Single-channel speech enhancement using spectral subtraction in the short-time modulation domain

Kuldip Paliwal, Kamil Wójcicki and Belinda Schwerin

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

Abstract

In this paper we investigate the modulation domain as an alternative to the acoustic domain for speech enhancement. More specifically, we wish to determine how competitive the modulation domain is for spectral subtraction as compared to the acoustic domain. For this purpose, we extend the traditional analysis-modification-synthesis framework to include modulation domain processing. We then compensate the noisy modulation spectrum for additive noise distortion by applying the spectral subtraction algorithm in the modulation domain. Using an objective speech quality measure as well as formal subjective listening tests, we show that the proposed method results in improved speech quality. Furthermore, the proposed method achieves better noise suppression than the MMSE method. In this study, the effect of modulation frame duration on speech quality of the proposed enhancement method is also investigated. The results indicate that modulation frame durations of 180–280 ms, provide a good compromise between different types of spectral distortions, namely musical noise and temporal slurring. Thus given a proper selection of modulation frame duration, the proposed modulation spectral subtraction does not suffer from musical noise artifacts typically associated with acoustic spectral subtraction. In order to achieve further improvements in speech quality, we also propose and investigate fusion of modulation spectral subtraction with the MMSE method. The fusion is performed in the short-time spectral domain by combining the magnitude spectra of the above speech enhancement algorithms. Subjective and objective evaluation of the speech enhancement fusion shows consistent speech quality improvements across input SNRs.

Key words: Speech enhancement, modulation spectral subtraction, speech enhancement fusion, analysis-modification-synthesis (AMS), musical noise

1. Introduction

Speech enhancement aims at improving the quality of noisy speech. This is normally accomplished by reducing the noise (in such a way that the residual noise is not annoying to the listener), while minimising the speech distortion introduced during the enhancement process. In this paper we concentrate on the single-channel speech enhancement problem, where the signal is derived from a single microphone. This is especially useful in mobile communication applications, where only a single microphone is available due to cost and size considerations.

Many popular single-channel speech enhancement methods employ the analysis-modification-synthesis (AMS) framework (Allen, 1977; Allen and Rabiner, 1977; Crochiere, 1980; Portnoff, 1981; Griffin and Lim, 1984; Quatieri, 2002) to perform enhancement in the acoustic spectral domain (Loizou, 2007). The AMS framework consists of three stages: 1) the analysis stage, where the input speech is processed using the short-time Fourier transform (STFT) analysis; 2) the modification stage, where the noisy spectrum undergoes some kind of modification; and 3) the synthesis stage, where the inverse STFT is followed by the overlap-add synthesis to reconstruct the output signal. In this paper, we investigate speech enhancement in the modulation spectral domain by extending the acoustic AMS framework to include modulation domain processing.

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 standard (acoustic) frequency. More recently, Atlas et al. (2004) defined acoustic frequency as the axis of the first STFT of the input signal and modulation frequency as the independent variable of the second STFT transform. We therefore 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.

There is growing psychoacoustic and physiological evidence to support the significance of the modulation domain in the analysis of speech signals. Experiments of Bacon and Grantham (1989), for example, showed that there are channels in the auditory system which are tuned for the detection of modulation frequencies. Sheft and Yost (1990) showed that our perception of temporal dynamics corresponds to our perceptual filtering into modulation frequency channels and that faithful representation of these modulations is critical to our perception of speech. Experiments of Schreiner and Urbas (1986) showed that a neural representation of amplitude modulation is preserved through all levels of the mammalian auditory system, including the highest level of audition, the auditory cortex. Neurons in the auditory cortex are thought to decompose the acoustic spectrum into spectro-temporal modulation content (Mesgarani and Shamma, 2005), and are best driven by sounds that combine both spectral and temporal modulations (Kowalski et al., 1996; Shamma, 1996; Depireux et al., 2001).

Low frequency modulations of sound have been shown to be the fundamental carriers of information in speech (Atlas and Shamma, 2003). Drullman et al. (1994b,a), for example, investigated the importance of modulation frequencies for intelligibility by applying low-pass and high-pass filters to the temporal envelopes of acoustic frequency subbands. They showed frequencies between 4 and 16 Hz to be important for intelligibility, with the region around 4-5 Hz being the most significant. In a similar study, Arai et al. (1996) showed that applying band-pass filters between 1 and 16 Hz does not impair speech intelligibility.

While the envelope of the acoustic magnitude spectrum represents the shape of the vocal tract, the modulation spectrum represents how the vocal tract changes as a function of time. It is these temporal changes that convey most of the linguistic information (or intelligibility) of speech. In the above intelligibility studies, the lower limit of 1 Hz stems from the fact that the slow vocal tract changes do not convey much linguistic information. In addition, the lower limit helps to make speech communication more robust, since the majority of noises occurring in nature vary slowly as a function of time and hence their modulation spectrum is dominated by modulation frequencies below 1 Hz. The upper limit of 16 Hz is due to the physiological limitation on how fast the vocal tract is able to change with time.

Modulation domain processing has grown in popularity finding applications in areas such as speech coding (Atlas and Vinton, 2001; Thompson and Atlas, 2003; Atlas, 2003), speech recognition (Hermansky and Morgan, 1994; Nadeu et al., 1997; Kingsbury et al., 1998; Kanedera et al., 1999; Tyagi et al., 2003; Xiao et al., 2007; Lu et al., 2010), speaker recognition (Vuuren and Hermansky, 1998; Malayath et al., 2000; Kinnunen, 2006; Kinnunen et al., 2008), objective speech intelligibility evaluation (Steeneken and Houtgast, 1980; Payton and Braida, 1999; Greenberg and Arai, 2001; Goldsworthy and Greenberg, 2004; Kim, 2004) as well as speech enhancement. In the latter category, a number of modulation filtering methods have emerged. For example, Hermansky et al. (1995) proposed the band-pass filtering of the time trajectories of cubic-root compressed short-time power spectrum for enhancement of speech corrupted by additive noise. More recently in (Falk et al., 2007; Lyons and Paliwal, 2008), similar band-pass filtering was applied to the time trajectories of the short-time power spectrum for speech enhancement.

There are two main limitations associated with typical modulation filtering methods. First, they use a filter design based on the long-term properties of the speech modulation spectrum, while ignoring the properties of noise. As a consequence, they fail to eliminate noise components present within the speech modulation regions. Second, the modulation filter is fixed and applied to the entire signal, even though the properties of speech and noise change over time. In the proposed method, we attempt to address these limitations by processing the modulation spectrum on a frame-by-frame basis. In our approach, we assume the noise to be additive in nature and enhance noisy speech by applying spectral subtraction algorithm, similar to the one proposed by Berouti et al. (1979), in the modulation domain.

In this paper, we evaluate how competitive the modulation domain is for speech enhancement as compared to the acoustic domain. For this purpose, objective and subjective speech enhancement experiments were carried out. The results of these experiments demonstrate that the modulation domain is a useful alternative to the acoustic domain. We also investigate fusion of the proposed technique with the MMSE method for further speech quality improvements.

In the main body of this paper, we provide the 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 were found to be qualitatively similar. In order not to clutter the main body of this paper, we include the results for the coloured noises in Appendix C.

The rest of this paper is organised as follows. Section 2 details the traditional AMS-based speech processing. Section 3 presents details of the proposed modulation domain speech enhancement method along with the discussion of objective and subjective enhancement experiments and their results. Section 4 gives the details of the proposed speech enhancement fusion algorithm, along with experimental evaluation and results. Final conclusions are drawn in Section 5.

2. Acoustic analysis-modification-synthesis

Let us consider an additive noise model

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

where n is the discrete-time index, while $x(n)$, $s(n)$ and $d(n)$ denote discrete-time signals of noisy speech, clean speech and noise, respectively. Since speech can be assumed to be quasi-stationary, it is analysed frame-wise using the short-time Fourier analysis. The STFT of the corrupted speech signal $x(n)$ is given by

$$X(n,k) = \sum_{l=-\infty}^{\infty} x(l)w(n-l)\mathrm{e}^{-j2\pi kl/N}, \tag{2}$$

where k refers to the index of the discrete acoustic frequency, N is the acoustic frame duration (in samples) and $w(n)$ is an acoustic analysis window function.[1] In speech processing, the Hamming window with 20–40 ms duration is typically employed (Paliwal and Wójcicki, 2008). Using STFT analysis we can represent Eq. (1) as

$$X(n,k) = S(n,k) + D(n,k), \tag{3}$$

where $X(n,k)$, $S(n,k)$, and $D(n,k)$ are the STFTs of noisy speech, clean speech, and noise, respectively. Each of these can be expressed in terms of acoustic magnitude spectrum and acoustic phase spectrum. For instance, the STFT of the noisy speech signal can be written in polar form as

$$X(n,k) = \big|X(n,k)\big|\mathrm{e}^{j\angle X(n,k)}, \tag{4}$$

where $\big|X(n,k)\big|$ denotes the acoustic magnitude spectrum and $\angle X(n,k)$ denotes the acoustic phase spectrum.[2]

Traditional AMS-based speech enhancement methods modify, or enhance, only the noisy acoustic magnitude spectrum while keeping the noisy acoustic phase spectrum unchanged. The reason for this is that for Hamming-windowed frames (of 20–40 ms duration) the phase spectrum is considered unimportant for speech enhancement (Wang and Lim, 1982; Shannon and Paliwal, 2006). Such algorithms attempt to estimate the magnitude spectrum of clean speech. Let us denote the enhanced magnitude spectrum as $\big|\widehat{S}(n,k)\big|$, then the modified spectrum is constructed by combining $\big|\widehat{S}(n,k)\big|$ with the noisy phase spectrum, as follows

$$Y(n,k) = \big|\widehat{S}(n,k)\big|\mathrm{e}^{j\angle X(n,k)}. \tag{5}$$

[1] Note that in principle, Eq. (2) could be computed for every acoustic sample, however, in practice it is typically computed for each acoustic frame (and acoustic frames are progressed by some frame shift). We do not show this decimation explicitly in order to keep the mathematical notation concise.

[2] In our discussions, when referring to the magnitude, phase or (complex) spectra, the STFT modifier is implied unless otherwise stated. Also, wherever appropriate, we employ the acoustic and modulation modifiers to disambiguate between acoustic and modulation domains.

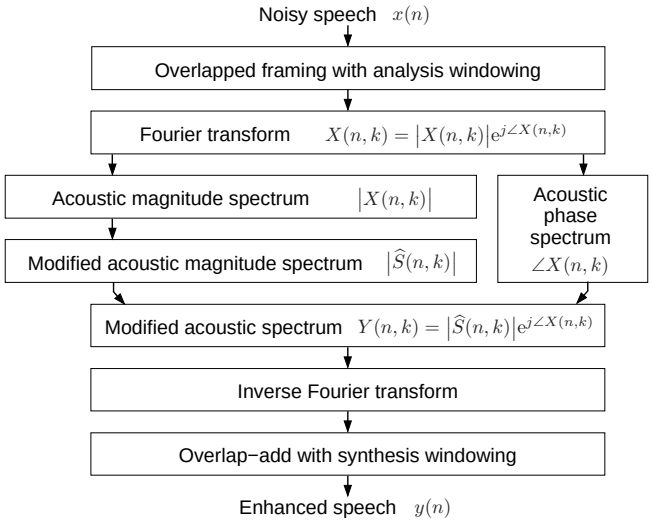

Fig. 1: *Block diagram of a traditional AMS-based acoustic domain speech enhancement procedure.*

The enhanced speech signal, $y(n)$, is constructed by taking the inverse STFT of the modified acoustic spectrum followed by least-squares overlap-add synthesis (Griffin and Lim, 1984; Quatieri, 2002):

$$y(n) = \frac{1}{W_0(n)} \sum_{l=-\infty}^{\infty} \left[\left(\frac{1}{N} \sum_{k=0}^{N-1} Y(l,k)\mathrm{e}^{j2\pi nk/N} \right) w_s(l-n) \right], \tag{6}$$

where $w_s(n)$ is the synthesis window function, and $W_0(n)$ is given by

$$W_0(n) = \sum_{l=-\infty}^{\infty} w_s^2(l-n). \tag{7}$$

In the present study, as the synthesis window we employ the modified Hanning window (Griffin and Lim, 1984), given by

$$w_s(n) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi(n+0.5)}{N}\right), & 0 \le n < N \\ 0, & \text{otherwise} \end{cases}. \tag{8}$$

Note that the use of the modified Hanning window means that $W_0(n)$ in Eq. (7) is constant (*i.e.*, independent of n).

A block diagram of a traditional AMS-based speech enhancement framework is shown in Fig. 1.

3. Modulation spectral subtraction

3.1. Introduction

Classical spectral subtraction (Boll, 1979; Berouti et al., 1979; Lim and Oppenheim, 1979) is an intuitive and effective speech enhancement method for the removal of additive noise. Spectral subtraction does, however, suffer from perceptually annoying spectral artifacts refered to as musical noise. Many approaches that attempt to address this problem have been investigated in the literature (*e.g.*,

$$\left|\widehat{\mathcal{S}}(\eta,k,m)\right| = \begin{cases} \left(\left|\mathcal{X}(\eta,k,m)\right|^{\gamma} - \rho\left|\widehat{\mathcal{D}}(\eta,k,m)\right|^{\gamma}\right)^{\frac{1}{\gamma}}, & \text{if } \left|\mathcal{X}(\eta,k,m)\right|^{\gamma} - \rho\left|\widehat{\mathcal{D}}(\eta,k,m)\right|^{\gamma} \geq \beta\left|\widehat{\mathcal{D}}(\eta,k,m)\right|^{\gamma} \\ \left(\beta\left|\widehat{\mathcal{D}}(\eta,k,m)\right|^{\gamma}\right)^{\frac{1}{\gamma}}, & \text{otherwise} \end{cases} \tag{11}$$

Vaseghi and Frayling-Cork, 1992; Cappe, 1994; Virag, 1999; Hasan et al., 2004; Hu and Loizou, 2004; Lu, 2007).

In this section, we propose to apply the spectral subtraction algorithm in the short-time modulation domain. Traditionally, the modulation spectrum has been computed as the Fourier transform of the intensity envelope of a band-pass filtered signal (*e.g.*, Houtgast and Steeneken, 1985; Drullman et al., 1994a; Goldsworthy and Greenberg, 2004). The method proposed in our study, however, uses the short-time Fourier transform (STFT) instead of band-pass filtering. In the acoustic STFT domain, the quantity closest to the intensity envelope of a band-pass filtered signal is the magnitude-squared spectrum. However, in the present paper we use the time trajectories of the short-time acoustic magnitude spectrum for the computation of the short-time modulation spectrum. This choice is motivated from more recently reported papers dealing with modulation-domain processing based speech applications (Falk et al., 2007; Kim, 2005), and is also justified empirically in Appendix B. Once the modulation spectrum is computed, spectral subtraction is done in the modulation magnitude-squared domain. Empirical justification for use of modulation magnitude-squared spectra is also given in Appendix B.

The proposed approach is then evaluated through both objective and subjective speech enhancement experiments as well as through spectrogram analysis. We show that given a proper selection of modulation frame duration, the proposed method results in improved speech quality and does not suffer from musical noise artifacts.

