Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 282–305

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator

Kuldip Paliwal, Belinda Schwerin *, Kamil Wójcicki

*Signal Processing Laboratory, Griffith School of Engineering, Griffith University, Nathan, QLD 4111, Australia*

Received 15 December 2010; received in revised form 7 September 2011; accepted 14 September 2011
Available online 24 September 2011

**Abstract**

In this paper we investigate the enhancement of speech by applying MMSE short-time spectral magnitude estimation in the modulation domain. For this purpose, the traditional analysis-modification-synthesis framework is extended to include modulation domain processing. We compensate the noisy modulation spectrum for additive noise distortion by applying the MMSE short-time spectral magnitude estimation algorithm in the modulation domain. A number of subjective experiments were conducted. Initially, we determine the parameter values that maximise the subjective quality of stimuli enhanced using the MMSE modulation magnitude estimator. Next, we compare the quality of stimuli processed by the MMSE modulation magnitude estimator to those processed using the MMSE acoustic magnitude estimator and the modulation spectral subtraction method, and show that good improvement in speech quality is achieved through use of the proposed approach. Then we evaluate the effect of including speech presence uncertainty and log-domain processing on the quality of enhanced speech, and find that this method works better with speech uncertainty. Finally we compare the quality of speech enhanced using the MMSE modulation magnitude estimator (when used with speech presence uncertainty) with that enhanced using different acoustic domain MMSE magnitude estimator formulations, and those enhanced using different modulation domain based enhancement algorithms. Results of these tests show that the MMSE modulation magnitude estimator improves the quality of processed stimuli, without introducing musical noise or spectral smearing distortion. The proposed method is shown to have better noise suppression than MMSE acoustic magnitude estimation, and improved speech quality compared to other modulation domain based enhancement methods considered.


*Keywords:* Modulation domain; Analysis-modification-synthesis (AMS); Speech enhancement; MMSE short-time spectral magnitude estimator (AME); Modulation spectrum; Modulation magnitude spectrum; MMSE short-time modulation magnitude estimator (MME)

## 1. Introduction

Speech enhancement methods aim to improve the quality of noisy speech by reducing noise, while at the same time minimising any speech distortion introduced by the enhancement process. Many enhancement methods are based on the short-time Fourier analysis-modification-synthesis framework. Some examples of these are the spectral subtraction method (Boll, 1979), the Wiener filter method (Wiener, 1949), and the MMSE short-time spectral amplitude estimation method (Ephraim and Malah, 1984).

Spectral subtraction is perhaps one of the earliest and most extensively studied methods for speech enhancement. This simple method enhances speech by subtracting a spectral estimate of noise from the noisy speech spectrum in either the magnitude or energy domain. Though this method is effective at reducing noise, it suffers from the problem of musical noise distortion, which is very annoying to listeners. To overcome this problem, Ephraim and Malah (1984) proposed the MMSE short-time spectral amplitude estimator, referred to throughout this work as the acoustic magnitude estimator (AME). In the literature (e.g., Cappe, 1994; Scalart and Filho, 1996), it has been suggested that the good performance of the AME can be largely attributed to the use of the decision-directed

* Corresponding author. Tel.: +61 7 3735 3754; fax: +61 7 3735 5198.
*E-mail address:* belinda.schwerin@griffithuni.edu.au (B. Schwerin).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.09.003

approach for estimation of the *a priori* signal-to-noise ratio (*a priori* SNR). The AME method, even today, remains one of the most effective and popular methods for speech enhancement.

Recently, the modulation domain has become popular for speech processing. This has been in part due to the strong psychoacoustic and physiological evidence, which supports the significance of the modulation domain for the analysis of speech signals.[1] Zadeh (1950) was perhaps the first to propose a two-dimensional bi-frequency system, where the second dimension for frequency analysis was the transform of the time variation of the magnitudes at each standard (acoustic) frequency. Atlas et al. (2004) more recently defines the acoustic frequency as the axis of the first short-time Fourier transform (STFT) of the input signal and the modulation frequency as the independent variable of the second STFT transform.

Early efforts to utilise the modulation domain for speech enhancement assumed speech and noise to be stationary, and applied fixed filtering on the trajectories of the acoustic magnitude spectrum. For example, Hermansky et al. (1995) proposed band-pass filtering the time trajectories of the cubic-root compressed short-time power spectrum to enhance speech. Falk et al. (2007) and Lyons and Paliwal (2008) applied similar band-pass filtering to the time trajectories of the short-time magnitude (power) spectrum for speech enhancement.

However, speech and possibly noise are known to be nonstationary. To capture this nonstationarity, one option is to assume speech to be quasi-stationary, and process the trajectories of the acoustic magnitude spectrum on a short-time basis. At this point it is useful to differentiate the acoustic spectrum from the modulation spectrum as follows. The acoustic spectrum is the STFT of the speech signal, while the modulation spectrum at a given acoustic frequency is the STFT of the time series of the acoustic spectral magnitudes at that frequency. The short-time modulation spectrum is thus a function of time, acoustic frequency and modulation frequency.

This type of short-time processing in the modulation domain has been used in the past for automatic speech recognition (ASR). Kingsbury et al. (1998), for example, applied a modulation spectrogram representation that emphasised low-frequency amplitude modulations to ASR for improved robustness in noisy and reverberant conditions. Tyagi et al. (2003) applied mel-cepstrum modulation features to ASR to give improved performance in the presence of non-stationary noise. Short-time modulation domain processing has also been applied to objective quality. For example, Kim and Oct (2004, 2005) as well as Falk and Chan (2008) used the short-time modulation magnitude spectrum to derive objective measures that characterise the quality of processed speech.

[1] A review of the significance of the modulation domain for human speech perception can be found in (Atlas and Shamma, 2003).

For speech enhancement, short-time modulation domain processing was recently applied in the modulation spectral subtraction method (ModSSub) of Paliwal et al. (2010). Here, the spectral subtraction method was extended to the modulation domain, enhancing speech by subtracting the noise modulation energy spectrum from the noisy modulation energy spectrum in an analysis-modification-synthesis (AMS) framework. In ModSSub method, the frame duration used for computing the short-time modulation spectrum was found to be an important parameter, providing a trade-off between quality and level of musical noise. Increasing the frame duration reduced musical noise, but introduced a slurring distortion. A somewhat long frame duration of 256 ms was recommended as a good compromise. The disadvantages of using longer modulation domain analysis window are as follows. Firstly, we are assuming stationarity which we know is not the case. Secondly, quite a long portion is needed for the initial estimation of noise, and thirdly, as shown by Paliwal et al. (2011), speech quality and intelligibility is higher when the modulation magnitude spectrum is processed using short frame durations and lower when processed using longer frame durations. For these reasons, we aim to find a method better suited to the use of shorter modulation analysis window durations.

Since the AME method has been found to be more effective than spectral subtraction in the acoustic domain, in this paper, we explore the effectiveness of this method in the short-time modulation domain. For this purpose, the traditional analysis-modification-synthesis framework is extended to include modulation domain processing, then the noisy modulation spectrum is compensated for additive noise distortion by applying the MMSE short-time spectral magnitude estimation algorithm. The advantage of applying a MMSE-based method is that it does not introduce musical noise and hence can be used with shorter frame durations in the modulation domain. The proposed approach, referred to as the modulation magnitude estimator (MME), is demonstrated to give better noise removal than the AME approach, without the musical noise of the spectral subtraction type approach, or the spectral smearing of the ModSSub method. In the body of this paper, we provide enhancement results for the case of speech corrupted by additive white Gaussian noise (AWGN). We have also investigated enhancement performance for various coloured noises and the results, included in the Appendices, are shown to be qualitatively similar.

The rest of the paper is organised as follows. Section 2 details an AMS-based framework for enhancement in the short-time modulation domain. In Section 3 we describe the proposed MME approach, then in Section 4 we give details of the experiments used to tune the parameters of the MME method. In Section 5, the performance of the MME method is evaluated by comparison to a number of different speech enhancement approaches. In Section 6, we consider the effect of speech presence uncertainty and log-domain processing on the performance of the MME

method. In Sections 7 and 8, we compare the quality of the proposed MME method to a wider range of enhancement methods, including different acoustic domain MMSE formulations and a number of modulation domain based speech enhancement methods. Final conclusions are drawn in Section 9.

## 2. AMS-based framework for speech enhancement in the short-time spectral modulation domain

As mentioned previously, many frequency domain speech enhancement methods are based on the (acoustic) short-time Fourier AMS framework (e.g., Lim and Oppenheim, 1979; Berouti et al., 1979; Ephraim and Malah, 1984; Ephraim and Malah, 1985; Martin, 1994; Sim et al., 1998; Virag, 1999; Cohen, 2005; Loizou, 2005). A traditional acoustic AMS procedure for speech enhancement consists of three stages: (1) the analysis stage, where the noisy speech is processed using the STFT analysis; (2) the modification stage, where the noisy spectrum is compensated for noise distortion to produce the modified spectrum; and (3) the synthesis stage, where an inverse STFT operation is followed by overlap-add synthesis to reconstruct the enhanced signal. The above framework has recently been extended to facilitate enhancement in the short-time spectral modulation domain (Paliwal et al., 2010). For this purpose, a secondary AMS procedure was utilized for framewise processing of the time series of each frequency component of the acoustic magnitude spectra. In this section, the details of the AMS-based framework for speech enhancement in the short-time spectral modulation domain are briefly reviewed.

Let us assume an additive noise model in which clean speech is corrupted by uncorrelated additive noise to produce noisy speech as given by

$$x(n) = s(n) + d(n), \tag{1}$$

where $x(n)$, $s(n)$, and $d(n)$ are the noisy speech, clean speech, and noise signals, respectively, and $n$ denotes a discrete-time index. The noisy speech signal is then processed using the running STFT analysis (Vary and Martin, 2006) given by

$$X_l(k) = \sum_{n=0}^{N-1} x(n + lZ)v(n)e^{-j2\pi nk/N}, \tag{2}$$

where $l$ refers to the acoustic frame index, $k$ refers to the index of the acoustic frequency, $N$ is the acoustic frame duration[2] (AFD) in samples, $Z$ is the acoustic frame shift (AFS) in samples, and $v(n)$ is the acoustic analysis window function. In speech processing, an AFD of 20–40 ms along with an AFS of 10–20 ms and the Hamming analysis window are typically employed (e.g., Picone, 1993; Huang et al., 2001; Loizou, 2007; Paliwal and Wójcicki, 2008; Rabiner and Schafer, 2010).

In polar form, the STFT of the speech signal can be expressed as

$$X_l(k) = |X_l(k)|e^{j\angle X_l(k)}, \tag{3}$$

where $|X_l(k)|$ denotes the acoustic magnitude spectrum and $\angle X_l(k)$ denotes the acoustic phase spectrum. The time trajectories for each frequency component of the acoustic magnitude spectra are then processed framewise using a second AMS procedure as outlined below. The running STFT is used to compute the modulation spectrum from the acoustic magnitude spectrum as follows

$$\mathcal{X}_\ell(k,m) = \sum_{l=0}^{\mathcal{N}-1} |X_{l+\ell\mathcal{Z}}(k)|u(l)e^{-j2\pi lm/\mathcal{N}}, \tag{4}$$

where $\ell$ is the modulation frame index, $k$ is the index of the acoustic frequency, $m$ refers to the index of the modulation frequency, $\mathcal{N}$ is the modulation frame duration (MFD) in terms of acoustic frames, $\mathcal{Z}$ is the modulation frame shift (MFS) in terms of acoustic frames, and $u(l)$ is the modulation analysis window function. The modulation spectrum can be written in polar form as

$$\mathcal{X}_\ell(k,m) = |\mathcal{X}_\ell(k,m)|e^{j\angle\mathcal{X}_\ell(k,m)}, \tag{5}$$

where $|\mathcal{X}_\ell(k,m)|$ is the modulation magnitude spectrum, and $\angle\mathcal{X}_\ell(k,m)$ is the modulation phase spectrum. In the present work, the modulation magnitude spectrum of clean speech is estimated from the noisy modulation magnitude spectrum, while the noisy modulation phase spectrum is left unchanged.[3] The modified modulation spectrum is then given by

$$\mathcal{Y}_\ell(k,m) = \left|\widehat{\mathcal{S}}_\ell(k,m)\right|e^{j\angle\mathcal{X}_\ell(k,m)}, \tag{6}$$

where $\left|\widehat{\mathcal{S}}_\ell(k,m)\right|$ is an estimate of the clean modulation magnitude spectrum. Eq. (6) can also be written in terms of spectral gain function, $\mathcal{G}_\ell(k,m)$, applied to the modulation spectrum of noisy speech as follows

$$\mathcal{Y}_\ell(k,m) = \mathcal{G}_\ell(k,m)\mathcal{X}_\ell(k,m), \tag{7}$$

where

$$\mathcal{G}_\ell(k,m) = \frac{\left|\widehat{\mathcal{S}}_\ell(k,m)\right|}{|\mathcal{X}_\ell(k,m)|}. \tag{8}$$

The inverse STFT operation, followed by least-squares overlap-add synthesis (Quatieri, 2002), are then used to compute the modified acoustic magnitude spectrum as given by

[2] Note that frame duration and window duration mean the same thing and we use these two terms interchangeably in this paper.

[3] The relative importance of the modulation phase spectrum with respect to the modulation magnitude spectrum depends on the MFD. For example, the results of a recent study by Paliwal et al. (2011) suggest that for short MFDs ($\leqslant$64 ms) the modulation phase spectrum does not significantly contribute towards speech intelligibility or quality.

