Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 465–494
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

The importance of phase in speech enhancement

Kuldip Paliwal, Kamil Wójcicki *, Benjamin Shannon [1]

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

Received 24 March 2010; received in revised form 30 November 2010; accepted 6 December 2010
Available online 24 December 2010

Abstract

Typical speech enhancement methods, based on the short-time Fourier analysis-modification-synthesis (AMS) framework, modify only the magnitude spectrum and keep the phase spectrum unchanged. In this paper our aim is to show that by modifying the phase spectrum in the enhancement process the quality of the resulting speech can be improved. For this we use analysis windows of 32 ms duration and investigate a number of approaches to phase spectrum computation. These include the use of matched or mismatched analysis windows for magnitude and phase spectra estimation during AMS processing, as well as the phase spectrum compensation (PSC) method. We consider four cases and conduct a series of objective and subjective experiments that examine the importance of the phase spectrum for speech quality in a systematic manner. In the first (oracle) case, our goal is to determine maximum speech quality improvements achievable when accurate phase spectrum estimates are available, but when no enhancement is performed on the magnitude spectrum. For this purpose speech stimuli are constructed, where (during AMS processing) the phase spectrum is computed from clean speech, while the magnitude spectrum is computed from noisy speech. While such a situation does not arise in practice, it does provide us with a useful insight into how much a precise knowledge of the phase spectrum can contribute towards speech quality. In this first case, matched and mismatched analysis window approaches are investigated. Particular attention is given to the choice of analysis window type used during phase spectrum computation, where the effect of spectral dynamic range on speech quality is examined. In the second (non-oracle) case, we consider a more realistic scenario where only the noisy spectra (observable in practice) is available. We study the potential of the mismatched window approach for speech quality improvements in this non-oracle case. We would also like to determine how much room for improvement exists between this case and the best (oracle) case. In the third case, we use the PSC algorithm to enhance the phase spectrum. We compare this approach with the oracle and non-oracle matched and mismatched window techniques investigated in the preceding cases. While in the first three cases we consider the usefulness of various approaches to phase spectrum computation within the AMS framework when noisy magnitude spectrum is used, in the fourth case we examine the usefulness of these techniques when enhanced magnitude spectrum is employed. Our aim (in the context of traditional magnitude spectrum-based enhancement methods) is to determine how much benefit in terms of speech quality can be attained by also processing the phase spectrum. For this purpose, the minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimates are employed instead of noisy magnitude spectra. The results of the oracle experiments show that accurate phase spectrum estimates can considerably contribute towards speech quality, as well as that the use of mismatched analysis windows (in the computation of the magnitude and phase spectra) provides significant improvements in both objective and subjective speech quality – especially, when the choice of analysis window used for phase spectrum computation is carefully considered. The mismatched window approach was also found to improve speech quality in the non-oracle case. While the improvements were found to be statistically significant, they were only modest compared to those observed in the oracle case. This suggests that research into better phase spectrum estimation algorithms, while a challenging task, could be worthwhile. The results of the PSC experiments indicate that the PSC method achieves better speech quality improvements than the other non-oracle methods considered. The results of the MMSE experiments suggest that accurate phase spectrum estimates have a potential to significantly improve performance of existing magnitude spectrum-based methods. Out of the non-oracle approaches

* Corresponding author.
E-mail address: kamil.wojcicki@ieee.org (K. Wójcicki).
[1] In memoriam Benjamin James Shannon (1978–2010).

0167-6393/$ - see front matter
doi:10.1016/j.specom.2010.12.003

considered, the combination of the MMSE STSA method with the PSC algorithm produced significantly better speech quality improvements than those achieved by these methods individually.