3.2. Procedure

The proposed speech enhancement method extends the traditional AMS-based acoustic domain enhancement to the modulation domain. To achieve this, each frequency component of the acoustic magnitude spectra, obtained during the analysis stage of the acoustic AMS procedure outlined in Section 2, is processed frame-wise across time using a secondary (modulation) AMS framework. Thus the modulation spectrum is computed using STFT analysis as follows

$$\mathcal{X}(\eta,k,m) = \sum_{l=-\infty}^{\infty} \left|X(l,k)\right| v(\eta-l) \mathrm{e}^{-j2\pi ml/M}, \tag{9}$$

where η is the acoustic frame number,[3] k refers to the index of the discrete acoustic frequency, m refers to the index of the discrete modulation frequency, M is the modulation frame duration (in terms of acoustic frames) and $v(\eta)$ is a modulation analysis window function. The resulting spectra can be expressed in polar form as

$$\mathcal{X}(\eta,k,m) = \left|\mathcal{X}(\eta,k,m)\right| \mathrm{e}^{j\angle\mathcal{X}(\eta,k,m)}, \tag{10}$$

where $\left|\mathcal{X}(\eta,k,m)\right|$ is the modulation magnitude spectrum and $\angle\mathcal{X}(\eta,k,m)$ is the modulation phase spectrum.

We propose to replace $\left|\mathcal{X}(\eta,k,m)\right|$ with $\left|\widehat{\mathcal{S}}(\eta,k,m)\right|$, where $\left|\widehat{\mathcal{S}}(\eta,k,m)\right|$ is an estimate of clean modulation magnitude spectrum obtained using a spectral subtraction rule similar to the one proposed by Berouti et al. (1979) and given by Eq. (11). In Eq. (11), ρ denotes the subtraction factor that governs the amount of over-subtraction; β is the spectral floor parameter used to set spectral magnitude values falling below the spectral floor, $\left(\beta\left|\widehat{\mathcal{D}}(\eta,k,m)\right|^{\gamma}\right)^{\frac{1}{\gamma}}$, to that spectral floor; and γ determines the subtraction domain, *e.g.*, for γ set to unity the subtraction is performed in the magnitude spectral domain, while for $\gamma = 2$ the subtraction is performed in the magnitude-squared spectral domain.

The estimate of the modulation magnitude spectrum of the noise, denoted by $\left|\widehat{\mathcal{D}}(\eta,k,m)\right|$, is obtained based on a decision from a simple voice activity detector (VAD) (Loizou, 2007), applied in the modulation domain. The VAD classifies each modulation domain segment as either 1 (*speech present*) or 0 (*speech absent*), using the following binary rule

$$\Phi(\eta,k) = \begin{cases} 1, & \text{if } \phi(\eta,k) \geq \theta \\ 0, & \text{otherwise} \end{cases}, \tag{12}$$

where $\phi(\eta,k)$ denotes a modulation segment SNR computed as follows

$$\phi(\eta,k) = 10\log_{10}\left(\frac{\sum_{m}\left|\mathcal{X}(\eta,k,m)\right|^2}{\sum_{m}\left|\widehat{\mathcal{D}}(\eta-1,k,m)\right|^2}\right) \tag{13}$$

[3] Note that in principle, Eq. (9) could be computed for every acoustic frame, however, in practice we compute it for every modulation frame. We do not show this decimation explicitly in order to keep the mathematical notation concise.

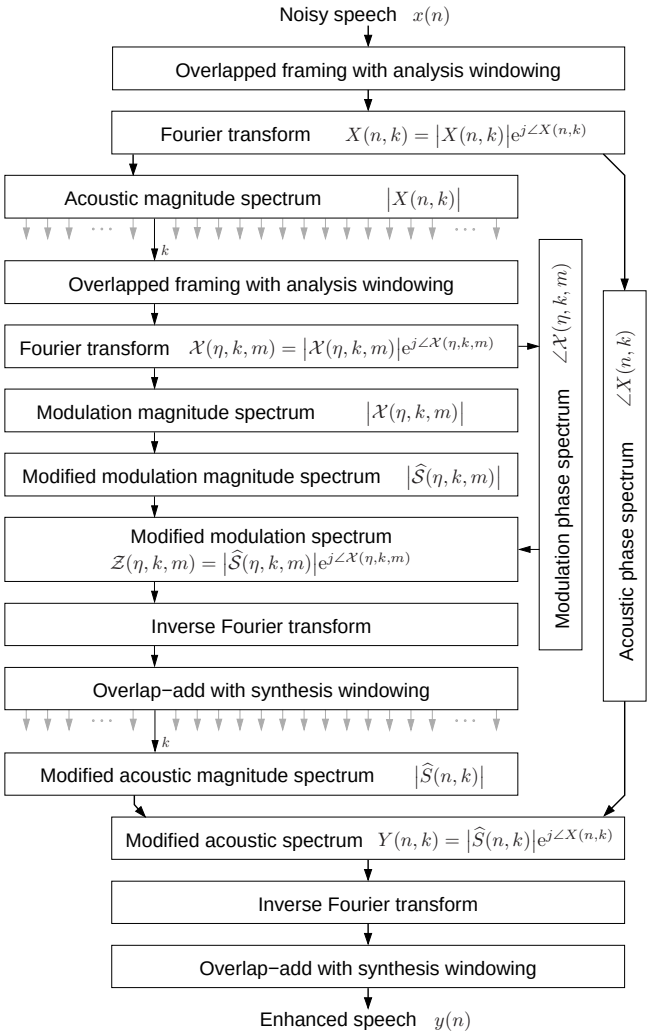

Fig. 2: *Block diagram of the proposed AMS-based modulation domain speech enhancement procedure.*

and θ is an empirically determined speech presence threshold. The noise estimate is updated during speech absence using the following averaging rule (Virag, 1999)

$$\left|\widehat{\mathcal{D}}(\eta,k,m)\right|^{\gamma} = \lambda\left|\widehat{\mathcal{D}}(\eta-1,k,m)\right|^{\gamma} + (1-\lambda)\left|\mathcal{X}(\eta,k,m)\right|^{\gamma}, \tag{14}$$

where λ is a forgetting factor chosen depending on the stationarity of the noise.[4]

The modified modulation spectrum is produced by combining $\left|\widehat{\mathcal{S}}(\eta,k,m)\right|$ with the noisy modulation phase spectrum as follows

$$\mathcal{Z}(\eta,k,m) = \left|\widehat{\mathcal{S}}(\eta,k,m)\right|e^{j\angle\mathcal{X}(\eta,k,m)}. \tag{15}$$

Note that unlike the acoustic phase spectrum, the modulation phase spectrum does contain useful information (Hermansky et al., 1995). In the present work, we keep $\angle\mathcal{X}(\eta,k,m)$ unchanged, however, future work will investigate approaches that can be used to enhance it. In the present study, we obtain the estimate of the modified acoustic magnitude spectrum $\left|\widehat{S}(n,k)\right|$, by taking the inverse STFT of $\mathcal{Z}(\eta,k,m)$ followed by overlap-add with synthesis windowing. A block diagram of the proposed approach is shown in Fig. 2.

3.3. Experiments

In this section we detail objective and subjective speech enhancement experiments that assess the suitability of modulation spectral subtraction for speech enhancement.

3.3.1. Speech corpus

In our experiments we employ the Noizeus speech corpus (Loizou, 2007; Hu and Loizou, 2007).[5] Noizeus is composed of 30 phonetically-balanced sentences belonging to six speakers, three males and three females. The corpus is sampled at 8 kHz and filtered to simulate receiving frequency characteristics of telephone handsets. Noizeus comes with non-stationary noises at different SNRs. For our experiments we keep the clean part of the corpus and generate noisy stimuli by degrading the clean stimuli with additive white Gaussian noise (AWGN) at various SNRs. The noisy stimuli are constructed such that they begin with a noise only section long enough for (initial) noise estimation in both acoustic and modulation domains (approx. 500 ms).

3.3.2. Stimuli types

Modulation spectral subtraction (ModSpecSub) stimuli were constructed using the procedure detailed in Section 3.2. The acoustic frame duration was set to 32 ms, with an 8 ms frame shift and the modulation frame duration was set to 256 ms, with a 32 ms frame shift. Note that modulation frame durations between 180 ms and 280 ms were found to work well. However, at shorter durations the musical noise was present, while at longer durations a slurring effect was observed. The duration of 256 ms was chosen as a good compromise. A more detailed look at the effect of modulation frame duration on speech quality of ModSpecSub stimuli is presented in Appendix A. The Hamming window was used for both the acoustic and modulation analysis windows. The FFT-analysis length was set to $2N$ and $2M$ for the acoustic and modulation AMS frameworks, respectively. The value of the subtraction parameter ρ was selected as described in (Berouti et al., 1979). The spectral floor parameter β was set to 0.002. Magnitude-squared spectral subtraction was used in the modulation domain, *i.e.*, $\gamma=2$. The speech presence threshold θ was set to 3 dB. The forgetting factor λ was set to 0.98. Griffith and Lim's method for windowed

[4]Note that due to the temporal processing over relatively long frames, the use of VAD for noise estimation will not achieve truly adaptive noise estimates. This is one of the limitations of the proposed method as discussed in Section 3.4.

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

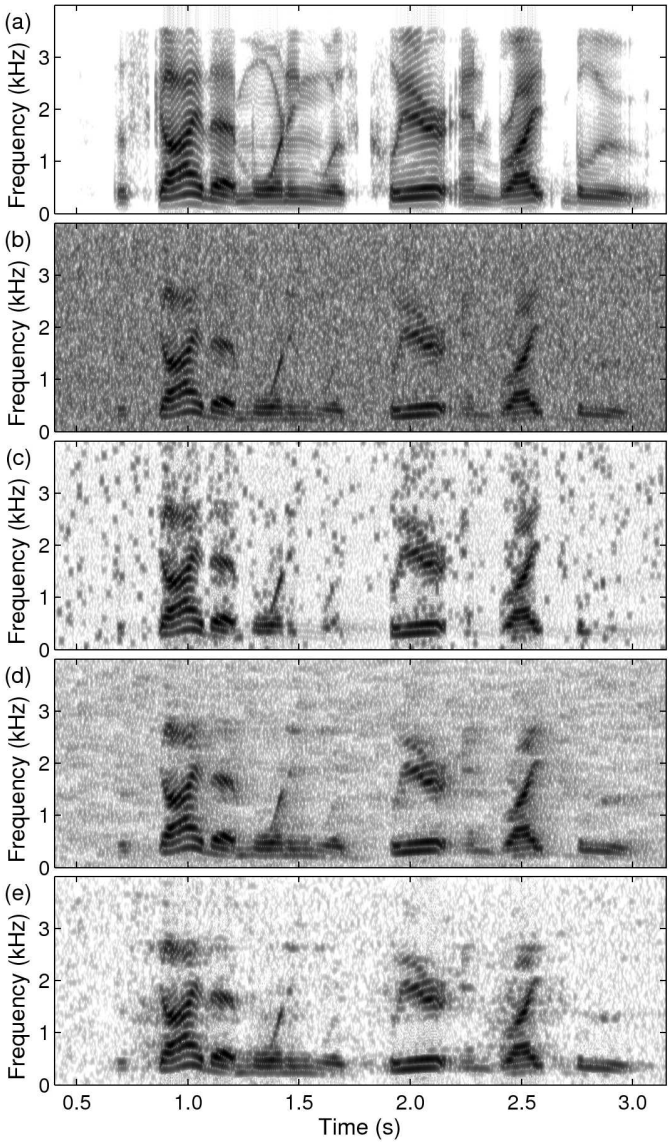

Fig. 3: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by AWGN at* 5 *dB SNR (PESQ:* 1.80*); as well as the noisy speech enhanced using: (c) acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) (PESQ:* 2.07*); (d) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.26*); and (e) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.42*).*

overlap-add synthesis (Griffin and Lim, 1984) was used for both acoustic and modulation syntheses.

For our experiments we have also generated stimuli using two popular speech enhancement methods, namely the acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) and the MMSE method (Ephraim and Malah, 1984). Publicly available reference implementation of these methods (Loizou, 2007) was employed in our study. In the SpecSub method, the subtraction was performed in the magnitude-squared spectral domain, with the noise spectrum estimates obtained through recursive averaging of non-speech frames. Speech presence or absence was determined using a voice activity detection (VAD) algorithm, based on a simple segmental SNR measure (Loizou, 2007). In the MMSE method, optimal estimates (in the minimum mean square error sense) of the short-time spectral amplitudes were computed. The decision-directed approach was used for the *a priori* SNR estimation, with the smoothing factor α set to 0.98.[6] In the MMSE method, noise spectrum estimates were computed from non-speech frames using recursive averaging with speech presence or absence determined using a log-likelihood ratio based VAD (Loizou, 2007). Further details on the implementation of both methods are given in (Loizou, 2007).

In addition to the ModSpecSub, SpecSub, and MMSE stimuli, clean and noisy speech stimuli were also included in our experiments. Example spectrograms for the above stimuli are shown in Fig. 3.[7,8]

3.3.3. Objective experiment

The objective experiment was carried out over the Noizeus corpus for AWGN at 0, 5, 10 and 15 dB SNR. Perceptual evaluation of speech quality (PESQ) (Rix et al., 2001) was used to predict mean opinion scores for the stimuli types outlined in Section 3.3.2.

3.3.4. Subjective experiment

The subjective evaluation was in a form of AB listening tests that determine method preference. Two Noizeus sentences (sp10 and sp27) belonging to male and female speakers were included. AWGN at 5 dB SNR was investigated. The stimuli types detailed in Section 3.3.2 were included. Fourteen English speaking listeners participated in this experiment. None of the participants reported any hearing defects. The listening tests were conducted in a quiet room. The participants were familiarised with the task during a short practice session. The actual test consisted of 40 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 labeled options on a digital 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

[6] Please note that in the decision-directed approach for the *a priori* SNR estimation, the smoothing parameter α has a significant effect on the type and intensity of the residual noise present in the enhanced speech (Cappe, 1994). While the MMSE stimuli used in the experiments presented in the main body of this paper were constructed with α set to 0.98, a supplementary examination of the effect of α on speech quality of the MMSE stimuli is provided in Appendix D.

[7] Note that all spectrograms, presented in this study, have the dynamic range set to 60 dB. The highest spectral peaks are shown in black, while the lowest spectral valleys ($\geq$60 dB below the highest peaks) are shown in white. Shades of gray are used in-between.

[8] The audio stimuli files are available on-line from the following url: `http://maxwell.me.gu.edu.au/spl/research/modspecsub/`.