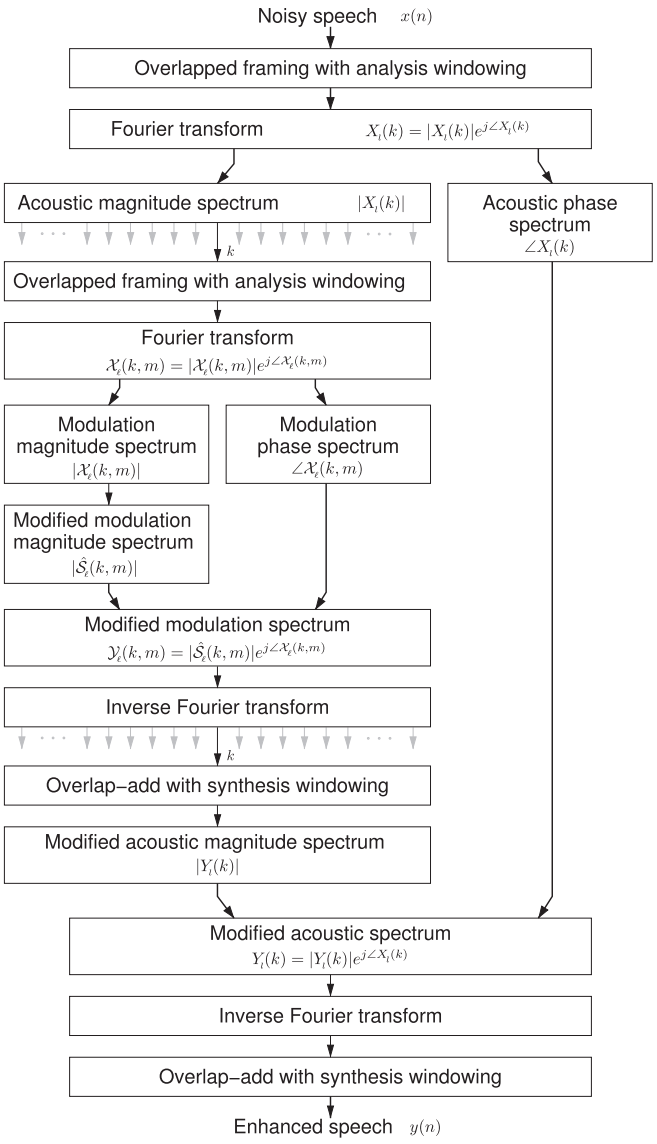


Fig. 1. Block diagram of the AMS–based framework for speech enhancement in the short-time spectral modulation domain.

$$|Y_l(k)| = \sum_{\ell}\left\{w(l-\ell\mathcal{Z})\sum_{m=0}^{\mathcal{N}-1}\mathcal{Y}_\ell(k,m)e^{j2\pi(l-\ell\mathcal{Z})m/\mathcal{N}}\right\}, \quad (9)$$

where $w(l)$ is a synthesis window function. The modified acoustic magnitude spectrum is combined with the noisy acoustic phase spectrum,[4] to produce the modified acoustic spectrum as follows

$$Y_l(k) = |Y_l(k)|e^{j\angle X_l(k)}. \quad (10)$$

The enhanced speech signal is constructed by applying the inverse STFT operation, followed by least-squares overlap-add synthesis, to the modified acoustic spectrum as given by

$$y(n) = \sum_{l}\left\{w(n-lZ)\sum_{k=0}^{N-1}Y_l(k)e^{j2\pi(n-lZ)k/N}\right\}. \quad (11)$$

A block diagram of the AMS-based framework for speech enhancement in the short-time spectral modulation domain is shown in Fig. 1.

## 3. Minimum mean-square error short-time spectral modulation magnitude estimator

The minimum mean-square error short-time spectral amplitude estimator of Ephraim and Malah (1984) has been employed in the past for speech enhancement in the acoustic frequency domain with much success. In the present work we investigate its use in the short-time spectral modulation domain. For this purpose the AMS-based framework detailed in Section 2 is used. In the following discussions we will refer to the original method by Ephraim and Malah (1984) as the MMSE acoustic magnitude estimator (AME), while the proposed modulation domain approach will be referred to as the MMSE modulation magnitude estimator (MME). The details of the MME are presented in the remainder of this section.

In the MME method, the modulation magnitude spectrum of clean speech is estimated from noisy observations. The proposed estimator minimises the mean-square error between the modulation magnitude spectra of clean and estimated speech

$$\epsilon = \mathrm{E}\left[\left(|\mathcal{S}_\ell(k,m)| - \left|\widehat{\mathcal{S}}_\ell(k,m)\right|\right)^2\right], \quad (12)$$

where E[·] denotes the expectation operator. Closed form solution to this problem in the acoustic spectral domain has been reported by Ephraim and Malah (1984) under the assumptions that speech and noise are additive in the time domain, and that their individual short-time spectral components are statistically independent, identically distributed, zero-mean Gaussian random variables. In the present work we make similar assumptions, namely that (1) speech and noise are additive in the short-time acoustic spectral magnitude domain, i.e.,

$$|X_l(k)| = |S_l(k)| + |D_l(k)| \quad (13)$$

and (2) the individual short-time modulation spectral components of $\mathcal{S}_\ell(k,m)$ and $\mathcal{D}_\ell(k,m)$ are independent, identically distributed Gaussian random variables.

The reasoning for the first assumption is that at high SNRs the phase spectrum remains largely unchanged by additive noise distortion (Loizou, 2007). For the second assumption, we can apply an argument similar to that of Ephraim and Malah (1984), where the central limit theorem is used to justify the statistical independence of spectral components of the Fourier transform. For the STFT, this assumption is valid only in the asymptotic sense, that

[4] Typically, AMS-based speech enhancement methods modify only the acoustic magnitude spectrum while keeping the acoustic phase spectrum unchanged. One reason for this is that for Hamming-windowed frames of 20–40 ms duration, the phase spectrum is considered unimportant for speech enhancement (e.g., Wang and Lim, 1982; Shannon and Paliwal, 2006).

is, when the frame duration is large. However, Ephraim and Malah have used an acoustic frame duration of 32 ms in their formulation to get good results. In our use of the MMSE approach in the modulation domain, we should also make the modulation frame duration to be as large as possible, however it must not be so large as to be adversely affected by the nonstationarity of the magnitude spectral sequence as mentioned in the introduction. Keeping Ephraim and Malah's 32 ms acoustic frame duration in mind, we want to find a compromise between these two competing requirements. For this, we investigate in this paper the performance of our method as a function of modulation frame duration.

With the above assumptions in mind, the modulation magnitude spectrum of clean speech can be estimated from the noisy modulation spectrum under the MMSE criterion (following Ephraim and Malah, 1984) as

$$\left|\widehat{\mathcal{S}}_\ell(k,m)\right| = \mathrm{E}[|\mathcal{S}_\ell(k,m)||\mathcal{X}_\ell(k,m)] \tag{14}$$

$$= \mathcal{G}_\ell(k,m)|\mathcal{X}_\ell(k,m)| \tag{15}$$

where $\mathcal{G}_\ell(k,m)$ is the MMSE-MME spectral gain function given by

$$\mathcal{G}_\ell(k,m) = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_\ell(k,m)}}{\gamma_\ell(k,m)}\Lambda[v_\ell(k,m)], \tag{16}$$

in which $v_\ell(k,m)$ is defined as

$$v_\ell(k,m) \triangleq \frac{\xi_\ell(k,m)}{1+\xi_\ell(k,m)}\gamma_\ell(k,m) \tag{17}$$

and $\Lambda[\cdot]$ is the following function

$$\Lambda[\theta] = \exp\left(-\frac{\theta}{2}\right)\left[(1+\theta)I_0\left(\frac{\theta}{2}\right) + \theta I_1\left(\frac{\theta}{2}\right)\right], \tag{18}$$

where $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively. In the above equations $\xi_\ell(k,m)$ and $\gamma_\ell(k,m)$ are interpreted (after McAulay and Malpass, 1980) as the *a priori* SNR, and the *a posteriori* SNR. These quantities are defined as

$$\xi_\ell(k,m) \triangleq \frac{\mathrm{E}[|\mathcal{S}_\ell(k,m)|^2]}{\mathrm{E}[|\mathcal{D}_\ell(k,m)|^2]} \tag{19}$$

and

$$\gamma_\ell(k,m) \triangleq \frac{|\mathcal{X}_\ell(k,m)|^2}{\mathrm{E}[|\mathcal{D}_\ell(k,m)|^2]}, \tag{20}$$

respectively.

Since in practice only noisy speech is observable, the $\xi_\ell(k,m)$ and $\gamma_\ell(k,m)$ parameters have to be estimated. For this task we apply the decision-directed approach (Ephraim and Malah, 1984) in the short-time spectral modulation domain. In the decision-directed method the *a priori* SNR is estimated by recursive averaging as follows

$$\hat{\xi}_\ell(k,m) = \alpha\frac{\left|\widehat{\mathcal{S}}_{\ell-1}(k,m)\right|^2}{\hat{\lambda}_{\ell-1}(k,m)} + (1-\alpha)\max[\hat{\gamma}_\ell(k,m)-1,0], \tag{21}$$

where $\alpha$ controls the trade-off between noise reduction and transient distortion (Cappe, 1994; Ephraim and Malah, 1984), $\hat{\lambda}_\ell(k,m)$ is an estimate of $\lambda_\ell(k,m) \triangleq \mathrm{E}[|\mathcal{D}_\ell(k,m)|^2]$, and the *a posteriori* SNR estimate is obtained by

$$\hat{\gamma}_\ell(k,m) = \frac{|\mathcal{X}_\ell(k,m)|^2}{\hat{\lambda}_\ell(k,m)}. \tag{22}$$

Note that limiting the minimum value of the *a priori* SNR has a considerable effect on the nature of the residual noise (Ephraim and Malah, 1984; Cappe, 1994). For this reason, a lower bound $\xi_{min}$ is typically used to prevent *a priori* SNR estimates falling below its prescribed value, i.e.,

$$\hat{\xi}_\ell(k,m) = \max\left[\hat{\xi}_\ell(k,m), \xi_{min}\right]. \tag{23}$$

Many approaches have been employed in the literature for noise power spectrum estimation in the acoustic spectral domain (e.g., Scalart and Filho, 1996; Martin, 2001; Cohen and Berdugo, 2002; Loizou, 2007). In the present work, spectral modulation domain estimates are needed. For this task a simple procedure is employed, where an initial estimate of modulation power spectrum of noise is computed from six leading silence frames.[5] This estimate is then updated during speech absence using a recursive averaging rule (e.g., Scalart and Filho, 1996; Virag, 1999), applied in the modulation spectral domain as follows

$$\hat{\lambda}_\ell(k,m) = \varphi\hat{\lambda}_{\ell-1}(k,m) + (1-\varphi)|\mathcal{X}_\ell(k,m)|^2, \tag{24}$$

where $\varphi$ is a forgetting factor chosen depending on the stationarity of the noise. The speech presence or absence is determined using a statistical model-based voice activity detection (VAD) algorithm by Sohn et al. (1999), applied in the modulation spectral domain.[6]

## 4. Subjective tuning of MME parameters

One of the reasons for the good performance of the AME method of Ephraim and Malah (1984) is that its parameters have been well tuned. In the current work, this MMSE estimator is applied in the spectral modulation domain. Consequently, the parameters of the proposed MME method need to be retuned.

The adjustable parameters of the MME approach include the acoustic frame duration (AFD), acoustic frame shift (AFS), modulation frame duration (MFD), modulation frame shift (MFS), as well as the smoothing parameter

[5] Using six non-overlapped frames in the modulation domain for initial noise estimation, around 220 ms of leading silence is required.

[6] More specifically, the decision-directed decision rule without hangover (Sohn et al., 1999) is used.

$\alpha$ and the lower bound $\xi_{min}$ used in *a priori* SNR estimation. Tuning of some of these parameters can be done qualitatively from our knowledge of speech processing, and can be fixed without further investigation. For example, speech can be assumed to be approximately stationary over short durations, and therefore acoustic frameworks typically use a short AFD of around 20–40 ms (e.g., Picone, 1993; Huang et al., 2001; Loizou, 2007; Paliwal and Wójcicki, 2008), which at the same time is long enough to provide reliable spectral estimates. Based on these qualitative reasons, an AFD of 32 ms was selected in this work. We have also chosen to use a 1 ms AFS to facilitate experimentation with a wide range of frame sizes and shifts in the modulation domain, and to increase the adaptability of the proposed method to changes in signal characteristics. For other parameters, subjective listening tests were conducted to determine values that maximise the subjective quality of stimuli enhanced using the MME method.

In the remainder of this section, we first describe details common to subsequent experiments. These include the speech corpus, settings used for stimuli generation and the listening test procedure. We then present experiments, results, and discussions. This section is concluded with a summary of the tuned parameters.

### 4.1. Speech corpus

The Noizeus speech corpus (Loizou, 2007; Hu and Loizou, 2007)[7] was used for the experiments presented in this section. The corpus contains 30 phonetically-balanced sentences belonging to six speakers (three males and three females), and each having an average length of around 2.6 s. The recorded speech was originally sampled at 25 kHz. The recordings were then downsampled to 8 kHz and filtered to simulate the receiving frequency characteristics of telephone handsets. The corpus includes stimuli with non-stationary noises at different SNRs. For our experiments only the clean stimuli were used. Corresponding noisy stimuli were generated by degrading the clean stimuli with additive white Gaussian noise (AWGN) at 5 dB SNR. Since use of the entire corpus was not feasible for human listening tests, in our experiments four sentences were employed. Of these, two (sp20 and sp22 belonging to a male and female speaker) were used for parameter tuning, while the other two (sp10 and sp26 also belonging to a male and female speaker) were used in subjective testing.

### 4.2. Stimuli

The settings used for the construction of MME stimuli are as follows. The Hamming window was used as both the acoustic and modulation analysis window functions. The FFT-analysis length was set to $2N$ and $2\mathcal{N}$ for acoustic and modulation domain processing, respectively. Least-squares overlap-add synthesis (Quatieri, 2002) was used for both acoustic and modulation syntheses. The threshold for the statistical voice activity detector (Sohn et al., 1999) was set to 0.15, and the forgetting factor $\varphi$ for noise estimate updates was set to 0.98. The AFD was set to 32 ms and the AFS was set to 1 ms. Other parameters used in the construction of MME stimuli for experiments presented in this section, are as defined in the description of each experiment.

### 4.3. Listening test procedure

Subjective testing was done in the form of AB listening tests that determined parameter preference. For each subjective experiment, listening tests were conducted in a quiet room. Participants were familiarised with the task during a short practice session. The actual test consisted of stimuli pairs played back in randomised order over closed circumaural headphones at a comfortable listening level. For each stimuli pair, the listeners were presented with three labelled options on a computer and asked to make a subjective preference. The first and second options were used to indicate a preference for the corresponding stimuli, while the third option was used to indicate a similar preference for both stimuli. The listeners were instructed to use the third option only when they did not prefer one stimulus over the other. Pair-wise scoring was used, with a score of +1 awarded to the preferred treatment, and 0 to the other. For the similar preference response, each treatment was awarded a score of +0.5. Participants could re-listen to stimuli if required.

### 4.4. Parameter tuning: modulation frame duration

Typical modulation domain methods use modulation frame durations (MFDs) of around 250 ms (Greenberg and Kingsbury, 1997; Thompson and Atlas, 2003; Kim, 2005; Falk and Chan, 2008; Wu et al., 2009; Falk and Chan, 2010; Falk et al., 2010; Paliwal et al., 2010). However, recent experiments (Paliwal et al., 2011) suggest that shorter MFDs may be better suited (in the context of intelligibility and quality) when processing the modulation magnitude spectrum. Paliwal et al. (2011) also showed that objective quality decreased for increasing MFDs. In this experiment we evaluate the effect of MFD on the quality of stimuli enhanced using the MME method.

Enhanced stimuli were created by applying the MME method (see Section 3) to noisy speech (see Section 4.1). Using a MFS of 2 ms, $\alpha = 0.998$, and $\xi_{min} = -25$ dB, MFD values of 32, 48, 64, 128 and 256 ms were investigated. The quality of the resulting stimuli was assessed through subjective listening tests using the procedure given in Section 4.3. Five subjects participated in this experiment. Each was presented with 40 comparisons. The session lasted approximately 10 min.

Mean subjective preference scores as a function of MFD are given in Fig. 2. The results show that use of long MFDs

[7] The Noizeus speech corpus is publicly available on-line at the following url: http://www.utdallas.edu/~loizou/speech/noizeus.

