. This approach is modified in [18] which fits a 2-component Gaussian mixture model to the histogram of log energy and assumes that the lower component represents the noise. A similar approach is used by [61] in which an adaptive threshold is used in each frequency bin to eliminate speech frames and the peak of the histogram of recent noise frames is used as an estimate of the noise power in that bin. The reported accuracy of this approach, which is an extension of [62], is much greater than that of the VAD approach. The Hirsch histogram methods is shown in (2.10) below,

$$\hat{N}(t,k) = \lambda \hat{N}(t-1,k) + (1-\lambda)\, O_{\text{mode}}(t,k) \tag{2.10}$$

where λ is the smoothing constant, and $O_{\text{mode}}(t,k)$ is the mode of the distribution of the noisy speech spectrum histogram $O(t,k)$ over the past L frames.

A similar approach is also used by Ying [131] to train a sequential Gaussian Mixture Model (SGMM) to track noise power in log-spectral domain. Ris [104] proposed a new approach based on the harmonic filtering, where the speech periodicity property was used to update the noise level estimate during voiced parts of speech, and thus can track fast modulations of the noise energy. However, in his study, he found that minimum statistics performed better than the Hirsch histogram [61].

[132] and [45] demonstrate that the discrete Fourier transform (DFT) or Karhunen

Lòeve Transform (KLT) coefficients of speech signals follow a Laplacian distribution rather than the more commonly assumed Gaussian or Gamma distributions. They therefore propose a voice activity detector [44] that models the noisy signal subband coefficients as the sum of zero-mean Laplacian and Gaussian random variables respectively. They find that using the KLT rather than the DFT gives marginally better performance.

Although histogram-based algorithms can estimate the noise without any implicit or explicit VAD, they only track the most frequent occurrence of energy values as the noise signal, whereas other types of noise component with lower probability of presence will be ignored. Hence they generally do not work well with a mixture of different types of highly non-stationary noises.

2.2.5 Overview

A well-presented evaluation of several noise spectrum estimation techniques is given in [120] where it was found that the best performance of the tested algorithms was given by MMSE-Hendriks [58], and that this outperformed minimum statistics [90] and MCRA [16, 15] on a wide range of noises in terms of the mean and variance of the log estimation errors.

Most of the noise estimation algorithms described above have implicitly or explicitly detected whether speech is absent and estimated the average noise power during these intervals. During speech activity, the noise normally will not be updated, and it is assumed that the noise characteristic is unchanged during these intervals. A two-component noise model [31], which is used to account for a slowly evolving component and a random component has been proposed to have a better noise estimate. However, if the noise is intermittent or highly non-stationary, such as media interference and unwanted co-talkers, describing such noise requires a richer model.

2.3 Noise Estimator Performance Estimation

In many speech enhancement applications, the objective is to estimate the clean speech as closely as possible. Much attention has been given to evaluating the enhanced speech, but little has been done on noise estimator assessment. There are mainly two ways to assess the noise estimator, one is to compare the improvement of both quality and intelligibility of the enhanced speech through the proposed speech enhancement of interest, the other is to compare the noise power spectrum with its estimate. We will discuss briefly both methods and their drawbacks.

2.3.1 Evaluating the Performance of Noise Estimators

Currently, there are very few methods or algorithms that have been developed specifically for noise estimator assessment. Many performance evaluations of noise estimators are done by judging indirectly from the enhanced speech in the specific application of interest, and showing that incorporating a better noise estimator would further improve the speech enhancement algorithm. For example, Deng et. al. [23] presented a non-stationary noise estimator using iterative stochastic approximation, and evaluated using AURORA-2 noisy digit recognition, and quantitatively showed it is better than MMSE noise estimator [28]. Such approaches are specific to specific to one application rather than a general assessment of the noise estimator.

Another method is to compare the similarity between the original noise power spectrum N_k and its approximation $\hat{N}_k$, where k is the frequency index in each STFT time block. A number of quantitative spectral distortion measures have been developed for measuring the closeness between two signal spectrum. One of them is log spectral distance, there are a few variations to assess the distortion between two spectra. Rangachari [100] used an MCRA based method to estimate the highly non-stationary noise and assessed using mean square error $\frac{1}{K}\sum_k \frac{\left(N_k-\hat{N}_k\right)^2}{N_k}$ between the estimated and the true noise. Such a measure may give misleading results as it is sensitive to outliers in one or two frequency bins. Zhao [133] proposed an online noise estimation method and used log

spectral distance $\sqrt{\frac{1}{K}\sum_k \left(10\log_{10}\frac{N_k}{\hat{N}_k}\right)^2}$ to assess the goodness of the noise estimation, whereas Hendriks [59] used $\frac{1}{K}\sum_k \left|10\log_{10}\frac{N_k}{\hat{N}_k}\right|$ instead. Taghia [120] proposed similar error measures and further included the analysis of the variance of such error measurement, and claimed that if the log distance was similar, the one having the smaller variance of log distance would be preferred due to its lower tendency to produce musical noise.

The Itakura distortion d_I is a psycho-acoustically motivated distortion measure in the log spectral domain:

$$d_I\left(N,\hat{N}\right) = \log\left(\sum_k \frac{N_k}{\hat{N}_k}\right) - \sum_k \log\frac{N_k}{\hat{N}_k}$$

However the Itakura distortion only measures the closeness of the spectral shape rather than their absolute spectral difference. Since an accurate estimation of noise spectral characteristic is important, and any estimation error might degraded the quality of the enhanced speech, the Itakura measure is not a good choice for the evaluation of noise estimators. The Itakura-Saito distortion d_{IS} is a gain-dependent version of Itakura-Saito distortion and it given by

$$d_{IS}\left(N,\hat{N}\right) = \sum_k (\frac{N_k}{\hat{N}_k} - \log\frac{N_k}{\hat{N}_k}) - 1$$

Owing to its asymmetric nature, the Itakura-Saito distortion will give more emphasis to noise underestimation (i.e. $\frac{N_k}{\hat{N}_k} > 1$) than noise overestimation (i.e. $\frac{N_k}{\hat{N}_k} < 1$). The COSH distortion d_{COSH} is the symmetric version of Itakura-Saito distortion, which weights noise overestimation and underestimation equally. We have used the COSH measure in our evaluation below because, as mentioned above, both overestimation and underestimation of the noise have serious effects.

$$d_{COSH}\left(N,\hat{N}\right) = \frac{1}{2}\left[d_{IS}\left(N,\hat{N}\right) + d_{IS}\left(\hat{N},N\right)\right] = \sum_k \cosh(\log\frac{N_k}{\hat{N}_k}) - 1$$

Gray [52] evaluated a number of spectral distortion measures and showed that the COSH and log spectral distortion measures are identical for small errors but recommended the use of the COSH measure because it gives a greater penalty to large errors.

2.4 Review and Summary

Among different classes of speech enhancement algorithms, the MMSE based methods have been found to perform the best with the highest quality and least speech distortion [85]. Furthermore, it is also computationally efficient and commonly used. In the rest of the thesis, we will use the MMSE method [30] as our speech enhancement algorithm.

Various noise estimation methods have been discussed in Sec. 2.2, histogram-based methods [61] are usually computational expensive, and might ignore intermittent noise that occurred infrequently. Minimum statistics [90] showed a good performance when estimating non-stationary noise [104], but the adaptation time can exceed T when the noise level is increasing [101]. Minima controlled recursive averaging method [17, 16] relaxed the requirement of a VAD, and provide an elegant way to update the noise in frequency bins where the speech is absent. The minimum power within a specified time frame serves as an indicator of speech presence probability mask. A recent evaluation [120] found that MMSE-Hendriks [58] gives the best performance over the noise estimation methods mentioned above. The MMSE-Hendriks was later extended by Gerkmann [46], and we will use this unbiased MMSE-based noise power estimator as one of our reference algorithms for performance evaluation.

The noise estimator can be judged indirectly from the enhanced speech when incorporated into a specific speech enhancement system. Alternatively, the goodness of the noise estimation can also be assessed by the spectral distortion between the true noise spectrum. COSH distance is identical to log spectral distortion measures for small errors, and showed a greater penalty to large errors [52]. It will be used for spectral distance measure in the rest of this thesis.

Chapter 3

Multi-state HMM Noise Model

3.1 Overview of a HMM

The hidden Markov Model(HMM) is a powerful multi-state model that can be characterised by an underlying process generating an sequence of observation. The underlying process is assumed to be a Markov process with unobserved (hidden) states, where the conditional probability distribution of future states of the process depends only upon the present state. Within each state, all possible observations are emitted with a finite probability. Thus given a observation sequence and the HMM, we can determine the most likely state sequence to produce the observations. Fig. 3.1 shows a 3-state HMM, and the arrows indicate the transition from one state to another.

3.2 Introduction

Many of the noise estimation methods described in Chapter 2, such as minimum statistics [90], minima controlled recursive averaging [17, 16] and unbiased MMSE-based noise estimator [46], are based on first-order recursive averaging, which effectively assumes a one-state noise model. Thus they might not follow a rapid change of the noise

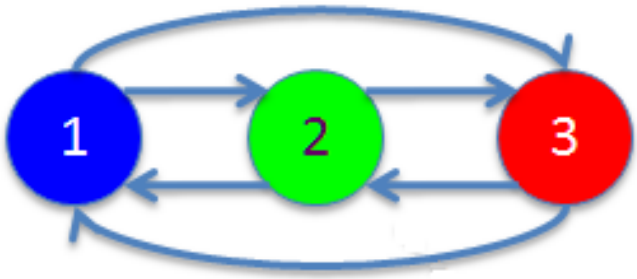


Figure 3.1: Illustration of the 3-state HMM.

characteristics, especially intermittent noise. In this chapter, we propose a multi-state model for estimating highly non-stationary noise. There are many situations in which the nature of the interfering noise will change over time. In some cases, the characteristics (and hence power spectrum) of a source may change gradually. When this occurs, we would like to adapt the noise characteristics associated with the corresponding state so that it tracks the changes of the source. In other cases, the noise may change abruptly due to the introduction of a new noise source or a change in the operation of an existing one. Such abrupt changes should be represented in the noise model by the creation of new model states. Over time, the occupancy of some states will fall to almost zero and they can be removed from the model; it may however be advantageous to retain their characteristics in a library so that, if the noise source reappears, its model does not need to be retrained from scratch. Inspired by the human hearing ability to learn and adapt to a new noisy environment [8], non-stationary noise can therefore be better modelled as a set of discrete states which capture the characteristics of noise sources encountered in the past. In this approach, each state corresponds to a distinct noise power spectrum and the evolution of the state sequence over time is conveniently represented by a hidden Markov model (HMM).

Our aim in this chapter is to develop an on-line noise HMM that can track and update the distinct noise characteristics represented by each state, and create a new state if a novel noise source is detected. For simplicity at the start, in this chapter, we assume that there is a perfect voice activity detector (VAD) that can identify exactly when the speech is present, which enables us to train and update the noise model using noise

signals only without the presence of speech. The structure of the rest of this chapter is as follows. We first give a brief literature review of multi-state noise model. Next we develop an on-line noise HMM recursive update algorithm for estimating slowly evolving noise environment, followed by the extension of the HMM model to accommodate any abruptly changing noises. Finally, the performance of the HMM model is evaluated both in estimating the noise spectrum and when used with a speech enhancement algorithm.

3.3 Literature Review

In this section, we will briefly review different noise models that have been proposed.

3.3.1 Stationary Spectral Model

The most common model of the noise is that it is a Gaussian process with a slowly changing power spectrum. The spectrum is normally represented by individual spectral, mel-spectral or cepstral coefficients. As an example, an all-pole spectrum model is used for the noise process and an approximation to the maximum likelihood estimate of the auto-regressive (AR) parameters is provided by conventional Linear Predictive Coding (LPC) analysis [31, 107]. The distance between two AR signals can be expressed as the dot product between their autocorrelations [71].

3.3.2 Two-component model

Rennie et. al [102] propose to model the noise as the sum of a slowly evolving component and a random component in the Mel log-power spectral domain, and claim that this model is both more realistic and allows better tracking of the evolving component. The power of the continuously evolving component is modelled as a first order Gaussian AR process in each frequency bin, while the random component is zero-mean Gaussian. In order to account for abrupt changes in noise level, there is a small but non-zero

probability that the continuously evolving component may revert to a prior mixture of diagonal-covariance Gaussians; they also suggest reverting to a minimum-statistics noise estimate as an alternative. The paper gives update procedures for the mean and variance of the noise level components under the assumption of a fixed Gaussian mixture model for the speech. A similar noise model was also used implicitly by [89].

3.3.3 Hidden Markov Model

A number of authors present their noise model as a multi-state HMM, but in most cases they actually use only a single state (albeit with multiple mixture components). A brief review of ways of estimating a noise HMM can be found in [49]. Sameti et al. [107] use a 3-state Gaussian mixture model (GMM) model for each noise type. During non-speech intervals, a library of noise types is searched and is used for any subsequent speech spurt. In order to reduce the danger that fricatives will be interpreted as noise, only non-speech intervals longer than 100 ms are used. Srinvansan [113, 114] uses a codebook of auto-regressive (AR) coefficient sets for both speech and noise spectra and explicitly finds the MMSE choice of AR coefficients and maximum likelihood gains for each frame; the codebook used is mostly predefined from training noise data but also includes the estimate from minimum statistics [90]. A similar codebook model was used by Zhao [133] who used 10th order LPC, eight 16-mixture states for speech and five states for noise; however their system is different in how the noise states are updated. The system updates the noise model at each frame using an expectation–maximization (EM) procedure with a forgetting factor, to update noise states and noise gains. They show that this method tracks noise amplitude changes better than minimum statistics, and gives a better improvement of segmental SNR. They also note that the LPC order of 6 is inadequate for noise spectra with many harmonics.

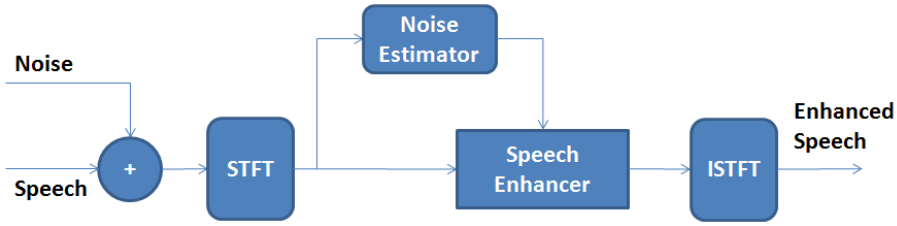


Figure 3.2: Overview of a typical speech enhancement system.

3.4 HMM noise Modelling

In this chapter, we assume that there is a perfect voice activity detector (VAD) that can identify exactly when the speech is present, which enables us to train and update the noise model using noise signals only without the presence of speech. This assumption will be removed in Chapter 4 where we extend the algorithm to allow operation in the presence of speech. Within each state of the HMM, there is only one mixture component to represent a distinct type of noise power spectrum. In the following sections, we will first consider how the noise model can be updated effectively when the characteristics of the noise change slowly, then how it can adapt to abrupt changes in the noise.

3.4.1 Frame-based processing

In many speech enhancement systems, the short time Fourier transform (STFT) is used to analyse the characteristics of signals which change over time. Fig. 3.2 shows a typical speech enhancement system, where the noisy speech signal is assumed to be the sum of the noise and speech signal. It is first converted from the time domain to the frequency domain using the STFT, where the noise can be estimated and used by the speech enhancer to obtain the enhanced speech. It is then converted back into the time domain by using inverse short time Fourier transform (ISTFT). In the remaining sections of this chapter, we assume that the observed noisy speech signal is decomposed into overlapping frames, which are windowed and transformed into the frequency domain using the Discrete Fourier Transform (DFT). We have used a 50% overlap of frames and applied a square-root Hanning window in both the analysis and synthesis stages to give per-

fect reconstruction in the absence of any frequency domain processing. The observed complex-valued frequency-domain signal in time frame t is defined as $o_t(k)$ with frequency index $k \in \{1 \ldots K\}$. Assuming the noise is additive, the observed signal model is defined as $o_t(k) = s_t(k) + n_t(k)$, where $s_t(k)$ and $n_t(k)$ are clean speech and noise respectively.

3.4.2 Model Structure

In this chapter, we assumed there is a perfect VAD to identify when speech is absent, i.e. $o_t(k) = n_t(k)$ for frames we will use for training the noise model. We first initialize the Hidden Markov model on the basis of the first T_0 observed frames, i.e. $O^{(T_0)} = \{O_t : t \in [1, T_0]\}$[3]. The model parameter set for an HMM with H states is $\zeta = \{\pi, A, B\}$ [98], where $\pi = \{\pi_i\}$ is the set of initial state probabilities, $A = \{a_{ij}\}$ is the set of state transition probabilities and $B = \{b_j(O_t)\}$ is the set of observation probabilities within each state j.

Following [29] we assume that in time frame t, the spectral component of the noise in frequency bin k, $o_t(k)$, is Gaussian distributed with uncorrelated real and imaginary parts. Under this assumption, the power spectral components $O_t(k) = |o_t(k)|^2$ will follow a negative exponential distribution or, equivalently, a χ^2 distribution with 2 degrees of freedom,

$$p(O_t(k)) = \frac{1}{E\{O_t(k)\}} \exp\left(-\frac{O_t(k)}{E\{O_t(k)\}}\right) \tag{3.1}$$

where $E\{\}$ denotes expectation. Under the assumption of a perfect VAD, the speech is absent in the observed signal $O_t(k)$, i.e. $O_t(k) = N_t(k)$, where $N_t(k)$ is the noise power spectrum. Hence, with respect to a mean noise power spectrum vector μ_0, the log observation probability $\log b(O_t)$ is given by

$$\log b(O_t \mid \mu_0) = \log\left(\prod_k \frac{1}{\mu_0(k)} \exp\left(-\frac{O_t(k)}{\mu_0(k)}\right)\right) = \sum_k (-\log \mu_0(k) - \frac{O_t(k)}{\mu_0(k)}) \tag{3.2}$$

under the assumption that the frequency components of O_t are conditionally independent given μ_0.

If we do not have any information about the initial state probabilities, π, we can assume them to be the steady state probabilities of the HMM, i.e. π is taken to be the eigenvector satisfying $A^T\pi = \pi$. Furthermore, the observation probabilities B are determined from (3.2) using $\mu_0 = \mu_j$, where μ_j is the mean power spectrum of state j. Therefore, we can simplify the noise model as $\zeta = \{\mu, A\}$ where $\mu = \{\mu_j : j \in [1, H]\}$.

In the following section we will develop an adaptive algorithm for estimating the HMM. Since our noise model is adaptive, we will denote the model at time T by $\zeta^{(T)}$, and, in general, we will use the $^{(T)}$ superscript to denote the model parameters estimated from the available observations $O^{(T)} = \{O_t : t \in [1, T]\}$. The superscript will normally be omitted if all quantities in an equation are from the same model.

3.4.3 Model Initialization

In order to initialize the noise model, we first cluster the initial T_0 frames into H states using the k-means algorithm where $T_0 \gg H$. We then use Viterbi decoding [97] to obtain the maximum likelihood sequence of states $i(t)$. The mean spectrum in state j, μ_j, is then taken to be the average of all frames assigned to state j and the transition probabilities are calculated as

$$a_{ij} = \frac{c_{ij}(1, T_0)}{\sum_j c_{ij}(1, T_0)} \tag{3.3}$$

where $c_{ij}(1, T_0)$ is the total transition count from state i to state j based on the maximum likelihood state sequence $\{i(t) : t \in [1, T_0]\}$. It can happen that within this state sequence $i(t)$, there is not any transition from state i to state j, i.e. $c_{ij}(1, T_0) = 0$. If this happens, the transition will be forbidden and a_{ij} will remain permanently at zero. Under the assumption that the prior probability of each state is $1/H$, Laplace's law of succession [48] suggests that the state probabilities should be estimated by including

one additional "pseudo-count" for each state. Dividing this pseudo-count equally between the H possible next states results in the following estimate for the state transition probabilities which is used instead of (3.3):

$$a_{ij} = \frac{\frac{1}{H} + c_{ij}\left(1, T_0\right)}{1 + \sum_j c_{ij}\left(1, T_0\right)} \tag{3.4}$$

Thus the constant terms in the numerator $1/H$ and denominator 1 ensure that the probability of any state transition that has not yet been observed is initialized to a small positive value. If $\sum_j c_{ij}\left(1, T_0\right)$ is large, then (3.4) is approximately the same as (3.3). If$\sum_j c_{ij}\left(1, T_0\right)$ is 0, then the transition probabilities to state j will be equal. Thus the initial model is created as $\zeta^{(T_0)} = \left\{\mu^{(T_0)}, A^{(T_0)}\right\}$.