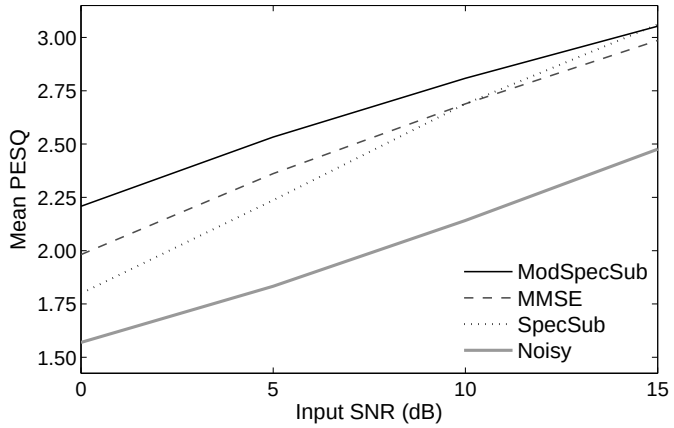

Fig. 4: *Speech enhancement results for the objective experiment detailed in Section 3.3.3. The results are in terms of mean PESQ scores as a function of input SNR (dB) for AWGN over the Noizeus corpus.*

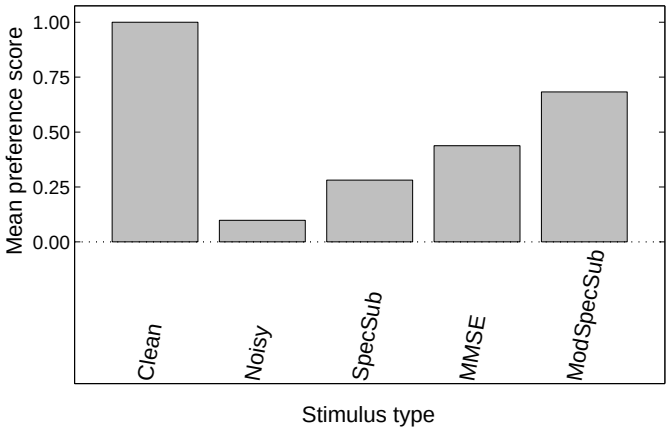

Fig. 5: *Speech enhancement results for the subjective experiment detailed in Section 3.3.4. The results are in terms of mean preference scores for AWGN at 5 dB SNR for two Noizeus utterances (sp10 and sp17).*

not prefer one stimulus over the other. Pairwise scoring was employed, with a score of +1 awarded to the prefered method and +0 to the other. For a similar preference response each method was awarded a score of +0.5. The participants were allowed to re-listen to stimuli if required. The responses were collected via keyboard. No feedback was given.

3.4. Results and discussion

The results of the objective experiment, in terms of mean PESQ scores, are shown in Fig. 4. The proposed method performs consistently well across the SNR range, with particular improvements shown for stimuli with lower input SNRs. The MMSE method showed the next best performance, with all enhancement methods achieving comparable results at 15 dB SNR.

The results of the subjective experiment are shown in Fig. 5. The subjective results are in terms of average preference scores. A score of one for a particular stimuli type, indicates that the stimuli type was always preferred. On the other hand, a score of zero means that the stimuli type was never preferred. Subjective results show that the clean stimuli were always preferred, while the noisy stimuli were the least preferred. Of the enhancement methods tested, ModSpecSub achieved significantly better preference scores ($p < 0.01$) than MMSE and SpecSub, with SpecSub being the least preferred. Notably, the subjective results are consistent with the corresponding objective results (AWGN at 5 dB SNR). More detailed subjective results, in the form of a method preference confusion matrix are shown in Table 1(a) of Appendix F.

The above results can be explained as follows. The acoustic spectral subtraction introduces spurious peaks scattered throughout the non-speech regions of the acoustic magnitude spectrum. At a given acoustic frequency bin, these spectral magnitude values vary over time (*i.e.*, from frame to frame) causing audibly annoying sounds referred to as the musical noise. This is clearly visible in the SpecSub spectrogram of Fig. 3(c). On the other hand, the proposed method subtracts the modulation magnitude spectrum estimate of the noise from the modulation magnitude spectrum of the noisy speech along each acoustic frequency bin. While some spectral magnitude variation is still present in the resulting acoustic spectrum, the residual peaks have much smaller magnitudes. As a result, ModSpecSub stimuli do not suffer from the musical noise audible in SpecSub stimuli (given a proper selection of modulation frame duration as discussed in Appendix A). This can be seen by comparing spectrograms in Fig. 3(c) and Fig. 3(e).

The MMSE method does not suffer from the problem of musical noise (Cappe, 1994; Loizou, 2007), however, it does not suppress background noise as effectively as the proposed method. This can be seen by comparing spectrograms in Fig. 3(d) and Fig. 3(e). In addition, listeners found the residual noise present after MMSE enhancement to be perceptually distracting. On the other hand, the proposed method uses larger frame durations in order to avoid musical noise (see Appendix A). As a result, stationarity has to be assumed over a larger duration. This causes temporal slurring distortion. This kind of distortion is mostly absent in the MMSE stimuli constructed with smoothing factor α set to 0.98. The need for longer frame durations in the ModSpecSub method also means that larger non-speech durations are required to update noise estimates. This makes the proposed method less adaptive to rapidly changing noise conditions. Finally, the additional processing involved in the computation of the modulation spectrum for each acoustic frequency bin, adds to the computational expense of the ModSpecSub method. In the next section, we propose to combine ModSpecSub and MMSE algorithms in the acoustic STFT domain in order to reduce some of their unwanted effects and to achieve further improvements in speech quality.

We would also like to emphasise that the phase spectrum

$$\left|\widehat{S}(n,k)\right| = \left(\Psi(\sigma_n)\left|Y_{MMSE}(n,k)\right|^{\gamma} + \left(1-\Psi(\sigma_n)\right)\left|Y_{ModSpecSub}(n,k)\right|^{\gamma}\right)^{\frac{1}{\gamma}} \quad (16)$$

plays a more important role in the modulation domain than in the acoustic domain (Hermansky et al., 1995). While in this preliminary study we keep the noisy modulation phase spectrum unchanged, in future work further improvements may be possible by also processing the modulation phase spectrum.

4. Speech enhancement fusion

4.1. Introduction

In the previous section, we have proposed the application of spectral subtraction in the short-time modulation domain. We have shown that modulation spectral subtraction (ModSpecSub) improves speech quality and does not suffer from musical noise artifacts associated with acoustic spectral subtraction. ModSpecSub does, however, introduce temporal slurring distortion. On the other hand, the MMSE method does not suffer from the slurring distortion, but it is less effective at removal of background noise. In this section, we attempt to exploit the strengths of the two methods, while trying to avoid their weaknesses, by combining (or fusing) them in the acoustic STFT domain. We then evaluate the proposed approach against methods investigated in Section 3.

4.2. Procedure

Let $\left|Y_{MMSE}(n,k)\right|$ denote the acoustic STFT magnitude spectrum of speech enhanced using the MMSE method (Ephraim and Malah, 1984) and $\left|Y_{ModSpecSub}(n,k)\right|$ be the acoustic STFT magnitude spectrum of speech enhanced using the ModSpecSub method. In the following discussions we will refer to these as the MMSE magnitude spectrum and the ModSpecSub magnitude spectrum, respectively. We propose to fuse ModSpecSub with the MMSE method by combining their magnitude spectra as given by Eq. (16), where $\Psi(\sigma_n)$ is the fusion-weighting function, σ_n is the *a posteriori* SNR (Ephraim and Malah, 1984) of the nth acoustic segment averaged across frequency, and γ determines the fusion domain (*i.e.*, for $\gamma=1$ the fusion is performed in the magnitude spectral domain, while for $\gamma=2$ the fusion is performed in the magnitude-squared spectral domain).

4.3. Fusion-weighting function

Empirically determined fusion-weighting function, employed in this study and shown in Fig. 6, is given by

$$\Psi(\sigma) = \begin{cases} 0, & \text{if } g(\sigma) \leq 2 \\ \frac{g(\sigma)-2}{14}, & \text{if } 2 < g(\sigma) < 16 \\ 1, & \text{if } g(\sigma) \geq 16 \end{cases}, \quad (17)$$

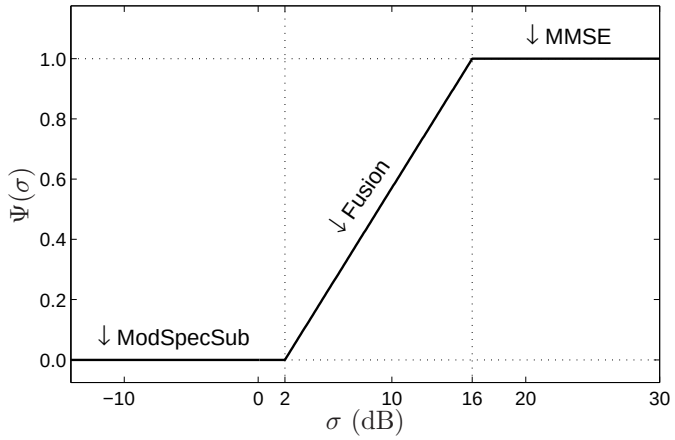

Fig. 6: *Fusion-weighting function, $\Psi(\sigma)$, as a function of average* a posteriori *SNR, σ,* as used in the construction of Fusion stimuli for experiments detailed in Section 4.4.

where $g(\sigma) = 10\log_{10}(\sigma)$. The above weighting favours the ModSpecSub method at low segment SNRs (*i.e.*, during speech pauses and low energy speech regions), while stronger emphasis is given to the MMSE method at high segment SNRs (*i.e.*, during high energy speech regions). Thus for $\Psi(\sigma)=0$ only ModSpecSub magnitude spectrum is used, for $0<\Psi(\sigma)<1$ a combination of ModSpecSub and MMSE magnitude spectra is employed, while for $\Psi(\sigma)=1$ only MMSE magnitude spectrum is used. This allows us to exploit the respective strengths of the two enhancement methods.

4.4. Experiments

Objective and subjective speech enhancement experiments were conducted to evaluate the performance of the proposed approach against methods investigated in Section 3. The details of these experiments are similar to those presented in Section 3.3, with the differences outlined below.

4.4.1. Stimuli types

Fusion stimuli were included in addition to the stimuli listed in Section 3.3.2. The Fusion stimuli were constructed using the procedure outlined in Section 4.2. The fusion was performed in magnitude-squared spectral domain, *i.e.*, $\gamma = 2$. Fusion-weighting function defined in Section 4.3 was employed. The settings used to generate MMSE and ModSpecSub magnitude spectra in the proposed fusion were the same as those used for their standalone counterparts.

Figure 7 gives a further insight into how the proposed algorithm works. Clean and noisy speech spectrograms

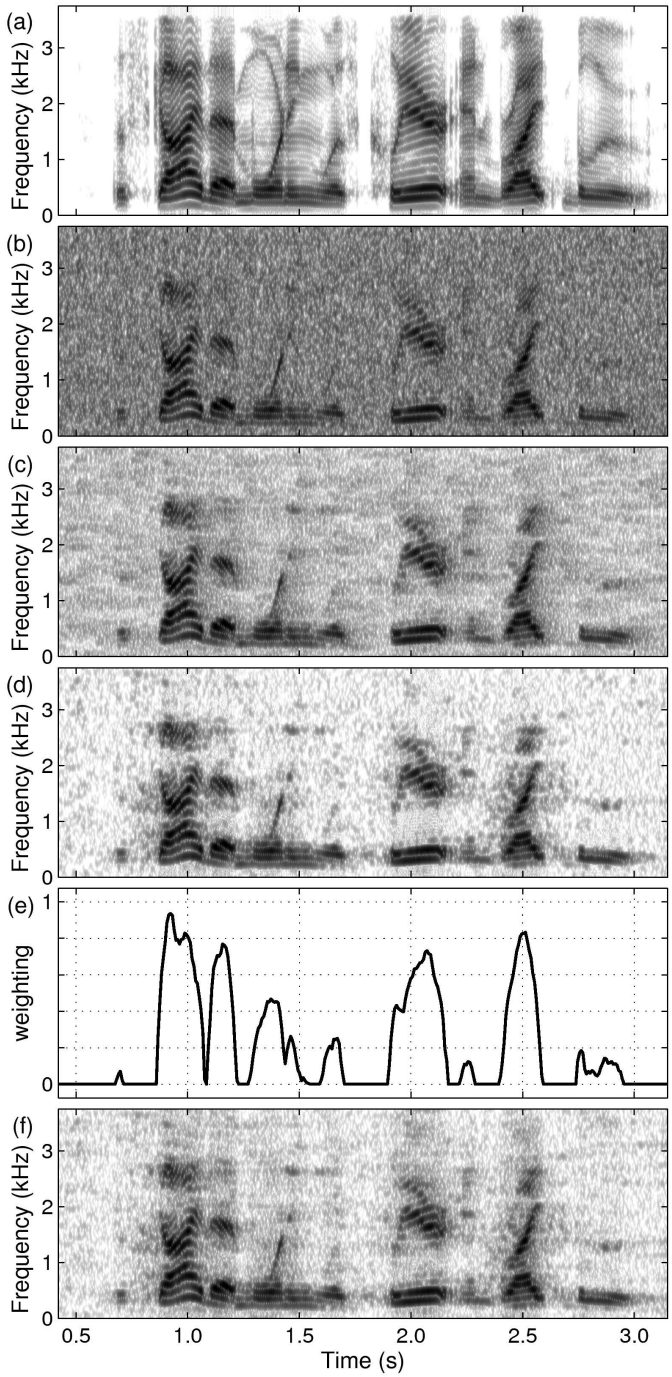

Fig. 7: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by AWGN at* 5 *dB SNR (PESQ:* 1.80*); as well as the noisy speech enhanced using: (c) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.26*); (d) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.42*); and (f) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.51*); as well as (e) fusion-weighting function $\Psi(\sigma_n)$ computed across time for the noisy utterance shown in the spectrogram of sub-plot (b).*

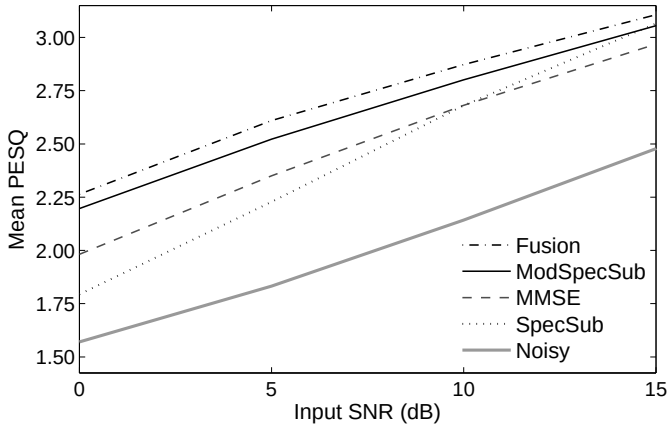

Fig. 8: *Speech enhancement results for the objective experiment detailed in Section 4.4.2. The results are in terms of mean PESQ scores as a function of input SNR (dB) for AWGN over the Noizeus corpus.*

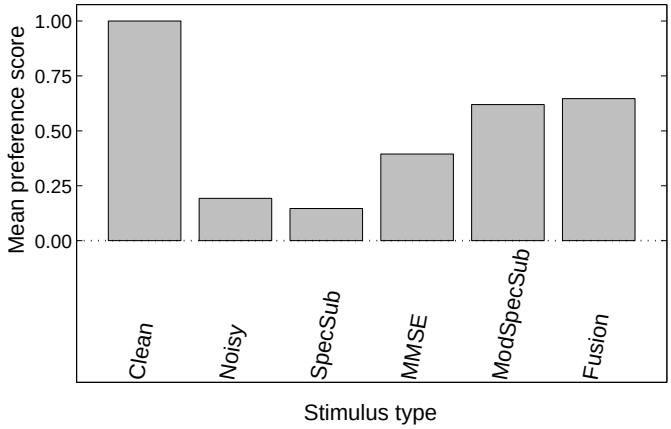

Fig. 9: *Speech enhancement results for the subjective experiment detailed in Section 4.4.3. The results are in terms of mean preference scores for AWGN at* 5 *dB SNR for two Noizeus utterances (sp10 and sp17).*

are shown in Fig. 7(a) and Fig. 7(b), respectively. Spectrograms of noisy speech enhanced using MMSE and ModSpecSub methods are shown in Fig. 7(c) and Fig. 7(d), respectively. Figure 7(e) shows the fusion-weighting function, $\Psi(\sigma_n)$, for the given utterance. As can be seen, $\Psi(\sigma_n)$ is near zero during low energy speech regions as well as during speech pauses. On the other hand, during high energy speech regions, $\Psi(\sigma_n)$ increases towards unity. The spectrogram of speech enhanced using the Fusion method is shown in Fig. 7(f).