Keywords: Speech enhancement; Analysis window; Short-time Fourier analysis; Analysis-modification-synthesis (AMS); Magnitude spectrum; Minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator; Phase spectrum; Phase spectrum compensation (PSC); MMSE PSC

1. Introduction

1.1. Background

In the field of speech enhancement we are interested in enhancing the quality of speech corrupted by additive noise distortion. Depending on the number of audio channels available, speech enhancement methods can be grouped into single-channel and multi-channel approaches. Various single-channel speech enhancement approaches have been proposed in the literature. These can be grouped into spectral subtraction (Boll, 1979; Berouti et al., 1979), minimum mean-square error (MMSE) estimation (Ephraim and Malah, 1984, 1985), Wiener filtering (linear MMSE) (Wiener, 1949), Kalman filtering (Paliwal and Basu, 1987) and subspace (Ephraim and Trees, 1995) methods. Several of these methods employ the short-time Fourier analysis-modification-synthesis (AMS) framework. We focus here on the AMS-based approach to speech enhancement.

1.2. AMS framework based speech enhancement

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 construct the output signal.

Let us consider an additive noise model

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

where $x(n)$, $s(n)$ 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

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

where k refers to the index of the discrete frequency, L is the length of frequency analysis, and $w(n)$ is an analysis window function.[2] In speech processing, the Hamming window with 20–40 ms duration is typically employed.

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 the STFT magnitude spectrum and the STFT phase spectrum. For instance, the STFT of the noisy speech signal can be written in polar form as

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

where $|X(n,k)|$ denotes the magnitude spectrum and $\angle X(n,k)$ denotes the phase spectrum.[3]

Traditional AMS-based speech enhancement methods (e.g., Lim and Oppenheim, 1979; Berouti et al., 1979; Ephraim and Malah, 1984; Ephraim et al., 1985; Martin, 1994; Sim et al., 1998; Virag, 1999), modify only the noisy magnitude spectrum, while keeping the noisy phase spectrum unchanged. Such algorithms attempt to estimate the magnitude spectrum of clean speech. Let us denote the enhanced magnitude spectrum as $|\widehat{S}(n,k)|$, then the modified spectrum is constructed by combining $|\widehat{S}(n,k)|$ with the complex exponential of the noisy phase spectrum, as follows:

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

The enhanced speech signal, $y(n)$, is constructed by taking the inverse STFT of $Y(n,k)$ followed by least-squares overlap-add synthesis (Griffin and Lim, 1984; Quatieri, 2002). A block diagram of a traditional AMS-based speech enhancement framework is shown in Fig. 1.

1.3. Earlier studies on the usefulness of the short-time phase spectrum in speech processing

As mentioned previously, the existing AMS-based speech enhancement algorithms modify (or enhance) the magnitude spectrum, but do not change the phase spectrum. The reason for this is twofold. Firstly, it has been shown that the noisy phase is an optimal estimator of the clean phase (Ephraim and Malah, 1984). Secondly, there has been a long standing belief among speech researchers that for small window durations (20–40 ms),[4] typically employed in speech processing, the short-time phase

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

[3] In our discussions when referring to the magnitude spectrum, phase spectrum and (complex) spectrum, the STFT (or short-time) modifier is implied unless otherwise stated.

[4] In this paper small window durations of 20–40 ms are implied if not explicitly stated.

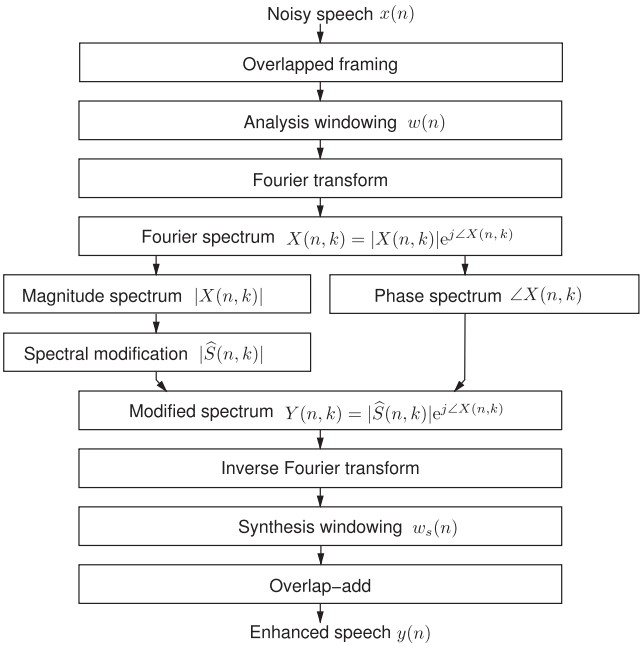

Fig. 1. Block diagram of a traditional AMS-based speech enhancement procedure.