The processing steps of the model initialization can be summarized as follows:

1. Cluster the initial T_0 frames into H states using the k-means algorithm
2. Apply Viterbi decoding to obtain the maximum likelihood state sequence $i\left(t\right)$
3. Compute the mean $\mu^{(T_0)}$ and the state transition probability $A^{(T_0)}$ based on $i\left(t\right)$

3.4.4 Model Update Equations

From the standard Baum-Welch equations [97], we obtain a recursive expression for the forward probability α and backward probability β for the model, $\zeta^{(T)}$, based on the observations, $O^{(T)}$,

$$\alpha_i^{(T)}(t) = \sum_j \alpha_j^{(T)}(t-1) a_{ji} b_i(O_t) \qquad \text{with} \quad \alpha_i^{(T)}(0) = \pi_i \tag{3.5}$$

$$\beta_i^{(T)}(t) = \sum_j a_{ij} b_j(O_{t+1}) \beta_j^{(T)}(t+1) \qquad \text{with} \quad \beta_i^{(T)}(T) = \pi_i \tag{3.6}$$

where $b_j\left(O_t\right)$ is the observation probability of O_t belonging to the state j as given by (3.2). In addition, we introduce the total probability density, P, of the observation $O^{(T)}$ and it will be used as a normalization constant in (3.9) - (3.11) below

$$P^{(T)} = \sum_i \alpha_i^{(T)}(T)\beta_i^{(T)}(T) \tag{3.7}$$

Thus the model $\zeta^{(T)} = \left\{\mu^{(T)}, A^{(T)}\right\}$ can be obtained

$$\begin{aligned}
\mu_i^{(T)} &= \frac{\sum_{t=1}^{T} \lambda^{T-t}\alpha_i^{(T)}(t)\beta_i^{(T)}(t)O_t}{\sum_{t=\tau_1}^{T} \lambda^{T-t}\alpha_i(t)\beta_i(t)} \\
a_{ij}^{(T)} &= \frac{\sum_{t=1}^{T-1} \lambda^{T-1-t}a_{ij}^{(T-1)}\alpha_i^{(T)}(t)b_j(O_{t+1})\beta_j^{(T)}(t+1)}{\sum_{t=\tau_1}^{T-1} \lambda^{T-1-t}\alpha_i^{(T)}(t)\beta_i^{(T)}(t)}
\end{aligned} \tag{3.8}$$

These are the standard Baum-Welch update equations except for the exponential "forgetting factor" λ^{T-t} which reduces the contribution of time frames that are in the distant past [76]. The choice of λ is a compromise between being able to track rapidly changing noise characteristics within a single state (where λ is small) and obtaining good parameter estimates (where λ is close to 1).

In order to simplify the development in the following sub-section, we define the following quantities:

$$U_i(\tau_1, \tau_2) = \frac{1}{P}\sum_{t=\tau_1}^{\tau_2} \lambda^{\tau_2-t}\alpha_i(t)\beta_i(t)O_t \tag{3.9}$$

$$Q_i(\tau_1, \tau_2) = \frac{1}{P}\sum_{t=\tau_1}^{\tau_2} \lambda^{\tau_2-t}\alpha_i(t)\beta_i(t) \tag{3.10}$$

$$R_{ij}(\tau_1, \tau_2) = \frac{1}{P}\sum_{t=\tau_1}^{\tau_2} \lambda^{\tau_2-t}\alpha_i(t)b_j(O_{t+1})\beta_j(t+1) \tag{3.11}$$

where the U, Q, R are the weighted sums of the state observations, state occupancies and transition prevalence respectively. With these definitions, the model can be expressed as

$$
\begin{aligned}
\mu_i^{(T)} &= \frac{U_i^{(T)}(1,T)}{Q_i^{(T)}(1,T)} \\
a_{ij}^{(T)} &= \frac{a_{ij}^{(T-1)} R_{ij}^{(T)}(1,T-1)}{Q_i^{(T)}(1,T-1)}
\end{aligned}
\tag{3.12}
$$

Notice that the quantities U, Q, R have been normalized by the total stationary probability density P, as described in (3.7). The model parameter μ_i and a_{ij} are unaffected since the normalization factor $\frac{1}{P}$ presents both in the nominator and denominator, so that (3.12) will be exactly the same as (3.8). However, the quantity P is important for the recursive update that will be introduced in Sec. 3.4.5.

3.4.5 Recursive noise model update

In this section, we derive a procedure for updating the noise model recursively so that it is able to track slowly varying noise sources. This will be extended in Sec. 3.4.6 to the tracking of rapidly changing noise spectra. We assume that we have already determined $\zeta^{(T-1)}$ and now wish to perform a time update on the model to obtain $\zeta^{(T)}$. Re-evaluating (3.5) - (3.12) directly would require us to save the entire set of observations $\{O_t\}$. To avoid this, we wish to retain only the L most recent observations and assume that for sufficiently old frames, the state occupation probabilities are unchanged, i.e.

$$
\frac{\alpha_i^{(T)}(t)\beta_i^{(T)}(t)}{P^{(T)}} \approx \frac{\alpha_i^{(T-1)}(t)\beta_i^{(T-1)}(t)}{P^{(T-1)}} \quad \text{for} \quad t \leq (T-L) \triangleq T_L \tag{3.13}
$$

We will use T_L instead of $T-L$ for compactness in all the equations below. Noticing that the forward transition probability is independent of time T, i.e. $\alpha_i^{(T)}(t) = \alpha_i^{(T-1)}(t)$, the difference between the two quantities in (3.13) arises because β_i is calculated using (3.6), from a starting point of time T and $T-1$ respectively. The assumption in (3.13) will be valid provided that the state assignment at $t = T$ has a negligible effect on that at $t = T_L$ or, equivalently, that the second largest eigenvalue of A is $\ll 1-1/L$. Since the transition

probability matrix can be expressd as $\mathbf{A}^n = \mathbf{V}^{-1}\mathbf{D}^n\mathbf{V}$, where the columns of $\mathbf{V}$ contain the eigenvectors of $\mathbf{A}$, and $\mathbf{D}$ is a diagonal matrix with respective eigenvalue, the rate of convergence will depend on non-unity eigenvalues of the transition probability matrix. With this assumption, the normalized partial accumulated sum of the mean U can be expressed as below:

$$\begin{aligned} U_i^{(T)}(1, T_L) &= \frac{\sum_{t=1}^{T_L} \lambda^{T_L - t} \alpha_i^{(T)}(t) \beta_i^{(T)}(t) O_t}{P^{(T)}} \\ &\approx \frac{\sum_{t=1}^{T_L} \lambda^{T_L - t} \alpha_i^{(T-1)}(t) \beta_i^{(T-1)}(t) O_t}{P^{(T-1)}} = U_i^{(T-1)}(1, T_L) \end{aligned}$$

Similarly, we find that

$$\begin{aligned} Q_i^{(T)}(1, T_L - 1) &\approx Q_i^{(T-1)}(1, T_L - 1) \\ R_i^{(T)}(1, T_L) &\approx R_i^{(T-1)}(1, T_L) \end{aligned}$$

These equations show that we can assume the accumulated sum of each quantity remains unchanged up to time T_L given the arrival of a new frame T. Thus we can write the update equations as

$$\mu_i^{(T)} \approx \frac{\lambda^L U_i^{(T-1)}(1, T_L) + U_i^{(T)}(T_L + 1, T)}{\lambda^L Q_i^{(T-1)}(1, T_L) + Q_i^{(T)}(T_L + 1, T)} \tag{3.14}$$

and

$$a_{ij}^{(T)} \approx \frac{a_{ij}^{(T-1)} \left(\lambda^L R_{ij}^{(T-1)}(1, T_L - 1) + R_{ij}^{(T)}(T_L, T - 1) \right)}{\lambda^L Q_i^{(T-1)}(1, T_L - 1) + Q_i^{(T)}(T_L, T - 1)}. \tag{3.15}$$

The advantage of these expressions is that the first terms in the numerator and denominator of both (3.14) and (3.15) can be calculated recursively without reference to past observations and the sums implicit in the second terms extend over only the past

L observations. To determine the time updated model $\zeta^{(T)}$, we first initialise it using

$$\begin{aligned}
\mu_i^{(T)} &= \mu_i^{(T-1)} \\
a_{ij}^{(T)} &= a_{ij}^{(T-1)} \\
\alpha_i^{(T)}(T-1) &= \alpha_i^{(T-1)}(T-1)
\end{aligned}$$

we then calculate $\alpha_i(T)$ from (3.5), $\beta_j(t)$ for $t \in [T_L+1, T]$ from (3.6). We can then calculate all the remaining quantities in (3.14) and (3.15) and update the model. Finally, in preparation for the next time step, we update the first terms in the numerator and denominator of (3.14) and (3.15) using

$$\begin{aligned}
U_i^{(T)}(1, T_L+1) &= \lambda U_i^{(T-1)}(1, T_L) + \frac{\alpha_i^{(T)}(T_L+1)\beta_i^{(T)}(T_L+1)O_{T_L+1}}{P^{(T)}} \\
Q_i^{(T)}(1, T_L+1) &= \lambda Q_i^{(T-1)}(1, T_L) + \frac{\alpha_i^{(T)}(T_L+1)\beta_i^{(T)}(T_L+1)}{P^{(T)}} \\
R_{ij}^{(T)}(1, T_L) &= \lambda R_{ij}^{(T-1)}(1, T_L-1) + \frac{\alpha_i^{(T)}(T_L-1)b_j^{(T)}(O_{T_L})\beta_i^{(T)}(T_L)}{P^{(T)}}
\end{aligned} \tag{3.16}$$

Although it is not needed for updating the model, we also want to accumulate the total transition count $c_{ij}\,(1, T_L+1)$, since it will be required in Sec. 3.4.6

$$c_{ij}\,(1, T_L+1) = \begin{cases} c_{ij}\,(1, T_L) + 1 & \text{for } i = \arg\max_i \alpha_i^{(T-1)}(T_L)\beta_i^{(T-1)}(T_L), \\ & \quad j = \arg\max_j \alpha_j^{(T)}(T_L+1)\beta_j^{(T)}(T_L+1) \\ c_{ij}\,(1, T_L) & \text{otherwise} \end{cases} \tag{3.17}$$

3.4.6 Adapting to rapidly changing noise characteristics

In order to accommodate an abrupt change to the noise characteristics as might, for example, arise from the introduction of an entirely new noise source, we need to create a new state to model the newly observed noise spectrum. Fig. 3.3 shows a example of the spectrogram of an antique clock, there is a constant "tic-toc" sound in the background, with a sudden arrival of "chime" sound from the chime. To avoid increasing the

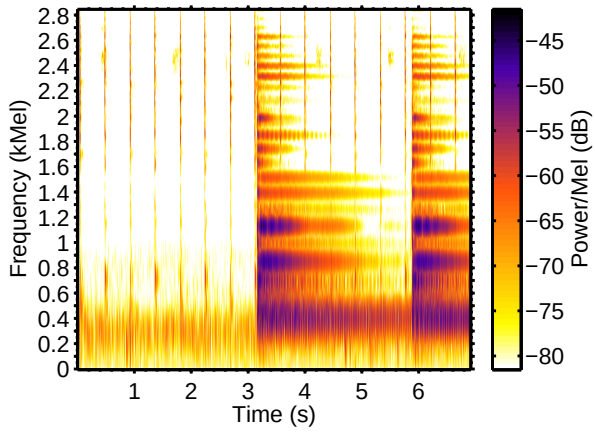


Figure 3.3: Spectrogram of an antique chiming clock.

complexity of the model with the repeated introduction of new states, we merge two of the existing states whenever we create a new state.

Goodness of fit test

In order to decide when to introduce a new state, we calculate a measure $Z^{(T)}$ that indicates how well the most recent L frames of observed data fit the current model, $\zeta^{(T)}$. From (3.2), it is straightforward to show that if $E\{O_t\} = \mu$, then the mean and variance of the observation log probability density are given by

$$\begin{aligned}
\mathbf{E}\{\log b(O_t \mid \mu)\} &= \mathbf{E}\left\{\sum_k (-\log \mu(k) - \frac{O_t(k)}{\mu(k)})\right\} = -\sum_k (\log \mu(k) + 1) \\
\mathrm{Var}\{\log b(O_t \mid \mu)\} &= \sum_k \mathbf{E}\left\{(\frac{O_t(k) - \mu(k)}{\mu(k)})^2\right\} = \sum_k \frac{\mu^2(k)}{\mu^2(k)} = K
\end{aligned}$$

Accordingly we define $Z^{(T)}$ as the normalized difference between the weighted log-likelihood of the most recent L frames and its expectation

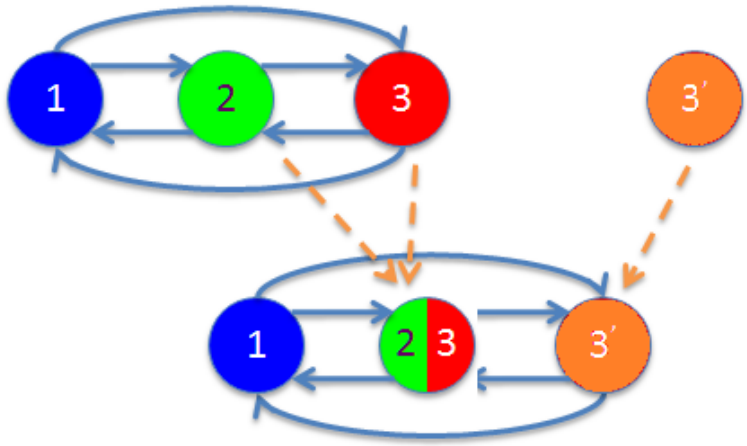


Figure 3.4: Illustration of the creation of a new noise state, where two states are merged in the new model thereby making room for the new state.

$$
\begin{aligned}
Z^{(T)} &= \frac{\sum_{t=T_L+1}^{T} \lambda^{T-t} \left(\log b\left(O_t\right) - E\left\{\log b\left(O_t\right)\right\}\right)}{\sqrt{\sum_{t=T_L+1}^{T} \left(\lambda^{T-t}\right)^2 Var\left\{\log b\left(O_t\right)\right\}}} \\
&= \frac{\sum_{t=T_L+1}^{T} \lambda^{T-t} \sum_k \left(1 - \frac{O_t(k)}{\mu_{i(t)}(k)}\right)}{\sqrt{K \sum_{t=T_L+1}^{T} \lambda^{2T-2t}}}
\end{aligned}
\tag{3.18}
$$

where $i(t)$ gives the state occupied at time t in the maximum likelihood state sequence. The normalization factor in the denominator is the standard deviation of the numerator under the assumption that the likelihoods of each frame are independent given the correct state sequence. With this assumption, $Z^{(T)}$ should be zero mean and unit variance. However, if the number of frequency bins, K, is large, the spectral components in adjacent frequency bins become more strongly correlated and we can no longer assume that they are independent. For this reason, the appropriate value of $Z^{(T)}$ must be determined empirically.

Creating a new state

If $\left|Z^{(T)}\right|$ exceeds an empirically determined threshold, θ_Z, then this indicates that $\zeta^{(T)}$ should be re-estimated and a new type of noise might be present. In this case, we there-

fore create a tentative model, $\hat{\zeta}^{(T)}$, in which two of the existing states have been merged and a new state created, such that the total number of states H is fixed. This is to avoid the over-fitting that would result from repeatedly introducing additional states. We require that the modelling improvement that results from introducing a new state must outweigh the degradation that results from merging two existing states. An example for 3-state HMM is illustrated in Fig. 3.4, which shows the original HMM, $\zeta^{(T)}$, at the top. To create $\hat{\zeta}^{(T)}$, states 2 and 3 are merged and a new state $3'$ is added.

For the tentative model $\hat{\zeta}^{(T)}$, we first determine the pair of states, $\{i_0,\, j_0\}$, whose merging will cause the least reduction in log-likelihood of the model, which is defined as, $Q_{i_0}(1,T_L)\, D(\mu_{i_0},\hat{\mu}_{i_0})+Q_{j_0}(1,T_L)\, D(\mu_{j_0},\hat{\mu}_{i_0})$, where $D(\mu_i,\hat{\mu}_j)=\sum_k\left(\frac{\mu_i(k)}{\hat{\mu}_j(k)}-\log\frac{\mu_i(k)}{\hat{\mu}_j(k)}-1\right)$ is the Itakura-Saito distance and equals the expected decrease in log likelihood of a frame whose original mean power spectrum μ is re-modelled by a new mean $\hat{\mu}$. We then initialize the state means for the model $\hat{\zeta}^{(T-1)}$ as

$$\hat{\mu}_r^{(T-1)}=\begin{cases} O_T & \text{for } r=j_0 \\ \frac{Q_{i_0}(1,T_L)\mu_{i_0}^{(T-1)}+Q_{j_0}(1,T_L)\mu_{j_0}^{(T-1)}}{Q_{i_0}(1,T_L)+Q_{j_0}(1,T_L)} & \text{for } r=i_0 \\ \mu_r^{(T-1)} & \text{otherwise} \end{cases} \tag{3.19}$$

The state j_0 models the new noise spectrum (which we assume is exemplified in frame T) and state i_0 is initialized as a weighted average of the previous states i_0 and j_0. The weights in (3.19) are taken to be the occupancy counts $Q_{i_0}(1,T_L)$ and $Q_{j_0}(1,T_L)$ from (3.10), where the most recent L frames which are excluded because they may contain examples of the new state. We also re-evaluate the accumulated transition counts of the new model from $c_{ij}\,(1,T_L)$ that have previously updated in (3.17),

$$\hat{c}_{ij}\,(1,T_L)=\begin{cases} 0 & \text{for } j=j_0 \\ c_{ij_0}\,(1,T_L)+c_{ii_0}\,(1,T_L) & \text{for } j=i_0 \\ c_{ij}\,(1,T_L) & \text{otherwise} \end{cases} \tag{3.20}$$

and re-estimate the transition probability $\hat{a}_{ij}$ using (3.4). We then re-train this initial model, $\hat{\zeta}^{(T-1)}$, using Viterbi decoding on the most recent L frames, $\{O_t : t \in [T_L + 1, T]\}$.

Update the new Baum-Welch

The final step in creating the new model is to perform a Baum-Welch update as detailed in section 3.4.5. In order to do this, we need the accumulated sums U, Q and R defined in section 3.4.4. However these sums were accumulated based on the old model which includes two states, i and j, that now have been merged. Accordingly we re-distribute the accumulated sums of each old state to the states of the new model. The ratio of re-distribution is based on ϕ_{ij}, which is the probability that a frame that was previously in state i of the old model belongs to state j of the new model: $\phi_{ij} = \frac{b\left(\mu_i^{(T-1)}|\hat{\mu}_j^{(T-1)}\right)}{\sum_j b\left(\mu_i^{(T-1)}|\hat{\mu}_j^{(T-1)}\right)}$.

Now, we re-calculate the accumulated sums by distributing them to each of the new states according to the new mean $\hat{\mu}^{(T-1)}$:

$$\begin{aligned}
\hat{U}_i^{(T-1)}(1, T_L) &= \sum_m \phi_{mi} U_m^{(T-1)}(1, T_L) \\
\hat{Q}_j^{(T-1)}(1, T_L) &= \sum_m \phi_{mj} Q_m^{(T-1)}(1, T_L) \\
\hat{R}_{ij}^{(T-1)}(1, T_L - 1) &= \sum_m \sum_n \phi_{mi} \phi_{nj} R_{mn}^{(T-1)}(1, T_L - 1)
\end{aligned} \tag{3.21}$$

By using the Expectation–maximization (EM) re-estimation algorithm from (3.14) & (3.16), $\hat{\zeta}^{(T)}$ is obtained.

Log-likelihood test

We only wish to use this revised model if it will result in an increase in log likelihood. Accordingly the increase, $I^{(T)}$, in the log-likelihood of the L most recent frames is estimated as

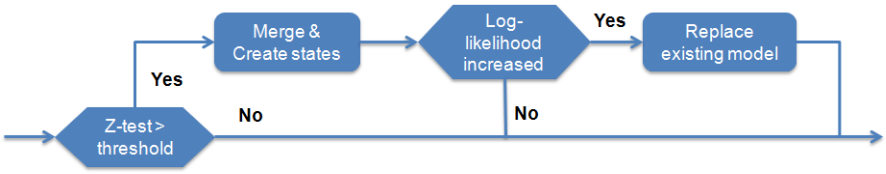


Figure 3.5: Flow diagram illustrating the criteria used to decide whether to create a new state.

$$
\begin{aligned}
I^{(T)} &= \left(\sum_{t=T_L+1}^{T} \lambda^{T-t} \sum_i \hat{Q}_i\left(t,t\right) \log b(O_t, \hat{\mu}_i) - Q_i\left(t,t\right) \log b(O_t, \mu_i) \right) \\
&\quad - \frac{\lambda^L}{1-\lambda} \sum_i \sum_j \phi_{ij} \pi_i D(\mu_i, \hat{\mu}_j)
\end{aligned} \tag{3.22}
$$

where $D(\mu_i, \hat{\mu}_j) = \sum_k \left(\frac{\mu_i(k)}{\hat{\mu}_j(k)} - \log \frac{\mu_i(k)}{\hat{\mu}_j(k)} - 1 \right)$ is the Itakura-Saito distance and equals the expected decrease in log likelihood of a frame whose true mean power spectrum is μ_i when it is modelled by a state with mean $\hat{\mu}_j$. The first term in (3.22) gives the log likelihood improvement of the most recent L frames while the second term approximates the decrease in log likelihood of the earlier frames. If $I^{(T)} > 0$, the model is updated by replacing ζ with $\hat{\zeta}$, replacing the accumulated sums with those calculated in (3.19) and (3.21).

3.4.7 Noise estimation algorithm overview

The criteria used to decide whether to create a new state are illustrated in Fig. 3.5. At each frame the Z-test is evaluated to determine how the current model fit to the past L frames . If the test indicated a poor fit, a tentative model is created in which two states are merged and a new state created. Finally if the new model gives a better fit to the observation, it replaces the existing model.

The processing steps of the proposed algorithm can be summarized as follows:

1. Compute the initialized model $\zeta^{(T_0)}$ using Viterbi training on observations $O^{(T_0)} = \{O_t : t \in [1, T_0]\}$ and set $T = T_0$.

2. Compute and update the model $\zeta^{(T)}$ from $\zeta^{(T-1)}$ using (3.14) - (3.15).

3. Compute the $Z^{(T)}$ using (3.18).

4. If $Z^{(T)} > \theta_Z$,

   (a) Create a tentative model $\hat{\zeta}^{(T-1)}$ using parameters described in (3.19) - (3.21).