4.4.2. Objective experiment

The objective experiment was again carried out over the Noizeus corpus using the PESQ measure.

4.4.3. Subjective experiment

Two Noizeus sentences were employed for the subjective tests. The first (sp10) belonged to a male speaker and second (sp17) to a female speaker. Fourteen English speaking listeners participated in this experiment. Five of them were the same as in the previous experiment, while the remaining nine were new. None of the listeners

reported any hearing defects. The participants were presented with 60 audio stimuli pairs for comparison.

4.5. Results and discussion

The results of the objective evaluation in terms of mean PESQ scores are shown in Fig. 8. The results show that the proposed fusion achieves small but consistent speech quality improvement across the input SNR range as compared to the ModSpecSub method.

This is confirmed by the results of the listening tests shown in terms of average preference scores in Fig. 9. The Fusion method achieves subjective preference improvements over the other speech enhancement methods investigated in this comparison. These improvements were found to be statistically significant at the 99% confidence level, except for the case of Fusion versus ModSpecSub, where the Fusion method was better on average but the improvement was not statistically significantly ($p = 0.0898$). More detailed subjective results, in the form of method preference confusion matrix, are shown in Table 1(b) of Appendix F.

Results of an objective intelligibility evaluation in terms mean speech-transmission index (STI) (Steeneken and Houtgast, 1980) scores have been provided in Fig. 25 of Appendix E. These results show that the Fusion, ModSpecSub and SpecSub methods achieve similar performance, while being consistently better than the MMSE method.

5. Conclusions

In this study, we have proposed to compensate noisy speech for additive noise distortion by applying the spectral subtraction algorithm in the modulation domain. To evaluate the proposed approach, both objective and subjective speech enhancement experiments were carried out. The results of these experiments show that the proposed method results in improved speech quality and it does not suffer from musical noise typically associated with spectral subtractive algorithms. These results indicate that the modulation domain processing is a useful alternative to acoustic domain processing for the enhancement of noisy speech. Future work will investigate the use of other advanced enhancement techniques, such as MMSE, Kalman filtering, etc., in the modulation domain.

We have also proposed to combine ModSpecSub and MMSE methods in the STFT magnitude domain to achieve further speech quality improvements. Through this fusion we have exploited the strengths of both methods while to some degree limiting their weaknesses. The fusion approach was also evaluated through objective and subjective speech enhancement experiments. The results of these experiments demonstrate that it is possible to attain some objective and subjective improvements through speech enhancement fusion in the acoustic STFT domain.

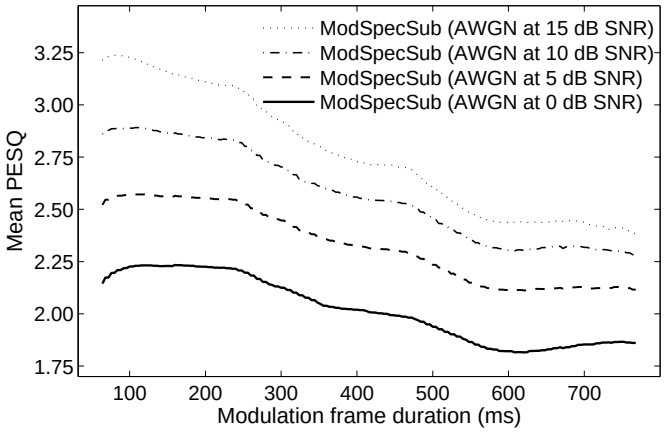

Fig. 10: *Speech enhancement results for the objective experiment detailed in Appendix A. The results are in terms of mean PESQ scores as a function of modulation frame duration (ms) for AWGN over the Noizeus corpus.*

A. Effect of modulation frame duration on speech quality of modulation spectral subtraction stimuli

In order to determine a suitable modulation frame duration, for the modulation spectral subtraction method proposed in Section 3, we have conducted an objective speech enhancement experiment as well as informal subjective listening tests and spectrogram analysis. The details of these are briefly described in this appendix.

In the objective experiment, different modulation frame durations were investigated. These ranged from 64 ms to 768 ms. Mean PESQ scores were computed for ModSpecSub stimuli over the Noizeus corpus for each frame duration. AWGN at 0, 5, 10 and 15 dB SNR was considered.

The results of the objective experiment are shown in Fig. 10. In general, modulation frame durations between 64 ms and 280 ms yielded best PESQ improvements. At higher input SNRs (10 and 15 dB) shorter frame durations of approx. 80 ms produced highest PESQ scores, while at lower input SNRs (0 and 5 dB) the improvement peak was much broader, with highest PESQ scores achieved for durations of 64–280 ms.

Figure 11(c,d,e) shows the spectrograms of the ModSpecSub stimuli, constructed using the following modulation frame durations: 64, 256 and 512 ms, respectively. The frame duration of 64 ms resulted in the introduction of strong musical noise, which can be seen in the spectrogram of Fig. 11(c). On the other hand, a frame duration of 512 ms resulted in temporal slurring distortion as well as somewhat poorer noise suppression. This can be observed in the spectrogram of Fig. 11(e). Modulation frame durations between 180 ms and 280 ms were found to work well. A good compromise between musical noise and temporal slurring was achieved with 256 ms frame duration as shown in the spectrogram of Fig. 11(d). While at the 256 ms duration some slurring is still present, this

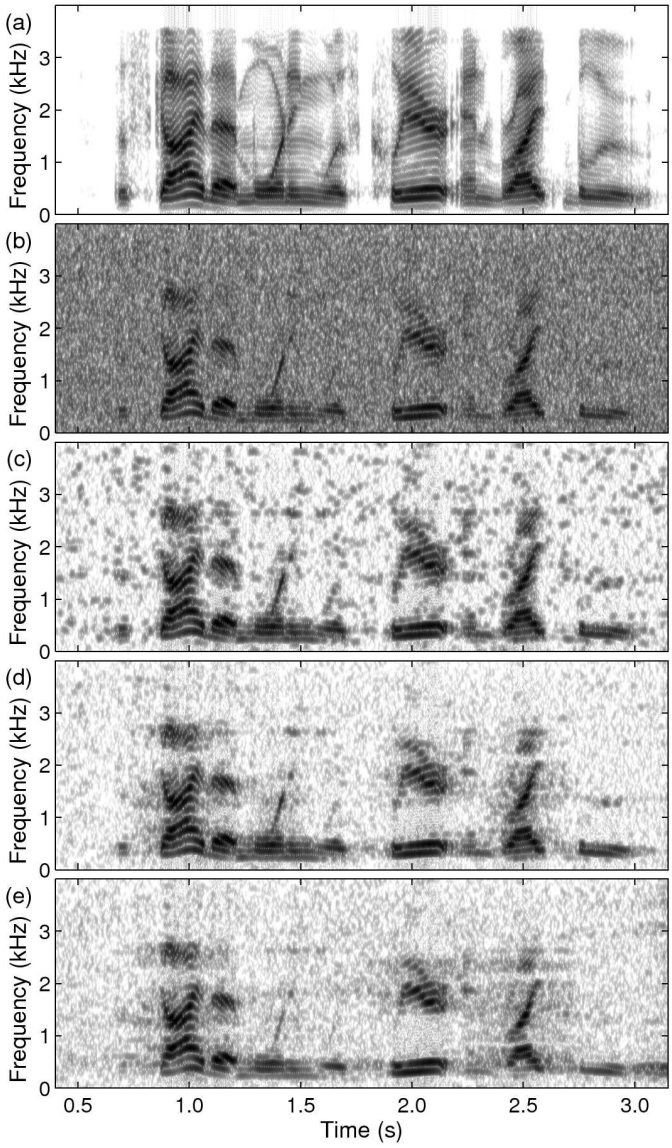

Fig. 11: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by AWGN at* 5 *dB SNR (PESQ:* 1.80*); as well as the noisy speech enhanced using modulation spectral subtraction (ModSpecSub) with the following modulation frame durations: (c)* 64 *ms (PESQ:* 2.38*); (d)* 256 *ms (PESQ:* 2.42*); and (e)* 512 *ms (PESQ:* 2.16*).*

effect is much less perceptually distracting (as determined through informal listening tests) than the musical noise. Thus, when analysis window is too short, the enhanced speech has musical noise, while for long frame durations, lack of temporal localization results in temporal slurring (Thompson and Atlas, 2003).

We have also investigated the effect of the modulation window duration on speech intelligibility using the speech-transmission index (STI) (Steeneken and Houtgast, 1980) as an objective measure. A brief description of the STI measure is included in Appendix E. The window durations between 128 ms and 256 ms were found to have highest intelligibility.

B. Effect of acoustic and modulation domain magnitude spectrum exponents on speech quality of modulation spectral subtraction stimuli

Traditional (acoustic domain) spectral subtraction methods (Boll, 1979; Berouti et al., 1979; Lim and Oppenheim, 1979) have been applied in the magnitude as well as magnitude-squared (acoustic) spectral domains, as clean speech and noise can be considered to be additive in these domains. Additivity in the magnitude domain has been justified by the fact that at high SNRs, the phase spectrum remains largely unchanged by additive noise distortion (Loizou, 2007). Additivity in the magnitude-squared domain has been justified by assuming the speech signal $s(n)$ and noise signal $d(n)$ (see Eq. (1)) to be uncorrelated; making the cross-terms (between clean speech and noise) in the computation of the autocorrelation function (or, the power spectrum) of the noisy speech to be zero.

In the present study, we propose to apply the spectral subtraction method in the short-time modulation domain. Since both the acoustic magnitude and magnitude-squared domains are additive, one can compute the modulation spectrum from either the acoustic magnitude or acoustic magnitude-squared trajectories. Using similar arguments to those presented for acoustic magnitude and magnitude-squared domain additivity, the additivity assumption can be extended to the modulation magnitude and magnitude-squared domains. Therefore, modulation domain spectral subtraction can be carried out on either the modulation magnitude or magnitude-squared spectra.

Thus, for the implementation of modulation domain spectral subtraction, the following two questions have to be answered. First, should the short-time modulation spectrum be derived from the time trajectories of the acoustic magnitude or magnitude-squared spectra? Second, in the short-time modulation spectral domain, should the subtraction be performed on the magnitude or magnitude-squared spectra? In this appendix, we try to answer these two questions experimentally by considering the following four combinations:

1. MAG-MAG: corresponding to acoustic magnitude and modulation magnitude;
2. MAG-POW: corresponding to acoustic magnitude and modulation magnitude-squared;
3. POW-MAG: corresponding to acoustic magnitude-squared and modulation magnitude; and
4. POW-POW: corresponding to acoustic magnitude-squared and modulation magnitude-squared.

Experiments were conducted to examine the effect of each choice on objective speech quality. The Noizeus speech corpus, corrupted by AWGN at 0, 5, 10 and 15 dB SNR, was used. Mean PESQ scores were computed over all 30 Noizeus sentences, for each of the four combinations and each SNR.

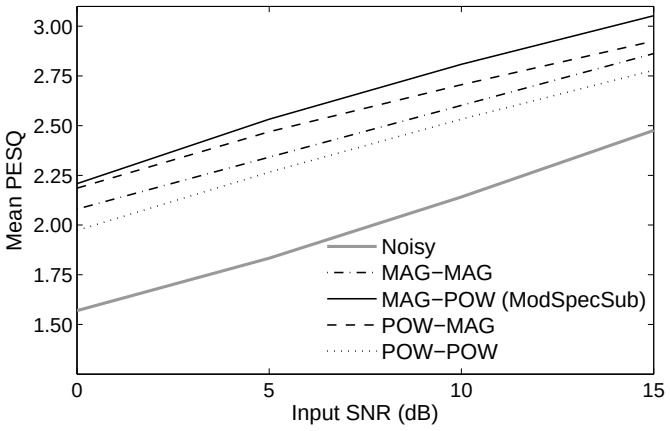

Fig. 12: *Speech enhancement results for the objective experiment detailed in Appendix B. Results for various magnitude spectrum exponent combinations are shown. The results are in terms of mean PESQ scores as a function of input SNR (dB) for AWGN over the Noizeus corpus.*

The objective results in terms of mean PESQ scores are shown in Fig. 12. The MAG-POW combination is shown to work best, with all other combinations achieving lower scores. Based on informal listening tests and analysis of spectrograms shown in Fig. 13, the following qualitative comments can be made about the quality of speech enhanced using the spectral subtraction method applied in the short-time modulation domain using each of the combinations described above. The MAG-MAG combination has improved noise suppression, but the speech content is overly suppressed. The effect is clearly visible in the spectrogram of Fig. 13(c). The MAG-POW combination (Fig. 13(d)) produces the best sounding speech. The POW-MAG combination (Fig. 13(e)) results in poorer noise suppression and the residual noise is musical in nature. The POW-POW combination (Fig. 13(f)) is by far the most audibly distracting to listen to, due to the presence of strong musical noise. The above observations affirm that out of the four choices investigated in our experiment, the MAG-POW combination is best suited for the application of the spectral subtraction algorithm in the short-time modulation domain.