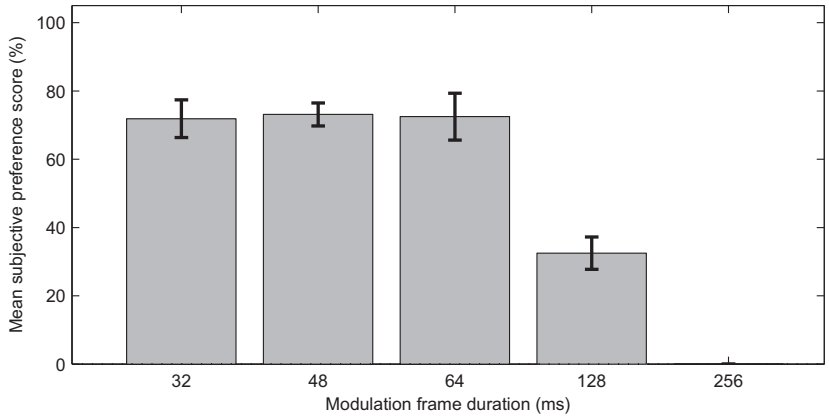


Fig. 2. Mean subjective preference scores (%) for stimuli generated using MME with 2 ms MFS, $\alpha = 0.998$, $\xi_{min} = -25$ dB, and MFD values of 32, 48, 64, 128, and 256 ms.

(such as 256 ms) reduce the quality of enhanced stimuli. The reason for this is that long frame durations cause spectral smearing, which can be heard as a reverberant type of distortion. On the other hand, use of short MFDs (such as 32–64 ms) produce stimuli with higher quality. Use of a 32 ms modulation frame duration is acceptable in the modulation domain for reasons similar to those used to justify a 32 ms acoustic frame duration by Ephraim and Malah in their MMSE formulation and discussed in Section 3. It is also noted that the results of this experiment are consistent with those reported by Paliwal et al. (2011), where shorter frame durations were found to work better for processing of the modulation magnitude spectrum. Based on the results of this experiment, a MFD of 32 ms was selected for use in the experiments in presented in later sections.

### 4.5. Parameter tuning: modulation frame shift

The modulation frame shift (MFS) affects the ability of the MME method to adapt to changes in the properties of the signal, with shorter shifts offering some reduction in the introduced distortion during more transient parts. However, smaller shifts also add to the computational cost of the method.

In this experiment, we evaluate the effect of MFS on the subjective quality of speech corrupted with 5 dB AWGN enhanced with the MME method. For this experiment, MFD is set to 32 ms, $\alpha = 0.998$, $\xi_{min} = -25$ dB, and MFS durations of 1, 2, 4, and 8 ms were investigated. The quality of the resulting stimuli were then compared via subjective human listening experiments, conducted under the same conditions as described in Section 4.3. Five listeners participated in the test. The listeners were presented with 24 comparisons in a session which lasted around 5 min.

Fig. 3 shows the mean subjective preference scores for each of the MFS settings investigated. Results show a clear preference by listeners for stimuli generated using shorter MFSs than for longer shifts, with a 2 ms shift being the most preferred. When longer MFSs such as 8 ms were used, stimuli sounded hollow and slurred due to the slowed response to changing speech characteristics. Using shorter MFSs, on the other hand, resulted in crisper speech, however for a 1 ms MFS, speech started to sound muffled with some low periodic tones present. Therefore a MFS of 2 ms was found to result in the best quality stimuli.

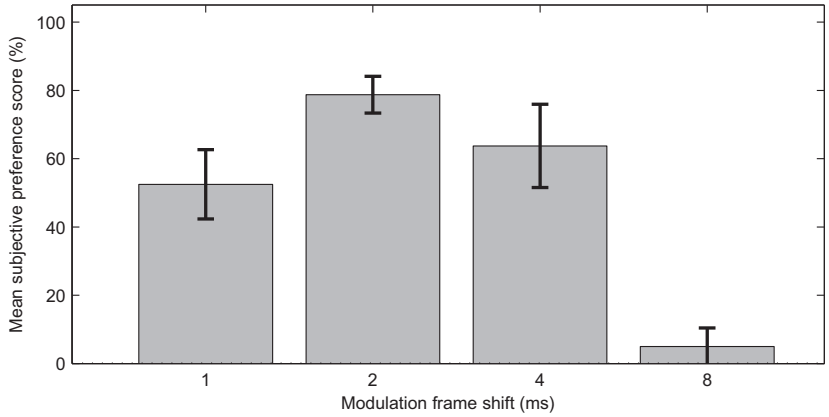


Fig. 3. Mean subjective preference scores (%) for stimuli generated using MME with 32 ms MFD, $\alpha = 0.998$, $\xi_{min} = -25$ dB, and MFS values of 1, 2, 4, and 8 ms.

### 4.6. Parameter tuning: smoothing parameter

For AME, the *a priori* SNR is commonly estimated using the decision-directed approach, where smoothing parameter $\alpha$ determines how much of the *a priori* SNR estimate is based on the current frame versus that which is based on the previous frame. The value selected for $\alpha$ has a significant effect on both the adaptability of the estimator to changes in signal characteristics, as well as the type of residual distortion present in the resulting stimuli. Cappe (1994) notes that an $\alpha$ value greater than 0.9 is required to suppress the fluctuations which introduce musical noise to stimuli. However if $\alpha$ is too high, low-energy components can be clipped making speech sound bottled (Breithaupt and Martin, 2011). Additionally, a higher $\alpha$ value makes the estimator slower to adapt to signal changes in more transient parts, resulting in a slurring type of distortion being introduced to processed stimuli. The value chosen for $\alpha$ in AME is thus a trade-off between reducing noise (by smoothing the *a priori* SNR estimate) and reducing the transient distortion introduced to the signal (Ephraim and Malah, 1984; Cappe, 1994).

As detailed in Section 3, the proposed MME method also uses the decision-directed approach to estimate the *a priori* SNR. Therefore in this experiment we evaluate the effect of $\alpha$ on the subjective quality of stimuli enhanced using MMSE magnitude estimation in the modulation domain. Enhanced stimuli were generated by applying the MME method to noisy speech (corrupted with 5 dB AWGN), using the settings given in Section 4.2 with MFD set to 32 ms, MFS set to 2 ms, and $\xi_{min} = -25$ dB. $\alpha$ values of 0.994, 0.996, 0.998, and 0.999 were investigated. Listening tests were then used to subjectively evaluate the quality of the resulting stimuli. Each test consisted of 24 comparisons, and took around 5 min to complete. Five listeners participated in the tests.

The mean subjective preference scores for each choice of $\alpha$ are shown in Fig. 4. Stimuli generated using $\alpha = 0.998$ were clearly preferred over other $\alpha$ values. This value for $\alpha$ is much higher than 0.98, which is typically used by AME implementations. For $\alpha < 0.998$, stimuli were clear but a musical-type distortion was also present. This musical noise increased in intensity as $\alpha$ was reduced. For $\alpha > 0.998$, enhanced speech sounded bottled. This type of distortion was found to be more annoying to some listeners than to others (as indicated by the larger variance bars on the preference scores for $\alpha = 0.999$). Thus, a similar trade-off was observed in the modulation domain as is seen in the acoustic domain. For MME, the subjectively preferred $\alpha$ value of 0.998 was found to be a good compromise between the above types of distortions.

### 4.7. Parameter tuning: lower bound of a priori estimate

As mentioned in Section 3, it has been shown for AME that limiting the minimum value of the *a priori* SNR also affects the type of distortion which is present in the resulting stimuli (Cappe, 1994). In (Ephraim and Malah, 1984) a limit of $\xi_{min} = -15$ dB was recommended, while in the reference AME implementation by Loizou (2007) $\xi_{min} = -25$ dB was used. For MME, informal listening experiments determined that $\xi_{min} = -25$ dB gives the best results. Using $\xi_{min} < -25$ dB, musical noise distortion (increasing in intensity with decreasing $\xi_{min}$) could be heard. Using $\xi_{min} > -25$ dB, the residual noise is whiter and increasing in intensity for increasing $\xi_{min}$. Therefore, a lower bound of $\xi_{min} = -25$ dB was chosen as the best compromise between the above distortions, resulting in stimuli with low-level non-musical residual noise.

### 4.8. Conclusion

In this section, listening tests were performed to determine values of various MME method parameters that maximise subjective speech quality of enhanced speech. While experiments showed MME to be quite sensitive to the values of MFD, MFS and $\alpha$, the preferred values for each were found to be around the same value for the different stimuli and the types of noise disturbances investigated. The tuned values for these parameters are a 32 ms MFD,

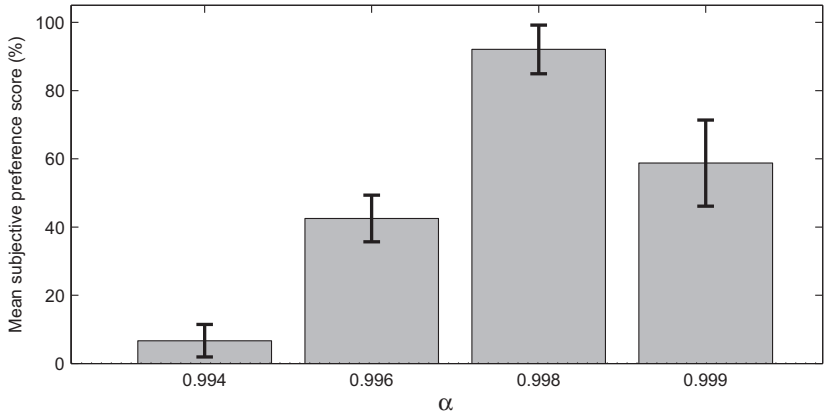


Fig. 4. Mean subjective preference scores (%) for stimuli generated using MME with a 32 ms MFD, 2 ms MFS, $\xi_{min} = -25$ dB, and $\alpha = 0.994$, 0.996, 0.998, and 0.999.

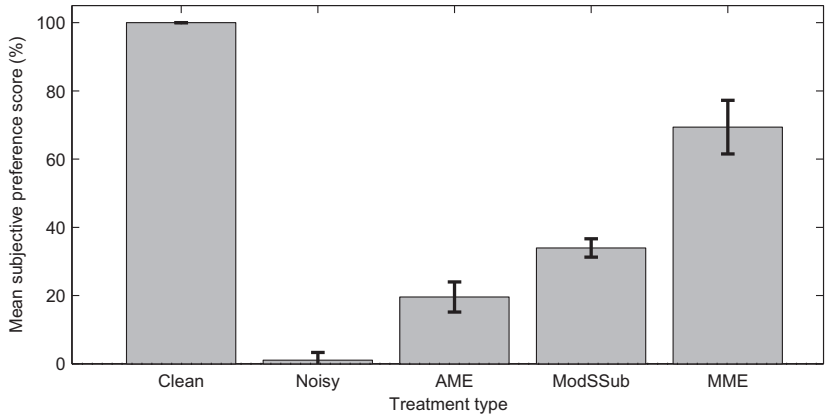


Fig. 5. Mean subjective preference scores (%) for (a) clean; (b) noisy (degraded at 5 dB AWGN); and stimuli generated using the following treatment types: (c) AME (Ephraim and Malah, 1984); (d) ModSSub (Paliwal et al., 2010); and (e) MME (proposed).

2 ms MFS, 0.998 for smoothing parameter $\alpha$, and $-25$ dB for $\xi_{min}$. The MME stimuli produced with the above settings were found to have good noise suppression without the introduction of musical noise or temporal slurring distortion.

## 5. Speech enhancement experiment

The performance of the MME method was evaluated by comparing its speech enhancement performance with that of the following two methods: (1) AME method of Ephraim and Malah (1984) and (2) ModSSub of Paliwal et al. (2010). The quality of stimuli enhanced by each of the above approaches was evaluated using formal listening tests.

The remainder of this section is organised as follows. We begin by giving details of the settings used for stimuli construction. The testing procedure for the subjective experiment is then outlined. Finally, the results are presented and discussed.

### 5.1. Stimuli

MME stimuli were constructed using the procedure detailed in Section 3. The parameters used were those described in Section 4.2, and those determined using subjective tests detailed in Section 4.8 (i.e., AFD = 32 ms, AFS = 1 ms, MFD = 32 ms, MFS = 2 ms, $\alpha = 0.998$, and $\xi_{min} = -25$ dB).

Stimuli enhanced using the AME method (Ephraim and Malah, 1984) were generated using a publicly available reference implementation (Loizou, 2007). Here, optimal estimates (in the minimum mean-square error sense) of the short-time acoustic spectral magnitudes were computed. The AMS procedure used a 20 ms AFD and a 10 ms AFS. The decision-directed approach was used for the *a priori* SNR estimation, with the smoothing factor set to 0.98, and the *a priori* SNR lower bound was set to $-25$ dB. Noise spectrum estimates were computed from non-speech frames using recursive averaging with speech presence or absence determined using a statistical model-based VAD (Sohn et al., 1999).

ModSSub stimuli were created used an AFD of 32 ms, with an 8 ms AFS, and MFD of 256 ms, and a 32 ms MFS. The spectral floor parameter $\beta$ was set to 0.002, and $\gamma$ was set to 2 for subtraction in the modulation magnitude-squared domain. Speech presence or absence was determined using a VAD algorithm based on segmental SNR. The speech presence threshold was set to 3 dB. The forgetting factor was set to 0.98.

In addition to the MME, AME, and ModSSub stimuli, clean and noisy speech stimuli were also included in our experiments. Example spectrograms for each of the stimuli types are shown in Fig. 6.[8]

### 5.2. Listening test

Experiments that compare the subjective quality of stimuli described in Section 5.1 were conducted using the procedure described in Section 4.3. Two Noizeus sentences belonging to one male and one female speaker and degraded with 5 dB AWGN were used. The complete test included 40 comparisons and lasted approximately 10 min. Twelve listeners participated in the test.

### 5.3. Results and discussion

The mean subjective preference scores for each enhancement method are shown in Fig. 5. MME scores are significantly higher than those of ModSSub and AME methods, indicating that listeners consider MME stimuli to have a higher quality than those of both AME and ModSSub types. This is an important finding as it demonstrates the efficiency of short-time modulation processing for speech enhancement, and demonstrates that the

[8] Examples of audio stimuli are available for the interested reader at the following url: http://maxwell.me.gu.edu.au/spl/research/modmmse/index.html.