   (b) Compute $I^{(T)}$ using (3.22).

   (c) If $I^{(T)} > 0$, update the model $\zeta^{(T)} = \hat{\zeta}^{(T)}$.

5. Increment $T = T + 1$, and go back to step 2 for the next time frame.

3.5 Noise Estimation during Speech Activity

In this chapter, we are assuming that an external voice activity detector (VAD) is available and we only update the noise model when speech is absent. During speech presence we freeze the noise model ζ, and use it to estimate the noise state for each frame. In the speech enhancement experiments described below, we assume that the clean speech power spectrum may be approximated as $\gamma_t \bar{\mu}$ where $\bar{\mu}$ is the Long-Term Average Speech Spectrum (LTASS) [69] and γ_t is the speech level at time t. For each noise state, j, we evaluate the likelihood $b\left(O_t \mid \mu_j + \gamma_t \bar{\mu}\right)$ and select the maximum likelihood estimate of the speech level as $\gamma_t\left(j\right) = \arg\max_{\gamma_t} b\left(O_t \mid \mu_j + \gamma_t \bar{\mu}\right)$, thus the observation probabilities are given by $b\left(O_t \mid \mu_j + \gamma_t\left(j\right) \bar{\mu}\right)$. Once we have evaluated the observation probabilities we can use the Viterbi algorithm to determine the most likely noise state sequence. Given the noise state sequence, we use the corresponding state means, μ_j, as the *a priori* noise estimates within speech enhancement algorithms.

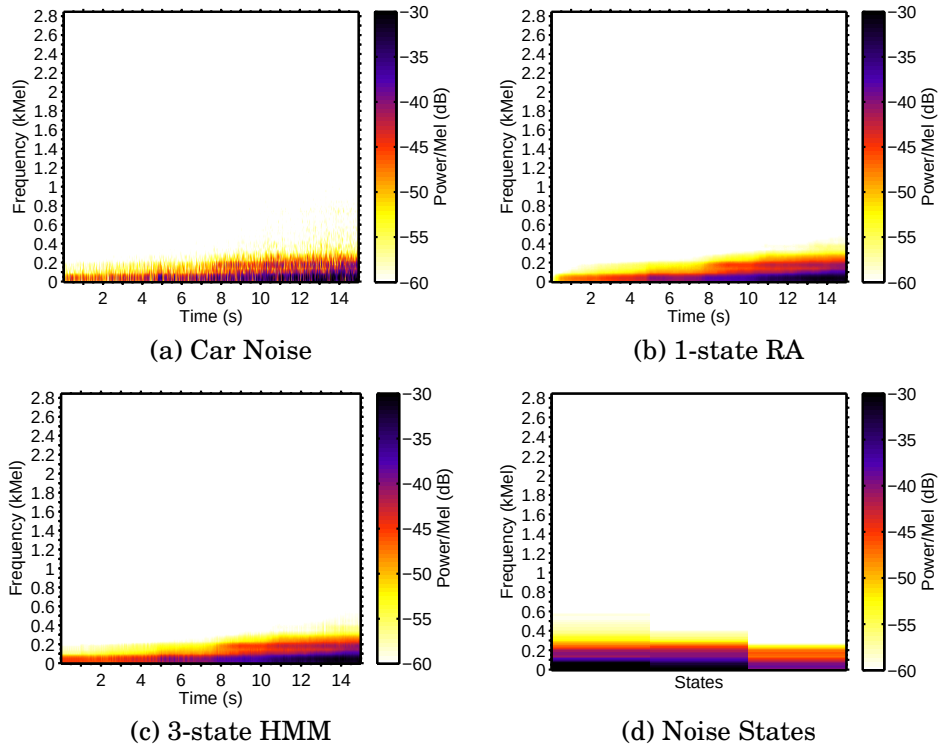
3.6 Experimental Results

As discussed in 3.2, a good noise estimator should be able not only to track slowly evolving noise spectra, but also to detect and update any abrupt change in noise characteristics. In this section, we first demonstrate the noise tracking abilities of our proposed multi-state HMM noise estimation algorithm. Then in the context of speech enhancement, we compare the performance of our noise algorithm with other noise estimation algorithms.

For all the experiments, the signals are sampled at a frequency of 16 kHz and decomposed into overlapping frames. The DFT is then used to determine the power spectrum of each frame. Using the frame settings recommended in [90], the time-frames have a length of 32 ms with a 50% overlap resulting in $K = 257$ frequency bins. The window length L should be long enough for the HMM re-estimation, but short enough to follow follow non-stationary noise variations. A suitable search window is typically 0.5 to 1.5 seconds [17]. In our experiment setting, we retain the most recent $L = 30$ frames (480 ms), and also set the initial training time $T_0 = 30$ frames. The forgetting factor is chosen to be $\lambda = 1 - 1/(2L)$, which gives a time constant of $2L = 960$ ms. The other noise estimation methods used for comparison are the minimum statistics estimator [90, 92], unbiased MMSE-based noise estimator [58, 46] and 1-state recursive averaging. The 1-state recursive averaging model (1-state RA) is defined as $\mu^{(T)} = (1-\lambda)\,\mu^{(T-1)} + \lambda O_T$, where the same value of λ is used as above. This 1-state RA is representative of noise estimation methods based on temporal averaging when speech is absent, for instance, the Minima Controlled Recursive Averaging [16]. The threshold θ_Z defined in Sec. 3.4.6 is determined to be 30 empirically. The noise signals will be used below are from a library of special sound effects and NOISEX database [124].

3.6.1 Noise Tracking

In this section, we evaluate the performance of the 1-state RA and 3-state HMM noise estimation models on three types of noise (a) slowly evolving (b) Non-stationary and



(a) Car Noise (b) 1-state RA

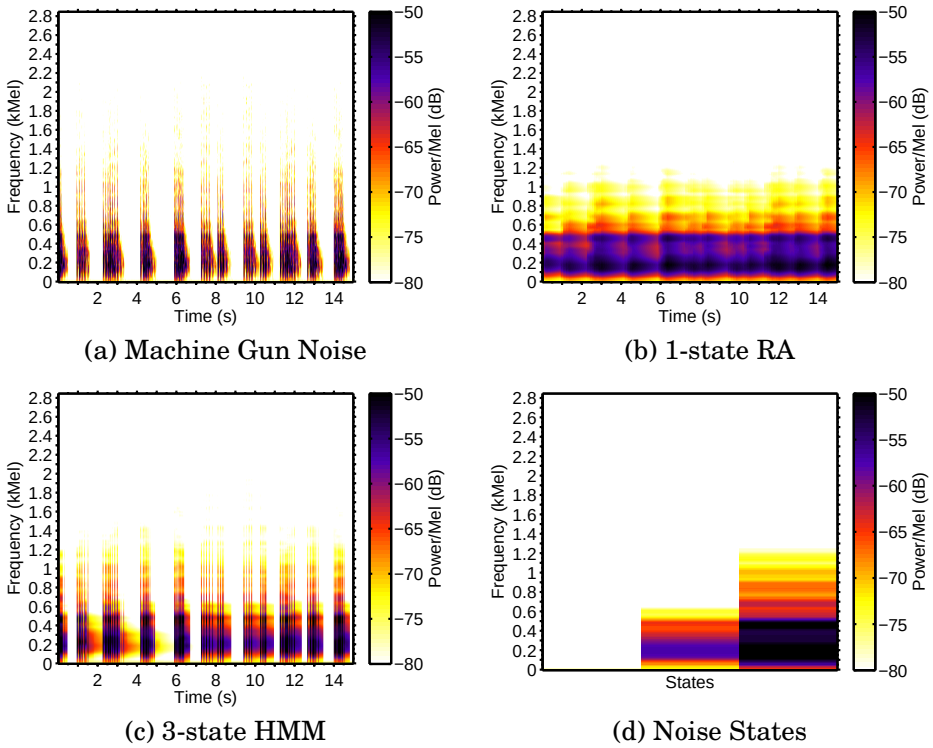
(c) 3-state HMM (d) Noise States

Figure 3.6: Spectrogram of (a) increasing car noise, with its estimation using (b) 1-state recursive averaging (c) a 3-state HMM; (d) Spectrum of estimated noise states at $t = 15$ s.

(c) abruptly changing. We evaluate the performance of the algorithms using COSH distance between the true noise spectrum and its estimates.

Slowly evolving noise

A good noise estimator should be able to track and update gradual changes in the noise characteristics. Fig. 3.6(a) shows the spectrogram of car noise with increasing amplitude at the rate of 2 dB/sec. The spectrogram of the estimated noise using the 1-state recursive averaging method and the 3-state HMM method are shown in Fig. 3.6(b) and Fig. 3.6(c) respectively, where both of them show a good representation of noise. It can be seen that the 3-state HMM performs slightly better as it is a richer model, and, as will be seen in Table 3.1 below, the 3-state HMM results in a lower COSH error. Fig. 3.6(d) shows the spectrogram of the updated noise states of the HMM at the end



(a) Machine Gun Noise (b) 1-state RA

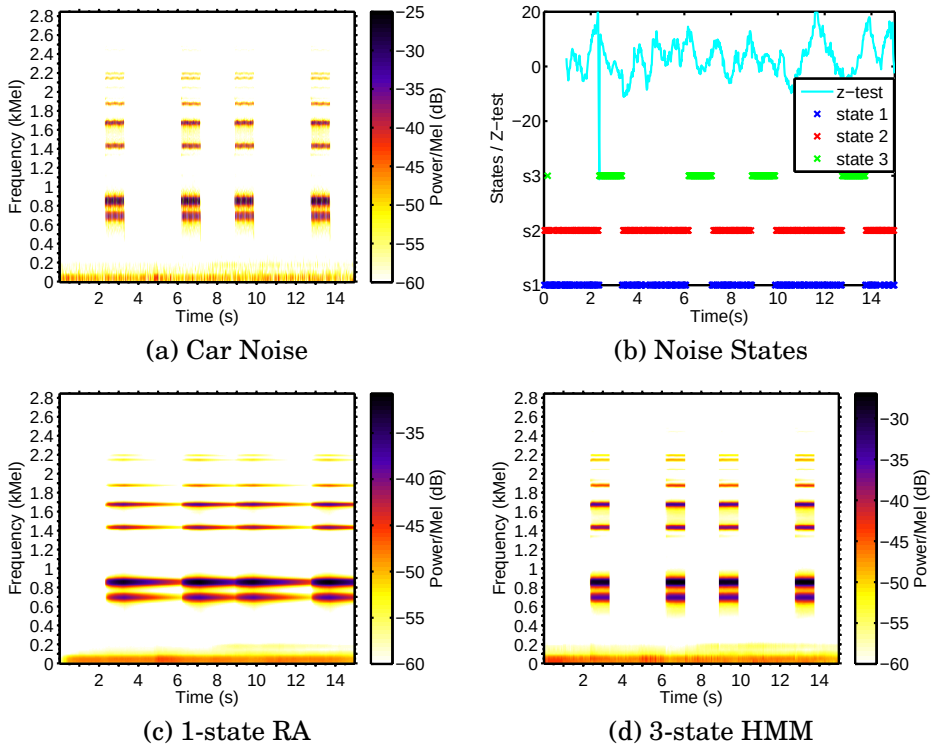
(c) 3-state HMM (d) Noise States

Figure 3.7: Spectrogram of (a) machine gun noise, with its estimation using (b) 1-state recursive averaging (c) a 3-state HMM; (d) Spectrum of estimated noise states at $t = 15$ s.

of the signal. We can see that between the three states we have a good description of the recent evolution of the signal and that the second state corresponds with the most recent frames.

Non-stationary noise

Fig. 3.7 (a) shows the spectrogram of a machine gun noise. The noise consists of impulsive sounds separated by silent intervals. The spectrogram of the estimated noise using 1-state recursive averaging method and 3-state HMM method are shown in Fig. 3.7(b) and Fig. 3.7(c) respectively. The 1-state RA model fails to follow the rapid change of noise characteristics and converges to an average spectrum. In contrast, the HMM has assigned separate states to model the silence and gun fire, as can be seen from Fig. 3.7(d). By comparing Fig. 3.7(c) with Fig. 3.7(a), we see that even with only three states,



(a) Car Noise (b) Noise States

(c) 1-state RA (d) 3-state HMM

Figure 3.8: Spectrogram of (a) car+phone noise, with its estimation using (c) 1-state recursive averaging (d) a 3-state HMM; (b) Mean power of the three noise states together with the value of the Z-test defined in (3.18).

the HMM is able to model the noise signal well.

Abrupt noise detection

In this experiment, the noise of a ringing phone is added to a background car engine noise which is predominantly low frequency. Fig. 3.8(a) shows the spectrogram of this composite noise and it can be seen that the noise spectrum changes abruptly whenever the phone rings. The spectrogram of the estimated noise using 1-state recursive averaging method is shown in Fig. 3.8(c). As would be expected this model is unable to track the rapidly changing noise and smears the spectrum in the time direction. A 3-state HMM is used to estimate this noise, and the state assignment is shown in Fig. 3.8(b), and the Z-test value $Z^{(T)}$ is plotted above, which measures how well the L most recent observations fit the model. We see that when the first phone ring occurs, at approxi-

	Car	Gun	Phone
1-state RA	17.4	36769.0	6458.5
2-state HMM	13.1	443.0	25.0
3-state HMM	**13.0**	366.1	11.6
4-state HMM	13.1	**287.2**	**10.8**

Table 3.1: COSH distance of different noise estimations using 1-state RA and 3-state HMM.

mately 2.3 s, there is an abrupt fall in $Z^{(T)}$ which indicates the arrival of a novel noise spectrum. Since state 3 has very low occupancy count before the merge, two of the existing states, state 2 and 3, are therefore merged and state 3 is reallocated to model the new noise spectrum. The corresponding spectrogram for our proposed model is shown in Fig. 3.8(d) in which the estimated noise spectrum follows the state mean of the maximum likelihood state sequence. We see that the abrupt changes in noise spectrum are perfectly tracked and well modelled.

COSH errors

The average COSH distances between the true noise signal and its estimates using 1-state RA model and multi-state HMMs are shown in Table 3.1. The results confirm our observations for Fig. 3.6 to 3.8. For slowly varying car noise, both noise estimators work well and have a low COSH distance for the true noise spectrum. The 3-state HMM is a richer model than the 1-state RA estimator and so is able to achieve slightly lower error. The 1-state RA model is unable to track abrupt changes in noise characteristics, and shows large COSH errors when estimating non-stationary noise such as “Gun” and “Phone” noise. On the other hand, the 3-state HMM always shows a better noise estimation than 1-state RA method. The COSH error for the “Gun” noise is larger than for the other signals as the echo from the firing of the machine gun varies depends on the interval between each gunfire. For stationary white noise, which can be modelled precisely by a single state, the COSH errors for different number of the states stay roughly the same. For other two types of noise, the COSH errors decrease as number of state increases, but the improvements are small, as compared to the RA method.

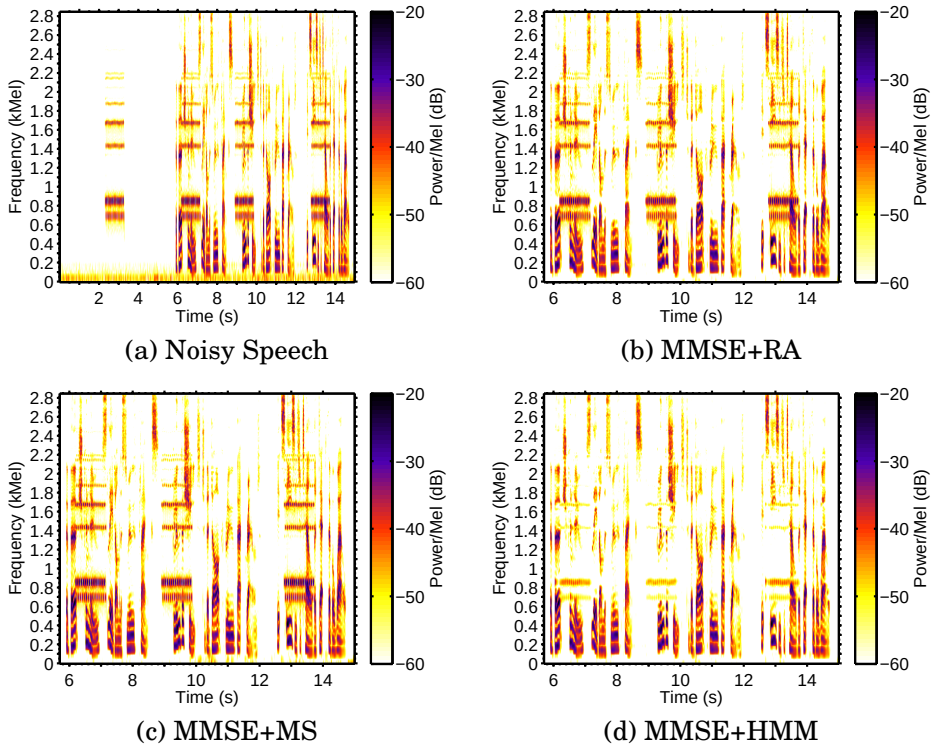


Figure 3.9: Spectrogram of (a) the unenhanced noisy speech corrupted by the car+phone noise, and the MMSE enhanced speech using different noise estimator (b) RA (c) MS (d) HMM.

3.6.2 Speech Enhancement

In this section, we incorporate our HMM noise estimator into a speech enhancer to evaluate whether our noise estimator improves the quality of speech as compared to other noise estimation methods. We will first demonstrate an example of how well the noise can be suppressed using our method, then we will run a set of experiments to show the improvements in terms of PESQ and segmental SNR of the enhanced speech. All the clean speech signals were taken from the IEEE sentence database [106] by concatenating three sentences to give an average duration about 10 seconds.

MMSE speech enhancer

Fig. 3.9(a) shows an example of a speech signal corrupted by a ringing phone noise at 0 dB SNR, shown in Fig. 3.8 (a). We assume that there will be non-speech segment at

	Unenhanced	RA	MS	HMM
PESQ	2.18	1.91	2.15	2.44
ΔPESQ	0	-0.27	-0.03	0.26

Table 3.2: PESQ scores and improvements of the enhanced speech with car+phone noise.

the beginning of the signal, roughly 5 seconds in this case, and it is used to train our noise estimation model, and the rest of the signal forms the speech active segment. The noise characteristics are assumed to remain stationary while the speech is active.

The speech active segments of the given noisy speech signal is then enhanced by the MMSE algorithm [30] using different noise estimators. Fig. 3.9 shows the enhanced speech signals using respectively (b) 1-state recursive averaging (RA), (c) minimum statistics (MS) [11] and (d) multi-state hidden Markov model (HMM) respectively. The noise-only segment is not included in the spectrograms for the enhanced speech. The RA and HMM estimators are trained on the initial noise-only segment and frozen at approximately $t = 5$ seconds, while the MS estimator is allowed to adapt continuously throughout the signal. We see that the stationary low frequency noise component is effectively removed using all three methods but only with the HMM method in Fig 3.8(d), is the phone ringing largely eliminated. As seen previously in Fig. 3.8(c), the noise estimate from the RA method is blurred in time and so, with this estimate, distortion is introduced in the gaps between rings. Even though the MS method tracks the variation of noise level during speech presence, it cannot respond quickly enough to eliminate the phone noise. Although the training segment includes only one phone ring, this is sufficient for the HMM method to learn its characteristics and to attenuate it greatly when it subsequently occurs. We assess the quality of the speech by means of the PESQ (Perceptual Evaluation of Speech Quality) score [70]. The PESQ score for the unenhanced noisy speech is 2.18, and the PESQ scores and the improvements for the enhanced speech signals when using the RA, MS and HMM methods to estimate the noise are shown in Table 3.2. We see that as measured by PESQ, the RA and MS methods actually degraded the speech whereas the HMM method improve it. This indicates that our proposed HMM method gives a noticeably greater quality improvement than the other methods.

white/gSNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	-12.38	-7.38	-2.39	2.61	7.60	12.60
RA	**-0.74**	**2.54**	**5.68**	**8.92**	**12.41**	**16.17**
MS	-1.42	1.91	5.12	8.35	11.63	14.91
UM	-1.65	1.92	5.22	8.39	11.47	14.41
HMM	**-0.74**	**2.54**	**5.68**	**8.92**	**12.41**	**16.17**

Table 3.3: Segmental SNR of enhanced speech corrupted by white noise using different noise estimation methods.

gun/gSNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	2.19	7.19	12.19	17.18	22.18	27.18
RA	0.90	3.23	6.03	9.68	14.10	18.86
MS	**2.38**	**6.10**	**9.50**	**12.79**	**15.99**	**19.19**
UM	2.19	5.87	9.38	12.61	15.36	17.67
HMM	2.22	3.82	5.53	7.98	11.38	15.66

Table 3.4: Segmental SNR of enhanced speech corrupted by machine gun noise using different noise estimation methods.

Evaluation using Segmental SNR

A set of experiments was performed with noise+speech at different SNRs with a noise-only segment at the beginning of the noisy speech signal as before. 20 different clean speech signals were used, with an average duration of about 10s. Four different kinds of noise estimation algorithms were evaluated: (i) 1-state recursive averaging (RA), (ii) minimum statistics (MS), (iii) unbiased MMSE-based noise estimator (UM) [46] and (iv) our proposed multi-state hidden Markov model (HMM). The number of states used for the HMM is set to 3 for all the noisy speech signals below.