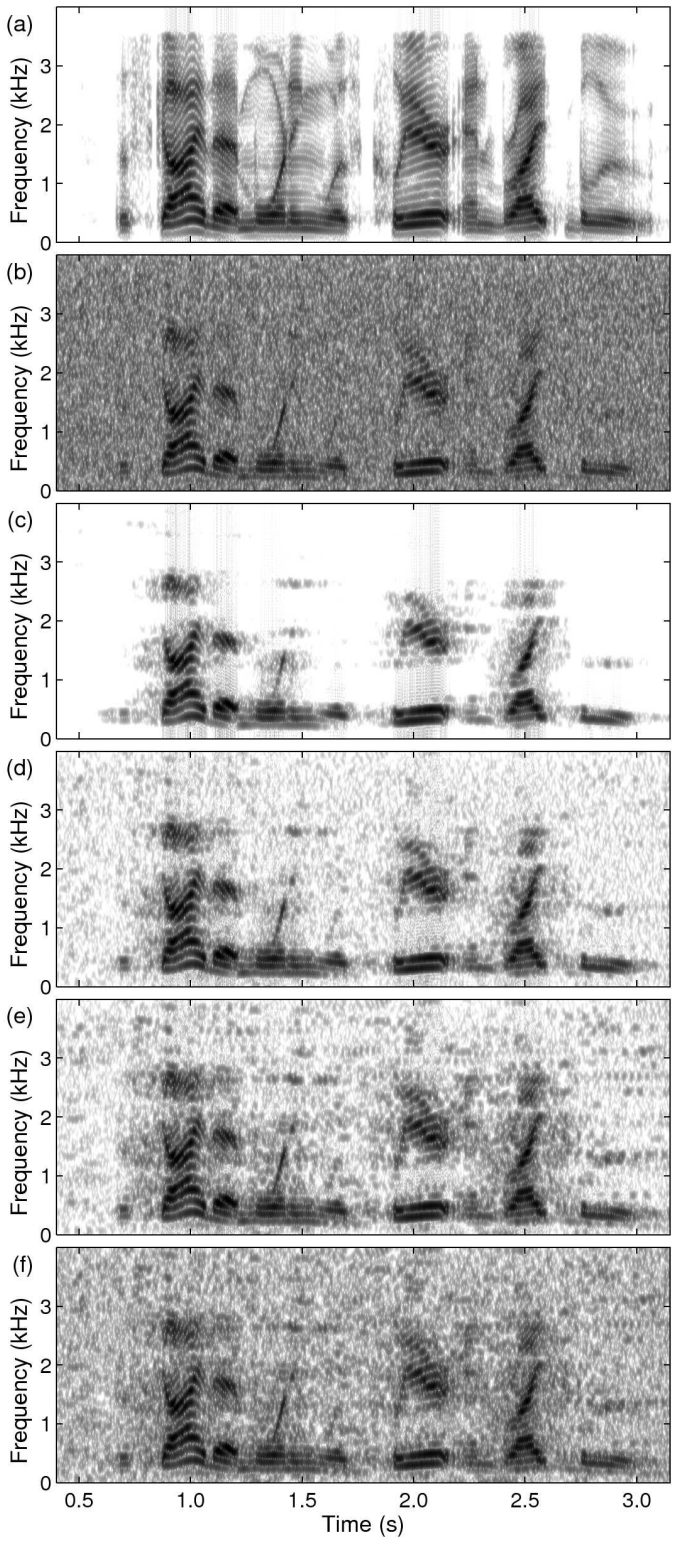

Fig. 13: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by AWGN at* 5 *dB SNR (PESQ:* 1.80*); as well as the noisy speech enhanced using modulation spectral subtraction with various exponents for acoustic and modulation spectra within the dual-AMS framework: (ModSpecSub) (c) MAG-MAG (PESQ:* 2.22*); (d) MAG-POW (PESQ:* 2.42*); (e) POW-MAG (PESQ:* 2.37*); and (f) POW-POW (PESQ:* 2.19*).*

C. Speech enhancement results for coloured noises

In this paper we have proposed to apply the spectral subtraction algorithm in the modulation domain. More specifically, we have formulated a dual-AMS framework where the classical spectral subtraction method (Berouti et al., 1979) is applied after the second analysis stage (*i.e.*, in the short-time modulation domain instead of the short-time acoustic domain employed in the original work of Berouti et al. (1979)). Since the effect of noise on speech is dependent on the frequency, and the SNR of noisy speech varies across the acoustic spectrum (Kamath and Loizou, 2002), it is reasonable to expect that the ModSpecSub method will attain better performance for coloured noises than the acoustic spectral subtraction. This is because one of the strengths of the proposed algorithm is that each subband is processed independently and thus it is the time trajectories in each subband that are important and not the relative levels in-between bands at a given time instant. It is also for this reason that the modulation spectral subtraction method avoids much of the musical noise problem associated with the acoustic spectral subtraction.

This appendix includes some additional results for various coloured noises, including airport, babble, car, exhibition, restaurant, street, subway and train. Mean PESQ scores for the different noise types are shown in Fig. 14. Both ModSpecSub and Fusion have generally achieved higher improvements than the other methods tested. The Fusion method showed best improvements for car, exhibition and train noise types, while for the remaining noises, both Fusion and ModSpecSub methods achieved comparable results.

Example spectrograms for the various noise types are shown in Figs. 15–22.

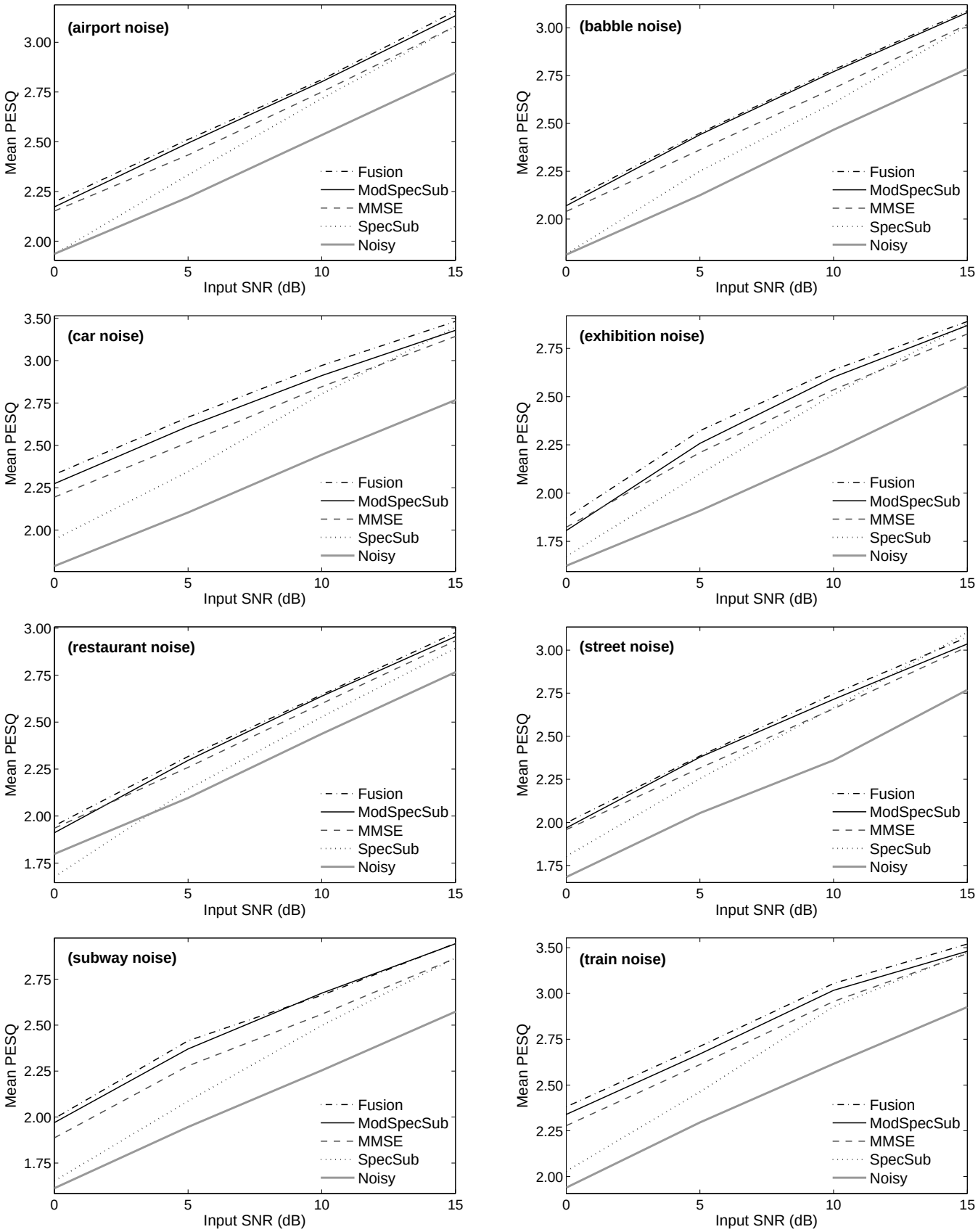

Fig. 14: *Speech enhancement results for the objective experiment detailed in Appendix C. The results are in terms of mean PESQ scores as a function of input SNR (dB) for various coloured noises over the Noizeus corpus.*

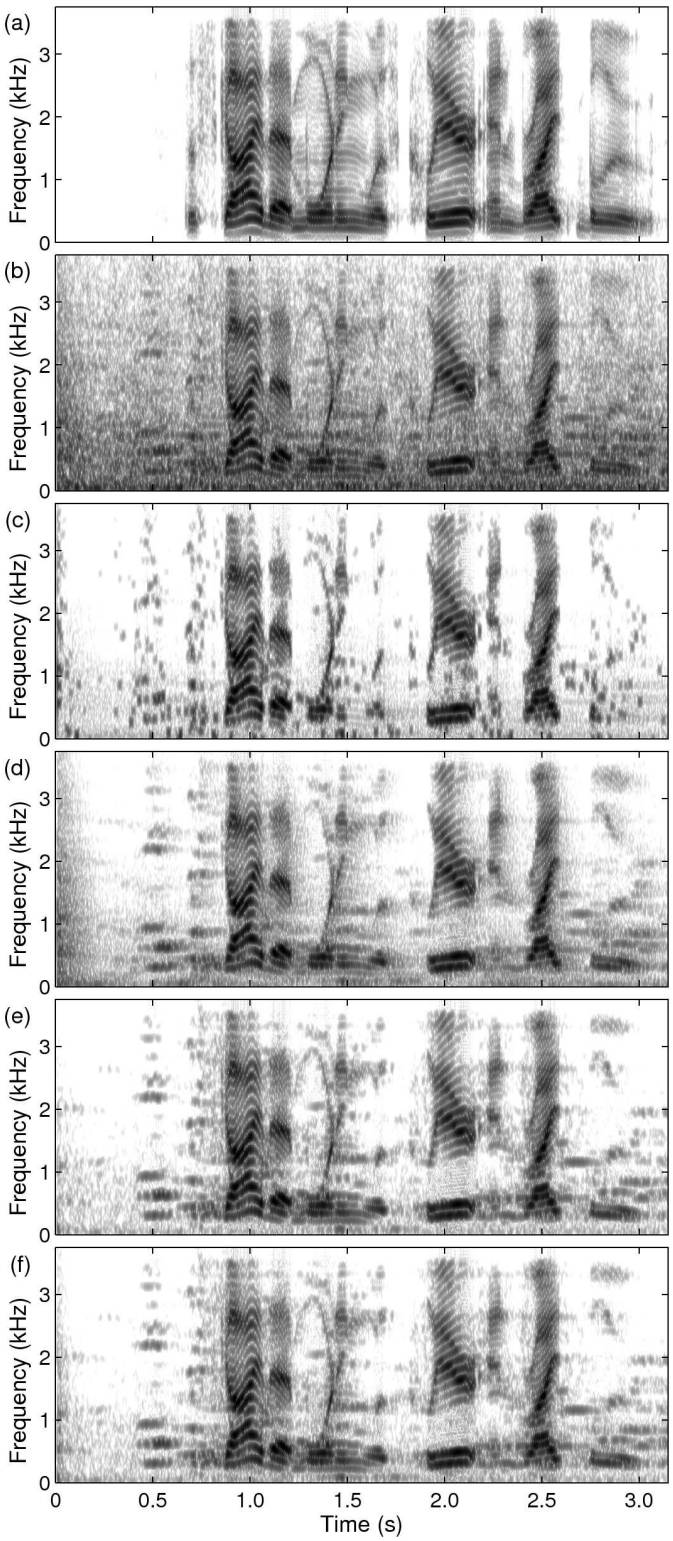

Fig. 15: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by* ***airport noise*** *at* 5 *dB SNR (PESQ:* 2.24*); as well as the noisy speech enhanced using: (c) acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) (PESQ:* 2.34*); (d) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.54*); (e) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.55*); and (f) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.59*).*

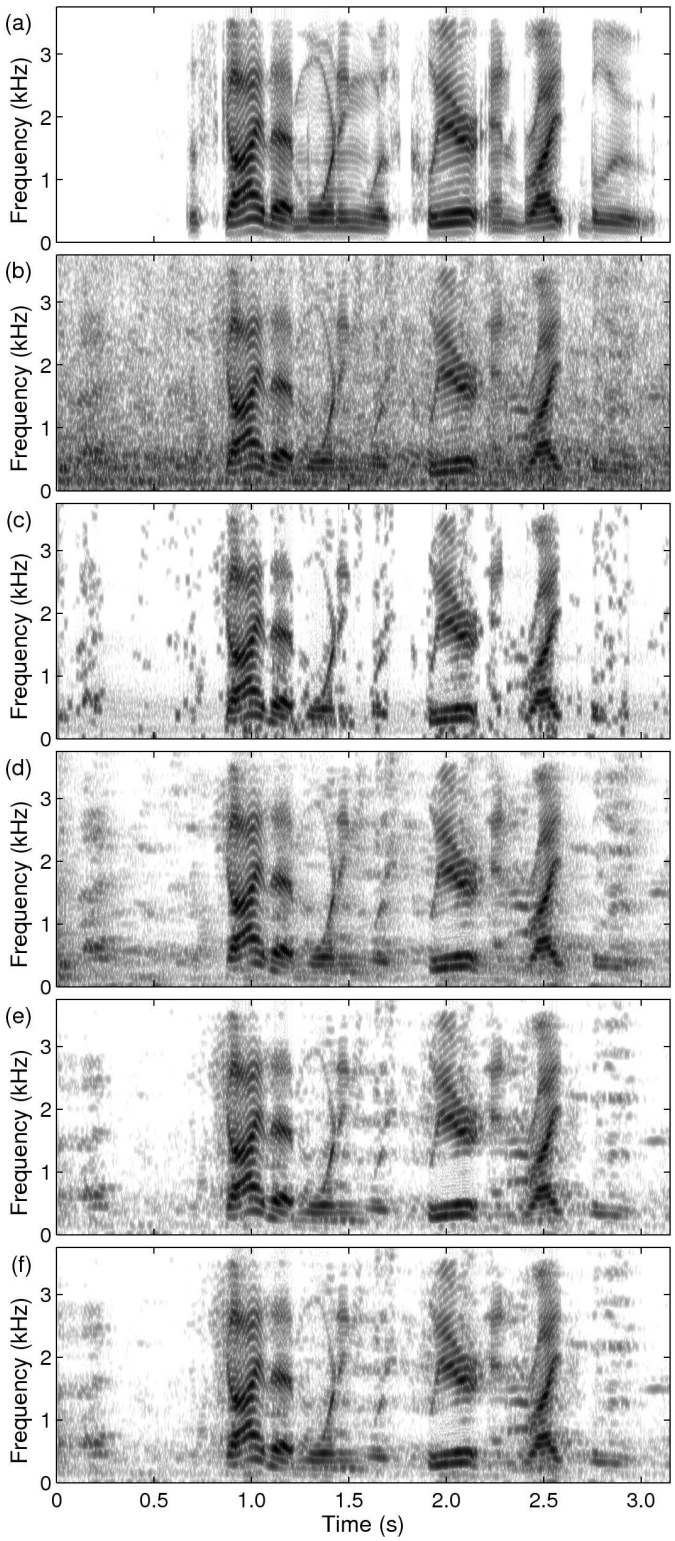

Fig. 16: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by* ***babble noise*** *at* 5 *dB SNR (PESQ:* 2.19*); as well as the noisy speech enhanced using: (c) acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) (PESQ:* 2.25*); (d) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.45*); (e) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.39*); and (f) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.46*).*