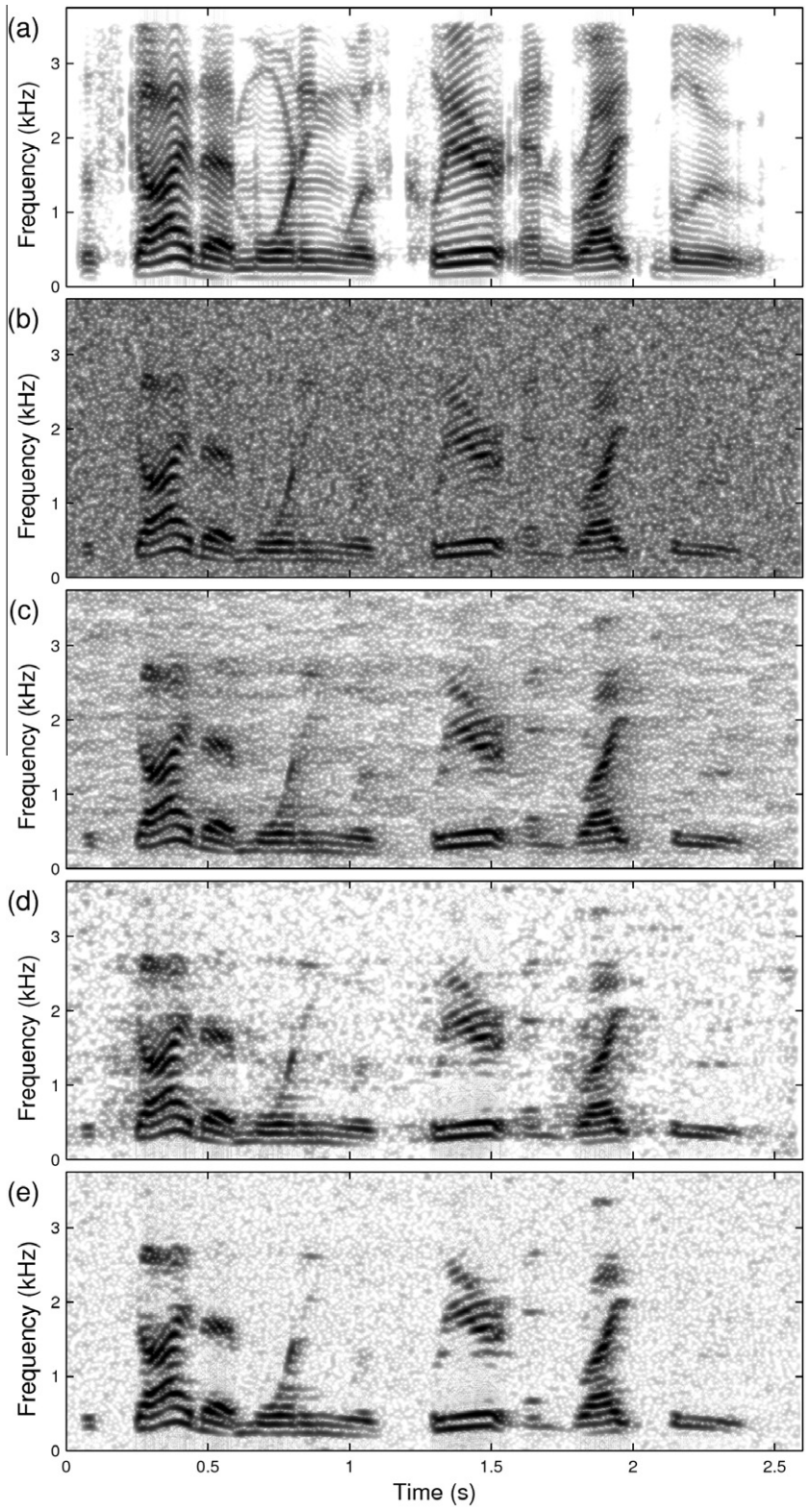


Fig. 6. Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech; (b) speech degraded by AWGN at 5 dB SNR; and noisy speech enhanced using: (c) AME (Ephraim and Malah, 1984); (d) ModSSub (Paliwal et al., 2010); and (e) MME (proposed).

performance of existing acoustic domain approaches can be potentially improved when these are applied in the short-time modulation domain.

Spectrograms of the utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus are shown in Fig. 6. Spectrograms

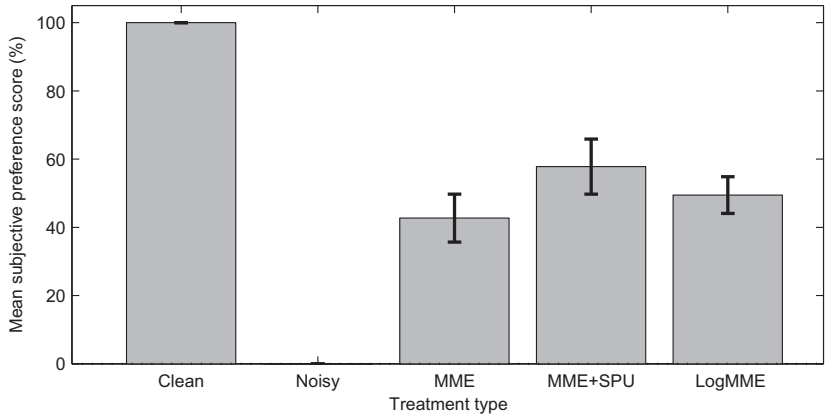


Fig. 7. Mean subjective preference scores (%) with standard error bars for (a) clean; (b) noisy (degraded at 5 dB AWGN); and stimuli generated using the following treatment types: (c) MME; (d) MME+SPU; and (e) LogMME.

for clean and noisy (degraded with 5 dB AWGN) stimuli are shown in Fig. 6(a) and (b) respectively. Fig. 6(c) shows the spectrogram of stimuli enhanced using AME, where much residual noise can be seen. The ModSSub stimuli, as shown by the spectrogram of Fig. 6(d), are much cleaner, but there is some spectral smearing and visible spectral artefacts. These distortions can be heard as a type of slurring and some low level musical-type noise. The spectrogram for the MME stimuli is shown in Fig. 6(e). As can be seen, MME stimuli have better noise suppression than AME without introducing the spectral artefacts visible in the ModSSub spectrogram. Informal listening confirms that speech quality of MME stimuli are improved without the introduction of the annoying residual distortions heard in the other stimuli types investigated.

## 6. Effect of using SPU and LogMMSE on MME formulation

In Ephraim and Malah's classical paper on acoustic magnitude estimation (AME), authors also proposed an AME formulation under the uncertainty of speech presence (Ephraim and Malah, 1984). Here, the quality of enhanced speech was shown to be further improved (compared to that generated by AME alone), without introducing any additional distortions. In a later paper, they went on to show that applying AME to the log-magnitude spectrum (Ephraim and Malah, 1985), which is more suited to speech processing (Gray et al., 1980), also results in improved enhanced speech quality. Motivated by these observations, we investigated the effect of applying speech presence uncertainty (SPU) and log-magnitude spectral processing to the MME formulation.[9] Findings of this investigation are presented in this section.

### 6.1. MME with speech presence uncertainty

Using SPU, the optimal estimate of the modulation magnitude spectrum is given by the relation

$$\left|\widehat{\mathcal{S}}_\ell(k,m)\right| = \phi(k,m)\mathcal{G}_\ell(k,m)|\mathcal{X}_\ell(k,m)|, \tag{25}$$

where $\mathcal{G}_\ell(k,m)$ is the MME spectral gain function given by Eq. (16), and $\phi(k,m)$ is given by

$$\phi(k,m) = \frac{\Lambda(k,m)}{1+\Lambda(k,m)} \tag{26}$$

with

$$\Lambda(k,m) = \frac{(1-q_m)}{q_m}\cdot\frac{\exp(v_\ell(k,m))}{1+\hat{\xi}_\ell(k,m)} \tag{27}$$

and $v_\ell(k,m)$ given by Eq. (17). Here $q_m$ is the probability of signal presence in the $m$th spectral component, and is a tunable parameter. Applying $\left|\widehat{\mathcal{S}}_\ell(k,m)\right|$ of Eq. (25) in the framework described in Section 2 produced stimuli denoted type MME+SPU.

The key parameters of MME+SPU were tuned subjectively using tuning stimuli[10] corrupted with 5 dB of AWGN, and a similar procedure as that described for tuning MME in Section 4. Preferred parameters for MME+SPU were an AFD of 32 ms, a 1 ms AFS, a 32 ms MFD, and a 2 ms MFS. For parameter $q_m$, a value of 0.3 was found to work best. $\xi_{min}$, as before, gave best results using $-25$ dB. The smoothing parameter used for decision-directed *a priori* SNR estimation as able to be reduced to 0.995, without introducing musical noise. Throughout this work, stimuli referred to as type MME+SPU are assumed to be constructed using these parameter values in the above described procedure.

[9] In addition to SPU and log-spectrum formulations, acoustic domain LogAME with SPU has been investigated in the literature but found to not work particularly well. This was also the case in the modulation domain and so is not included here.

[10] Stimuli used for tuning parameters were from the Noizeus speech corpus (see Section 4.1). For this purpose, speech files sp20 and sp22, belonging to a male and female speaker, and corrupted with 5 dB of AWGN were used.

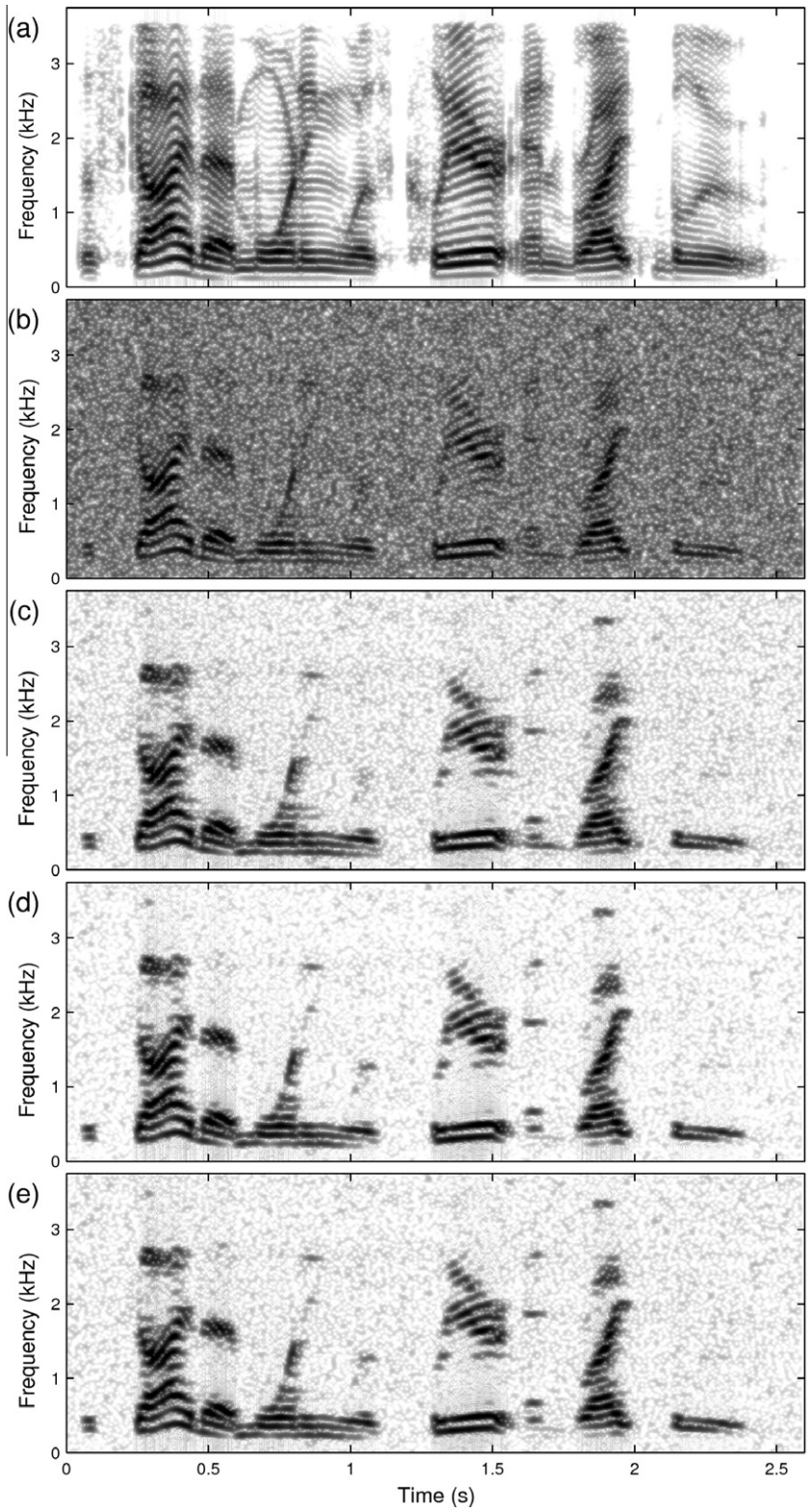


Fig. 8. Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech; (b) speech degraded by AWGN at 5 dB SNR; and noisy speech enhanced using: (c) MME; (d) MME+SPU; and (e) LogMME.

### 6.2. MME of the log-modulation magnitude spectrum

Minimising the mean-squared error of the log-modulation magnitude spectrum, the optimal estimate of the modulation magnitude spectrum is given by the relation

$$\left|\widehat{\mathcal{S}}_\ell(k,m)\right| = \mathcal{G}_\ell(k,m)|\mathcal{X}_\ell(k,m)|, \tag{28}$$

where $\mathcal{G}_\ell(k,m)$ is the spectral gain function given by

$$\mathcal{G}_\ell(k,m) = v_\ell(k,m)\exp\left(\frac{1}{2}\mathrm{Ei}\left[v_\ell(k,m)\right]\right), \tag{29}$$

Ei[·] is the exponential integral, and $v_\ell(k,m)$ (a function of *a priori* and *a posteriori* SNRs) is given by Eq. (17). Stimuli of type LogMME are then constructed by applying $\left|\widehat{\mathcal{S}}_\ell(k,m)\right|$ given by Eq. (28) in the framework described in Section 2.

Using tuning stimuli[10] and the procedure described in Section 4, parameters for LogMME were subjectively tuned. Preferred LogMME parameters were an AFD of 32 ms, an AFS of 1 ms, a 32 ms MFD, and a 2 ms MFS (as used for both MME and MME+SPU). $\xi_{min}$ again gave best results using −25 dB. The smoothing parameter gave preferred stimuli for 0.996. These parameters and the above procedure were used to generate test stimuli referred to as type LogMME.

### 6.3. Subjective evaluation and discussion

An experiment was conducted to subjectively compare the quality of stimuli enhanced using the MME, MME+SPU and LogMME methods. The subjective evaluation was in the form of AB listening tests to determine method preference. Test sentences (sp10 and sp26 from the Noizeus corpus) belonging to a male and female speaker were used. Stimuli types included in the test were clean, noisy (corrupted with 5 dB of AWGN), MME, MME+SPU and LogMME. Conditions for the test were as described in Section 4.3. Six listeners participated in the test, which involved selecting their preference in each of 40 stimuli pair comparisons.

Results in terms of mean subjective preference including standard error bars, are shown in Fig. 7. Here we see that MME+SPU was preferred by listeners over MME and LogMME stimuli types. It is noted that the difference between the MME, LogMME and MME+SPU stimuli types is relatively small compared to that observed in the acoustic domain, and mainly heard as an improvement in background noise attenuation. This can be seen in the spectrograms of Fig. 8, where Fig. 8(d) shows the spectrogram of MME+SPU has a background that is lighter but of similar appearance to that of MME and LogMME shown in Fig. 8(c) and (e). It was also observed through informal experimentation that the improvement in speech quality due to the use of SPU was more noticeable in some stimuli than in others. Informal experimentation using coloured noise stimuli also showed a small improvement in speech quality compared to MME, with MME+SPU and LogMME being quite similar in quality.