Tables 3.3 to 3.5 shows the segmental SNR (sSNR) at different global SNR (gSNR) of enhanced speech which is corrupted by (i) white noise, (ii) gun noise and (iii) "car+phone" noise respectively, and the sSNR improvement at different SNR for different noise are shown graphically in Fig. 3.10. For the white noise shown in Table 3.3 and Fig. 3.10(a), the HMM method shows almost identical sSNR scores to the RA method as white noise is stationary and the noise characteristics does not change over time. The UM and MS methods shows a slightly lower sSNR at low gSNR, as they both underestimate the noise power when the noise power and speech power are comparable. For the "car+phone" noise in Table 3.5 and Fig. 3.10(c), the HMM method improves the sSNR score at all

car+phone/gSNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	-6.04	-1.05	3.95	8.95	13.94	18.94
RA	-0.34	3.35	6.89	10.28	13.69	17.41
MS	-0.34	3.75	7.51	10.80	13.74	16.35
UM	-1.09	3.32	7.31	10.82	13.77	16.20
HMM	**4.24**	**7.05**	**9.19**	**12.27**	**16.48**	**19.12**

Table 3.5: Segmental SNR improvement of enhanced speech by "car+phone" noise using different noise estimation methods.

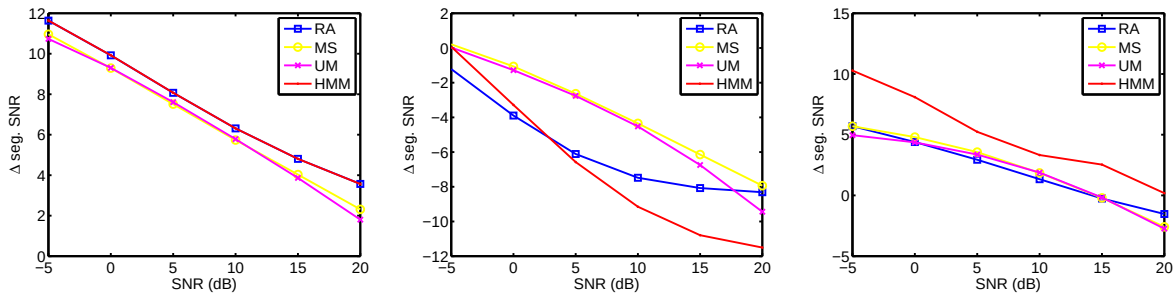
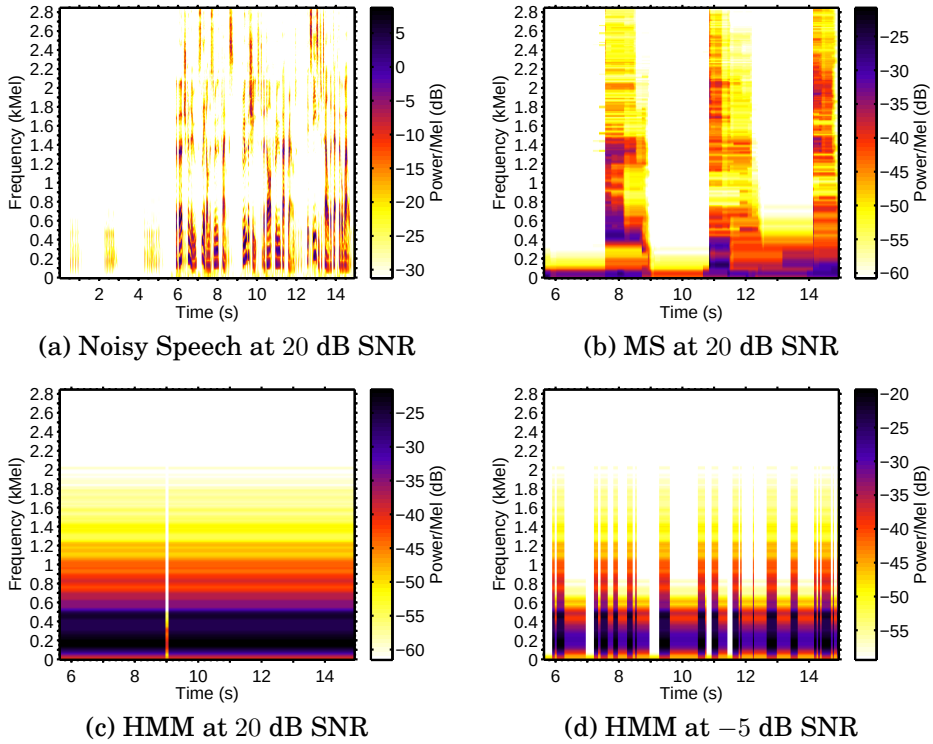


Figure 3.10: Improvement of Segmental SNR scores at different SNRs for (a) white noise (b) machine gun noise (c) "car+phone" noise.

SNRs and consistently outperforms the other methods by a large margin. We see that the UM and MS methods degrade the sSNR score at nearly all SNRs indicating their inability to track highly non-stationary noise. The noise estimate from the RA method is blurred in time and so, with this estimate, more speech distortion is introduced in the gaps between machine gun firing or phone rings. Thus it performs poorly at low SNR. For the machine gun noise in Table 3.4 and Fig. 3.10(b), all noise estimation methods failed to track this non-stationary noise, resulting in a decrease of sSNR. The MS method shows the least sSNR degradation, while UM method shows similar result. The RA methods perform poorly at low gSNR as expected, but the HMM method shows the worst performance at high gSNRs. Fig. 3.11(a) shows an example of a speech signal corrupted by machine gun noise at 20 dB SNR. Because machine gun noise power is much smaller than that of the speech, it cannot be easily differentiated from speech. Fig. 3.11(b) shows the estimated noise spectrum using the MS method. Comparing this with the actual noise spectrogram in Fig. 3.7(a), we see that the individual bursts of gun fires are smeared together and in consequence the sSNR is reduced. Although the HMM method correctly identifies the noise states in the training period (see Fig.



(a) Noisy Speech at 20 dB SNR

(b) MS at 20 dB SNR

(c) HMM at 20 dB SNR

(d) HMM at −5 dB SNR

Figure 3.11: Spectrogram of (a) the unenhanced noisy speech corrupted by the machine gun noise at 20 dB SNR, and the estimated noise spectrogram using (b) MS (c) HMM. The estimated noise spectrum using HMM and −5 dB SNR is shown in plot (d).

3.7(d)), it wrongly assigns almost all the frames to the "burst" state as can be seen from the estimated noise spectrogram in Fig. 3.11(c). In contrast, at a gSNR of −5 dB, the noise state assignment is much better as can be seen from Fig. 3.11(d), and as a result, the sSNR shows a small improvement.

Evaluation using PESQ

In order to evaluate the PESQ score of the enhanced speech, a similar set of experiments was performed as in the previous section. Tables 3.6 to 3.8 shows the PESQ score at different SNR of enhanced speech which is corrupted by (i) white noise, (ii) gun noise and (iii) "car+phone" noise respectively, and the PESQ improvement at different SNR for different noise are shown graphically in Fig. 3.12. For stationary noise, such as white noise, the HMM method shows almost identical PESQ scores to the RA method, while

white/gSNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	1.13	1.36	1.68	2.05	2.39	2.74
RA	**1.61**	**2.00**	**2.35**	2.63	**2.88**	**3.12**
MS	1.56	1.94	2.29	2.59	2.85	3.09
UM	1.53	1.93	2.30	2.62	2.88	3.10
HMM	**1.61**	**2.00**	**2.35**	**2.64**	**2.88**	**3.12**

Table 3.6: PESQ of enhanced speech corrupted by white noise using different noise estimation methods.

gun/gSNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	1.97	2.35	2.71	3.01	3.27	3.48
RA	1.99	2.45	2.78	**3.05**	**3.28**	3.48
MS	1.89	2.27	2.61	2.89	3.12	3.31
UM	1.89	2.29	2.63	2.91	3.15	3.33
HMM	**2.23**	**2.55**	**2.83**	3.04	3.26	**3.51**

Table 3.7: PESQ of enhanced speech corrupted by machine gun noise using different noise estimation methods.

the UM and MS method shows a slightly poorer PESQ score especially at low gSNRs. For the "car+phone" noise, the HMM method improves the PESQ score at all SNRs and consistently outperforms the other methods. We see that the other methods degrade the PESQ score at nearly all SNRs indicating their inability to track highly non-stationary noise. All these observations confirm our results from the previous section using relative segmental SNR. However, for the machine gun noise, the situation is different. The MS and UM methods degrade the PESQ score at all SNRs since they do not estimate this intermittent noise at all well as we can see in Fig. 3.11(b). The HMM method has a good PESQ improvement at low global SNR, but at high gSNR the PESQ score is essentially unchanged from that of the unenhanced speech.. This confirms our previous results regarding the estimation of noise states illustrated in Fig. 3.11(b) & (c), namely that at low gSNR, the model is better able to distinguish between speech and noise and therefore better able to assign the correct noise state to each frame.

Summary of quality assessments

The improvement of the segmental SNR and PESQ scores averaged across all global SNRs for different noise types are shown in Tables 3.9 and 3.10 respectively. Fig. 3.13 shows the hammering noise at a construct site. We have included this “hammer” noise

car+phone/gSNR	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	1.91	2.18	2.50	2.75	2.96	3.16
RA	1.65	1.91	2.25	2.60	2.86	3.07
MS	1.95	2.15	2.37	2.60	2.79	2.98
UM	1.93	2.12	2.40	2.61	2.81	3.00
HMM	**2.28**	**2.44**	**2.64**	**2.93**	**3.15**	**3.21**

Table 3.8: PESQ improvement of enhanced speech by "car+phone" noise using different noise estimation methods.

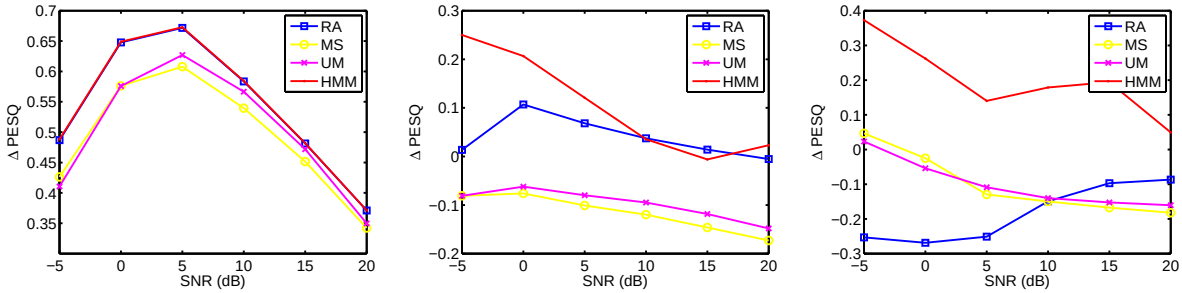


Figure 3.12: Improvement of PESQ scores at different SNRs for (a) white noise (b) gun noise (c) "car+phone" noise.

as one of the examples of non-stationary noise. When the noise is stationary, such as white and car noise, the improvement of the PESQ scores or segmental SNRs are about the same for all four methods. For the non-stationary noise, our proposed HMM method shows a much better PESQ improvement, indicating that our HMM method has a better noise estimation. However, in the case of the improvement of the segmental SNR, the HMM method perform well except for the machine gun noise. Although we have a good machine gun noise estimation in Fig. 3.7(d), we are not able to identify the correct noise state sequence especially when the noise power is small compare to that of speech. This indicates that we might need a better speech model.

3.6.3 Listening Test

Given the results from previous two experiments, we conducted a further listening test to verify the performance of our proposed algorithm in comparison to other algorithms. The listeners are instructed to to state their preference between two enhanced speech signals with input global SNR of 0 dB where different enhancement algorithms have

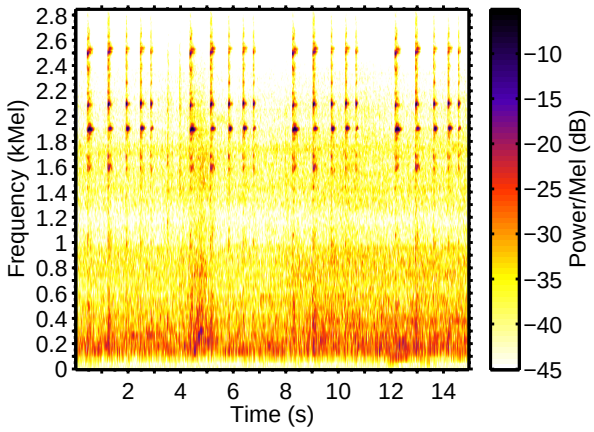


Figure 3.13: Spectrogram of hammering at a construction site.

Δ sSNR	white	car	gun	phone	hammer
RA	**7.39**	8.28	-5.88	2.10	3.89
MS	6.64	5.78	**-3.70**	2.19	3.58
UM	6.52	5.61	-4.17	1.94	3.45
HMM	**7.39**	**8.30**	-6.92	**4.94**	**4.40**

Table 3.9: mean segmental SNR improvement of enhanced speech signals using different noise estimation methods.

been used for the two signals: one is the proposed HMM algorithm, the other is one of the following: (i) unenhanced, (ii) RA or (iii) MS. The listeners do not know which algorithms are presented and the presentation order is random. Based on their preference, the following rating is assigned: "1" if they prefer the HMM algorithm, "0" if they prefers the other algorithm, or "0.5" if they are indifferent. A group of 10 listeners participated in this listening test, the mean rating scores of each comparison with HMM for different noise types are shown in Tables 3.11. As compared to the original (unenhanced) noisy speech, the proposed HMM is always preferred for all noise types except for the gun noise. This is possibly because although some of the gun noise is eliminated, some speech distortions are introduced due to incorrect state assignment. When compared to other enhancement methods, the HMM algorithm performs as well as RA, MS and UM methods for stationary white and car noise, but is preferred for non-stationary phone and hammer noise. The RA method for gun noise is not desirable because it introduced significant speech distortion. These results correlate well with our previous results for the improvements in segmental SNRs and PESQ scores.

Δ PESQ	white	car	gun	phone	hammer
RA	**0.54**	**0.16**	0.07	-0.18	0.29
MS	0.49	0.04	-0.05	-0.10	0.25
UM	0.50	0.05	-0.05	-0.10	0.26
HMM	**0.54**	**0.16**	**0.11**	**0.20**	**0.43**

Table 3.10: mean PESQ Improvement of enhanced speech signals using different noise estimation methods.

v.s. HMM	white	car	gun	phone	hammer
unenhanced	1.0	1.0	0.7	1.0	1.0
RA	0.5	0.5	1.0	1.0	1.0
MS	0.5	0.5	0.75	1.0	1.0
UM	0.5	0.5	0.7	1.0	1.0

Table 3.11: mean rating scores of enhanced speech signals using different noise estimation methods. A high score indicates that the HMM method was preferred.

3.7 Summary

In this chapter we have proposed an adaptive model for non-stationary noise signals based on a multi-state HMM in which each state describes a distinct noise power spectrum following a negative exponential distribution determined from its mean noise characteristics. We have described an update procedure that enables the model to track gradual changes in the amplitude or power spectrum of a noise source by adapting the mean power spectrum associated with each state. In addition, we have presented a method of detecting the presence of a noise source that does not match the existing model. When such a noise source is detected, our algorithm creates a new state and initializes the new state to represent the new source. At the same time, to avoid an ever-increasing number of model states, the two nearest states are merged and the state means and transition probabilities adjusted accordingly.

The noise modelling algorithm has been evaluated on noise examples that are stationary, gradually changing and highly non stationary. In all cases, the algorithm is able to create an accurate model of the noise and to track its changes over time. Its performance was compared with that of a recursive averaging approach typical of state-of-the-art estimators that use a VAD. It was found that the new algorithm almost always gave a better estimate of the noise, especially in the case of highly non-stationary noise.

The algorithm has also been evaluated by incorporating it into a speech enhancement system. For the purposes of this evaluation, the noise model was not adapted during speech presence and was combined with a very simple 1-state speech model in order to identify the correct noise state sequence during the presence of speech. It was found that, where the noise state sequence was correctly identified the new algorithm resulted in improvements in both segmental SNR and in quality as measured by PESQ. For one of the tested noise signals however, even though the noise model was accurately acquired, the noise state sequence was incorrectly identified when speech was present especially at high SNRs. In this case the speech enhancer performed poorly which resulted in a degraded segmental SNR. This indicates the need for an improved speech model in order to improve the discrimination between speech and noise.

In the next chapter, we extend our noise modelling algorithm so that it is able to track changing noise spectra and create new noise states even in the presence of speech.

Chapter 4

Noise Modelling in Speech Presence

4.1 Introduction

In Chapter 3, we developed an on-line HMM noise estimator that can work for a noise-only fragment, and we assumed the noise characteristics remained unchanged during the speech activity, i.e. we froze the model update once the speech is active. In order to detect and update the noise even during speech activity, there are mainly two problems we aim to solve: update slowly changing noise characteristics within each state during speech activity, and detect the advent of a new noise type which is different from either speech or an existing noise state. The first can be achieved by exploiting the fact that even during speech activity, the spectral power in some frequency bins will be dominated by the noise. Whenever the speech presence probability is low in some of the frequency bins, we can update the corresponding noise model states in those particular bins. In order to avoid the possible inclusion of any speech as a novel noise type, we introduced a multi-state speech model to be incorporated into the noise HMM described in Chapter 3, such that a new state is only created when the characteristics of the new noise is

significantly different from any combination of the states of both the speech and noise models.

Our aim in this chapter is to develop a robust HMM noise estimator that can track and update our model of highly non-stationary noise even during speech presence. The structure of the rest of this chapter is as follows. We first give a brief literature review of joint speech and noise modelling. Next we incorporate the speech model into the noise HMM to calculate the log likelihood of the observation probability using the joint speech+noise model. We propose a modified minima controlled recursive averaging method to update the mean power spectrum of each noise state especially during speech presence. We also propose an initial retraining scheme for use when a new noise type is detected. Finally, the performance of the HMM is evaluated both in estimating the noise spectrum and when used with a speech enhancement algorithm.

4.2 Noise Estimation using a Speech Model

Joint estimation of speech and noise from a combined speech and noise model has been widely used in speech recognition in which the probability of a speech state is determined by marginalising over all possible noise states [123]. It is later extended by Gales [40] and in subsequent papers [41, 42, 43]. These authors used HMMs to model both speech and noise in the mel-cepstral domain giving a combined model whose state count was the product of the speech and noise model counts. In practice, the noise model normally had very few states and often only one.

In [118] and [36], an EM approach is used to estimate the speech, noise and channel adaptively in the log spectrum domain. Each of these three components is represented with a Gaussian mixture model. In most of the examples they give, the noise model comprised only a single mixture but, for the case of aircraft noise at an airport, they investigated the use of up to 16 mixtures (the speech model, in contrast, used 256 mixtures). A first-order Taylor-series approximation is used to linearise the mapping between the log power domain and the linear power domain. In [37], the authors found

that their adaptive noise modelling reduced speech recognition word errors by about 15% compared to a non-adaptive model estimated from the beginning of the recording and that increasing the noise model from 1 to 4 mixtures gave a further improvement of up to 0.3%. A similar model (in the Mel log spectral domain) is used in [23] whose authors develop a recursive estimate of the parameters of the single-mixture noise model which was extended to a Bayesian formulation in [22].

A difficulty with the joint estimation approach when used for enhancement is that it is necessary to estimate the absolute speech energy; speech models developed for recognition generally ignore the overall speech level since it does not affect the speech state sequence. Subramanya [117] model speech using a 4-component GMM in the magnitude-normalised spectral domain rather than the more usual cepstral domain as this is the correct domain for adding noise and speech and avoids the difficulties that arise from the non-linear logarithmic transformation into the cepstral domain. They claim that applying magnitude normalisation significantly reduces the complexity required in the model although it entails modelling the overall speech energy separately.

Kristjansson [77] uses a noise GMM and found that selecting the maximum likelihood noise state performed similarly to marginalising over all noise states. Yao [129, 130] proposes a particle filter is used to represent the possible sequences of speech states, and the noise state may be estimated by marginalising over the speech states. In this application the speech model can be quite simple and only 18 states with 8 Gaussian mixtures per state in the log spectral domain is used. In a development of this work, Lee and Yao [80] estimate the noise characteristics in the log spectral domain using expectation–maximization (EM) but without a particle filter.

Zhao [133, 134] uses AR models for both speech (10th order) and noise (6th order) and has a fixed speech model with eight 16-mixture states (trained on TIMIT). The noise model uses five 1-mixture states together with an extra safety state derived from minimum statistics. At each frame the system updates the noise model using an EM procedure with a forgetting factor, to update noise states and noise gains. The system estimates a MMSE noise power spectrum by combining a Weiner filtered noisy speech spectrum

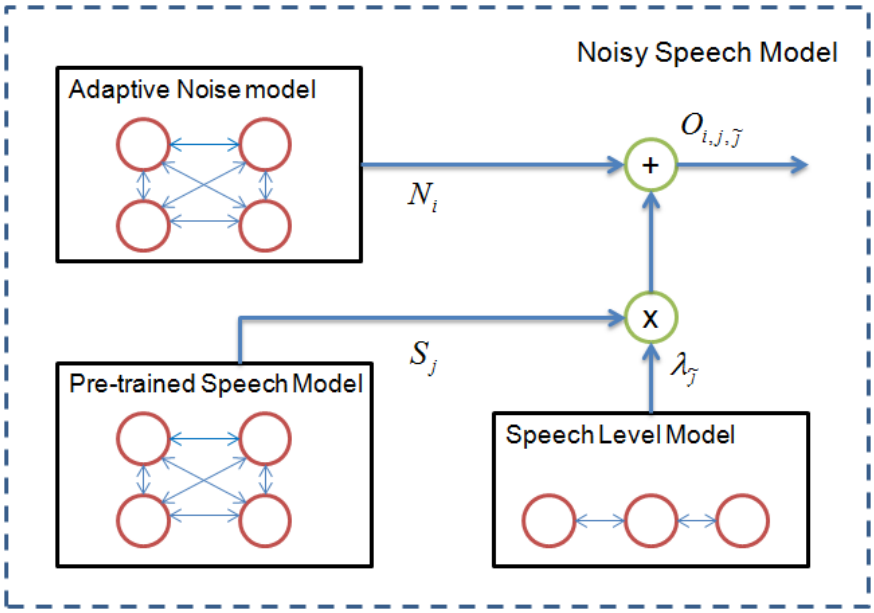


Figure 4.1: Overview of the noisy speech model.

with the spectrum of the noise state and taking a weighted average over all states.