spectrum contains little useful information (Oppenheim et al., 1979; Oppenheim and Lim, 1981; Vary, 1985; Liu et al., 1997) and is unimportant for speech enhancement (Lim and Oppenheim, 1979; Wang and Lim, 1982). Recent subjective studies (Paliwal, 2003; Paliwal and Alsteris, 2003; Alsteris and Paliwal, 2004; Paliwal and Alsteris, 2005; Alsteris and Paliwal, 2006; Shi et al., 2006; Wójcicki and Paliwal, 2007) suggest, however, that the phase spectrum does contain significant amount of useful information for speech intelligibility.

When Hamming window (having spectral dynamic range[5] of 43 dB) is used for short-time Fourier analysis, the magnitude spectrum contributes significantly more towards speech intelligibility than the phase spectrum (Liu et al., 1997; Paliwal, 2003). However, the phase spectrum becomes comparable to magnitude spectrum in terms of speech intelligibility when rectangular window (having spectral dynamic range of 13 dB) is used for spectral analysis. While it is generally accepted that the Hamming window is well suited for estimation of the short-time magnitude spectrum (Picone, 1993; Hayes, 1996), it may not necessarily be the best choice for the estimation of the short-time phase spectrum (e.g., Reddy and Swamy, 1985; Paliwal and Alsteris, 2005; Wójcicki and Paliwal, 2007; Wójcicki and Paliwal, 2008; Loveimi and Ahadi, 2010). For example, in (Wójcicki and Paliwal, 2007) the Chebyshev window function, characterised by adjustable equi-ripple side-lobes (Harris, 1978), was employed as the analysis window in the phase spectrum computation and the effect of its spectral dynamic range on the subjective intelligibility of stimuli constructed from only the short-time phase spectrum was studied. It was observed that lower dynamic range settings for the Chebyshev window resulted in higher subjective intelligibility. More recently, the short-time phase spectrum has also been shown to contain a significant amount of speaker dependent information (Wójcicki and Paliwal, 2008).

Current state-of-the-art speech processing systems use 20–40 ms Hamming window for speech analysis and almost exclusively rely on the processing of the magnitude spectrum to carry out a given speech application. For example, the automatic speech and speaker recognition systems derive cepstral features from the magnitude spectrum alone (e.g., Davis and Mermelstein, 1980; Hermansky, 1990). However, there is an increasing interest in utilising the phase spectrum for automatic speech recognition (Schlüter and Ney, 2001; Alsteris and Paliwal, 2005), speaker recognition (Nakagawa et al., 2007; Wang et al., 2009), speech coding (McAulay and Quatieri, 1995; Skoglund et al., 1997; Pobloth and Kleijn, 1999; Kim, 2003) and speech enhancement (Shi et al., 2006; Wójcicki et al., 2008; Lu and Loizou, 2008).

1.4. Aims of the paper and its organisation

In their classic paper, Wang and Lim (1982) have shown that the short-time phase spectrum is unimportant for speech enhancement. They used the title, "*the unimportance of phase in speech enhancement*", to emphasise this point. Their paper is perhaps the most cited work used to justify the use of noisy phase spectrum in the speech enhancement literature. However, we believe that the short-time phase spectrum can significantly contribute towards speech quality.

The aim of this study is to demonstrate the importance of the short-time phase spectrum for speech enhancement in a systematic manner.[6] For this we use analysis windows of 32 ms duration and investigate a number of approaches for phase spectrum computation. These include the use of matched or mismatched analysis windows for magnitude and phase spectra estimation during AMS processing, as well as the phase spectrum compensation (PSC) method (Stark et al., 2008). We consider four cases and conduct a series of objective and subjective experiments. In the first (oracle) case, the clean phase spectrum is assumed to be known during AMS processing, while the magnitude spectrum is computed directly from noisy speech. Matched and mismatched analysis window approaches are studied. Our goal – in this oracle case – is to determine the maximum speech quality improvements attainable when an estimate of the undistorted phase spectrum is available. While such a situation does not arise in practice, it does provide us

[5] The spectral dynamic range (or, side-lobe attenuation) of a window function is defined as the difference (in dB) in height between the main-lobe and the highest side-lobe.