The remainder of this work will now compare the performance of the MME+SPU approach with other speech enhancement methods.

## 7. Comparison of MME+SPU with different AME formulations

In this section, we aim to evaluate the effect of modulation domain processing on enhanced speech quality, by comparing MME+SPU to the different acoustic domain formulations, including AME (Ephraim and Malah, 1984), AME under the uncertainty of speech presence (AME+SPU) (Ephraim and Malah, 1984) and estimation of the log-acoustic magnitude spectrum (LogAME) (Ephraim and Malah, 1985). Here, quality is compared by way of a subjective listening experiment.

### 7.1. Stimuli

The subjective experiment investigated the enhancement of noisy speech corrupted with 5 dB of AWGN. Test stimuli (sp10 and sp26), by a male and female speaker, were from the Noizeus corpus. Type MME+SPU stimuli were constructed by processing noisy stimuli as described in Section 6.1. AME, AME+SPU and LogAME stimuli were generated by processing noisy stimuli using publicly available reference implementations (Loizou, 2007) of each method. Details of the AME implementation are provided in Section 5.1. The parameters of the LogAME implemen-

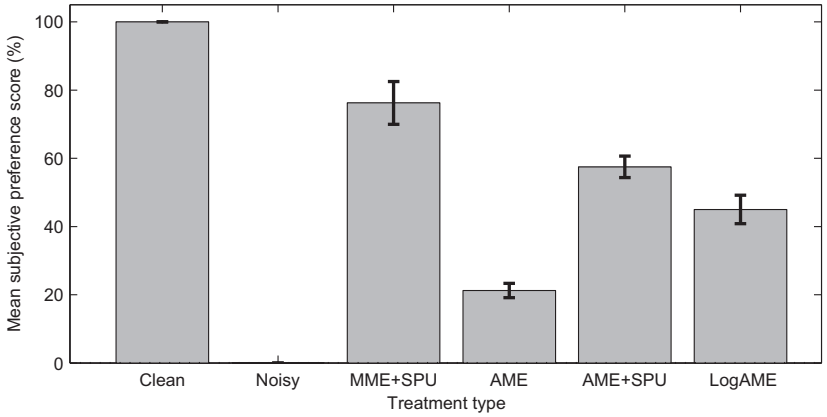


Fig. 9. Mean subjective preference scores (%) with standard error bars for (a) clean; (b) noisy (degraded at 5 dB AWGN); and stimuli generated using the following treatment types: (c) MME+SPU; (d) AME (Ephraim and Malah, 1984); (e) AME+SPU (Ephraim and Malah, 1984); and (f) LogAME (Ephraim and Malah, 1985).

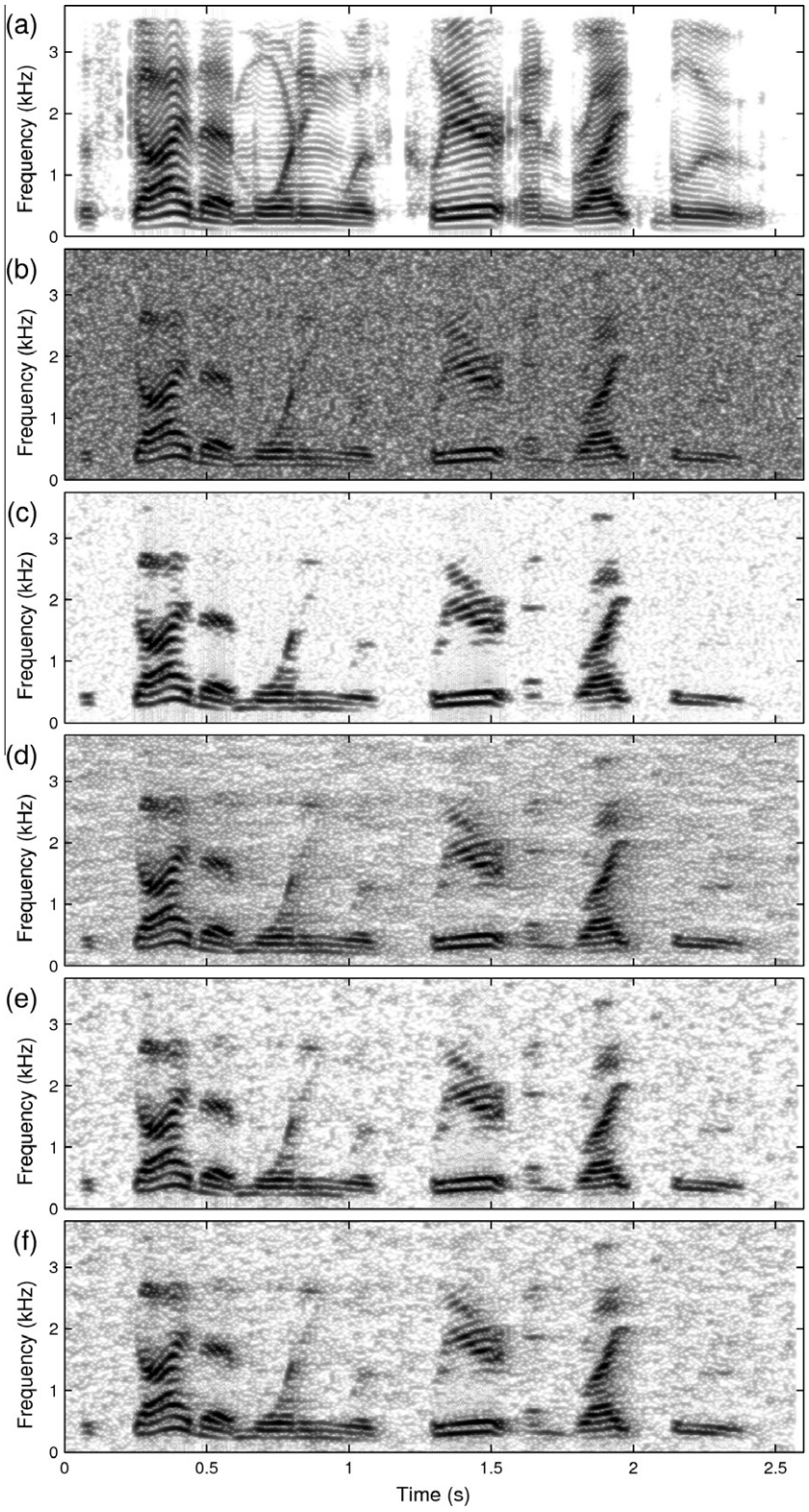


Fig. 10. Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech; (b) speech degraded by AWGN at 5 dB SNR; and noisy speech enhanced using: (c) MME+SPU; (d) AME (Ephraim and Malah, 1984); (e) AME+SPU (Ephraim and Malah, 1984); and (f) LogAME (Ephraim and Malah, 1985).

tation are as described for AME. For AME+SPU, the implementation is also as described for AME with the exception of the additional parameter $q_k$ which was set to 0.3. Clean and noisy stimuli were also included in the test.

### 7.2. Subjective evaluation and discussion

Subjective evaluation was in the form of AB human listening tests, in which listeners selected the stimuli they

preferred from 60 randomly played stimuli pairs. Conditions for the test are as described in Section 4.3. Six listeners participated in the test. Fig. 9 shows the resulting mean preference scores. MME+SPU is shown to be preferred by listeners over each of the acoustic domain methods included in the test. Of the acoustic domain methods, AME+SPU was the most preferred, though LogAME also scored well compared to AME. Example spectrograms for each treatment type are shown in Fig. 10. Fig. 10(c), corresponding to stimuli processed with MME+SPU, shows much better removal of noise than all other methods. The nature of the noise is otherwise quite similar and colourless.

Experimental results presented in this section use speech stimuli corrupted with AWGN. In Appendix A, results are also provided for similar experiments using stimuli corrupted with coloured noises (babble and street noise types). These results also show a preference for MME+SPU stimuli, suggesting that type MME+SPU stimuli are of higher quality than the acoustic domain formulations. Contrary to the results shown in Fig. 9 for AWGN, for coloured noises LogAME performed considerably better, having scores higher than both AME and AME+SPU. Therefore, it may be suggested that LogMME may work better than MME+SPU for coloured noise types. However, informal experimentation for coloured noise found the difference between each of the MME formulations much smaller than seen in the acoustic domain, with MME+SPU type stimuli generally preferred.

Appendix B provides results of objective experiments comparing MME+SPU, AME, AME+SPU and LogAME stimuli for noisy speech with additive white and coloured noise, for a range of SNRs. Segmental SNR and STI scores were shown to have the highest correlation to subjective preferences, scoring MME+SPU higher than the acoustic domain methods. Results for coloured noises again showed LogAME having a higher score than AME+SPU, while results for white noise showed AME+SPU having a higher score than LogAME. These scores are consistent with subjective results shown in Figs. 9, 13, and 14. However, PESQ scores were less consistent with LogAME and AME+SPU scores being better at low SNRs, and results for coloured noises showed MME+SPU better at higher SNRs (greater than around 10 dB).

In comparing the performance of the modulation and acoustic domain implementations of MMSE magnitude estimator, we have shown that the modulation domain implementation produces stimuli with improved noise suppression, with MME+SPU stimuli chosen by listeners as having better quality than acoustic domain MMSE formulations. However, a disadvantage of the method is its additional computational complexity. Since the STFT length used was 512, MMSE magnitude spectral estimation needs to be performed on each of the 257 acoustic bins, instead of just once. Consequently, this approach is not suited to applications where processing time is crucial. However, often it is the quality of enhanced speech which is of greater importance, in which case the MME+SPU approach is shown to offer improved performance.

## 8. Comparison of modulation domain enhancement methods

As previously mentioned, speech enhancement methods can generally be classified as either spectral subtraction type, statistical (MMSE) type, Wiener filtering type, Kalman filtering type or a subspace based method. In this work, we have proposed a MMSE-type formulation which operates in the modulation domain. In earlier work, we have also applied other speech enhancement methods in the modulation domain, including modulation domain based spectral subtraction (Paliwal et al., 2010) (ModSSub) and modulation domain Kalman filtering (So and Paliwal, 2011) (ModKalman). In this section, we compare the results of the proposed MME+SPU method with those of other previously reported modulation domain-based approaches. We have also included a modulation domain based implementation of Wiener filtering (ModWiener) for comparison.

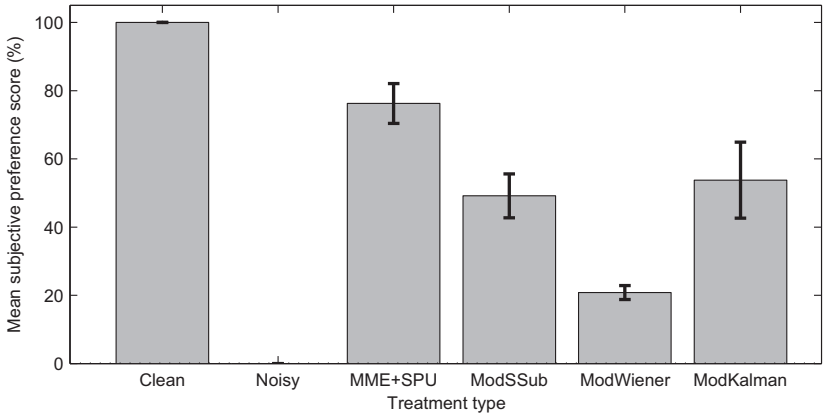


Fig. 11. Mean subjective preference scores (%) including standard error bars for (a) clean; (b) noisy (degraded at 5 dB AWGN); and stimuli generated using the following treatment types: (c) MME+SPU; (d) ModSSub (Paliwal et al., 2010); (e) ModWiener; and (f) ModKalman (So and Paliwal, 2011).

### 8.1. Stimuli

For the purpose of experiments presented in this section, stimuli from the Noizeus corpus and corrupted with 5 dB of AWGN, were enhanced by each method being investigated. For MME+SPU stimuli, noisy speech were enhanced using the procedure described in Section 6.1. ModSSub stimuli were constructed as described in Section 5.1.

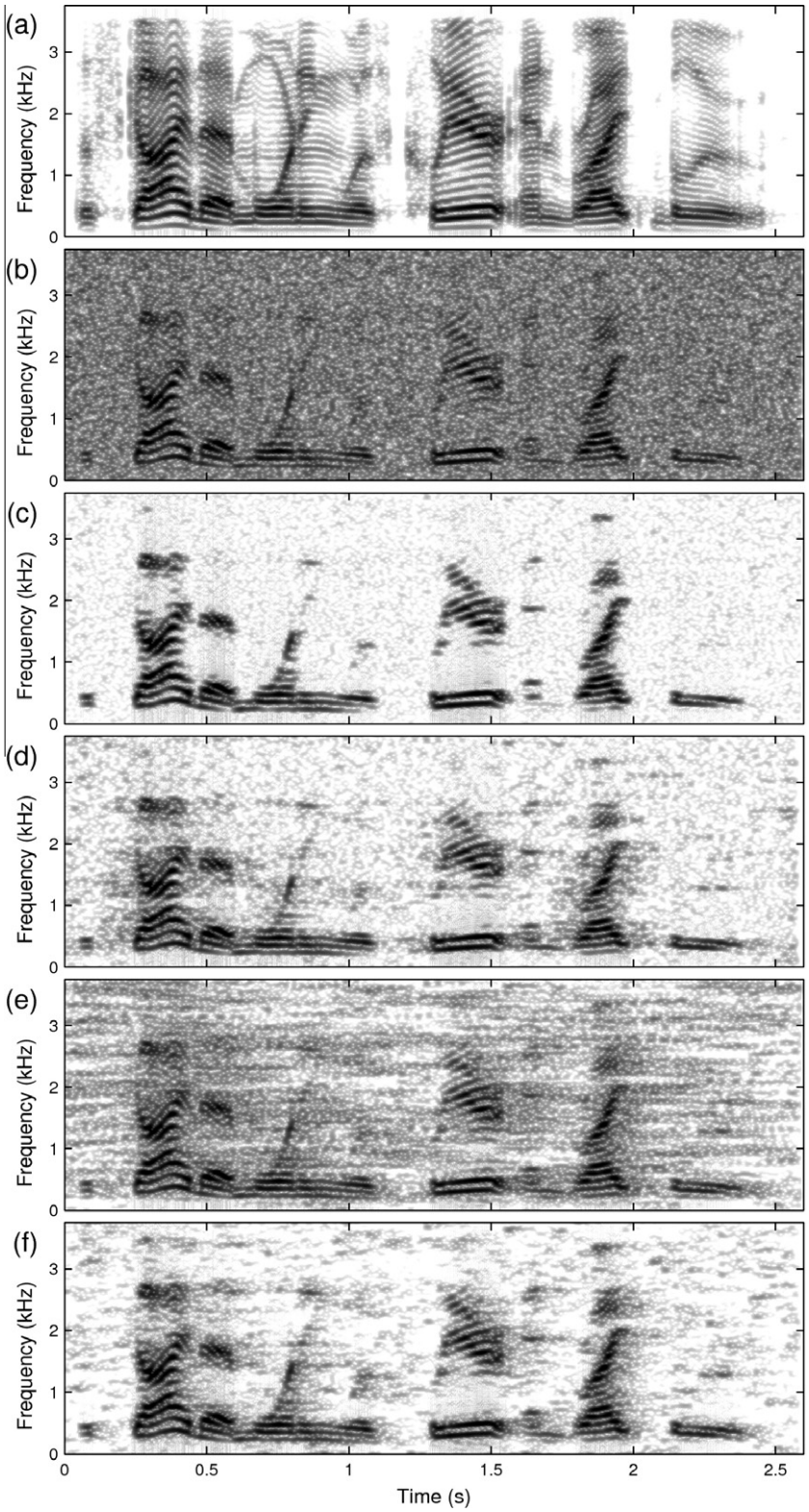


Fig. 12. Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech; (b) speech degraded by AWGN at 5 dB SNR; and noisy speech enhanced using: (c) MME+SPU; (d) ModSSub (Paliwal et al., 2010); and (e) ModWiener; (f) ModKalman (So and Paliwal, 2011).