As a summary, many methods has been proposed for joint estimation of speech and noise from a combined speech and noise model. They have been widely used in speech recognition task. Speech models developed for recognition often ignore the overall speech level. In the context of speech enhancement, the speech model can be comprised of two components: a speech level model, and a magnitude normalised speech model which can be pre-trained from a speech data base.

4.3 Noise Estimation During Speech Presence

4.3.1 Model overview

In this section we add a model of speech to our adaptive noise model and jointly estimate both the speech and noise. Since the speech signal is corrupted with uncorrelated additive noise, the observed noisy speech signal is given by $O_t(k) = S_t(k) + N_t(k)$ where t and k are the time and frequency indices respectively. In order to determine the noise

state during speech activity, we need to incorporate a speech model into our existing noise estimation model described in Sec. 3.4. An overview of the production model for noisy speech is shown in Fig. 4.1. It includes three components: the adaptive noise model developed in the previous section, a model for normalised speech and model for the overall speech level. The output, N_i, from the adaptive noise model is added to that of the speech and speech level model. The "normalised speech model" is trained on clean speech utterances that have been normalized to an active level of 0 dB as measured according to ITU P.56 [68]. Thus this model incorporates the spectral and level variations between different phones but not long term changes in speech level or amplifier gain. The speech model should also be trained using multiple speakers to ensure that it is speaker-independent. The output from the speech model is multiplied by that from the speech level model to give the speech power spectrum in each frame. The advantage of separating the speech model into these two components is two-fold: the number of states required in the "normalised speech model" is greatly reduced and the speech level model can enforce the long-term consistency of average speech power over periods of several seconds. The latter constraint is key to identifying abrupt changes in the noise when speech is present.

Since the speech level changes slowly over time, the estimated speech can be viewed as the product of the normalized speech power and the speech level. The speech model is a densely connected HMM and is pre-trained from a collection of clean speech signals with a normalized speech level. The complexity of the speech model is a compromise between accurate modelling of the speech and the computational requirement of the system. The speech level model is a sequential HMM, where the speech level $\gamma_{\widetilde{\jmath}}$ for the state $\widetilde{\jmath}$ is chosen from a discrete data set of possible speech levels. A fairly good estimation of speech level is required to distinguish abrupt changes of the noise when speech is present. The speech level HMM is sparsely connected with each state connected only to its immediate neighbours as illustrated in Fig. 4.1. The speech level model has a lower frame rate than the other two model and the combination of frame rate and level increment places a hard limit on the rate of change of speech level.

4.3.2 Log Mel-frequency domain

According to the hidden Markov model we introduced in Sec. 3.4, which is derived from the model $\zeta^{(T)}$ and the observations $O^{(T)}$ based on information available at time T, the forward and backward state occupation probabilities are given by:

$$\alpha_i(t) = \sum_j \alpha_j(t-1) a_{ji} b_i(O_t) \qquad \text{with} \quad \alpha_i(0) = \pi_i \tag{4.1}$$

$$\beta_i(t) = \sum_j a_{ij} b_j(O_{t+1}) \beta_j(t+1) \qquad \text{with} \quad \beta_i^{(T)}(T) = \pi_i \tag{4.2}$$

$$P^{(T)} = \sum_i \alpha_i^{(T)}(T) \beta_i^{(T)}(T) \tag{4.3}$$

where the power spectral components $O_t\,(k)$ are assumed to follow a negative exponential distribution, and $b_j\,(O_t)$ is taken to be the corresponding probability density from (3.2). The observation probabilities of a speech spectral model can be better represented using Gaussian pdfs in the Mel-frequency log power or cepstral domains [19]. For our noisy speech model, the first of these two is preferred because it preserves spectral locality when the speech energy and noise energy occupy predominantly different spectral regions. We therefore consider spectra in three different domains:

- the power domain indexed by k,
- the Mel-frequency power domain indicated by a subscript $[M]$ and indexed by m.
- the Mel-frequency log power domain indicated by a subscript $[L]$ and also indexed by m.

The Mel frequency scale [116] is defined by the nonlinear transformation of a frequency f Hz into Mel as [87],

$$\text{mel}(f) = 1000 \frac{\log\left(1 + \frac{f}{700}\right)}{\log\left(1 + \frac{1000}{700}\right)}. \tag{4.4}$$

If in a particular signal state, the mean and variance of the power spectrum are given by $\mu(k)$ and $\sigma^2(k)$, we can transform a mean spectrum, $\mu(k)$ into the Mel power domain by convolving it with a bank of triangular filters, $M_m(k)$, as in [19] to give

$$\mu_{[M]}(m) = \sum_k M_m(k) \star \mu(k). \tag{4.5}$$

If we assume that the spectral components are independent, the corresponding transformation for the variances is given

$$\sigma^2_{[M]}(m) = \sum_k M^2_m(k) \star \sigma^2(k). \tag{4.6}$$

The transformation into the Mel-frequency log power domain for an observed power spectrum $O(k)$ is likewise given by

$$O_{[L]}(m) = \log\left(O_{[M]}(m)\right) = \log\left(\sum_k M_m(k)O(k)\right). \tag{4.7}$$

Under the further assumption that the spectral components in the Mel-frequency power domain have a log-normal distribution, we have the following exact transformation [75, 41],

$$\begin{aligned} \sigma^2_{[L]} &= \log\left(1 + \frac{\sigma^2_{[M]}}{\mu^2_{[M]}}\right) \\ \mu_{[L]} &= \log\left(\mu_{[M]}\right) - \frac{1}{2}\log\left(1 + \frac{\sigma^2_{[M]}}{\mu^2_{[M]}}\right). \end{aligned} \tag{4.8}$$

The log observation probability in $[L]$ domain, $\vec{b}$, of an observation, O, is therefore given by

$$\log \vec{b}(O) = -\frac{1}{2}\sum_{m}\log\left(2\pi\sigma^2_{[L]}(m)\right) + \frac{\left(O_{[L]}(m) - \mu_{[L]}(m)\right)^2}{\sigma^2_{[L]}(m)} \tag{4.9}$$

where $\mu_{[L]}$ and $\sigma^2_{[L]}$ are obtained from $\mu_{[M]}$ and $\sigma^2_{[M]}$ using (4.8).

Incorporating the speech model

Given an observation of a noisy speech signal, we are interested in its log likelihood based on the noisy speech model illustrated in Fig. 4.1. The normalised speech model can be trained in the Mel-frequency power domain: $\zeta_s = \{\nu_j, \varsigma^2_j\}$, where ν_j and ς^2_j are the mean and the variance for the speech state j. The noise model is given as $\zeta = \{\mu_i, \sigma^2_i\}$ where the mean μ_i and the variance σ^2_i have been converted into Mel-frequency domain accordingly. Given the speech level $\gamma_{\widetilde{\jmath}}$ at state $\widetilde{\jmath}$ of the speech level model, the mean $\mu_{[M]}$ and variance $\sigma^2_{[M]}$ of the noisy speech model required in (4.8) can be expressed in the Mel-frequency power domain as the sum of components from the noise model state and level-adjusted speech model state:

$$\begin{aligned}
\sigma^2_{[M]}|\sigma^2_i, \varsigma^2_j, \gamma_{\widetilde{\jmath}} &= \sigma^2_i(m) + \gamma^2_{\widetilde{\jmath}}\varsigma^2_j(m) \\
\mu_{[M]}|\mu_i, \nu_j, \gamma_{\widetilde{\jmath}} &= \mu_i(m) + \gamma_{\widetilde{\jmath}}\nu_j(m)
\end{aligned}$$

given that the speech and noise signals are uncorrelated. Thus the log observation probability $\log \vec{b}(O)$, described in (4.9), can be expressed as a function of $\{\mu_i, \sigma^2_i, \nu_j, \varsigma^2_j, \gamma_{\widetilde{\jmath}}\}$. The computational complexity of implementing our noisy speech model can be substantially reduced by imposing the constraint that the transition probabilities of the normalized speech model depend only on the destination state. With this constraint, the maximum likelihood speech state, j, is independent of the previous state sequence. Thus for any given noise state, i, and speech level state, $\widetilde{\jmath}$, we can determine the most pro-

bable speech state, j, from (4.9), and the observation probability for any noise state can be expressed by

$$\log \vec{b}_i \left(O_t \mid \mu_i, \sigma_i^2, \gamma_{\tilde{\jmath}.}\right) = \max_j \left\{\log \vec{b}_{i,j} \left(O_t \mid \mu_i, \sigma_i^2, \nu_j, \varsigma_j^2, \gamma_{\tilde{\jmath}.}\right)\right\} \tag{4.10}$$

However, in our noise estimation, we do not have any prior knowledge of the speech level, and we have to estimate it from the observed noisy speech. In order to estimate the speech level, we perform a Viterbi algorithm over the most recent L frames to find the maximum likelihood sequence of noise states, $i(t)$, and speech level states, $\tilde{\jmath}(t)$. For $T_L + 1 < t \leq T$, the probability of a state sequence ending in states i and $\tilde{\jmath}$ is calculated recursively as

$$\phi_{i,\tilde{\jmath}}(t) = \left[\max_{i',\tilde{\jmath}'} \phi_{i',\tilde{\jmath}'}\left(t-1\right) a_{i'i} \widetilde{a}_{\tilde{\jmath}'\tilde{\jmath}}\right] \vec{b}_i \left(O_t \mid \mu_i, \sigma_i^2, \gamma_{\tilde{\jmath}(t)}\right) \tag{4.11}$$

where $\widetilde{a}_{\tilde{\jmath}'\tilde{\jmath}}$ is the transition probability of speech level from $\gamma_{\tilde{\jmath}'}$ to $\gamma_{\tilde{\jmath}}$, the initial values $\phi_{i,\tilde{\jmath}}(T_L)$ are saved from the previous iteration and $\vec{b}_i \left(O_t \mid \mu_i, \sigma_i^2, \gamma_{\tilde{\jmath}}\right)$ is defined in (4.10). From this, $i(T)$ and $\tilde{\jmath}(T)$ are taken as $\arg\max \phi_{i,\tilde{\jmath}}(T)$.

Since the speech level of a particular speech remains constant most of the time, we define the speech level state transition probabilities $\widetilde{a}_{\tilde{\jmath}'\tilde{\jmath}}$ as below

$$\widetilde{a}_{\tilde{\jmath}'\tilde{\jmath}} = \begin{cases} \kappa & \text{for } \tilde{\jmath} = \tilde{\jmath}' \\ \frac{(1-\kappa)}{2} & \text{for } \tilde{\jmath} = \tilde{\jmath}' \pm 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

where κ is the frame rate of the speech level can change.

From the Viterbi decoding algorithm, a most probable sequence of speech levels $\gamma_{\tilde{\jmath}(t)}$ is obtained by backtracking, thus the log observation probability of noisy speech is given as

$$\log \vec{b}_i \left(O_t \mid \mu_i, \sigma_i^2\right) = \max_j \left\{ \log \vec{b}_{i,j} \left(O_t \mid \mu_i, \sigma_i^2, \nu_j, \varsigma_j^2, \gamma_{\tilde{\jmath}(t)}\right) \right\} \tag{4.13}$$

Since both the $\mu_i(m)$ and $\sigma_i^2(m)$ can be calculated from $\mu_i(k)$, we will, for clarity, write $\log \vec{b}_i(O_t)$ instead of $\log \vec{b}_i\left(O_t \mid \mu_i, \sigma_i^2\right)$ in the remainder of this section.

Overview

The calculation of log observation probability can be summarised as below,

1. convert the mean spectrum of each noise model state in the frequency domain, $\mu_i(k)$, into the Mel-frequency domain to give mean and variance $\mu_i(m)$ and $\sigma_i^2(m)$ using (4.5) and (4.6)

2. convert the observed power spectrum in the frequency domain, $O_t(k)$, into the log Mel-frequency domain $O_{[L]}(m)$ using (4.7)

3. given a noise state, for every speech level, select a speech state that maximises the log-likelihood calculated in (4.10)

4. find the best sequence of speech level states $\tilde{\jmath}(t)$ from the modified Viterbi procedure described in (4.11)

5. the observation probability for a given noise state, $\vec{b}_i(O_t)$, is calculated from $\gamma_{\tilde{\jmath}(t)}$ with associated speech state determined in step 3 using (4.13)

We note that the Mel-frequency log power domain is used only for calculating $\log \vec{b}_i(O_t)$. Unless otherwise stated, the expressions in the following sections for estimating the mean and variance of the noise spectral components all operate in the linear-frequency power domain.

4.3.3 Time-Update

In this section, we present the noise estimation algorithm and the update procedures used for slowly evolving noise environments. From Sec. 3.4, in order to update the noise model parameters recursively, the accumulated mean power spectrum for state i is calculated recursively as

$$U_i^{(T)}(1, T_L + 1) = \lambda U_i^{(T-1)}(1, T_L) + \frac{\alpha_i^{(T)}(T_L + 1)\beta_i^{(T)}(T_L + 1)O_{T_L+1}}{P^{(T)}} \tag{4.14}$$

where speech is assumed to be absent, i.e. $O_t(k) = N_t(k)$. However, in the presence of speech, $O_t(k) = S_t(k) + N_t(k)$, we only wish to update those frequency bins in which the speech is absent. To do this, we determine a speech presence mask $\eta_i(k)$, where $\eta_i(k) = 1$ indicates the speech is present at frequency k given the noise estimate is $\mu_i(k)$.

The speech presence mask in each frequency bin is obtained using a minimum statistics approach presented in [16]. However, instead of tracking the global minimum spectral power, we have to track the minimum $\varpi_i(k)$ for each individual noise state. Each of the observations, O_t, is first assigned to the noise state with the highest observation probability, $\arg\max_i b_i(O_t)$. The observations assigned to any particular state are then smoothed using $\overline{O}_{i,t}(k) = \varepsilon\overline{O}_{i,t-1}(k) + (1-\varepsilon)O_t(k)$, where ε is a smoothing factor. Minimum tracking is performed over the past L frame estimates of $\overline{O}_{i,t}(k)$ to obtain $\varpi_i(k)$. The speech presence mask $\eta_i(k)$ is then determined by comparing the spectral power of the observation, $\overline{O}_{i,t}(k)$, within the minimum $\varpi_i(k)$,

$$\eta_{i,t}(k) = \begin{cases} 1 & \text{if } \frac{\overline{O}_{i,t}(k)}{\varpi_i(k)} > \Gamma \\ 0 & \text{otherwise} \end{cases}$$

where Γ is a decision threshold used to identify whether the speech is present in this time-frequency bin. Similar to [16], we use $\Gamma = 5$ for all frequency bins in Sec. 4.4 below.

Thus the update equation for the weighted state observation sum is

$$U_i^{(T)}(1, T_L+1; k) = \begin{cases} U_i^{(T-1)}(1, T_L; k) & \text{if } \eta_{i,T_L+1} = 1 \\ \lambda U_i^{(T-1)}(1, T_L; k) + \frac{\alpha_i^{(T)}(T_L+1)\beta_i^{(T)}(T_L+1)O_{T_L+1}(k)}{P^{(T)}} & \text{otherwise} \end{cases}$$

By defining $\hat{\lambda}_i\,(T_L+1; k) = \lambda + (1-\lambda)\,\eta_{i,T_L+1}\,(k)$ the expression above can be simplified as

$$\begin{aligned} U_i^{(T)}(1, T_L+1) &= \hat{\lambda}_i\,(T_L+1)\,U_i^{(T-1)}(1, T_L) \\ &\quad + (1-\eta_{i,T_L+1})\,\frac{\alpha_i^{(T)}(T_L+1)\beta_i^{(T)}(T_L+1)O_{T-L+1}}{P^{(T)}} \end{aligned} \tag{4.15}$$

The remaining update equations only require the occupation probability of each state, which depends on the observation probability given in (4.13), and thus remain unchanged from the previous model, which is given below,

$$Q_i^{(T)}(1, T_L+1) = \lambda Q_i^{(T-1)}(1, T_L) + \frac{\alpha_i^{(T)}(T_L+1)\beta_i^{(T)}(T_L+1)}{P^{(T)}} \tag{4.16}$$

$$R_{ij}^{(T)}(1, T_L) = \lambda R_{ij}^{(T-1)}(1, T_L-1) + \frac{\alpha_i^{(T)}(T_L-1)\vec{b}_j^{(T)}(O_{T_L})\beta_i^{(T)}(T_L)}{P^{(T)}} \tag{4.17}$$

and the means and transition probabilities are now calculated as

$$\mu_i^{(T)} \approx \frac{\lambda^L U_i^{(T-1)}(1, T_L) + U_i^{(T)}(T_L+1, T)}{\lambda^L Q_i^{(T-1)}(1, T_L) + Q_i^{(T)}(T_L+1, T)} \tag{4.18}$$

and

$$a_{ij}^{(T)} \approx \frac{a_{ij}^{(T-1)}\left(\lambda^L R_{ij}^{(T-1)}(1, T_L-1) + R_{ij}^{(T)}(T_L, T-1)\right)}{\lambda^L Q_i^{(T-1)}(1, T_L-1) + Q_i^{(T)}(T_L, T-1)}. \tag{4.19}$$

4.3.4 Adapting to rapidly changing noise characteristics

In situations where the noise characteristics evolve slowly with time, they will be tracked by the update procedure described in Sec. 4.3.3 above. However when an abrupt change occurs such as, for example, the introduction of an entirely new noise source, it is necessary to create an entirely new noise state. The procedure is similar to that described in Sec. 3.4.6 but needs to be modified to take account of the possible presence of speech.

We assume that the maximum number of noise states is fixed in advance and so it is necessary to merge the two closest states before creating a new one; this process was illustrated for a three-state noise model in Fig. 3.4 in Sec. 3.4.6. The criteria used to decide whether to create a new state is the same as illustrated in Fig. 3.5. Similar to Sec. 3.4.6, a "Z-test" is used to assess how well the most recent L frames match the existing noise model. However, this needs to be done in the log Mel-frequency domain. If the test indicates a poor fit, a tentative model is created by merging the closest two states and creating a new one. Only if this tentative model provides an improved fit to recent observation frames is it substituted for the existing model.

In order to decide when to introduce a new state, we calculate a measure $Z^{(T)}$ that indicates how well the most recent L frames of observed data fit the current model, $\zeta^{(T)}$. From (4.9) and (4.13), it is straightforward to show that, given its mean and variance and assuming the spectral components are independent, the log-likelihood of an observed frame, O_t, has the following mean and variance in Mel-frequency log power domain:

$$
\begin{aligned}
E\left\{\log \vec{b}\left(O_t \mid \mu, \sigma^2\right)\right\} &= E\left\{-\frac{1}{2}\sum_m \log\left(2\pi\sigma_{[L]}^2(m)\right) + \frac{\left(O_{[L]}(m) - \mu_{[L]}(m)\right)^2}{\sigma_{[L]}^2(m)}\right\} \\
&= -\frac{1}{2}\sum_m \left(\log\left(2\pi\sigma_{[L]}^2(m)\right) + 1\right) \\
Var\left\{\log \vec{b}\left(O_t \mid \mu, \sigma^2\right)\right\} &= E\left\{-\frac{1}{4}\sum_m \left(\frac{\left(O_{[L]}(m) - \mu_{[L]}(m)\right)^2}{\sigma_{[L]}^2(m)} - 1\right)^2\right\} \\
&= \frac{1}{4}\sum_m E\left\{\frac{\left(O_{[L]}(m) - \mu_{[L]}(m)\right)^4 - 2\sigma_{[L]}^2(m)\left(O_{[L]}(m) - \mu_{[L]}(m)\right)^2 + \sigma_{[L]}^4(m)}{\sigma_{[L]}^4(m)}\right\} \\
&= \frac{1}{4}\sum_m \frac{3\sigma_{[L]}^4(m) - 2\sigma_{[L]}^4(m) + \sigma_{[L]}^4(m)}{\sigma_{[L]}^4(m)} = \frac{M}{2}
\end{aligned}
$$

Accordingly we define $Z^{(T)}$ as the normalized difference between the weighted log-likelihood of the most recent L frames and its expectation

$$
Z^{(T)} = \frac{\frac{1}{2}\sum_{t=T_L+1}^{T} \lambda^{T-t} \sum_m \left(1 - \frac{\left(O_t(m) - \mu_{[L]}(m)\right)^2}{\sigma_{[L]}^2(m)}\right)}{\sqrt{\frac{M}{2}\sum_{t=T_L+1}^{T}\left(\lambda^{T-t}\right)^2}} \tag{4.20}
$$

where $i(t)$ gives the state occupied at time t in the maximum likelihood state sequence.

If $\left|Z^{(T)}\right|$ exceeds an empirically determined threshold, θ_Z, then this indicates that $\zeta^{(T)}$ should be re-estimated and a new type of noise might be present. In this case, we therefore create a tentative model, $\hat{\zeta}^{(T)}$, in which two of the existing states are merged and a new state created.