[6] Note that the aim here is different from our earlier studies (Paliwal, 2003; Paliwal and Alsteris, 2003, 2005; Alsteris and Paliwal, 2004, 2006; Wójcicki and Paliwal, 2007), which investigated the importance of the short-time phase spectrum for speech intelligibility and not speech quality.

with a useful insight into the upper bound of performance attainable through the short-time phase spectrum and thus its importance. The results of the oracle experiments show that accurate phase spectrum estimates considerably improve objective and subjective speech quality. Especially significant improvements were observed for the mismatched window approach, when low dynamic range analysis windows were employed during phase spectrum computation.

In the second (non-oracle) case, we consider a more realistic scenario, where only the noisy spectra (observable in practice) is available during the AMS processing. We study the potential of the mismatched window approach for speech quality improvements in this non-oracle case. We would also like to determine how much room for improvement exists between this case and the best (oracle) case. The results of non-oracle and oracle experiments are consistent, in that the use of low dynamic range analysis windows during phase spectrum estimation improves speech quality. However, only modest gains were observed in the non-oracle case compared to those achieved in the oracle case, showing that there exists room for further improvement and suggesting that research into more accurate phase spectrum estimation algorithms could be worthwhile.

In the third case, we investigate a recent speech enhancement method called phase spectrum compensation (PSC) (Wójcicki et al., 2008; Stark et al., 2008). In the PSC approach the short-time phase spectrum undergoes modification, while the short-time magnitude spectrum is kept unchanged during the enhancement process. The third case experiments compare the performance of the PSC algorithm against the techniques investigated in the oracle and non-oracle experiments. The results of these experiments show that the PSC method outperforms the other non-oracle methods considered.

In the fourth case, we again examine the performance of the oracle and non-oracle matched and mismatched window techniques as well as that of the PSC method, this time – however – enhanced magnitude spectrum is employed instead of the noisy one. Our goal is to determine if the performance of the traditional magnitude spectrum-based enhancement methods can be improved by also processing the phase spectrum. To achieve this, the minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator by Ephraim and Malah (1984) is employed for magnitude spectrum estimation. The results of the MMSE experiments indicate that given accurate phase spectrum estimates, the performance of the existing magnitude spectrum-based enhancement methods can be significantly improved. For the non-oracle approaches, the fusion of the MMSE STSA method (Ephraim and Malah, 1984) with the PSC algorithm (Stark et al., 2008) worked particularly well, producing significantly higher speech quality improvements than those achieved by these methods individually.

The remainder of this paper is organised as follows. Section 2 presents the oracle experiments (case 1). Section 3 details the non-oracle tests (case 2). The PSC experiments (case 3) are given in Section 4, while the MMSE experiments (case 4) are presented in Section 5. A final summary along with concluding remarks is given in Section 6.

2. Case 1: oracle phase spectrum experiments with matched or mismatched analysis windows

2.1. Introduction

In this section, our aim is to determine the maximum benefit in terms of speech quality that can be attained due to the phase spectrum alone. That is, given very precise estimates of the phase spectrum, we would like to determine how much the quality of the noisy speech can be improved during AMS processing when noisy magnitude spectrum is employed. Evaluating this scenario should tell us how much the phase spectrum can contribute towards speech quality. To achieve this, we conduct oracle-style experiments in which stimuli are constructed using a modified AMS procedure, where the phase spectrum is computed from clean speech, while the magnitude spectrum is computed from noisy speech. In these experiments we investigate the use of matched and mismatched analysis windows in the computation of the magnitude and phase spectra. We begin with an objective experiment the goal of which is to determine the dynamic range of the Chebyshev analysis window, used in the phase spectrum computation, that produces highest objective speech quality improvements.[7] We then conduct formal subjective listening tests, where we use the Chebyshev window with the dynamic range that produced highest objective speech quality improvements. We conclude this section with spectrogram analysis of example stimuli employed in the subjective experiment.