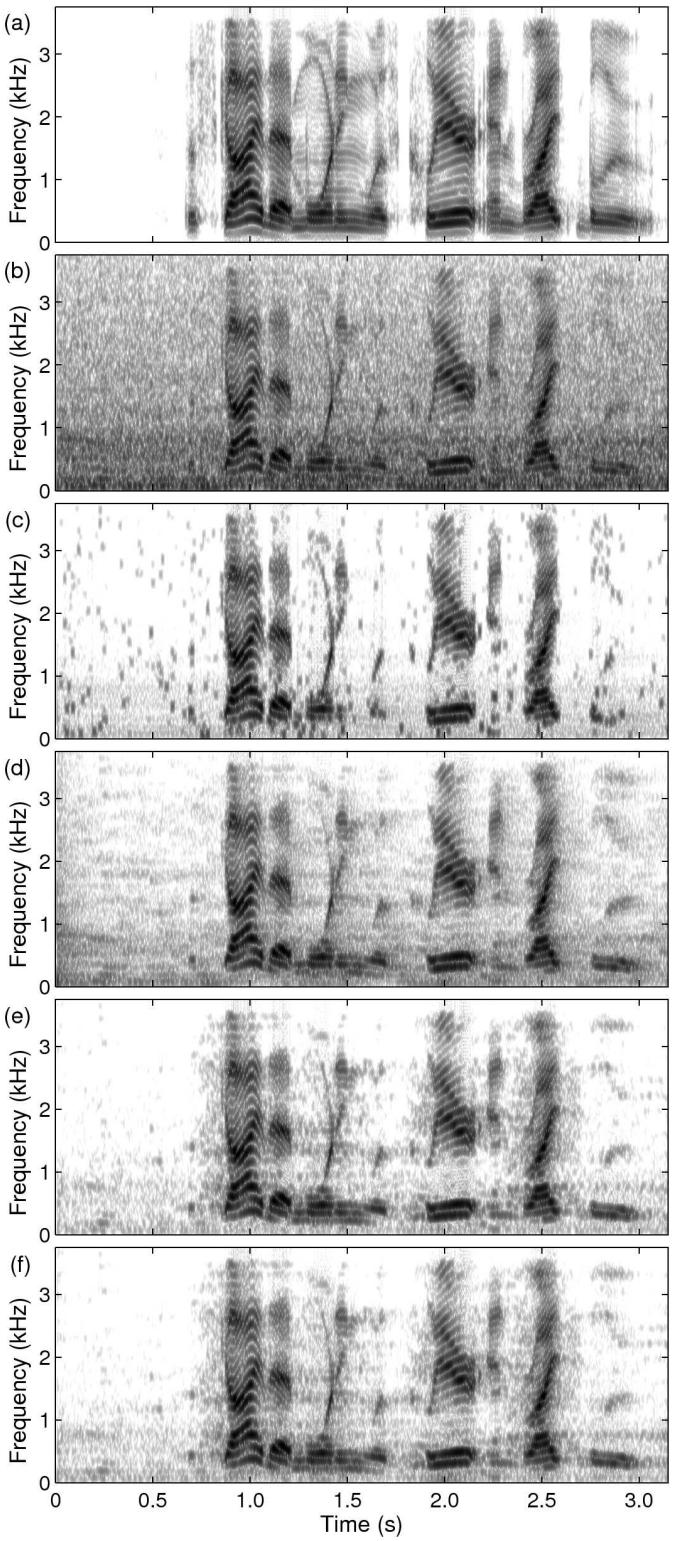

Fig. 17: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by* ***car noise*** *at* 5 *dB SNR (PESQ:* 2.13*); as well as the noisy speech enhanced using: (c) acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) (PESQ:* 2.41*); (d) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.66*); (e) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.60*); and (f) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.67*).*

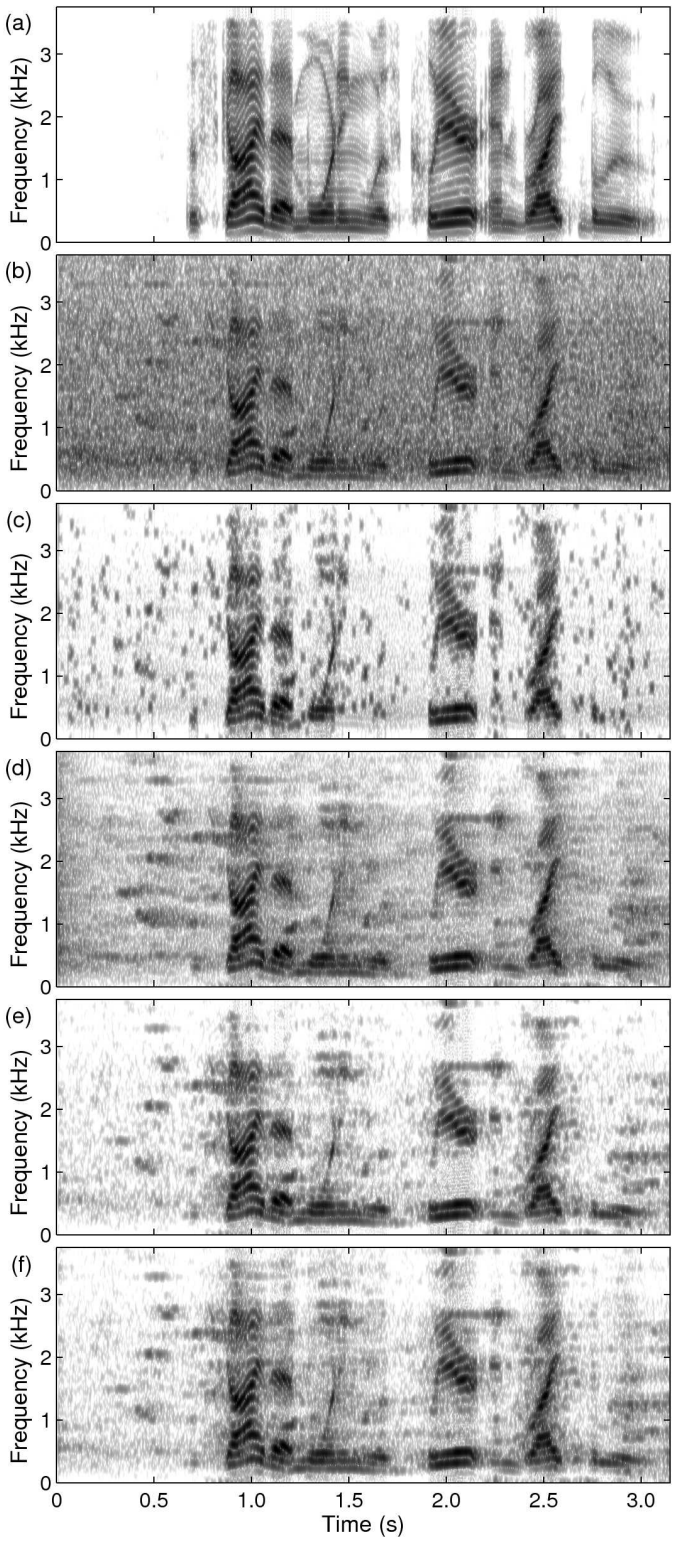

Fig. 18: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by* ***exhibition noise*** *at* 5 *dB SNR (PESQ:* 1.85*); as well as the noisy speech enhanced using: (c) acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) (PESQ:* 1.93*); (d) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.19*); (e) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.27*); and (f) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.33*).*

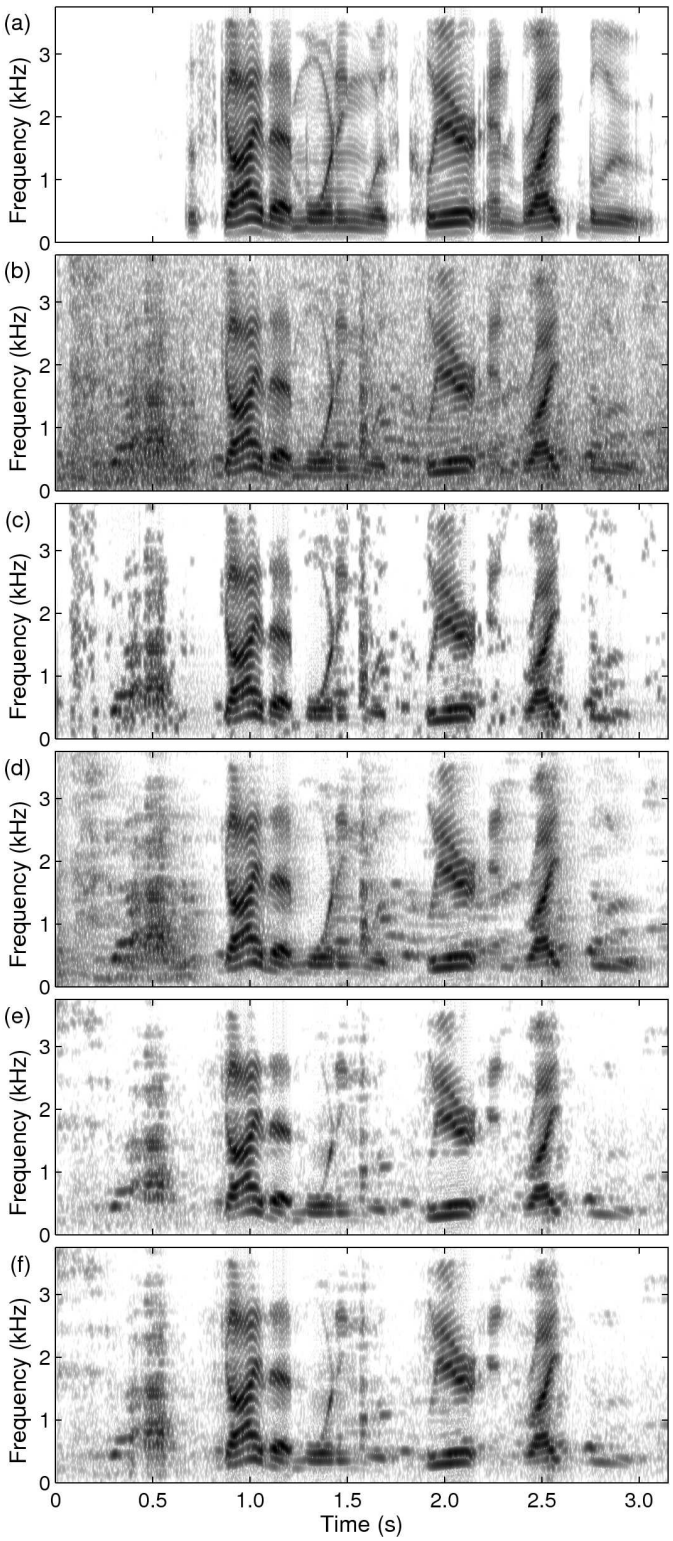

Fig. 19: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by* ***restaurant noise*** *at* 5 *dB SNR (PESQ:* 2.23*); as well as the noisy speech enhanced using: (c) acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) (PESQ:* 2.02*); (d) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.32*); (e) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.26*); and (f) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.37*).*

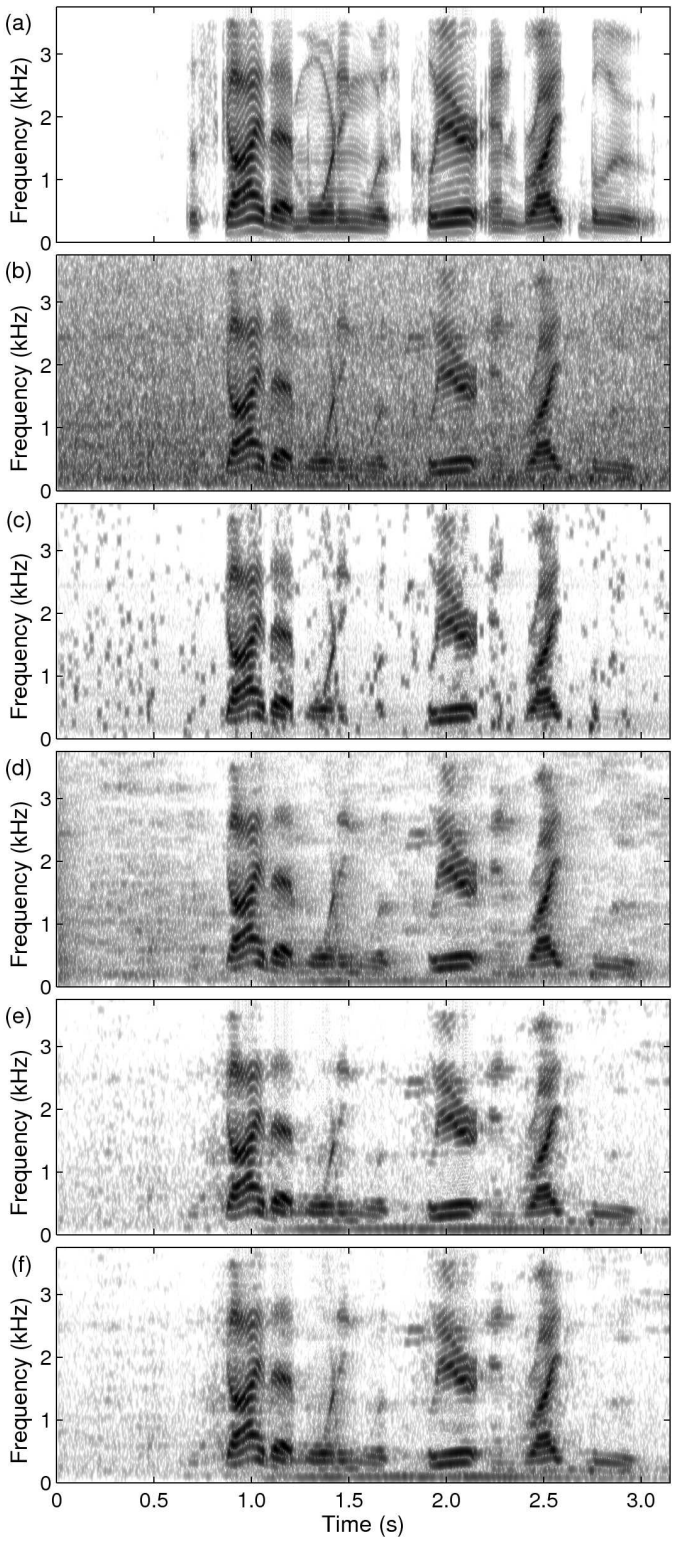

Fig. 20: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by* ***street noise*** *at* 5 *dB SNR (PESQ:* 2.00*); as well as the noisy speech enhanced using: (c) acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) (PESQ:* 2.24*); (d) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.40*); (e) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.39*); and (f) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.50*).*