Initialising the new state

As for the speech absent procedure in Sec. 3.4.6, we first create an initial model $\hat{\zeta}^{(T-1)}$ and then perform the time update from Sec. 4.3.3 to determine $\hat{\zeta}^{(T)}$. For the tentative model $\hat{\zeta}^{(T-1)}$, we first determine the pair of states, $\{i_0, j_0\}$, whose merging will cause the least reduction in likelihood. In contrast to the speech absent case, we cannot initialise

the new state to O_T because O_T might be corrupted by speech. Accordingly, a robust initial estimate for the mean power spectrum, Θ, of the new state is obtained by taking the median in each frequency bin of the $L^{'}$ frames out of the most recent L that have the lowest likelihood under the current noise model, i.e. $\log \vec{b}_{i(t)}\left(O_t \mid \mu_{i(t)}\right)$. The choice of $L^{'}$ is a compromise; it needs to be large enough to provide a robust initial estimate of the new state's power spectrum but small enough that the majority of included frames include examples of the new noise source, (currently we set $L^{'} = L/3$). The motivation for this is that the low likelihood frames are those most likely to include examples of any new noise source and that in each frequency bin the noise will be dominant in at least some of them. Therefore, we initialize the state means for the model $\hat{\zeta}^{(T-1)}$ to be

$$\hat{\mu}_r^{(T-1)} = \begin{cases} \Theta & \text{for } r = j_0 \\ \frac{Q_{i_0}(1,T_L)\mu_{i_0}^{(T-1)} + Q_{j_0}(1,T_L)\mu_{j_0}^{(T-1)}}{Q_{i_0}(1,T_L)+Q_{j_0}(1,T_L)} & \text{for } r = i_0 \\ \mu_r^{(T-1)} & \text{otherwise} \end{cases} \tag{4.21}$$

where the state j_0 models the new noise spectrum, and state i_0 is initialized as a weighted average of the previous states i_0 and j_0.

We now re-train the initial model, $\hat{\zeta}^{(T-1)}$, using Viterbi decoding on the most recent L frames by backtracking, $\{O_t : t \in [T_L + 1, T]\}$,

$$\hat{\varphi}_{j,\widetilde{\jmath}}(t) = \left[\max_{i,\widetilde{\imath}} \varphi_{i,\widetilde{\imath}}(t-1)\, a_{ij}\widetilde{a}_{\widetilde{\imath}\widetilde{\jmath}}\right] \vec{b}_j\left(O_t \mid \hat{\mu}_j, \sigma_j^2, \gamma_{\widetilde{\jmath}}\right)$$

where the maximum likelihood sequence of the noise state and speech level can be obtained. In order to update the mean of the new state $\hat{\mu}_j$, we are only interested in the frames that have been assigned to noise state j in our previous Viterbi decoding. We first define the set of frames, Ω_j, for which this is true:

$$\Omega_j = \{t : \ t \in [T_L + 1, T]\,;\ \text{frame } t \text{ assigned to noise state } j\} \tag{4.22}$$

It is possible that some of the frames within Ω_j might contain speech energy in addition

to noise, and so, when determining the initial new state mean $\hat{\mu}_j$, we need to mask out any time-frequency bins that might be dominated by speech energy. Thus the new state mean $\hat{\mu}_j(k)$ can be updated using the recursive expression shown below,

$$\hat{\mu}_j(k) = \text{median}\left\{O_t(k) : \ t \in \Omega_j; \ O_t(k) < \Gamma\hat{\mu}_j(k)\right\} \tag{4.23}$$

where the median is used to avoid extreme value. In rare cases, the subset of Ω_j defined in (4.23) might be empty since all the available frames might be masked by high energy of speech in certain frequency bins. If this is true, then we set $\hat{\mu}_j(k) = \min\left\{O_t(k) : \ t \in \Omega_j\right\}$. The process is repeated until $\hat{\mu}_j$ converges. For this newly created state mean $\hat{\mu}_j$, we will repeat the Viterbi decoding until $Z^{(T)}$ is minimized.

The initialization of the new state mean can be summarised as below,

1. Initialize the mean $\hat{\mu}_r^{(T-1)}$ as described in (4.21)
2. Apply the Viterbi decoding on the most recent L frame to obtain set Ω_j
3. For each frequency bin k, check whether $O_t(k) < \Gamma\hat{\mu}_j(k)$
4. Recalculate $\hat{\mu}_j(k)$ as described in (4.23)
5. Go to step 3 until $\hat{\mu}_j(k)$ converges
6. Recalculate $Z^{(T)}$ as described in (4.20)
7. Go to step 2 unless $Z^{(T)}$ does not decrease

Recalibrating the new model

The accumulated sums in (4.15) to (4.17) can now be re-calculated by distributing the existing sums between the new states accordingly,

$$
\begin{aligned}
\hat{U}_i^{(T-1)}(1,T_L) &= \sum_m \phi_{mj} U_m^{(T-1)}(1,T_L) \\
\hat{Q}_j^{(T-1)}(1,T_L) &= \sum_m \phi_{mj} Q_m^{(T-1)}(1,T_L) \\
\hat{R}_{ij}^{(T-1)}(1,T_L-1) &= \sum_m \sum_n \phi_{mi}\phi_{nj} R_{mn}^{(T-1)}(1,T_L-1)
\end{aligned} \tag{4.24}
$$

where $\phi_{ij} = \frac{b\left(\mu_i^{(T-1)}|\hat{\mu}_j^{(T-1)}\right)}{\sum_j b\left(\mu_i^{(T-1)}|\hat{\mu}_j^{(T-1)}\right)}$ estimates the probability of a frame that was previously in state i being in state j of the new model. As a final step, the time update of Sec. 4.3.3 are applied to update from $\hat{\zeta}^{(T-1)}$ to $\hat{\zeta}^{(T)}$.

However, we only wish to use this revised model if it will result in an increase in log likelihood. Accordingly the increase, $I^{(T)}$, in the log-likelihood is estimated as

$$
\begin{aligned}
I^{(T)} &= \sum_{t=T_L+1}^{T} \lambda^{T-t} \sum_i \hat{Q}_i(t,t)\log b(O_t,\hat{\mu}_i) - Q_i(t,t)\log b(O_t,\mu_i) \\
&\quad -\frac{\lambda^L}{1-\lambda}\sum_i\sum_j \phi_{ij}\pi_i D(\mu_i,\hat{\mu}_j)
\end{aligned} \tag{4.25}
$$

where $D(\mu_i,\hat{\mu}_j) = \sum_k \left(\frac{\mu_i(k)}{\hat{\mu}_j(k)} - \log\frac{\mu_i(k)}{\hat{\mu}_j(k)} - 1\right)$ is the Itakura-Saito distance and equals the expected increase in log likelihood of a frame whose true mean power spectrum is μ_i is modelled by a state with mean $\hat{\mu}_j$. The first two terms in (4.25) give the log likelihood improvement over the most recent L frames while the last term approximates the decrease in log likelihood of the earlier frames.

4.3.5 Safety-net state

In order to increase the robustness of our model, we define our last noise state, $i = H_n$, to be a "safety-net state". This safety-net state will be trained and updated as previously

described but with an exception: the mean of this state μ_{H_n} is determined using Minimum statistics (MS) [90, 11] instead of with (4.18). The introduction of this safety-net state prevents the noise model from diverging even if wrong state assignments are made during speech active intervals. However, the safety-net state was only used in the early stage of HMM algorithm development. With the latest HMM algorithm presented in this thesis, the safety-net state is never assigned in the most likely state sequence, and we will turn this safety-net state off for all the experiments below.

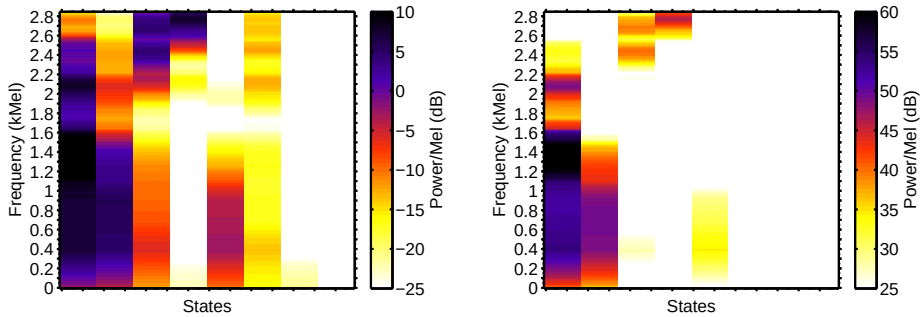
4.4 Experimental Results

For all the experiments, the signals are sampled at a frequency of 16 kHz, and the power spectrum calculated for overlapping frames using the STFT. Similar to the setting in Sec. 3.6, the time-frames have a length of 32 ms with a 50% overlap resulting in $K = 257$ frequency bins. We retain the most recent $L = 30$ frames (480 ms), and also set the initial training time $T_0 = 30$ frames. The forgetting factor is chosen to be $\lambda = 1 - 1/(2L)$, which gives a time constant of $2L = 960$ ms. Since the number of Mel-freqency bins M is small, $Z^{(T)}$ can be assumed to be normal distributed with mean of 0 and variance of 1, thus the threshold θ_Z defined in Sec. 4.3.4 is set to 1.645, i,e, reject the existing model at 5% significance level.

4.4.1 Training of the speech model

As described in Sec. 4.3.2, we need to train our speech model in the Mel-frequency domain. The number of states, H_s, in the speech model is set to 8; this was found to be the smallest number of states that gave a reasonable representation of the normalized power spectra encountered in speech. The transition probability from any state to another state is set to be $1/H_s$, i.e. it is initialised as equally likely to go from any state to any other state. For the speech training set we chose 10 sentences from IEEE sentence database [106]. We first normalize the active level of each sentence to 0 dB using [68], then convert the speech power spectrum $S(k)$ into the Mel-frequency power spectrum $S(m) = (M * S(k))$ as described in Sec. 4.3.2. Using a K-means clustering algorithm [11], we partitioned the speech into $H_\mathrm{s} - 1$ states, then we have added the H_sth state as a silence state, with the mean and variance in each frequency bin equal to 0. The mean power and its variance of each state are shown in Fig. 4.2.

For the speech level γ, we define a discrete set from -20 dB to 0 dB relative to the mean energy level of the noisy speech signal with 2 dB increments, this corresponds to an SNR range of -20 to $+\infty$ dB. The speech level state transition probabilities $\widetilde{a}_{\widetilde{i}\widetilde{j}}$ defined in (4.12) are given by,



(a) mean (b) variance

Figure 4.2: Spectrogram of the (a) mean (b) variance of different speech states.

$$\widetilde{a}_{\widetilde{\imath}\widetilde{\jmath}} = \begin{cases} 0.8 & \text{for } \widetilde{\jmath} = \widetilde{\imath} \\ 0.1 & \text{for } \widetilde{\jmath} = \widetilde{\imath} \pm 1 \\ 0 & \text{otherwise} \end{cases}$$

4.4.2 Noise Tracking

In this section, we evaluate the performance of the MS and 3-state HMM noise estimation models on three types of noise (a) slowly evolving (b) non-stationary and (c) abruptly changing. We evaluate the performance of the algorithms using COSH distance between the true noise signal and its estimates. In this section, we have turned off the safety-net state, which uses the UM method to determine the mean as one of the state of our noise estimation.

Slowly evolving noise

A good noise estimator should be able to track and update gradual changes in the noise characteristics. Fig. 4.3 (a) shows the spectrogram of noisy speech at an overall level of 0 dB SNR, corrupted by a car noise with increasing amplitude of power, where the active level is the same as the average power of car noise, and shown in Fig. 4.3(b). The noise level increases by roughly 7 dB over 10 seconds. The spectrogram of the estimated noise

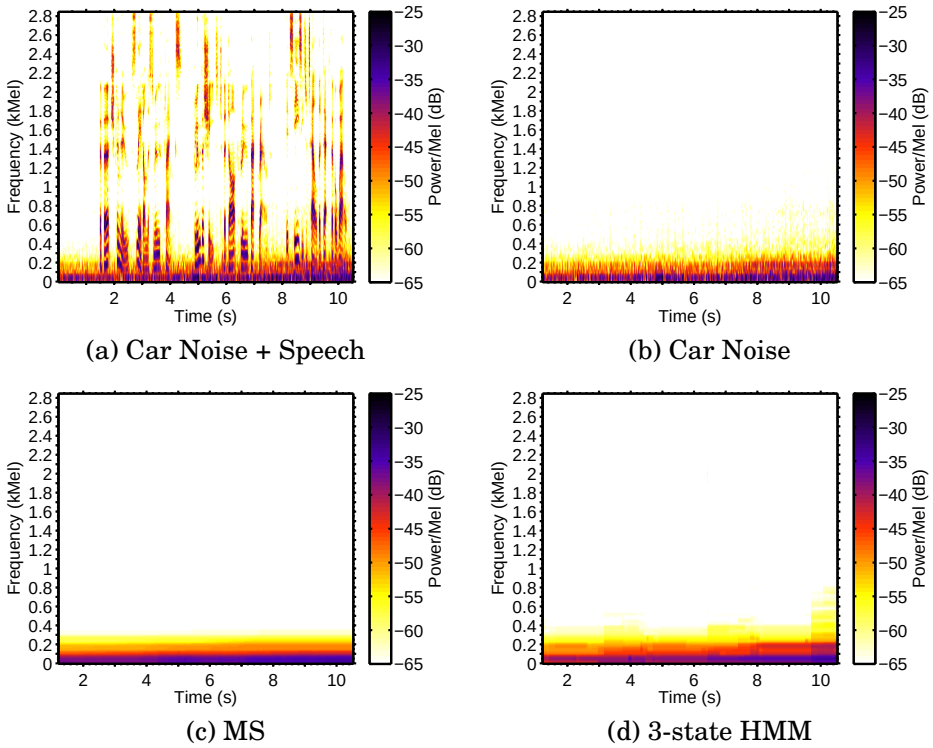


Figure 4.3: Spectrogram of (a) noisy speech corrupted by (b) increasing car noise with its estimation using (c) MS (d) a 3-state HMM.

using MS method and 3-state HMM method are shown in Fig. 4.3(c) and (d) respectively. From both the figures, we can see that they have modelled the noise well, although 3-state HMM performs slightly better visually. In this, and subsequent experiments, we assume the first 1 second of the signal contains no speech; this interval is used to initialised the noise model and is omitted from the other plots shown in Fig. 4.3(c) and (d).

Non-stationary noise

Fig. 4.4(a) shows the spectrogram of a speech signal corrupted by the machine gun noise shown in 4.4(b) at 0 dB. The spectrogram of the estimated noise using the MS method and the 3-state HMM method are shown in Fig. 4.4(c) and Fig. 4.4(d) respectively. The MS model is unable to follow the rapid change of noise characteristics and the noise estimates remains close to zero throughout. The HMM performs much better

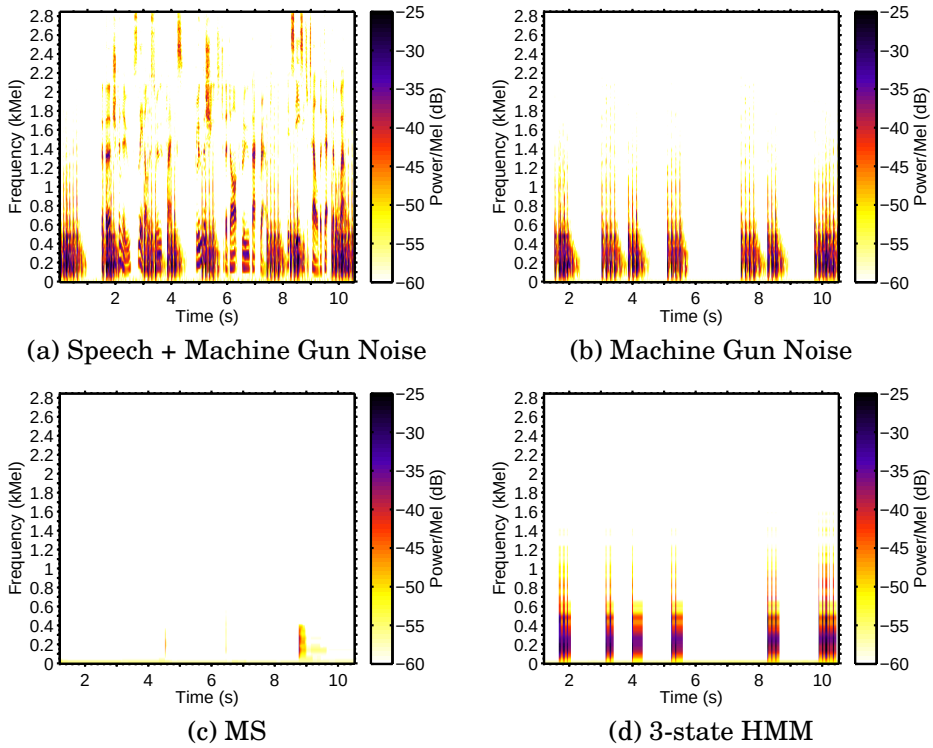
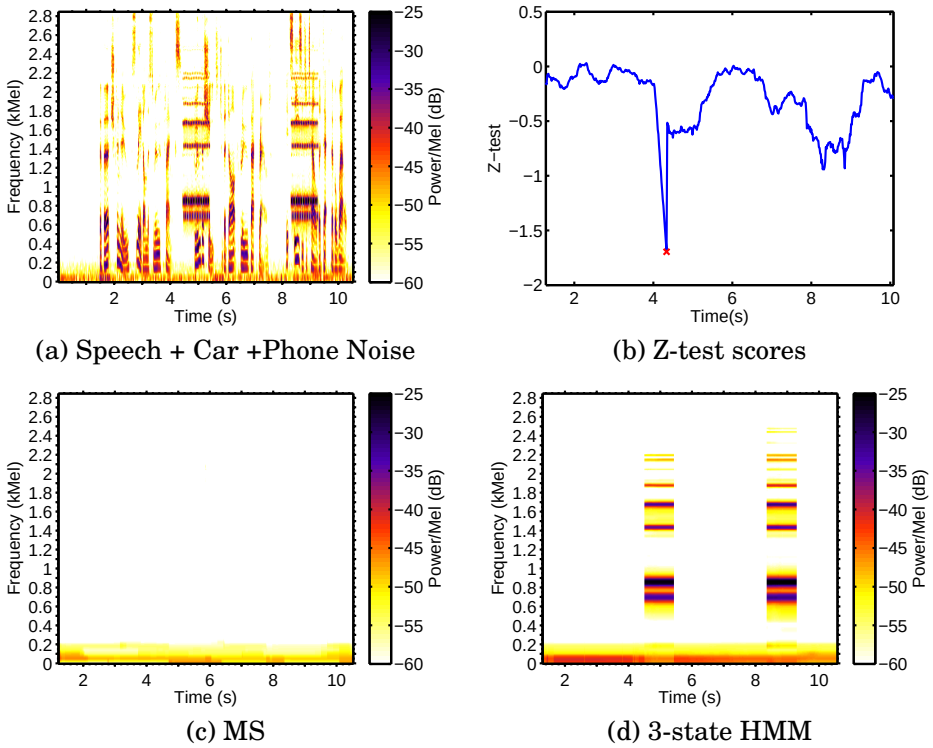


Figure 4.4: Spectrogram of (a) noisy speech corrupted by (b) machine gun noise with its estimation using (c) MS and (d) a 3-state HMM.

and assigns the correct state to the machine gun bursts that occur within the speech. Because the levels are quite similar, some machine gun noise frames are incorrectly classified as speech (e.g. between 7s and 8s). Note that in this example, the initial training interval (0 to 1s) included an example of the machine gun noise and so was included in the initial HMM noise model.

Abrupt noise detection

In this experiment, the noise of a ringing phone is added to a background car engine noise which is predominantly low frequency. Fig. 4.5(a) shows the spectrogram of a speech signal corrupted by this composite noise at 0 dB and it can be seen that the noise spectrum changes abruptly whenever the phone rings. In this example, the initial training interval (0 to 1s) does not include any example of the ringing noise, so the new noise must be acquired during speech presence. The spectrogram of the estimated noise



(a) Speech + Car +Phone Noise (b) Z-test scores

(c) MS (d) 3-state HMM

Figure 4.5: Spectrogram of (a) noisy speech corrupted by car+phone noise with its estimation using (c) MS (d) a 3-state HMM; (b) Z-test values.

using the MS method is shown in Fig. 4.5(c). As would be expected this model is unable to track the rapidly changing noise and fails to include any phone noise. Fig. 4.5(b) shows the the value of $Z^{(T)}$, which measures how well the L most recent observations fit the model. We see that when the first phone ring occurs, at approximately 4.3 s, there is an abrupt fall in $Z^{(T)}$ which indicates the arrival of a novel noise spectrum. The red cross at $t = 4.3$ s indicates where the new state is created. Two of the existing states are merged and a new state is reallocated to model the new noise spectrum. The corresponding noise spectrogram for the HMM is shown in Fig. 4.5(d) in which the estimated noise spectrum tracked the abrupt changes in noise characteristics, and shows a much better noise estimation than MS method. It is worth noting that the car noise component of ringing state is lower at first (at about 4.3 to 5.4 s), but gradually adapt to the actual energy level later (at about 8.3 to 9.4 s). This is due to the fact that during initial creation of the new state, the low frequency component that corresponds to the car noise is partially masked by the speech, and it takes some time for it to update

	Car	Gun	Phone
MS	20.3	2006.4	6458.0
2-state HMM	**8.8**	1032.1	62.1
3-state HMM	**8.8**	849.6	31.2
4-state HMM	**8.8**	**788.5**	**28.9**

Table 4.1: COSH distance of different noise estimations.

to the actual level.