Stimuli denoted ModKalman are generated using a modulation domain Kalman filter with LPCs estimated from AME enhanced speech, as described in (So and Paliwal, 2011). Here LPCs of order 2 were used for modelling modulating signals, and LPCs of order 4 were used for modelling the noise. The AMS framework used a 32 ms AFD, 4 ms AFS, 20 ms MFD, and no overlap (i.e., MFS of 20 ms) in the modulation domain.

ModWiener stimuli were generated by implementing an *a priori* based approach to Wiener filtering (Scalart and Filho, 1996) in the modulation domain with key parameters tuned subjectively to a 32 ms AFD, 1 ms AFS, 128 ms MFD, 16 ms MFS, and smoothing parameter for *a priori* SNR estimation of 0.998.

In addition to each of the enhanced stimuli, clean and noisy stimuli were also included in the test.

### 8.2. Subjective evaluation and discussion

AB listening tests were used to subjectively compare the quality of stimuli generated using different modulation based methods. These tests randomly played 60 stimuli pairs, with listeners asked to choose the one they preferred. Tests were conducted under the same conditions as described in Section 4.3. Six listeners participated in the experiment. The resulting mean preference scores are shown in Fig. 11. MME+SPU was clearly preferred by listeners over other stimuli types. ModKalman was the next most preferred. Some listeners had similar preference for ModSSub and ModKalman, while others preferred ModKalman over ModSSub. The good performance of the MME+SPU and ModKalman methods is partly attributed to their use small modulation frame durations, which is consistent with the findings reported in (Paliwal et al., 2011). ModWiener was the least preferred of the investigated enhancement methods.

The preferences shown in Fig. 11 are well explained by looking at the spectrograms of each stimuli type. Spectrograms of the utterance, "The sky that morning was clear and bright blue", by a male speaker are shown in Fig. 12. For type MME+SPU stimuli (shown in Fig. 12(c)), we can see there is good background noise removal, less residual noise, no musical-type noise, and no visible spectral smearing. ModKalman stimuli (Fig. 12(f)) also have good background noise removal and no visible spectral smearing, but has clear dark spots throughout the background heard as a musical type noise. ModSSub stimuli (Fig. 12(d)), on the other hand, has less musical type noise than ModKalman but also contains spectral smearing due to the use of longer frame durations, causing distortion in the processed speech. ModWiener, which was the least effective method, had considerable distortion in the stimuli, seen as darkness in the background of its spectrogram (Fig. 12(e)). The poor performance of the ModWiener was in part due to difficulty tuning, where parameters working better for one stimuli was considerably different for another.

Appendix C provides results of similar subjective experiments for stimuli corrupted with coloured noise types such as babble and street noise. Results shown there are consistent with those for AWGN, with MME+SPU stimuli indicated to be of higher quality than those enhanced by any of the other modulation domain methods investigated. The MME+SPU method was found to demonstrate good noise removal, without the introduction of spectral smearing or musical noise. Results also showed that ModKalman works well for more stationary noise types such as street noise, where (as was shown for AWGN) it scored higher than ModSSub. However, in the presence of more nonstationary noise such as babble, ModKalman and ModSSub were similarly preferred.

Appendix D has also been included to show an objective evaluation of these enhancement methods. Results presented compared the segmental SNR (Quackenbush et al., 1988), PESQ (Rix et al., 2001), and STI intelligibility scores (Drullman et al., 1994) of enhanced stimuli, for different noise types and input SNR. As found in Appendix B, the segmental SNR and STI measures demonstrated the highest correlation with subjective results, with MME+SPU generally scoring higher than other methods. While for white noise stimuli, objective scores for MME+SPU, ModSSub and ModKalman were quite close, there was a bigger difference in scores when considering coloured noise stimuli. Here, MME+SPU scored somewhat higher than other methods. For babble noise, ModSSub and ModKalman were very close with ModSSub scoring a little higher, while for street noise, ModKalman did much better. This is consistent with the findings of the coloured noise subjective experiments. In all cases ModWiener scored poorly.

## 9. Conclusion

In this paper we have proposed the MMSE modulation magnitude estimation method for speech enhancement. We have evaluated the effect of speech presence uncertainty and log-domain processing, and shown that the proposed approach works better with speech presence uncertainty. We have compared the performance of the proposed approach (with speech presence uncertainty) against that of different acoustic magnitude estimator formulations (Ephraim and Malah, 1984; Ephraim and Malah, 1985) and modulation domain speech enhancement methods such as modulation domain based spectral subtraction (Paliwal et al., 2010) and modulation domain Kalman filter (So and Paliwal, 2011). The results of our experiments show that the proposed method (when using speech presence uncertainly) achieves significantly higher subjective preference than the other methods investigated. This is because it uses a shorter modulation frame duration than modulation domain based spectral subtraction and thus does not suffer from slurring distortion, while not introducing any musical noise distortion as done in modulation

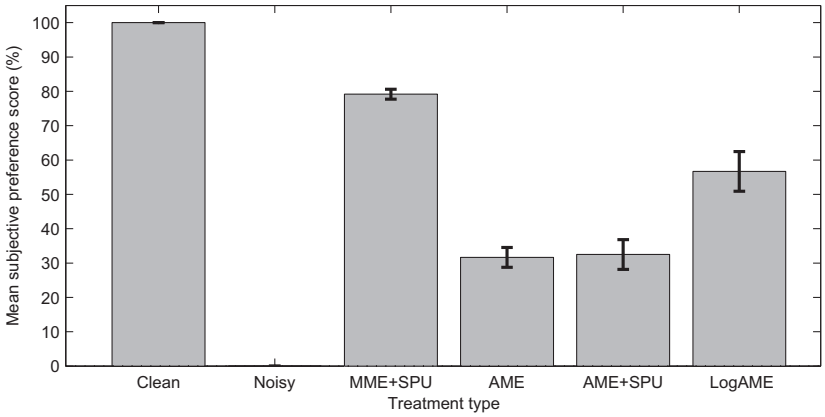


Fig. 13. Mean subjective preference scores (%) including standard error bars for: (a) clean; (b) noisy (degraded with babble noise at 5 dB); and stimuli generated using the following treatment types: (c) MME+SPU; (d) AME; (e) AME+SPU; and (f) LogAME.

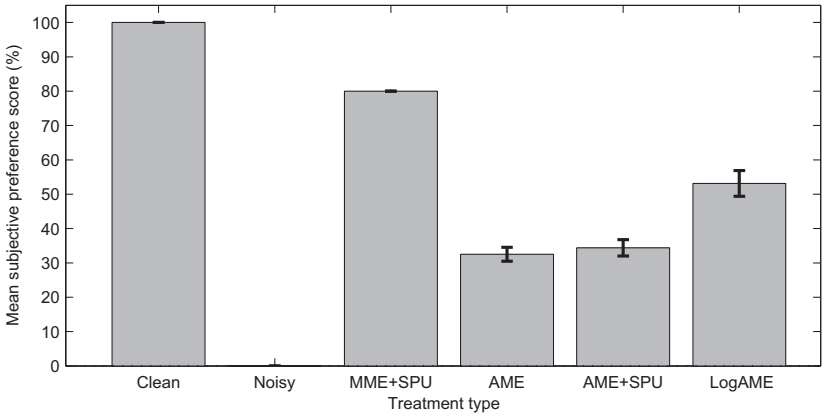


Fig. 14. Mean subjective preference scores (%) including standard error bars for: (a) clean; (b) noisy (degraded with street noise at 5 dB); and stimuli generated using the following treatment types: (c) MME+SPU; (d) AME; (e) AME+SPU; and (f) LogAME.

domain Kalman filtering. In addition, it achieves considerably better noise reduction than that associated with the acoustic domain magnitude estimators. Results for coloured and less stationary noises also showed the proposed approach to be similarly effective in reducing noise. However we do acknowledge that the average length of stimuli investigated was less than 3 s, which is too short to fully test the performance under non-stationary conditions (ITU-T P.835, 2007). Evaluation with much longer stimuli subjected to a range of non-stationary noise distortions can be investigated in future.

## Appendix A. Comparison of MME to AME formulations in the presence of coloured noise

In this experiment we extend the comparison of MME+SPU to acoustic domain formulations done in Section 7 with AWGN, to the processing of stimuli corrupted with additive coloured noise. Here, babble and street type noises at 5 dB are investigated. The subjective experiments were in the form of AB listening tests, and were conducted under the conditions previously described in Section 4.3. Listeners were asked to select the stimuli they preferred from 60 different randomly played stimuli pairs. Tests for each noise type were conducted in separate sessions, with five listeners participating in each test.

Results of tests using stimuli corrupted with babble and street noise, in terms of mean subjective preference, are shown in Figs. 13 and 14. Despite using very different types of noises, figures show similar preference scores. In both cases MME+SPU was preferred over all AME formulations. LogAME was the next most preferred, while both AME and AME+SPU had similarly low preference scores.

## Appendix B. Objective evaluation comparing MME+SPU to acoustic domain formulations

In Section 7 and Appendix A we have used subjective experiments to compare the quality of the proposed MME+SPU method with each of the acoustic domain MMSE formulations. In this appendix, we provide results

of objective experiments comparing the quality of stimuli processed using each enhancement method being investigated. Noisy stimuli with additive white, babble and street noise, added at SNRs of 0, 5, 10, 15 and 20 dB were considered. All 30 stimuli of the Noizeus corpus were used in the experiments.

Noisy and enhanced stimuli were compared to clean stimuli via popular objective quality measures such as segmental SNR (Quackenbush et al., 1988) and PESQ (Rix et al., 2001). The STI intelligibility measure (Drullman et al., 1994) was also included for an indication of the effect of processing on intelligibility, an important aspect of quality.

The mean scores across the 30 sentences of the Noizeus corpus for each treatment type, SNR and objective measure were calculated. These are shown in Figs. 15–17 for AWGN, Figs. 18–20 for babble noise, and Figs. 21–23 for street noise.

Results using both the segmental SNR and STI measures, suggest that MME+SPU stimuli is of better quality than other AME-based methods. AME+SPU was indicated to be the next best for stimuli corrupted with white noise, but LogAME was shown to be better than other AME-based methods for stimuli corrupted with coloured noise. PESQ scores, on the other hand, were less consistent with LogAME and AME+SPU scores being higher at low SNRs. At higher SNRs (greater than around 10 dB), results for coloured noises showed MME+SPU to score higher.

## Appendix C. Comparison of modulation domain based enhancement methods in the presence of coloured noise

Experiments presented in Section 8 subjectively compared stimuli generated using modulation-domain based enhancement methods including MME+SPU, ModSSub, ModWiener and ModKalman. There we investigated processing of stimuli corrupted with AWGN. In this section, we extend those experiments to consider stimuli corrupted with additive coloured noises (at an SNR of 5 dB) such as babble and street noise. The quality of stimuli constructed by each method was again compared subjectively using AB listening tests, conducted under the same conditions as previously described. Listeners selected stimuli they preferred from 60 randomly played stimuli pairs. Experiments for each noise type were conducted in separate sessions. Five listeners participated in each test.

Results in terms of mean preference scores for babble and street noise are shown in Figs. 24 and 25, respectively. As can be seen for both noise types, MME+SPU stimuli were preferred over other modulation domain formulations, and ModWiener stimuli were the least preferred. For babble noise, the subjective preferences for ModKalman and ModSSub were quite close, while there was a greater preference for ModKalman when street noise was processed.

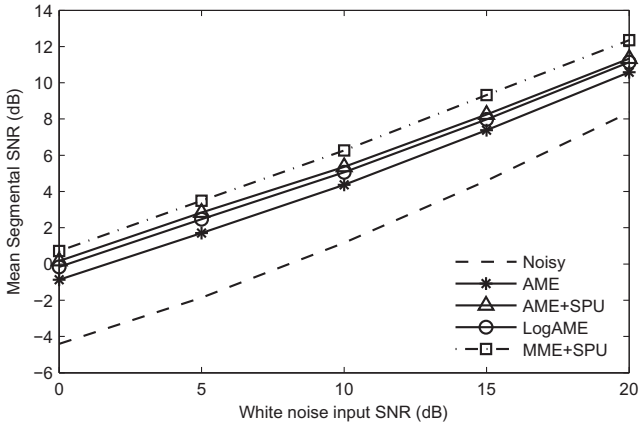


Fig. 15. Mean segmental SNR (dB) for: (a) noisy (degraded with AWGN at 5 dB); and stimuli generated by processing noisy stimuli with following treatment types: (b) MME+SPU; (c) AME; (d) AME+SPU; and (e) LogAME.

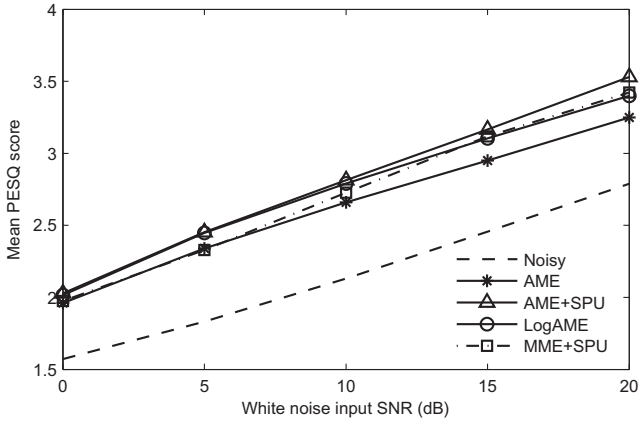


Fig. 16. Mean PESQ score for: (a) noisy (degraded with AWGN at 5 dB); and stimuli generated by processing noisy stimuli with following treatment types: (b) MME+SPU; (c) AME; (d) AME+SPU; and (e) LogAME.