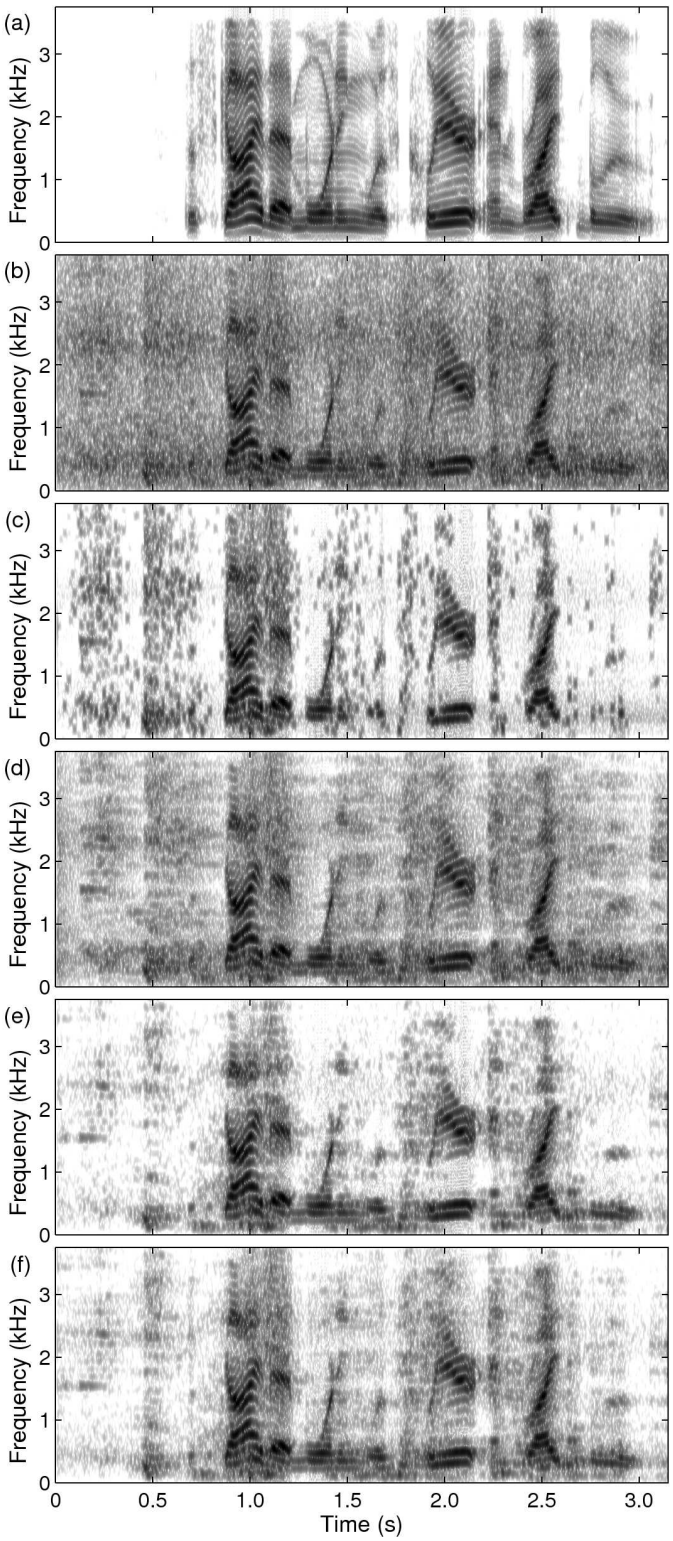

Fig. 21: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by* ***subway noise*** *at* 5 *dB SNR (PESQ:* 2.00*); as well as the noisy speech enhanced using: (c) acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) (PESQ:* 2.09*); (d) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.22*); (e) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.42*); and (f) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.45*).*

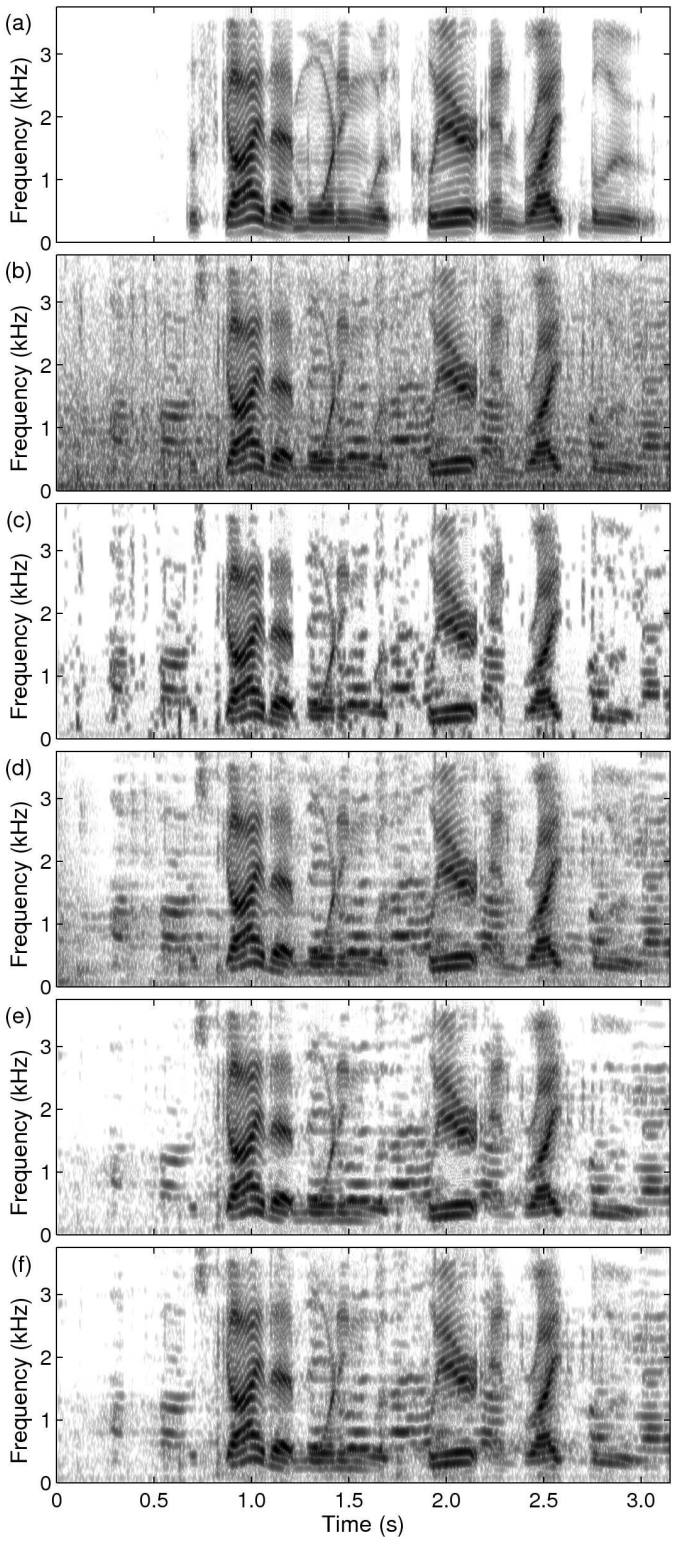

Fig. 22: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by* ***train noise*** *at* 5 *dB SNR (PESQ:* 2.13*); as well as the noisy speech enhanced using: (c) acoustic spectral subtraction (SpecSub) (Berouti et al., 1979) (PESQ:* 1.94*); (d) the MMSE method (Ephraim and Malah, 1984) (PESQ:* 2.25*); (e) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.30*); and (f) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.30*).*

D. Slurring versus musical noise distortion: a closer comparison of the modulation spectral subtraction algorithm with the MMSE method

The noise suppression in the MMSE method for speech enhancement (Ephraim and Malah, 1984, 1985) is achieved by applying a frequency dependent spectral gain function $G(p,\omega_k)$ to the short-time spectrum of the noisy speech $X(p,\omega_k)$ (Cappe, 1994).[9] The spectral gain function can be expressed in terms of the *a priori* and *a posteriori* SNRs, $\mathcal{R}_{\rm prio}(p,\omega_k)$ and $\mathcal{R}_{\rm post}(p,\omega_k)$, respectively. While $\mathcal{R}_{\rm post}(p,\omega_k)$ is a local SNR estimate computed from the current short-time frame, $\mathcal{R}_{\rm prio}(p,\omega_k)$ is an estimate computed from both the current and previous short-time frames.

Decision-directed approach is a popular method for the *a priori* SNR estimation. In the decision-directed approach the parameter of particular importance is α (Cappe, 1994). The parameter α is a weight which determines how much of the SNR estimate is based on the current frame and how much is based on the previous frame. The choice of α has a significant effect on the type and intensity of residual noise of the enhanced speech. For $\alpha \geq 0.9$, the musical noise is reduced. However, values of α very close to one result in temporal distortion during transient parts. This distortion is sometimes described as a *slurring* or *echoing* effect. On the other hand, for values of $\alpha < 0.9$ musical noise is introduced. The choice of α is thus a trade-off between introduction of the musical noise versus introduction of the temporal slurring distortion. The $\alpha = 0.98$ setting has been employed in the literature (Ephraim and Malah, 1984) and recommended as a good compromise for the above trade-off (Cappe, 1994).

Different types of residual noise distortion can have a different effect on the quality and intelligibility of enhanced speech. For example, the musical noise will typically be associated with somewhat reduced speech quality as compared to the temporal slurring. On the other hand, the musical noise distortion will not affect speech intelligibility as adversely as the temporal slurring.

In order to make the comparison of the methods proposed in this work with the MMSE method as fair as possible, in this appendix we compare the MMSE stimuli, constructed with various settings for the α parameter, with the ModSpecSub and Fusion stimuli. For this purpose an objective experiment was carried-out over all 30 utterances of the Noizeus corpus, each corrupted by AWGN at 0, 5, 10 and 15 dB SNR. Three α settings were considered: 0.80, 0.98 and 0.998. The results of the objective experiment, in terms of mean PESQ scores, are given in Fig. 23. The $\alpha = 0.98$ setting produced higher objective scores than the other α settings considered. The ModSpecSub and Fusion methods performed better than the MMSE method for all three MMSE α settings investigated.

[9] For the purposes of this appendix we adopt mathematical notation used by Cappe (1994).

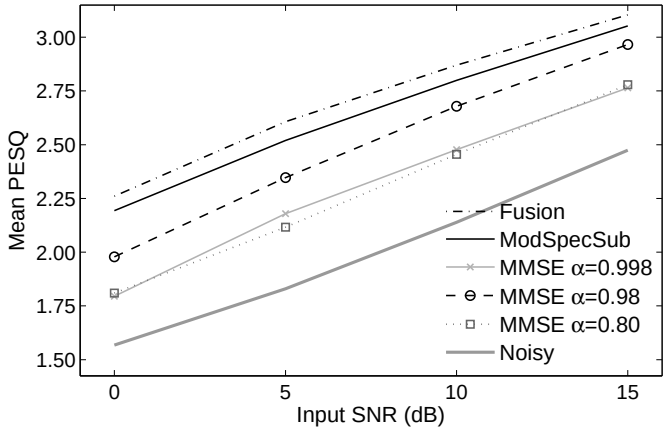

Fig. 23: *Speech enhancement results for the objective experiment detailed in Appendix D. The results are in terms of mean PESQ scores as a function of input SNR (dB) for AWGN over the Noizeus corpus. For the MMSE method, three settings for the parameter α were considered:* 0.8, 0.98 *and* 0.998.

Example spectrograms of the stimuli used in the above experiment are shown in Fig. 24. The spectrograms of MMSE enhanced speech are shown in Fig. 24(c,d,e) for α set to 0.998, 0.98 and 0.80, respectively. The $\alpha = 0.998$ (Fig. 24(c)) results in the best noise attenuation with the residual noise exhibiting little variance. However, during transients temporal slurring is introduced. For $\alpha = 0.98$ (Fig. 24(d)) the temporal slurring distortion has been reduced and the residual noise is not musical in nature, however, the variance and intensity of the residual noise have increased. For $\alpha = 0.80$ (Fig. 24(e)) the temporal slurring distortion has been eliminated, however, the enhanced speech suffers from poor noise reduction and a strong musical noise artefact. The results of informal subjective listening tests confirm the above observations.

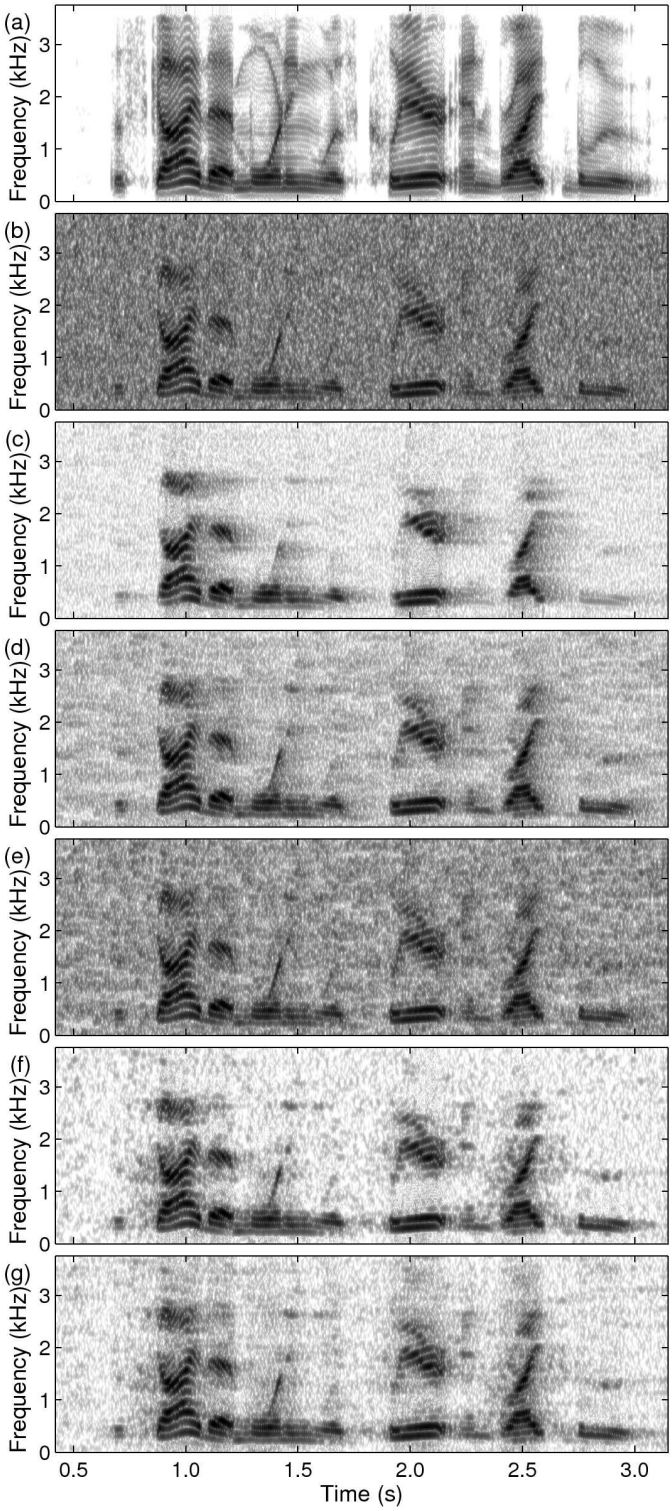

Fig. 24: *Spectrograms of sp10 utterance, "The sky that morning was clear and bright blue", by a male speaker from the Noizeus speech corpus: (a) clean speech (PESQ:* 4.50*); (b) speech degraded by AWGN at* 5 *dB SNR (PESQ:* 1.80*); as well as the noisy speech enhanced using the MMSE method (Ephraim and Malah, 1984) with: (c)* $\alpha = 0.998$ *(PESQ:* 2.00*); (d)* $\alpha = 0.98$ *(PESQ:* 2.26*); (e)* $\alpha = 0.80$ *(PESQ:* 2.06*). Also included are the following: (f) modulation spectral subtraction (ModSpecSub) (PESQ:* 2.42*); and (g) ModSpecSub fusion with MMSE (Fusion) (PESQ:* 2.51*).*

E. Objective intelligibility results

In speech enhancement we are primarily interested in the suppression of noise from noise corrupted speech so that the quality can be improved. Speech quality is a measure which quantifies how nice speech sounds and includes attributes such as intelligibility, naturalness, roughness of noise, etc. In the main body of this paper we have solely concentrated on the overall quality aspect of enhanced speech. However, in some speech processing applications (*e.g.*, automatic speech recognition), it is the intelligibility attribute that is perhaps the most important. By intelligibility we mean understanding (or recognition) of the individual linguistic items spoken (such as phonemes, syllables, words).