COSH errors

The average COSH distances between the true noise signal and its estimate using the MS model and using the HMM model with 2, 3 and 4 states are shown in Table 4.1. The results confirm our observations in Fig. 4.3 to 4.5. The MS model gives a low error for the car noise but is unable to track abrupt changes in noise characteristics, and shows large COSH errors when estimating non-stationary noise such as "Gun" and "Phone" noise. For the "gun" noise, both the HMM method also gives a large COSH error. The reason for this is that in some frames the state assignment is incorrect and the noise is under-estimated. Since the slow evolving car noise is quasi-stationary, there is no significant modelling improvement when the number of HMM states is varied from 2 to 4. In contrast, for the highly non-stationary car and phone noises, the COSH error continues to improve as the number of HMM states is increased. The improvement between 3 and 4 states is, however, very much less than that between 2.

4.4.3 Speech Enhancement

In this section, we incorporate our HMM noise estimator into a speech enhancer to assess whether our noise estimator improves the quality of speech as compared to other noise estimation methods. We will first demonstrate an example of how well the noise can be suppressed using our method, then we will run a set of experiments to show the improvements in terms of PESQ and segmental SNR of the enhanced speech. All the clean speech signals were taken from the IEEE sentence database [106] by concatenating three sentences to give an average duration of about 10 seconds.

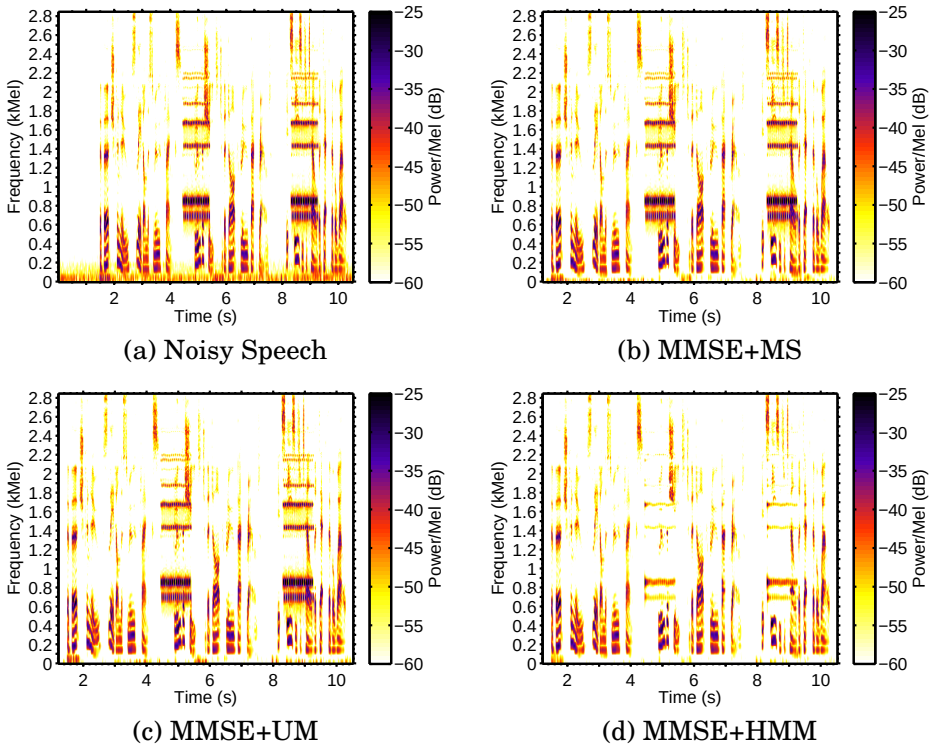


Figure 4.6: Spectrogram of (a) the unenhanced noisy speech corrupted by the car+phone noise, and the MMSE enhanced speech using different noise estimator (b) MS (c) UM (d) HMM.

Fig. 4.6(a) shows an example of a speech signal corrupted by a ringing phone noise at 0 dB SNR, shown in Fig. 4.5 (a). It is assumed that there will be non-speech segment at the beginning of the signal, roughly 1.5 s in this case, and it is used to initialize our noise estimation model, and the rest of the signal forms the speech active segment.

The speech active segments of the given noisy speech signal is then enhanced by the MMSE algorithm [30] using different noise estimators. Fig. 4.6 shows the enhanced speech signals using respectively (b) minimum statistics (MS) [90, 11], (c) unbiased MMSE estimator (UM) [46, 11] and (d) multi-state hidden Markov model (HMM) respectively. The noise-only segment is not included in the spectrogram for the enhanced speech. We see that the stationary low frequency noise component is effectively removed using all three methods but only with the HMM method is the phone ringing largely eliminated. Even though the MS and UM methods track the variation of noise level during speech presence, they cannot respond quickly enough to eliminate the phone noise.

	Unenhanced	MS	UM	HMM
PESQ	2.25	2.28	2.29	**2.44**
ΔPESQ	0	0.03	0.04	**0.19**

Table 4.2: PESQ scores and improvements of the enhanced speech.

white	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	-12.35	-7.35	-2.36	2.64	7.64	12.63
MS	-1.23	1.97	5.13	8.33	11.60	14.85
UM	-1.58	1.94	5.23	8.38	11.45	14.39
HMM	**-0.766**	**2.61**	**5.70**	**8.83**	**12.25**	**15.97**

Table 4.3: Segmental SNR of enhanced speech corrupted by white noise using different noise estimation methods.

We can assess the quality of the speech by means of the PESQ score [70]. The PESQ score for the unenhanced noisy speech is 2.25, and the PESQ score for the enhanced speech signals when using the MS, UM and HMM methods to estimate the noise are shown in Table 4.2. It can be seen that for this example, the MS and UM methods give a negligible improvement in PESQ, where as the HMM method results in a significant improvement.

Evaluation using Segmental SNR

A set of experiments was performed with noise+speech at different SNRs with a speech absent segment at the beginning of the noisy speech signal as before. 20 different clean speech signals were used, with an average duration of about 10s. Three different noise estimation algorithms were evaluated: (i) minimum statistics (MS), (ii) unbiased MMSE-based noise estimator (UM) and (iii) our proposed multi-state hidden Markov model (HMM). The number of states used for HMM is set to 3 for all the noisy speech signals below.

Tables 4.3 to 4.5 show the segmental SNR (sSNR) at different global SNRs (gSNR) of enhanced speech which has been corrupted by (i) white noise, (ii) gun noise and (iii) "car+phone" noise respectively, and the sSNR improvement at different SNR for different noise are shown graphically in Fig. 4.8. As we can see for the white noise, the HMM performs the best, whereas the UM and MS methods show a slightly lower sSNR

gun	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	2.07	7.07	**12.06**	17.06	**22.06**	**27.05**
MS	2.28	6.00	9.42	12.71	15.91	19.11
UM	2.10	5.79	9.30	12.54	15.31	17.61
HMM	**4.75**	**8.90**	12.02	**17.27**	21.51	25.77

Table 4.4: Segmental SNR of enhanced speech corrupted by machine gun noise using different noise estimation methods.

car+phone	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	-4.15	0.84	5.84	10.84	15.83	20.83
MS	1.70	6.02	10.00	13.44	16.15	18.51
UM	1.06	5.65	9.79	13.50	16.11	18.11
HMM	**5.49**	**9.08**	**13.31**	**15.95**	**18.60**	**21.18**

Table 4.5: Segmental SNR improvement of enhanced speech by "car+phone" noise using different noise estimation methods.

at low gSNR, as they underestimate the noise power when the noise power and speech power are comparable. For the "car+phone" noise, the HMM method improves the sSNR score at all SNRs and consistently outperforms the other methods. We see that the UM and MS methods degrade the sSNR score at nearly all SNRs indicating their inability to track highly non-stationary noise. For the machine gun noise, both MS and UM methods failed to track this non-stationary noise, resulting in a decrease of sSNR. At high values of gSNR, the HMM method degrades the sSNR but as the gSNR is decreased, it becomes easier for the algorithm to identify the frames containing machine gun noise. For gSNR≤ 0 dB the HMM method therefore successfully improves the sSNR.

Evaluation using PESQ

For evaluation of the PESQ scores of the enhanced speech signals, a similar set of experiments was performed as in the last section . Tables 4.6 to 4.8 show the PESQ score at different SNRs of enhanced speech which has been corrupted by (i) white noise, (ii) gun noise and (iii) “car+phone” noise respectively, and the PESQ improvement at different SNR for different noise are shown in Fig. 4.8. For the stationary noise white noise, the HMM method shows similar PESQ scores to the UM and MS methods. For the non-stationary machine gun noise, all three methods degrade the PESQ score although the degradation is lowest at all SNRs for the HMM method which results in PESQ scores

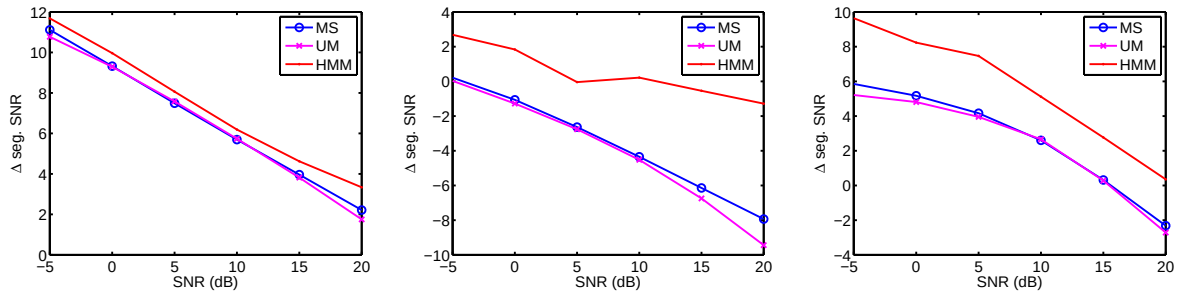


Figure 4.7: Improvement of Segmental SNR scores at different SNRs for (a) white noise (b) machine gun noise (c) "car+phone" noise.

white	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	1.13	1.36	1.68	2.05	2.40	2.74
MS	1.56	1.95	2.30	2.59	2.85	3.08
UM	1.55	1.95	2.32	2.63	2.88	3.10
HMM	**1.59**	**2.02**	**2.38**	**2.65**	**2.88**	**3.11**

Table 4.6: PESQ of enhanced speech corrupted by stationary white noise using different noise estimation methods.

very similar to those of the unenhanced noisy signal. For the abrupt "car+phone" noise, the HMM method improves the PESQ score at all SNRs and consistently outperforms the other methods. We see that the other methods degrade the PESQ score at nearly all SNRs indicating again their inability to track highly non-stationary noise. All these observations confirm the results obtained for segmental SNR.

Summary

The average improvement of the segmental SNR and PESQ scores across all global SNRs for different noise types are shown in Tables 4.9 and 4.10 respectively. When the noise is stationary, such as white and car noise, the improvement of the PESQ scores or segmental SNRs are similar for all three methods although the HMM method is consistently the best by a small margin. For the non-stationary noise, our proposed HMM method shows PESQ and sSNR improvements that are positive for all noises and substantially better than the other two methods. In the case of the improvement of the segmental SNR, the HMM method shows a good improvement over all SNRs.

gun	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	**1.97**	2.35	2.71	**3.01**	**3.27**	**3.48**
MS	1.89	2.27	2.61	2.89	3.12	3.31
UM	1.89	2.29	2.63	2.91	3.15	3.33
HMM	1.96	**2.36**	**2.72**	2.98	3.21	3.40

Table 4.7: PESQ of enhanced speech corrupted by machine gun noise using different noise estimation methods.

car+phone	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB
unenhanced	2.15	2.50	2.82	3.07	3.26	3.45
MS	2.22	2.48	2.72	2.90	3.07	3.24
UM	2.17	2.45	2.69	2.88	3.09	3.26
HMM	**2.56**	**2.70**	**3.07**	**3.23**	**3.41**	**3.50**

Table 4.8: PESQ improvement of enhanced speech by "car+phone" noise using different noise estimation methods.

Comparison with previous HMM

For the HMM noise estimation we have used in Sec. 3.5, we froze the HMM update during the speech activity and used Long-Term Average Speech Spectrum (LTASS) as our speech model to determine the noise state sequence. The performance of the average improvement of the quality of enhanced speech for different noise types are shown in Table 4.11, where “HMM-LTASS” denotes the noise model used in previous chapter and “HMM” denotes the noise model of our current proposed algorithm. It can be seen that for stationary noise, both methods achieved identical improvements in terms of segmental SNR and PESQ scores. For the non-stationary noise, the HMM-LTASS has states that already have included the different noise characteristics that will arise in the noisy speech, while the current HMM method does not have any prior knowledge thus has to identify the new arrivals of noise characteristics in the presence of speech. Despite the more challenging task, the HMM method performs nearly as well as HMM-LTASS method. There is an improvement in sSNR for the car+phone noise for the HMM method, since it has richer speech model, thus can better differentiate noise states in the presence of speech. For the machine gun noise, there is a significant improvement of sSNR but a decrease in PESQ scores. This is mainly because the HMM method did not correctly identify the silence intervals and the gun firing due to the similarity between the characteristics of proposed speech model and the machine gun noise.

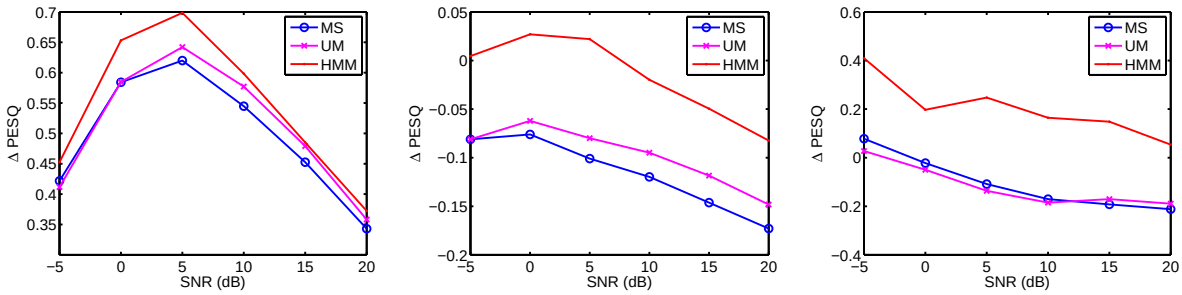


Figure 4.8: Improvement of PESQ scores at different SNRs for (a) white noise (b) gun noise (c) "car+phone" noise.

Δ sSNR	white	car	gun	phone	hammer
MS	6.64	5.78	-3.65	2.63	3.58
UM	6.52	5.61	-4.12	2.37	3.45
HMM	**7.39**	**8.30**	**0.48**	**5.60**	**4.40**

Table 4.9: mean segmental SNR improvement of enhanced speech using different noise estimation methods.

4.4.4 Listening Test

Given the results from previous two experiments, we conducted a further listening test to verify the performance of our proposed algorithm in comparison to other algorithm. A listening test was conducted as in Sec. 3.6.3, where a rating is assigned based on their preference: “1” if they prefer the HMM algorithm, “0” if they prefers the other algorithm, or “0.5” if they are indifferent. A group of 10 listeners participated in this listening test, the mean rating scores of each comparison with HMM for different noise types are shown in Tables 4.12. As compared to the original (unenhanced) noisy speech, the proposed HMM is preferred for all noise types except for the gun noise. When comparing to other enhancement methods, the HMM algorithm performs as well as UM and MS methods for stationary white and car noise, but is preferred for non-stationary phone and hammer noise. For the HMM-LTASS in Sec. 3.5, the proposed HMM method performs similarly in white, car and hammer noise, but gun noise is less preferred since it has a higher residual noise but with less distortion, and the phone noise is slightly preferred as it gives a better state assignment. These results correlate well with our previous results for the improvements in segmental SNRs and PESQ scores.

Δ PESQ	white	car	gun	phone	hammer
MS	0.49	0.04	-0.11	-0.10	0.25
UM	0.50	0.05	-0.10	-0.12	0.26
HMM	**0.54**	**0.16**	**0.01**	**0.20**	**0.43**

Table 4.10: mean PESQ Improvement of enhanced speech using different noise estimation methods.

Δ sSNR	white	car	gun	phone	hammer
HMM-LTASS	**7.39**	**8.30**	-6.92	4.94	**4.40**
HMM	**7.39**	**8.30**	**0.48**	**5.60**	**4.40**
Δ PESQ	white	car	gun	phone	hammer
HMM-LTASS	**0.54**	**0.16**	**0.11**	**0.20**	**0.43**
HMM	**0.54**	**0.16**	0.01	**0.20**	**0.43**

Table 4.11: mean segmental SNR improvement of enhanced speech corrupted by white noise using different noise estimation methods.

4.5 Summary

In this chapter we have extended the adaptive noise model that was introduced in Chapter 3 to enable it to track changing noise spectra and to create new noise states when needed even in the presence of speech. To make this possible, we have incorporated two additional components in order to create a composite production model for noisy speech. The new components are a level-normalized speech model and a speech level model. The separation of the speech model into two components allows the easy imposition of strong constraints on the rate of change of speech level; these were found to be essential for the identification of new noise sources during speech presence.

In order to track gradually changing noise spectra when speech is present, we identify the noise state corresponding to each time frame and update the corresponding noise spectra only in those frequency bins in which the noise is dominant. All the frequency bins in a state will therefore be updated over time, but only when the speech power spectrum falls below that of the noise.

By far the most challenging goal is the identification and modelling of new noise sources even when speech is present. We detect the presence of a new noise source when a low probability is assigned by our composite speech+noise model to the spectrum of the observed signal. Having detected a new noise source in this way, we create a new state

v.s. HMM	white	car	gun	phone	hammer
unenhanced	1.0	1.0	0.65	1.0	1.0
MS	0.5	0.5	0.65	1.0	1.0
UM	0.5	0.5	0.65	1.0	1.0
HMM-LTASS	0.5	0.5	0.4	0.6	0.5

Table 4.12: mean rating scores of enhanced speech signals using different noise estimation methods. A score of 1 indicates that the the enhancer using the HMM model from this chapter was preferred.

to represent it and initialize the state's mean power spectrum using a robust procedure that takes into account the possible presence of speech.

The adaptive noise modelling procedure has been evaluated on noise examples that are gradually changing and that are highly non-stationary. We have demonstrated that the algorithm is able to track both types of noise and also to detect new noise sources even when speech is present. However, even with the more sophisticated speech model used in this chapter, we have found that there are some circumstances in which speech is wrongly interpreted as noise resulting in an incorrect noise state sequence.

The algorithm has also been evaluated by incorporating it into a speech enhancement system where its performance was compared with two state-of-the-art noise estimators as well as the HMM-LTASS estimator from Sec. 3 which was trained on a noise-only signal in the absence of speech. Despite its more demanding task, the performance of the new estimator was almost identical to to that of the HMM-LTASS estimator and, for all noise types, it resulted in an average improvement in both segmental SNR and PESQ. Except for the PESQ improvement of the machine gun noise for HMM-LTASS, it was found that for all the tested noise types at all SNR levels the average improvement in both segmental SNR and PESQ was greater when using the new noise estimation algorithm than with either of the competing noise estimators.

Chapter 5

Summary and Conclusions

5.1 Summary and discussion

The aim of this thesis was to propose and investigate robust noise estimation methods for speech enhancement systems under adverse noisy environments. The thesis describes the successful development of a robust noise model that can recursively track both gradual and abrupt changes in the acoustic noise in a signal. In Chapter 3, we proposed the use of an HMM as a model for non-stationary noise in which each of the HMM states is associated with a distinct mean noise power spectrum. To cope with noise characteristics that change gradually over time, a procedure is described for adaptively updating each state's mean power spectrum without requiring the noise model to be completely retrained after each frame. The procedure includes a forgetting factor so that a higher weight is given to more recent frames. The updating procedure was then extended to detect the occurrence of a previously unseen noise power spectrum and, in response, to create a state representing the new noise source. In order to preserve the same total number of states, the procedure also merges together the two existing noise states that are closest to each other. The adaptation of the noise model is suspended whenever speech is present. By combining the model with a fixed LTASS model of speech, the maximum likelihood sequence of noise states is estimated during a speech

interval and the corresponding mean noise power spectra are used as the noise estimate for a speech enhancer. In Chapter 4, the model updating procedure is extended so that noise can be tracked and new noise states introduced where appropriate even during intervals when speech is present. To achieve this, an extended speech model was used which combined a pre-trained model of level-normalized speech together with a separate HMM representing the overall level of the speech. The factoring of the speech model in this way allowed long term temporal constraints to be placed on the speech level which were essential for reliably distinguishing between speech and noise. Both versions of the noise estimator were evaluated using an MMSE speech enhancement algorithm and it was found that the use of the multi-state HMM noise model resulted in consistent improvements in quality (as measured by PESQ) compared to conventional techniques that estimate only a single, quasi-stationary, noise power spectrum.

In summary, we have developed a noise HMM that can track and update fast-changing noise characteristics in a noisy speech signal without any prior training. The model parameters comprise the mean power in each state and the transition probability between states. The mean power within each noise state is only updated if the speech presence probability in individual frequency bin is low. A log-likelihood based measure is proposed to assess the goodness of fit of our existing model, such that a novel noise characteristic can be detected and a new state is created accordingly. In our experiments, we showed that the noise HMM is capable of robustly tracking both stationary and highly non-stationary noise, and that when it is incorporated into a standard speech enhancement algorithm, it gives a better performance, in terms of the enhanced speech quality improvement, than other state-of-the-art noise estimation methods.

5.2 Conclusion and Future Directions

In this thesis, robust noise estimation for speech enhancement was studied. We proposed the on-line adaptive noise HMM, which can effectively track any highly non-stationary noise even during speech activity. In the following some future work arising from this thesis is discussed.