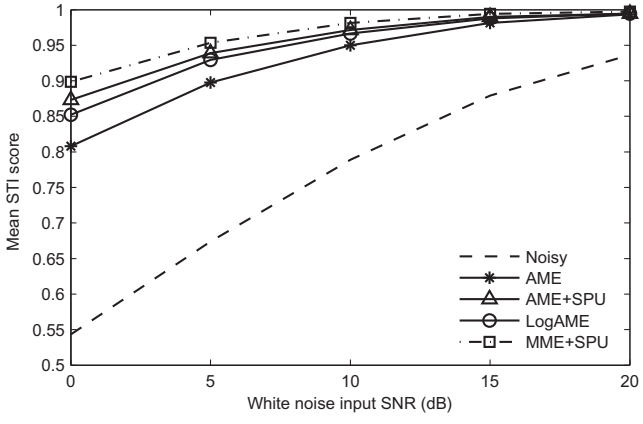


Fig. 17. Mean STI score for: (a) noisy (degraded with AWGN at 5 dB); and stimuli generated by processing noisy stimuli with following treatment types: (b) MME+SPU; (c) AME; (d) AME+SPU; and (e) LogAME.

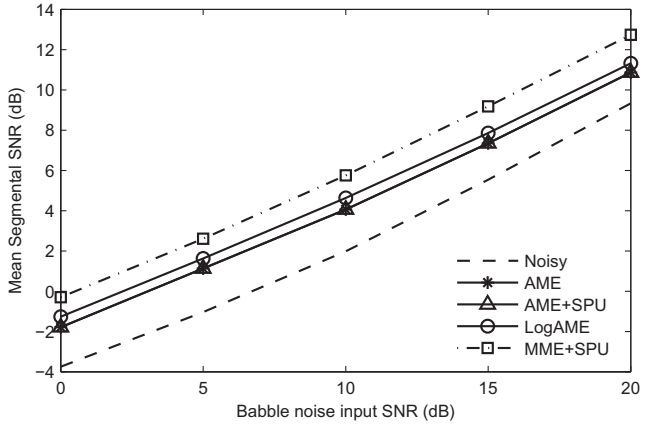


Fig. 18. Mean segmental SNR (dB) for: (a) noisy (degraded with babble noise at 5 dB); and stimuli generated by processing noisy stimuli with following treatment types: (b) MME+SPU; (c) AME; (d) AME+SPU; and (e) LogAME.

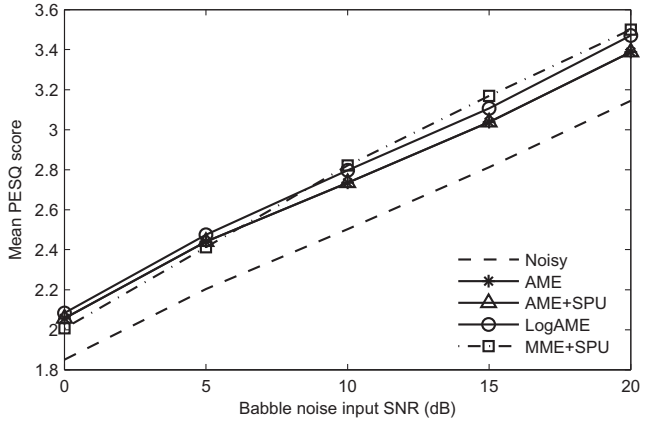


Fig. 19. Mean PESQ score for: (a) noisy (degraded with babble noise at 5 dB); and stimuli generated by processing noisy stimuli with following treatment types: (b) MME+SPU; (c) AME; (d) AME+SPU; and (e) LogAME.

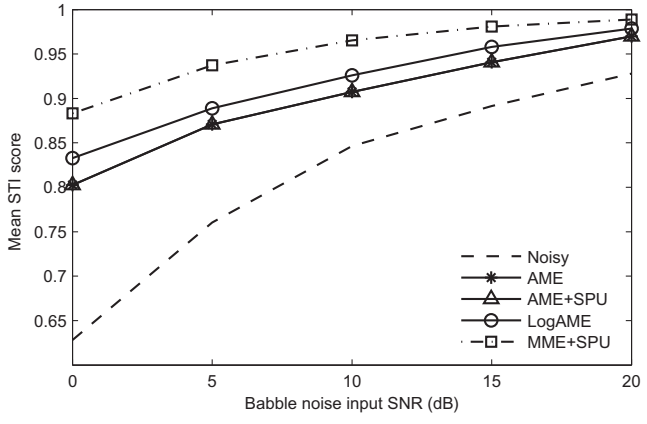


Fig. 20. Mean STI score for: (a) noisy (degraded with babble noise at 5 dB); and stimuli generated by processing noisy stimuli with following treatment types: (b) MME+SPU; (c) AME; (d) AME+SPU; and (e) LogAME.

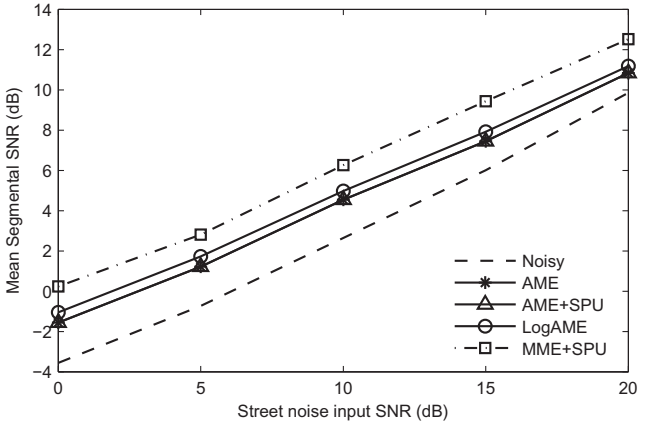


Fig. 21. Mean segmental SNR (dB) for: (a) noisy (degraded with street noise at 5 dB); and stimuli generated by processing noisy stimuli with following treatment types: (b) MME+SPU; (c) AME; (d) AME+SPU; and (e) LogAME.

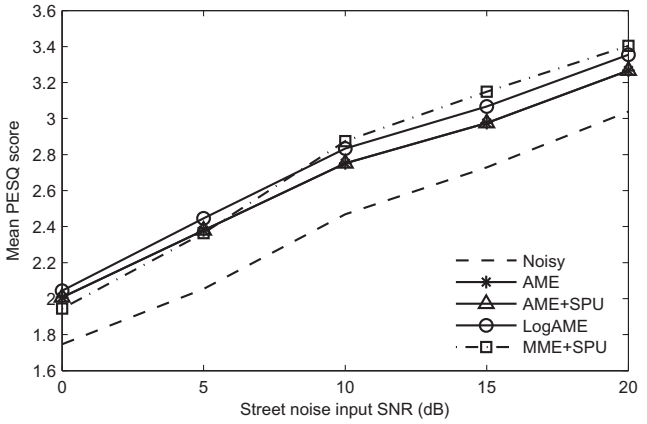


Fig. 22. Mean PESQ score for: (a) noisy (degraded with street noise at 5 dB); and stimuli generated by processing noisy stimuli with following treatment types: (b) MME+SPU; (c) AME; (d) AME+SPU; and (e) LogAME.

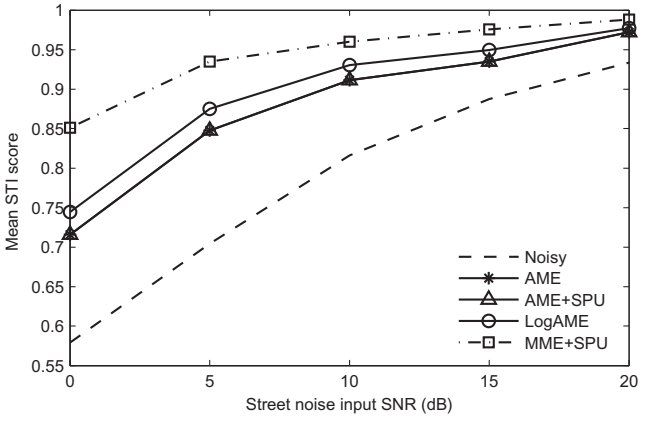


Fig. 23. Mean STI score for: (a) noisy (degraded with street noise at 5 dB); and stimuli generated by processing noisy stimuli with following treatment types: (b) MME+SPU; (c) AME; (d) AME+SPU; and (e) LogAME.

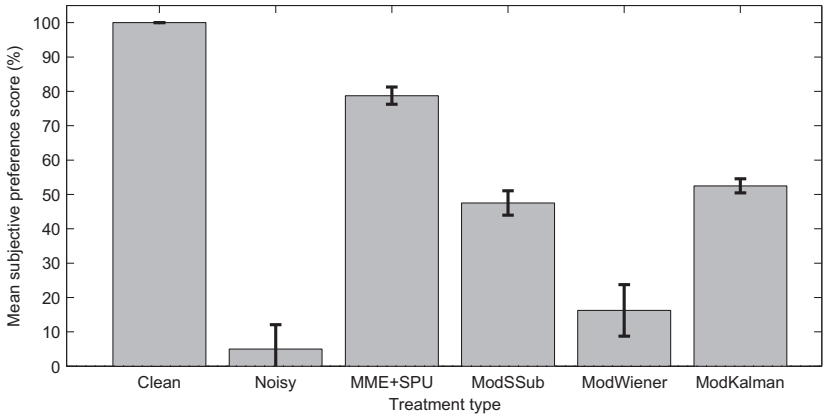


Fig. 24. Mean subjective preference scores (%) for (a) clean; (b) noisy (degraded with babble noise at 5 dB); and stimuli generated using the following treatment types: (c) MME+SPU; (d) ModSSub; (e) ModWiener; and (f) ModKalman.

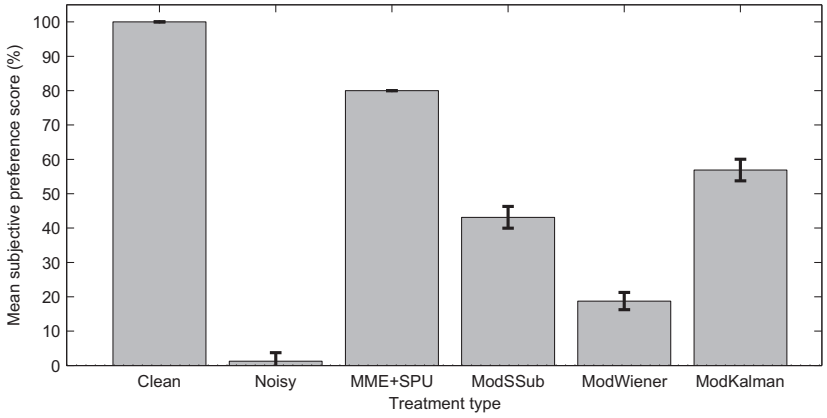


Fig. 25. Mean subjective preference scores (%) for (a) clean; (b) noisy (degraded with street noise at 5 dB); and stimuli generated using the following treatment types: (c) MME+SPU; (d) ModSSub; (e) ModWiener; and (f) ModKalman.

## Appendix D. Objective evaluation comparing MME+SPU to modulation domain formulations

An objective experiment was conducted to compare the quality of stimuli generated using MME+SPU to those generated using ModSSub, ModWiener, and ModKalman, for noisy stimuli corrupted with additive white, babble and street noise. All 30 stimuli of the Noizeus corpus were included in the test. Noisy stimuli were constructed by adding white, babble and street noises at SNRs of 0, 5, 10, 15, and 20 dB.

Again, the objective measures used to evaluate the quality of stimuli enhanced by each method were segmental SNR (Quackenbush et al., 1988), and PESQ (Rix et al., 2001), and the STI intelligibility measure (Drullman et al., 1994). Scores were calculated by comparing each noisy or enhanced stimuli with the corresponding clean stimuli. Mean scores were calculated for each enhancement method, noise type, and input SNR, as shown in Figs. 26–34, with each figure showing mean scores for the indicated noise type and quality measure. Results for segmental SNR and STI showed that MME+SPU generally scored higher than all other methods. For white and babble noise, ModKalman and ModSSub scored similarly, with ModSSub scoring only a little higher than ModKalman. For street noise, ModKalman had scores higher than ModSSub. In each case, ModWiener scored much lower than other methods. Again, more significant differences in scores were seen for tests on coloured noises. PESQ again showed less consistent results, with ModSSub scoring better at low SNRs and MME+SPU scoring better for higher SNRs.

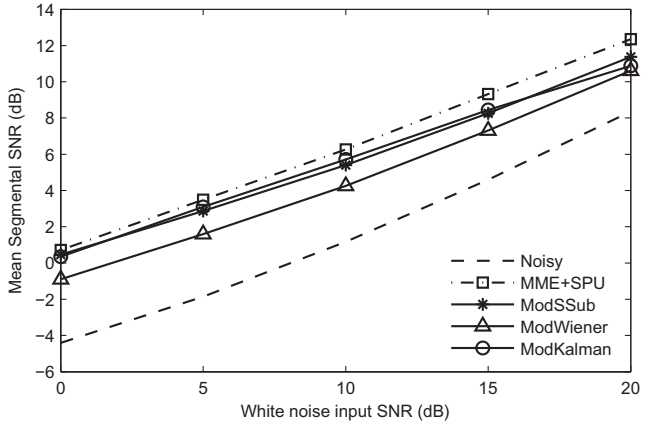


Fig. 26. Mean segmental SNR (dB) for: (a) noisy (degraded with AWGN at 5 dB); and stimuli generated by processing noisy stimuli with the following treatment types: (b) MME+SPU; (c) ModSSub; (d) ModWiener; and (e) ModKalman.

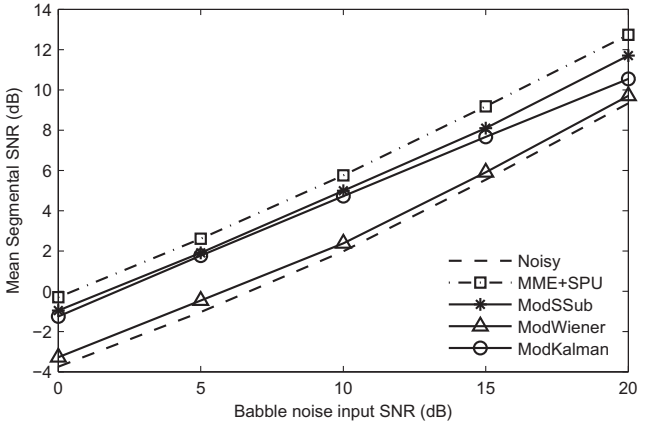


Fig. 29. Mean segmental SNR (dB) for: (a) noisy (degraded with babble noise at 5 dB); and stimuli generated by processing noisy stimuli with the following treatment types: (b) MME+SPU; (c) ModSSub; (d) ModWiener; and (e) ModKalman.

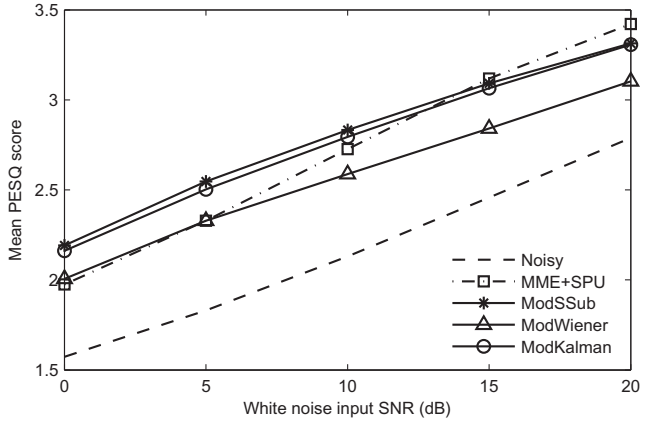


Fig. 27. Mean PESQ scores for: (a) noisy (degraded with AWGN at 5 dB); and stimuli generated by processing noisy stimuli with the following treatment types: (b) MME+SPU; (c) ModSSub; (d) ModWiener; and (e) ModKalman.