In this appendix, we provide some indication of the intelligibility of enhanced speech by using an objective intelligibility measure, namely the speech-transmission index (STI) (Steeneken and Houtgast, 1980). STI measures the extent to which slow temporal intensity envelope modulations are preserved in degraded listening environments (Payton and Braida, 1999). It is these slow intensity variations that are important for speech intelligibility. We employ the speech-based STI computation procedure where speech signal is used as a probe. Under this framework, the original and processed speech signals are passed separately through a bank of seven octave band filters. Each filtered signal is squared and low pass filtered (with cut-off frequency of 32 kHz) to derive the temporal intensity envelope. The power spectrum of the temporal intensity envelope is subjected to one-third octave band analysis. The components over each of the 14 one-third octave band intervals (with centres ranging from 0.63 Hz to 12.7 Hz) are summed, producing 98 modulation indices. The resulting modulation spectrum of the original speech, along with the modulation spectrum of the processed speech, can then be used to compute the modulation transfer function (MTF), which in turn is used to compute STI. We employ three different approaches for the computation of the MTF. The first approach is by Houtgast and Steeneken (1985), the second is by Drullman et al. (1994b) and the third is by Payton et al. (2002). The details of MTF and STI computations are given in Goldsworthy and Greenberg (2004).

An enhancement experiment was performed over all 30 Noizeus utterances, each corrupted by AWGN at 0, 5, 10 and 15 dB SNR. The results of the experiment, in terms of mean STI scores for Houtgast and Steeneken (1985), Drullman et al. (1994b) and Payton et al. (2002) methods, are shown in Fig. 25(a,b,c), respectively. The results suggest that the ModSpecSub and Fusion methods do not adversely affect speech intelligibility. On the contrary, significant intelligibility improvement was observed over that of noisy speech. The proposed methods achieved mean STI scores comparable to those produced by the acoustic spectral subtraction method and performed consistently

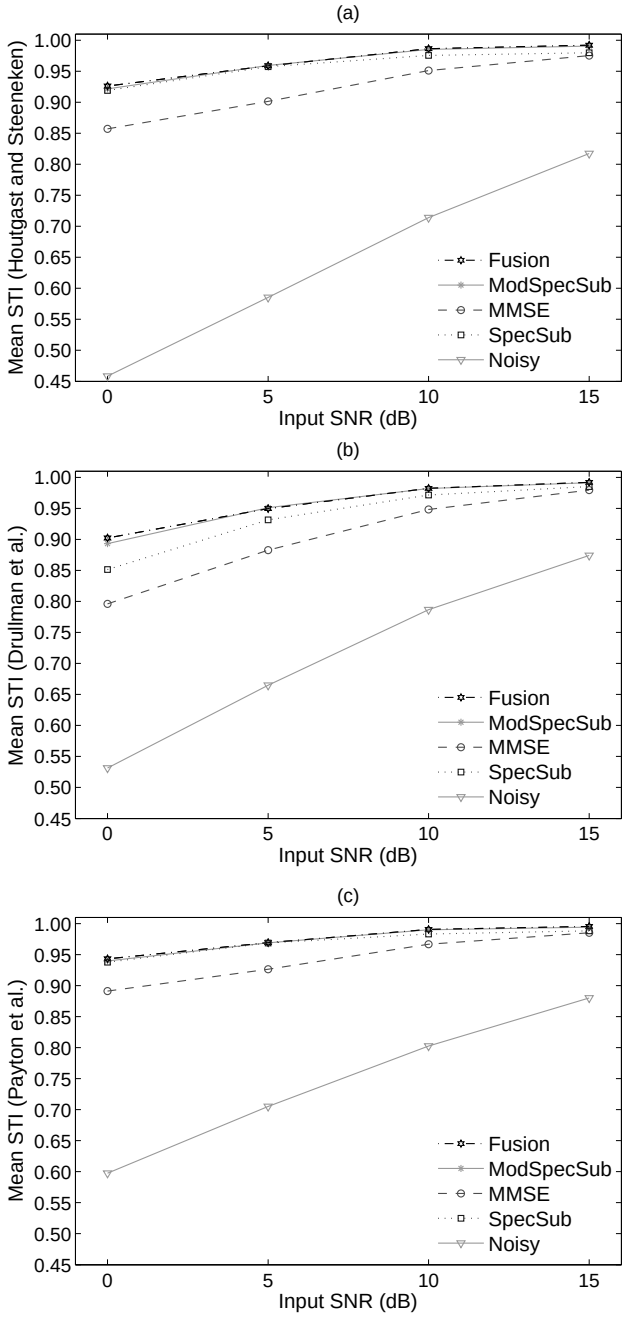

Fig. 25: *Speech enhancement results for the objective experiment detailed in Appendix E. The objective intelligibility results are in terms of mean speech-based STI as a function of input SNR (dB) for AWGN over the Noizeus corpus. The results for the following speech-based STI variants are shown: (a) Houtgast and Steeneken (1985) method; (b) Drullman et al. (1994b) method; and (c) Payton et al. (2002) method.*

better than the MMSE method.[10] The above observations are consistent across all three speech-based STI variants employed in our evaluation.

[10]Note that at low SNRs, the mean STI measure for the spectral subtraction method is better than that of the MMSE method. However, in terms of speech quality measured through listening tests and PESQ measure (Section 3.4), the opposite is the case.

F. Subjective method preference confusion matrices

Detailed results of the listening tests, described in Sections 3.3.4 and 4.4.3, are given in Tables 1(a) and 1(b), respectively. The results are in terms of subjective method preference confusion matrices, which can be read and interpreted as follows. For a given experiment, method labels are listed at the start of each row and at the top of each column. The scores shown in each row are mean subjective preferences for the method identified by the row label versus the methods identified by each column label. The scores are averaged across same method pairs. The minimum score is zero and the maximum score is one. A score of zero (one) means that the method identified by the row label was never (always) preferred over the method with the corresponding column label. For example, noisy speech was never preferred over the clean speech, since the Noisy versus Clean score is zero. Similarly, a score of 0.89 for ModSpecSub with respect to MMSE given in Table 1(b), indicates that ModSpecSub (89%) was generally preferred over MMSE (11%).

The results of Table 1(a) show that ModSpecSub was preferred over Noisy, SpecSub and MMSE. These improvements were found to be statistically significant ($p < 0.01$). The results of Table 1(b) show that Fusion was preferred over the Noisy, SpecSub, MMSE and ModSpecSub. Fusion achieved significantly higher preference over Noisy, SpecSub and MMSE ($p < 0.01$). However, while Fusion (58%) was found to perform better on average than ModSpecSub (42%), the improvement was not statistically significant ($p = 0.0898$).

A note on the consistency of the subjective results between experiments of Section 3 and Section 4. For the most part, the subjective results are consistent between the experiments. For both experiments, the ModSpecSub method achieves significantly higher preference scores over the SpecSub and MMSE methods in both experiments. However, some of the listeners in the second experiment have shown a stronger dislike towards the acoustic spectral subtraction – favouring Noisy over SpecSub stimuli. This is not at all surprising, as musical noise is very perceptually annoying, while AWGN can (for some listeners) be easier to "tune out".

Table 1: Confusion matrices: subjective method preferences for the listening tests detailed in (a) Section 3.3.4 and (b) Section 4.4.3.

(a)

	Clean	Noisy	SpecSub	MMSE	ModSpecSub
Clean	—	1.00	1.00	1.00	1.00
Noisy	0.00	—	0.27	0.05	0.07
SpecSub	0.00	0.73	—	0.30	0.09
MMSE	0.00	0.95	0.70	—	0.11
ModSpecSub	0.00	0.93	0.91	0.89	—

(b)

	Clean	Noisy	SpecSub	MMSE	ModSpecSub	Fusion
Clean	—	1.00	1.00	1.00	1.00	1.00
Noisy	0.00	—	0.62	0.12	0.14	0.07
SpecSub	0.00	0.38	—	0.17	0.07	0.12
MMSE	0.00	0.88	0.83	—	0.11	0.16
ModSpecSub	0.00	0.86	0.93	0.89	—	0.42
Fusion	0.00	0.93	0.88	0.84	0.58	—

References

Allen, J., Jun 1977. Short term spectral analysis, synthesis, and modification by discrete Fourier transform. IEEE Trans. Acoust., Speech, Signal Process. ASSP-25 (3), 235–238.

Allen, J., Rabiner, L., Nov 1977. A unified approach to short-time Fourier analysis and synthesis. Proc. IEEE 65 (11), 1558–1564.

Arai, T., Pavel, M., Hermansky, H., Avendano, C., Oct 1996. Intelligibility of speech with filtered time trajectories of spectral envelopes. In: Proc. Int. Conf. Spoken Language Process. (ICSLP). Philadelphia, PA, USA, pp. 2490–2493.

Atlas, L., 2003. Modulation spectral transforms: Application to speech separation and modification. Tech. Rep. 155, IEICE, Univ. Washington, Washington, WA, USA.

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

Atlas, L., Shamma, S., Jan 2003. Joint acoustic and modulation frequency. EURASIP Journal on Applied Signal Processing 2003 (7), 668–675.

Atlas, L., Vinton, M., 2001. Modulation frequency and efficient audio coding. In: Proc. SPIE The International Society for Optical Engineering. Vol. 4474. pp. 1–8.

Bacon, S., Grantham, D., Jun 1989. Modulation masking: Effects of modulation frequency, depth and phase. J. Acoust. Soc. Amer. 85 (6), 2575–2580.

Berouti, M., Schwartz, R., Makhoul, J., Apr 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Int. Conf. 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.

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

Crochiere, R., Feb 1980. A weighted overlap-add method of short-time Fourier analysis / synthesis. IEEE Trans. Acoust., Speech, Signal Process. ASSP-28 (1), 99–102.

Depireux, D., Simon, J., Klein, D., Shamma, S., Mar 2001. Spectrotemporal response field characterization with dynamic ripples in ferrect primary auditory cortex. Journal of Neurophysiology 85 (3), 1220–1234.

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

Drullman, R., Festen, J., Plomp, R., Feb 1994b. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Amer. 95 (2), 1053–1064.

Ephraim, Y., Malah, D., Dec 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., Apr 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., Stadler, S., Kleijn, W. B., Chan, W.-Y., Aug 2007. Noise suppression based on extending a speech-dominated modulation band. In: Proc. ISCA Conf. Int. Speech Commun. Assoc. (INTERSPEECH). Antwerp, Belgium, pp. 970–973.

Goldsworthy, R., Greenberg, J., Dec 2004. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J. Acoust. Soc. Amer. 116 (6), 3679–3689.

Greenberg, S., Arai, T., Sep 2001. The relation between speech intelligibility and the complex modulation spectrum. In: Proc. ISCA European Conf. Speech Commun. and Technology (EUROSPEECH). Aalborg, Denmark, pp. 473–476.

Griffin, D., Lim, J., 1984. Signal estimation from modified short-time Fourier transform. IEEE Trans. Acoust., Speech, Signal Process. ASSP-32 (2), 236–243.

Hasan, M., Salahuddin, S., Khan, M., Apr 2004. A modified a priori SNR for speech enhancement using spectral subtraction rules. IEEE Signal Process. Lett. 11 (4), 450–453.

Hermansky, H., Morgan, N., Oct 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2, 578–589.

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

Houtgast, T., Steeneken, H., Mar 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J. Acoust. Soc. Amer. 77 (3), 1069–1077.

Hu, Y., Loizou, P., Jan 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Speech Audio Process. 12 (1), 59–67.

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

Kamath, S., Loizou, P., May 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP). Orlando, FL, USA.

Kanedera, N., Arai, T., Hermansky, H., Pavel, M., May 1999. On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Communication 28 (1), 43–55.

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

Kim, D.-S., Sep 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., Aug 1998. Robust speech recognition using the modulation spectrogram. Speech Communication 25 (1–3), 117–132.

Kinnunen, T., May 2006. Joint acoustic-modulation frequency for speaker recognition. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP). Vol. 1. Toulouse, France, pp. 665–668.

Kinnunen, T., Lee, K., Li, H., Jan 2008. Dimension reduction of the modulation spectrogram for speaker verification. In: Proc. ISCA Speaker and Language Recognition Workshop (ODYSSEY). Stellenbosch, South Africa.

Kowalski, N., Depireux, D., Shamma, S., Nov 1996. Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses to moving ripple spectra. Journal of Neurophysiology 76 (5), 3503–3523.

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

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

Lu, C.-T., Aug 2007. Reduction of musical residual noise for speech enhancement using masking properties and optimal smoothing. Pattern Recog. Lett. 28 (11), 1300–1306.

Lu, X., Matsuda, S., Unoki, M., Nakamura, S., 2010. Temporal contrast normalization and edge-preserved smoothing of temporal modulation structures of speech for robust speech recognition. Speech Communication 52 (1), 1–11.

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

Malayath, N., Hermansky, H., Kajarekar, S., Yegnanarayana, B., Jan 2000. Data-driven temporal filters and alternatives to GMM in speaker verification. Digital Signal Processing 10 (1-3), 55–74.

Mesgarani, N., Shamma, S., Mar 2005. Speech enhancement based on filtering the spectrotemporal modulations. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP). Vol. 1. Philadelphia, PA, USA, pp. 1105–1108.

Nadeu, C., Pachs-Leal, P., Juang, B.-H., Sep 1997. Filtering the time sequences of spectral parameters for speech recognition,. Speech Communication 22 (4), 315–332.

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

Payton, K., Braida, L., Dec 1999. A method to determine the speech transmission index from speech waveforms. J. Acoust. Soc. Amer. 106, 3637–3648.

Payton, K. L., Braida, L. D., Chen, S., Rosengard, P., Goldsworthy, R., 2002. Computing the STI using speech as a probe stimulus. In: Past, Present and Future of the Speech Transmission Index. TNO Human Factors, Soesterberg, Netherlands, pp. 125–138.

Portnoff, M., Jun 1981. Short-time Fourier analysis of sampled speech. IEEE Trans. Acoust., Speech, Signal Process. ASSP-29 (3), 364–373.

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

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 codecs. ITU-T Recommendation P.862.

Schreiner, C., Urbas, J., 1986. Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). Hearing Research 21 (3), 227–241.

Shamma, S., Aug 1996. Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method. Network: Computation in Neural Systems 7 (3), 439–476.

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

Sheft, S., Yost, W., Aug 1990. Temporal integration in amplitude modulation detection. J. Acoust. Soc. Amer. 88 (2), 796–805.

Steeneken, H., Houtgast, T., Jan 1980. A physical method for

measuring speech-transmission quality. J. Acoust. Soc. Amer. 67 (1), 318–326.

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

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

Vaseghi, S., Frayling-Cork, R., Oct 1992. Restoration of old gramophone recordings. J. Audio Eng. 40 (10), 791–801.

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

Vuuren, S. V., Hermansky, H., Nov 1998. On the importance of components of the modulation spectrum for speaker verification. In: Proc. Int. Conf. Spoken Language Process. (ICSLP). Sydney, Australia, pp. 3205–3208.

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

Xiao, X., Chng, E., Li, H., Apr 2007. Normalization of the speech modulation spectra for robust speech recognition. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP). Vol. 4. Monolulu, Hawaii, USA, pp. 1021–1024.

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