The methods developed in this thesis give excellent results when no speech is present. Reliable identification of new noise sources when speech is present still, however, remains a challenge. In our model, the accurate estimation of the overall speech level is important for reliably distinguishing between the occurrence of a new noise source and an abrupt change in speech level. We therefore apply strong constraints to the rate at which our estimated speech level is permitted to change. Recent work [47] within our research group indicates that it is possible to obtain reliable estimations of the speech level even when the SNR is poor. Incorporating reliable external speech level estimation would potentially provide two benefits to our algorithm. First, the error in the estimated speech level would reduce and hence the accuracy of state assignment during speech presence would improve. Second, the algorithm would cope better with situations in which the true speech level changes rapidly because the constraints currently imposed by our algorithm would be removed.

As can be seen in Fig. 4.4, there are some occasions when, even though the estimated noise model is correct, our algorithm assigns incorrect noise states to some frames. These assignment errors have a serious effect on the resultant speech enhancement and arise because the model is not sufficiently able to distinguish between speech and noise. Drawing on research in speech recognition, it may be that incorporating delta coefficients in addition to static coefficients into the spectral models would improve the state assignment of the model.

Finally, other variations of HMM can be used for better noise estimation. For instance, in our HMM, we have assigned each noise state with a distinct characteristics, such that N different noise types will give 2^N different combinations, thus require 2^N states to fully describe the noise. A factorial HMM, with each state representing a distinct noise type, can be used to effectively reduce the number of states required. Noise can be estimated as any combination of N states, instead of a single state in our proposed model.

Bibliography

[1] Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms, November 2003.

[2] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, New York, 1972.

[3] J. Allen and L. Radiner. A unified approach to short-time Fourier analysis and synthesis. *Proc. IEEE*, 65(11):1558–1564, 1977.

[4] J. B. Allen. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. *IEEE Trans. Acoust., Speech, Signal Process.*, 25(3):235–238, June 1977.

[5] I. Andrianakis and PR White. Speech spectral amplitude estimators using optimally shaped gamma and chi priors. *Speech Communication*, 51(1):1–14, 2009.

[6] L. Arslan, A. McCree, and V. Viswanathan. New methods for adaptive noise suppression. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, 1995.

[7] H. Bai and H Wan. Two-pass quantile based noise spectrum estimation. *Center of Spoken Language Understanding, OGI School of Science and Engineering at OHSU*, 2003.

[8] T. Baldeweg. Repetition effects to sounds: evidence for predictive coding in the auditory system. *Trends in cognitive sciences*, 10(3):93–93, 2006.

[9] M. Berouti, R. Schwartz, and J. Makhoul. Enhancement of speech corrupted by acoustic noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 208–211, 1979.

[10] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-27(2):113–120, April 1979.

[11] D. M. Brookes. VOICEBOX: A speech processing toolbox for MATLAB. `http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html`, 1997.

[12] O. Cappe. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Trans. Speech Audio Process.*, 2(2):345–349, April 1994.

[13] B. Chen and P. C. Loizou. A Laplacian-based MMSE estimator for speech enhancement. *Speech Communication*, 49(2):134–143, February 2007.

[14] I. Cohen. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. *IEEE Signal Process. Lett.*, 9(4):113–116, April 2002.

[15] I. Cohen. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging. *IEEE Trans. Speech Audio Process.*, 11(5):466–475, September 2003.

[16] I. Cohen and B. Berdugo. Noise estimation by minima controlled recursive averaging for robust speech enhancement. *IEEE Signal Process. Lett.*, 9(1):12–15, January 2002.

[17] Israel Cohen and Baruch Berdugo. Speech enhancement for non-stationary noise environments. *Signal Processing*, 81(11):2403–2418, November 2001.

[18] Dirk Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3:151–167, 1989.

[19] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(4):357–366, August 1980.

[20] B. De Moor. The singular value decomposition and long and short spaces of noisy matrices. *IEEE Trans. Signal Process.*, 41(9):2826–2838, 1993.

[21] M. Dendrinos, S. Bakamidis, and G. Carayannis. Speech enhancement from noise: a regenerative approach. *Speech Communication*, 10(1):45–67, February 1991.

[22] Li Deng, J. Droppo, and A. Acero. Incremental Bayes learning with prior evolution for tracking nonstationary noise statistics from noisy speech data. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, April 2003.

[23] Li Deng, J. Droppo, and A. Acero. Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *IEEE Trans. Speech Audio Process.*, 11(6):568–580, November 2003.

[24] H. Ding, I.Y. Soon, S.N. Koh, and C.K. Yeo. A spectral filtering method based on hybrid wiener filters for speech enhancement. *Speech Communication*, 51(3):259–267, 2009.

[25] G. Doblinger. Computationally efficient speech enhancement by spectral minima tracking in subbands. In *Proc. European Conf. on Speech Communication and Technology*, pages 1513–1516, Madrid, September 1995.

[26] J. Egan. Articulation testing methods. *Laryngoscope*, 58(9):955–991, 1948.

[27] K. El-Maleh and P. Kabal. Comparison of voice activity detection algorithms for wirelesspersonal communications systems. In *IEEE Canadian Conference on Electrical and Computer Engineering*, volume 2, 1997.

[28] Y. Ephraim. Statistical-model-based speech enhancement systems. *Proc. IEEE*, 80(10):1526–1555, October 1992.

[29] Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 32(6):1109–1121, December 1984.

[30] Y. Ephraim and D. Malah. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Process.*, 33(2):443–445, 1985.

[31] Y. Ephraim, D. Malah, and B.-H. Juang. On the application of hidden Markov models for enhancing noisy speech. *IEEE Trans. Acoust., Speech, Signal Process.*, 37(12):1846–1856, December 1989.

[32] Y. Ephraim and H. L. Van Trees. A signal subspace approach for speech enhancement. *IEEE Trans. Speech Audio Process.*, 3(4):251–266, July 1995.

[33] J.S. Erkelens, R.C. Hendriks, R. Heusdens, and J. Jensen. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 15(6):1741–1752, 2007.

[34] H. Fletcher and J. Steinberg. Articulation testing methods. *Bell Syst. Tech. J.*, 8:806–854, 1929.

[35] D. K. Freeman, G. Cosier, C. B. Southcott, and I. Boyd. The voice activity detector for the pan-european digital cellular mobile telephone service. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 369–372, May 1989.

[36] Brendan J. Frey, Li Deng, Alex Acero, and Trausti Kristjansson. ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In *Proc. European Conf. on Speech Communication and Technology*, Aalborg, September 2001.

[37] Brendan J. Frey, Trausti T. Kristjansson, Li Deng, and Alex Acero. Learning dynamic noise models from noisy speech for robust speech recognition. In *Proc. Neural Information Processing Conf*, 2001.

[38] Z.H. Fu and J.F. Wang. Speech presence probability estimation based on integrated time-frequency minimum tracking for speech enhancement in adverse en-

vironments. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4258–4261. IEEE, 2010.

[39] M. Fujimoto, K. Ishizuka, and T. Nakatani. A voice activity detection based on the adaptive integration of multiple speech features and a signal decision scheme. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4441–4444. IEEE, 2008.

[40] M. J. F. Gales. *Model-based Techniques for Noise Robust Speech Recognition*. PhD thesis, Cambridge University, 1995.

[41] M. J. F. Gales and S. J. Young. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12:231–239, July 1993.

[42] M. J. F. Gales and S. J. Young. Robust speech recognition in additive and convolutional noise using parallel model combination. *Computer Speech and Language*, 9(4):289–307, October 1995.

[43] M. J. F. Gales and S. J. Young. Robust continuous speech recognition using parallel model combination. *IEEE Trans. Speech Audio Process.*, 4:352–359, September 1996.

[44] S. Gazor and Wei Zhang. A soft voice activity detector based on a Laplacian–Gaussian model. *IEEE Trans. Speech Audio Process.*, 11(5):498–505, September 2003.

[45] S. Gazor and Wei Zhang. Speech probability distribution. *IEEE Signal Process. Lett.*, 10(7):204–207, July 2003.

[46] T. Gerkmann and R.C. Hendriks. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. *IEEE Trans. Audio, Speech, Lang. Process.*, 20(4):1383 –1393, May 2012.

[47] Sira Gonzalez and Mike Brookes. Speech active level estimation in noisy conditions. In *Internal technical report, Imperial College London*, October 2012.

[48] I.J. Good. How to estimate probabilities. *IMA Journal of Applied Mathematics*, 2(4):364–383, 1966.

[49] M. Graciarena and H. Franco. Unsupervised noise model estimation for model-based robust speech recognition. In *Proc. ASRU IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 351–356, December 2003.

[50] V. Grancharov, J. Samuelsson, and B. Kleijn. On causal algorithms for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(3):764–773, May 2006.

[51] V. Grancharov, D. Y. Zhao, J. Lindblom, and W. B. Kleijn. Low-complexity, non-intrusive speech quality assessment. *IEEE Trans. Audio, Speech, Lang. Process.*, 14(6):1948–1956, November 2006.

[52] A. H. Gray, Jr. and J. D. Markel. Distance measures for speech processing. *IEEE Trans. Acoust., Speech, Signal Process.*, 24(5):380–391, October 1976.

[53] R. Gray, A. Buzo, A. Gray Jr, and Y. Matsuyama. Distortion measures for speech processing. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(4):367–376, 1980.

[54] R. Gray, A. Gray, G. Rebolledo, and J. Shore. Rate-distortion speech coding with a minimum discrimination information distortion measure. *IEEE Trans. Inf. Theory*, 27(6):708–721, November 1981.

[55] K. Grill-Spector, R. Henson, and A. Martin. Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, 10(1):14–23, 2006.

[56] S. Gustafsson, S. Nordholm, and I. Claesson. Spectral subtraction using reduced delay convolution and adaptive averaging. *IEEE Trans. Speech Audio Process.*, 9(8):799–807, November 2001.

[57] J. Hao, H. Attias, S. Nagarajan, T.W. Lee, and T.J. Sejnowski. Speech enhancement, gain, and noise spectrum adaptation using approximate bayesian estimation. *IEEE Trans. Audio, Speech, Lang. Process.*, 17(1):24–37, 2009.

[58] R.C. Hendriks, R. Heusdens, and J. Jensen. MMSE based noise PSD tracking with low complexity. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4266–4269, March 2010.

[59] R.C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems. Low complexity DFT-domain noise psd tracking using high-resolution periodograms. *EURASIP Journal on Applied Signal Processing*, 2009:55, 2009.

[60] R.C. Hendriks, J. Jensen, and R. Heusdens. Noise tracking using dft domain subspace decompositions. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(3):541–553, 2008.

[61] H. G. Hirsch and C. Ehrlicher. Noise estimation techniques for robust speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 153–156, 1995.

[62] Hans-Gunter Hirsch. Estimation of noise spectrum and its application to SNR-estimation and speech enhancement. Technical Report TR-93–012, ICSI Berkeley, Berkeley, 1993.

[63] Hans-Gunter Hirsch and David Pearce. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. ISCA Workshop on Automatic Speech Recognition*, pages 181–188, Paris, September 2000.

[64] Y. Hu and P. C. Loizou. A generalized subspace approach for enhancing speech corrupted by colored noise. *IEEE Trans. Speech Audio Process.*, 11(4):334–341, July 2003.

[65] Y. Hu and P. C. Loizou. Subjective comparison of speech enhancement algorithms. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 153–156, May 2006.

[66] Yi Hu and P. C. Loizou. A subspace approach for enhancing speech corrupted by colored noise. *IEEE Signal Process. Lett.*, 9(7):204–206, July 2002.

[67] Yi Hu and P. C. Loizou. Subjective comparison and evaluation of speech enhancement algorithms. *Speech Communication*, 49(7–8):588–601, July 2007.

[68] ITU-T. Objective measurement of active speech level, March 1993.

[69] ITU-T. Artificial voices, September 1999.

[70] ITU-T. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, February 2001.

[71] B. H. Juang and L. R. Rabiner. Mixture autoregressive hidden Markov models for speech signals. *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-33(6):1404–1413, December 1985.

[72] D. Kalikow, K. Stevens, and L. Elliott. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J. Acoust. Soc. Am.*, 61(5):1337–1351, 1977.

[73] S. Kamath and P. Loizou. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 4164–4164. Citeseer, 2002.

[74] S.M. Kay. *Fundamentals of statistical signal processing: estimation theory*. 1993.

[75] M.G. Kendall and A. Stuart. *The Advanced Theory of Statistics*, volume 1. Charles Griffin, 1977.

[76] V. Krishnamurthy and J. B. Moore. On-line estimation of hidden Markov model parameters based on the Kullback-Leibler information measure. *IEEE Trans. Signal Process.*, 41(8):2557–2573, 1993.

[77] T. Kristjansson, B. Frey, L. Deng, and A. Acero. Towards non-stationary model-based noise adaptation for large vocabulary speech recognition. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 337–340, May 2001.

[78] J.M. Kum and J.H. Chang. Speech enhancement based on minima controlled recursive averaging incorporating second-order conditional map criterion. *Signal Processing Letters, IEEE*, 16(7):624–627, 2009.

[79] J.M. Kum, Y.S. Park, and J.H. Chang. Speech enhancement based on minima controlled recursive averaging incorporating conditional maximum a posteriori criterion. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4417–4420. IEEE, 2009.

[80] Te-Won Lee and Kaisheng Yao. Speech enhancement by perceptual filter with sequential noise parameter estimation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 693–696, May 2004.

[81] J. S. Lim and A. V. Oppenheim. All-pole modeling of degraded speech. *IEEE Trans. Acoust., Speech, Signal Process.*, 26(3):197–210, June 1978.

[82] L. Lin, W. H. Holmes, and E. Ambikairajah. Adaptive noise estimation algorithm for speech enhancement. *IEE Electronics Lett.*, 39(9):754–755, 2003.

[83] L. Lin, W. H. Holmes, and E. Ambikairajah. Subband noise estimation for speech enhancement using a perceptual Wiener filter. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 80–83, 2003.

[84] P. Lockwood and J. Boudy. Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust recognition in cars. *Speech Communication*, 11:215–228, June 1992.

[85] P. C. Loizou. *Speech Enhancement Theory and Practice*. Taylor & Francis, 2007.

[86] Y. Lu and P.C. Loizou. A geometric approach to spectral subtraction. *Speech communication*, 50(6):453–466, 2008.

[87] J. Makhoul and L. Cosell. LPCW: An LPC vocoder with linear predictive spectral warping. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 466–469. IEEE, April 1976.

[88] D. Malah, R. V. Cox, and A. J. Accardi. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 789–792, March 1999.

[89] Kotta Manohar and Preeti Rao. Speech enhancement in nonstationary noise environments using noise properties. *Speech Communication*, 48(1):96–109, January 2006.

[90] R. Martin. Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Trans. Speech Audio Process.*, 9:504–512, July 2001.

[91] R. Martin. Speech enhancement based on minimum mean-square error estimation and supergaussian priors. *IEEE Trans. Speech Audio Process.*, 13(5):845–856, September 2005.

[92] R. Martin. Bias compensation methods for minimum statistics noise power spectral density estimation. *Signal Processing*, 86(6):1215–1229, June 2006.

[93] Rainer Martin. Spectral subtraction based on minimum statistics. In *Proc. European Signal Processing Conf*, pages 1182–1185, 1994.

[94] R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Trans. Acoust., Speech, Signal Process.*, 28(2):137–145, April 1980.

[95] Y.S. Park and J.H. Chang. A probabilistic combination method of minimum statistics and soft decision for robust noise power estimation in speech enhancement. *Research letters in signal processing*, 15:95–98, 2008.

[96] VF Pisarenko. The retrieval of harmonics from a covariance function. *Geophysical Journal of the Royal Astronomical Society*, 33(3):347–366, 1973.

[97] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, February 1989.

[98] L. Rabiner and B. Juang. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, January 1986.

[99] L. Rabiner and M. Sambur. Application of an LPC distance measure to the voiced-unvoiced-silence detection problem. *IEEE Trans. Acoust., Speech, Signal Process.*, 25:338–343, August 1977.

[100] S. Rangachari and P. C. Loizou. A noise-estimation algorithm for highly non-stationary environments. *Speech Communication*, 48(2):220–231, February 2006.

[101] S. Rangachari, P. C. Loizou, and Y. Hu. A noise estimation algorithm with rapid adaptation for highly nonstationary environments. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 305–308, May 2004.

[102] S. Rennie, T. Kristjansson, P. Olsen, and R. Gopinath. Dynamic noise adaptation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, May 2006.

[103] A. Rezayee and S. Gazor. An adaptive KLT approach for speech enhancement. *IEEE Trans. Speech Audio Process.*, 9(2):87–95, 2001.

[104] Christophe Ris and Stephane Dupont. Assessing local noise level estimation methods: Application to noise robust ASR. *Speech Communication*, 34(1–2):141–158, April 2001.

[105] A. Rix, J. Beerends, M. Hollier, and A. Hekstra. Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 749–752, 2001.

[106] E. H. Rothauser, W. D. Chapman, N. Guttman, M. H. L. Hecker, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock. IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust.*, 17(3):225–246, 1969.

[107] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan. HMM-based strategies for enhancement of speech signals embedded in nonstationary noise. *IEEE Trans. Speech Audio Process.*, 6(5):445–455, September 1998.

[108] M. Schwab, H.G. Kim, P. Noll, et al. Robust noise estimation applied to different speech estimators. In *Proc. Asilomar Conf. on Signals, Systems and Computers*, volume 2, pages 1904–1907. IEEE, 2003.

[109] J.W. Shin, J.H. Chang, and N.S. Kim. Voice activity detection based on statistical models and machine learning approaches. *Computer Speech & Language*, 24(3):515–530, 2010.

[110] J. Shore. Minimum cross-entropy spectral analysis. *IEEE Trans. Acoust., Speech, Signal Process.*, 29(2):230–237, April 1981.

[111] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. A statistical model-based voice activity detection. *IEEE Signal Process. Lett.*, 6(1):1–3, 1999.

[112] Jongseo Sohn and Wonyong Sung. A voice activity detector employing soft decision based noise spectrum adaptation. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 365–368, May 1998.

[113] S. Srinivasan, J. Samuelsson, and W.B. Kleijn. Codebook-based bayesian speech enhancement. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 1077–1080. Citeseer, 2005.

[114] S. Srinivasan, J. Samuelsson, and W.B. Kleijn. Codebook-based bayesian speech enhancement for nonstationary environments. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(2):441–452, 2007.

[115] V. Stahl, A. Fischer, and R. Bippus. Quantile based noise estimation for spectral subtraction and Wiener filtering. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 1875–1878, 2000.

[116] S.S. Stevens and J. Volkmann. The relation of pitch to frequency: A revised scale. *The American Journal of Psychology*, 53(3):329–353, 1940.

[117] Amarnag Subramanya, Zhengyou Zhang, Zicheng Liu, and Alex Acero. Speech modeling with magnitude-normalized complex spectra and its application to multisensory speech enhancement. Technical Report MSR-TR-2005–126, Microsoft Reseach, 2005.

[118] B. Frey T. Kristjansson and L. Deng. Joint estimation of noise and channel distortion in a generalized EM framework. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, Madonna di Campiglio, December 2001.

[119] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objecitve intelligibility measure for time-frequency weighted noisy speech. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4214–4217, 2010.

[120] J. Taghia, N. Mohammadiha, Jinqiu Sang, V. Bouse, and R. Martin. An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4640 –4643, May 2011.

[121] S. G. Tanyer and H. Ozer. Voice activity detection in nonstationary noise. *IEEE Trans. Speech Audio Process.*, 8(4):478–482, July 2000.

[122] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere. A study of complexity and quality of speech waveform coders. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, 1978.

[123] A. P. Varga and R. K. Moore. Hidden Markov model decomposition of speech and noise. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 845–848, April 1990.

[124] Andrew Varga and Herman J. M. Steeneken. 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*, 3(3):247–251, July 1993.

[125] Nathalie Virag. Single channel speech enhancement based on masking properties of the human auditory system. *IEEE Trans. Speech Audio Process.*, 7(2):126–137, March 1999.

[126] W. D. Voiers. Evaluating processed speech using the diagnostic rhyme test. *Speech Technology*, 1(4):30–39, 1983.

[127] S. Wang, A. Sekey, and A. Gersho. An objective measure for predicting subjective quality of speech coders. *IEEE J. Sel. Areas Commun.*, 10(5):819–829, 1992.

[128] N. Wiener. *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. John Wiley & Sons, Inc., New York, NY, USA, 1949.

[129] K. Yao and S. Nakamura. Sequential noise compensation by sequential Monte Carlo method. In *Advances in Neural Information Processing Systems*, volume 14, pages 1213–1220, 2002.

[130] Kaisheng Yao and Te-Won Lee. Speech enhancement with noise parameter estimated by a sequential Monte Carlo method. In *IEEE Workshop on Statistical Signal Processing*, pages 609–612, October 2003.

[131] D. Ying, Y. Yan, J. Dang, and F.K. Soong. Noise power estimation based on a sequential gaussian mixture model. In *Image and Signal Processing (CISP)*, volume 5, pages 2362–2365. IEEE, 2011.

[132] Wei Zhang and S. Gazor. Statistical modelling of speech signals. In *Proc. Intl. Conf. on Signal Processing*, volume 1, pages 480–483, August 2002.

[133] D. Y. Zhao, W. B. Kleijn, A. Ypma, and B. de Vries. Online noise estimation using stochastic-gain HMM for speech enhancement. *IEEE Trans. Audio, Speech, Lang. Process.*, 16(4):835–846, 2008.

[134] David Y. Zhao and W. Bastiaan Kleijn. HMM-based gain modeling for enhancement of speech in noise. *IEEE Trans. Audio, Speech, Lang. Process.*, 15(3):882–892, 2007.