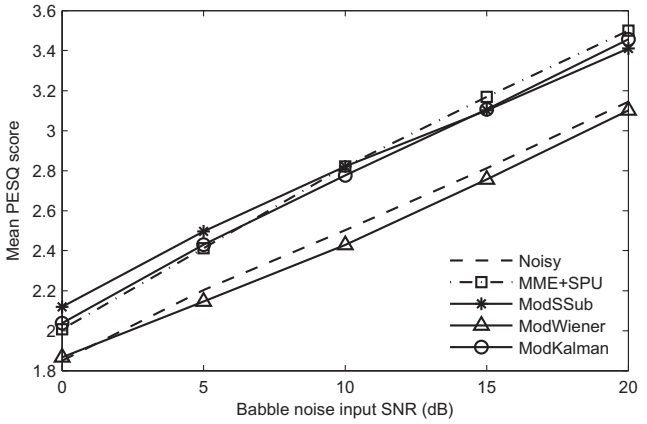


Fig. 30. Mean PESQ scores for: (a) noisy (degraded with babble noise at 5 dB); and stimuli generated by processing noisy stimuli with the following treatment types: (b) MME+SPU; (c) ModSSub; (d) ModWiener; and (e) ModKalman.

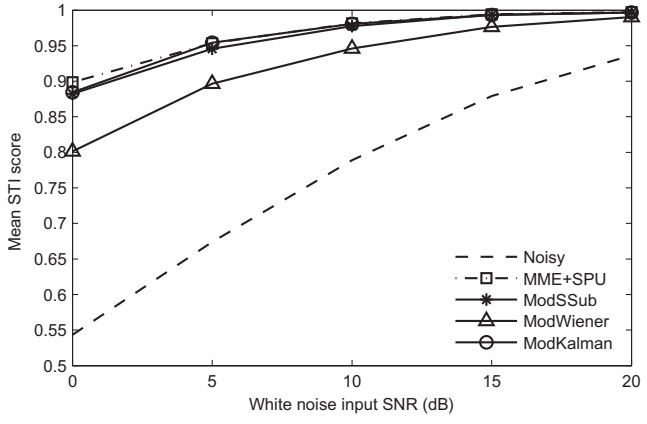


Fig. 28. Mean STI scores for: (a) noisy (degraded with AWGN at 5 dB); and stimuli generated by processing noisy stimuli with the following treatment types: (b) MME+SPU; (c) ModSSub; (d) ModWiener; and (e) ModKalman.

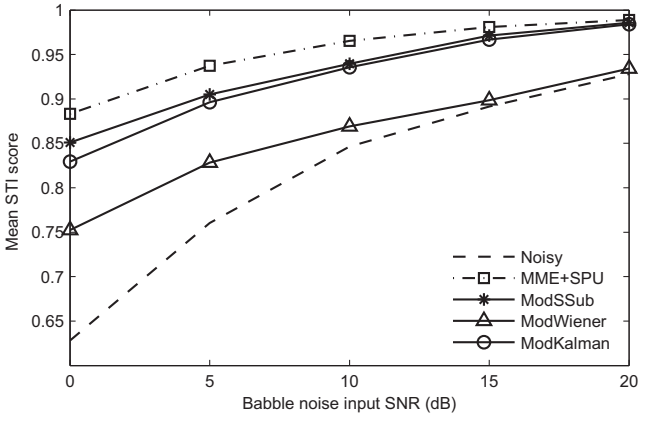


Fig. 31. Mean STI scores for: (a) noisy (degraded with babble noise at 5 dB); and stimuli generated by processing noisy stimuli with the following treatment types: (b) MME+SPU; (c) ModSSub; (d) ModWiener; and (e) ModKalman.

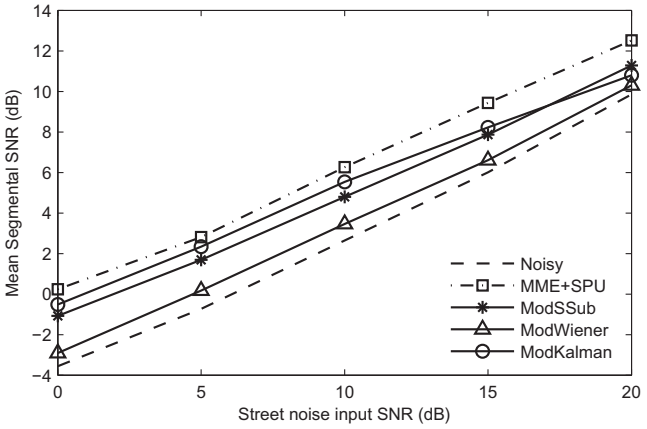


Fig. 32. Mean segmental SNR (dB) for: (a) noisy (degraded with street noise at 5 dB); and stimuli generated by processing noisy stimuli with the following treatment types: (b) MME+SPU; (c) ModSSub; (d) ModWiener; and (e) ModKalman.

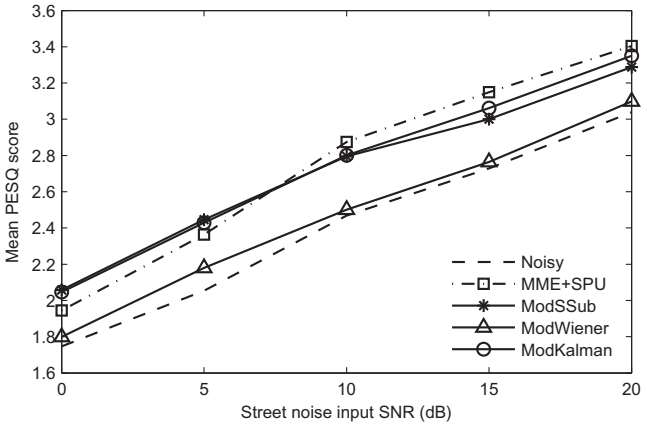


Fig. 33. Mean PESQ scores for: (a) noisy (degraded with street noise at 5 dB); and stimuli generated by processing noisy stimuli with the following treatment types: (b) MME+SPU; (c) ModSSub; (d) ModWiener; and (e) ModKalman.

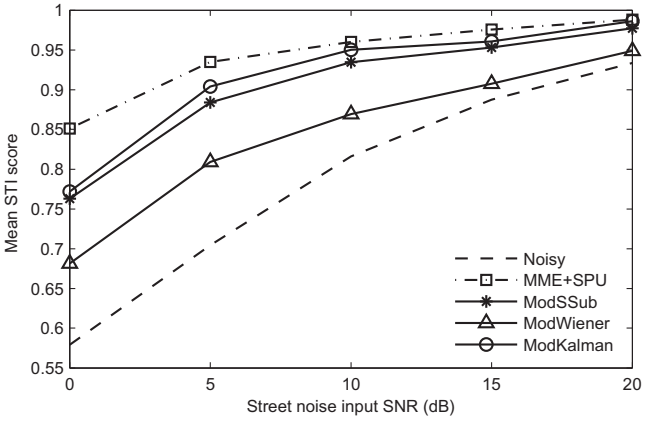


Fig. 34. Mean STI scores for: (a) noisy (degraded with street noise at 5 dB); and stimuli generated by processing noisy stimuli with the following treatment types: (b) MME+SPU; (c) ModSSub; (d) ModWiener; and (e) ModKalman.

## References


Atlas, L., Shamma, S., 2003. Joint acoustic and modulation frequency. EURASIP J. Appl. Signal Process. 2003 (7), 668–675.

Atlas, L., Li, Q., Thompson, J., 2004. Homomorphic modulation spectra. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process. (ICASSP), Vol. 2, Montreal, Quebec, Canada, pp. 761–764.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process (ICASSP), Vol. 4. Washington, DC, USA, pp. 208–211.

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

Breithaupt, C., Martin, R., 2011. Analysis of the decision-directed snr estimator for speech enhancement with respect to low-snr and transient conditions. IEEE Trans. Audio Speech Lang. Process. 19 (2), 277–289.

Cappe, O., 1994. Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor. IEEE Trans. Speech Audio Process. 2 (2), 345–349.

Cohen, I., 2005. Relaxed statistical model for speech enhancement and a priori SNR estimation. IEEE Trans. Speech Audio Process. 13 (5), 870–881.

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

Drullman, R., Festen, J., Plomp, R., 1994. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Amer. 95 (5), 2670–2680.

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

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

Falk, T.H., Chan, W.-Y., 2008. A non-intrusive quality measure of dereverberated speech. In: Proc. Internat. Workshop Acoust. Echo Noise Control.

Falk, T.H., Chan, W.-Y., 2010. Modulation spectral features for robust far-field speaker identification. IEEE Trans. Audio Speech Lang. Process. 18 (1), 90–100.

Falk, T., Stadler, S., Kleijn, W.B., Chan, W.-Y., 2007. Noise suppression based on extending a speech-dominated modulation band. In: Proc. ISCA Conf. Internat. Speech Commun. Assoc. (INTERSPEECH) Antwerp, Belgium, pp. 970–973.

Falk, T.H., Zheng, C., Chan, W.-Y., 2010. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans. Audio Speech Lang. Process. 18 (7), 1766–1774.

Gray, R., Buzo, A., Gray, A., Matsuyama, Y., 1980. Distortion measures for speech processing. IEEE Trans. Acoust. Speech Signal Process. ASSP-28 (4), 367–376.

Greenberg, S., Kingsbury, B., 1997. The modulation spectrogram: In persuit of an invariant representation of speech. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process. (ICASSP), Vol. 3. Munich, Germany, pp. 1647–1650.

Hermansky, H., Wan, E., Avendano, C., 1995. Speech enhancement based on temporal processing. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process (ICASSP), Vol. 1. Detroit, MI, USA, pp. 405–408.

Hu, Y., Loizou, P.C., 2007. Subjective comparison and evaluation of speech enhancement algorithms. Speech Comm. 49 (7–8), 588–601.

Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall, Upper Saddle River, New Jersey.

ITU-T P.835, 2007. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm:

Additional provisions for non-stationary noise suppressors. ITU-T P.835 Recommendation, Amendment 1.

Kim, D., 2004. A cue for objective speech quality estimation in temporal envelope representations. IEEE Signal Process. Lett. 11 (10), 849–852.

Kim, D., 2005. Anique: An auditory model for single-ended speech quality estimation. IEEE Trans. Speech Audio Process. 13 (5), 821–831.

Kingsbury, B., Morgan, N., Greenberg, S., 1998. Robust speech recognition using the modulation spectrogram. Speech Comm. 25 (1–3), 117–132.

Lim, J., Oppenheim, A., 1979. Enhancement and bandwith compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Loizou, P., 2005. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. IEEE Trans. Speech Audio Process. 13 (5), 857–869.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. Taylor and Francis, Boca Raton, FL.

Lyons, J., Paliwal, K., 2008. Effect of compressing the dynamic range of the power spectrum in modulation filtering based speech enhancement. In: Proc. ISCA Conf. Internat. Speech Commun. Assoc. (INTERSPEECH), Brisbane, Australia, pp. 387–390.

Martin, R., 1994. Spectral subtraction based on minimum statistics. In: Proc. EURASIP European Signal Process. Conf. (EUSIPCO), Edinburgh, Scotland, pp. 1182–1185.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9 (5), 504–512.

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

Paliwal, K., Wójcicki, K., 2008. Effect of analysis window duration on speech intelligibility. IEEE Signal Process. Lett. 15, 785–788.

Paliwal, K., Wójcicki, K., Schwerin, B., 2010. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Comm. 52 (5), 450–475.

Paliwal, K., Schwerin, B., Wójcicki, K., 2011. Role of modulation magnitude and phase spectrum towards speech intelligibility. Speech Comm. 53 (3), 327–339.

Picone, J., 1993. Signal modeling techniques in speech recognition. Proc. IEEE 81 (9), 1215–1247.

Quackenbush, S., Barnwell, T., Clements, M., 1988. Objective Measures of Speech Quality. Prentice Hall, Englewood Cliffs, NJ, USA.

Quatieri, T., 2002. Discrete-Time Speech Signal Processing: Principles and Practice. Prentice Hall, Upper Saddle River, NJ.

Rabiner, L.R., Schafer, R.W., 2010. Theory and Applications of Digital Speech Processing, first ed. Pearson Higher Education, Inc., Upper Saddle River, NJ, USA.

Rix, A., Beerends, J., Hollier, M., Hekstra, A., 2001. Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codes. ITU-T Recommendation P.862.

Scalart, P., Filho, J., 1996. Speech enhancement based on a priori signal to noise estimation. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process. (ICASSP), Vol. 2. Atlanta, Georgia, USA, pp. 629–632.

Shannon, B., Paliwal, K., 2006. Role of phase estimation in speech enhancement. In: Proc. Internat. Conf. on Spoken Language Process (ICSLP), Pittsburgh, PA, USA, pp. 1423–1426.

Sim, B.L., Tong, Y.C., Chang, J., Tan, C.T., 1998. A parametric formulation of the generalized spectral subtraction method. IEEE Trans. Speech Audio Process. 6 (4), 328–337.

So, S., Paliwal, K., 2011. Modulation-domain kalman filtering for single-channel speech enhancement. Speech Comm. 53 (6), 818–829.

Sohn, J., Kim, N.S., Sung, W., 1999. A statistical model-based voice activity detection. IEEE Signal Process. Lett. 6 (1), 1–3.

Thompson, J., Atlas, L., 2003. A non-uniform modulation transform for audio coding with increased time resolution. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Process (ICASSP), Vol. 5. Hong Kong, pp. 397–400.

Tyagi, V., McCowan, I., Bourland, H., Misra, H., 2003. On factorizing spectral dynamics for robust speech recognition. In: Proc. ISCA European Conf. on Speech Commun. and Technology (EUROSPEECH), Geneva, Switzerland, pp. 981–984.

Vary, P., Martin, R., 2006. Digital Speech Transmission: Enhancement, Coding And Error Concealment. John Wiley & Sons, Ltd., West Sussex, England.

Virag, N., 1999. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans. Speech Audio Process. 7 (2), 126–137.

Wang, D., Lim, J., 1982. The unimportance of phase in speech enhancement. IEEE Trans. Acoust. Speech Signal Process. ASSP-30 (4), 679–681.

Wiener, N., 1949. The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications. Wiley, New York.

Wu, S., Falk, T., Chan, W.-Y., 2009. Automatic recognition of speech emotion using long-term spectro-temporal features. In: Internat. Conf. on Digital Signal Process.

Zadeh, L., 1950. Frequency analysis of variable networks. Proc. IRE 38 (3), 291–299.