2.2. Modified AMS procedure

A modified AMS procedure, similar to the one proposed by Wang and Lim (1982), was used to generate speech stimuli for the oracle-style experiments. A block diagram of the modified AMS framework is shown in Fig. 2. The modified AMS procedure consists of two analysis branches. The first branch is used for analysis of noisy speech $x(n)$, while the second branch is used for analysis of clean speech $s(n)$. The dual-branch framework facilitates the use of matched or mismatched analysis windows, i.e., $w_a(n) = w_b(n)$ or $w_a(n) \neq w_b(n)$, respectively, where $w_a(n)$ is the analysis window employed in the first branch and $w_b(n)$ is the analysis window employed in the second branch. The first analysis branch is used for estimation of the magnitude spectrum from noisy speech, while the second branch is used for estimation of the corresponding

[7] Note that some preliminary results of a similar objective experiment have been previously reported by our group in a conference (Shannon and Paliwal, 2006).

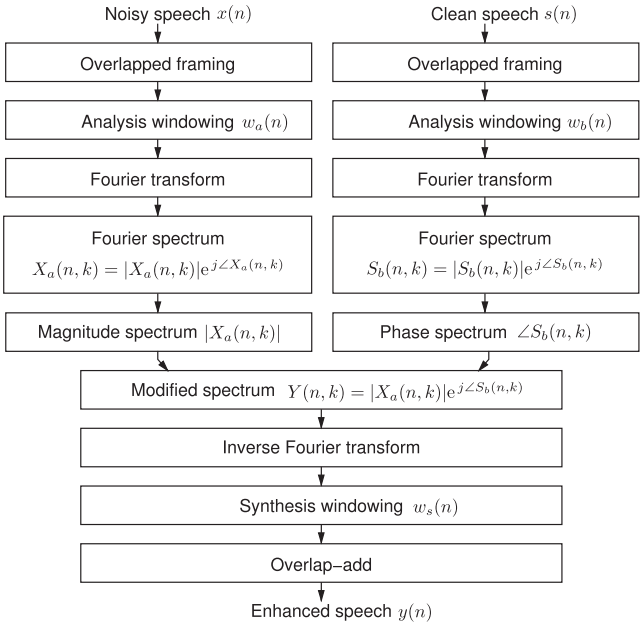

Fig. 2. Block diagram of a modified AMS procedure used for stimuli construction for the oracle (case 1) experiments. Both noisy speech and clean speech are used during processing. The magnitude spectrum is estimated from noisy speech, while the phase spectrum is estimated from clean speech. This framework facilitates the use of matched ($w_a(n) = w_b(n)$) or mismatched ($w_a(n) \neq w_b(n)$) analysis windows in the estimation of the magnitude and phase spectra.

phase spectrum from clean speech. We will refer to these representations as noisy magnitude spectrum and clean phase spectrum and we will denote them as $|X_a(n,k)|$ and $\angle S_b(n,k)$, respectively. In the spectral modification stage of the modified AMS procedure shown in Fig. 2, $|X_a(n,k)|$ and $\angle S_b(n,k)$ are combined to produce the modified spectrum as follows:

$$Y(n,k) = |X_a(n,k)|e^{j\angle S_b(n,k)}. \quad (6)$$

The speech stimulus, $y(n)$, is then constructed by taking the inverse STFT, followed by synthesis windowing and overlap-add reconstruction.

2.3. Speech corpus

In our experiments we employ the Noizeus speech corpus (Hu and Loizou, 2007).[8] Noizeus is composed of 30 phonetically-balanced sentences belonging to six speakers (three males and three females). The recorded speech was originally sampled at 25 kHz. The recordings were then downsampled to 8 kHz and filtered to simulate receiving frequency characteristics of telephone handsets. The corpus comes with non-stationary noises at different SNRs. For our experiments, we keep the clean part of the corpus and generate a corresponding set of stimuli corrupted by additive white Gaussian noise (AWGN) at two SNR levels, 0 and 10 dB.

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

2.4. Objective experiment

In the following subsection we present the details of our objective experiment, the goal of which was to determine the dynamic range of the Chebyshev analysis window that produces highest speech quality improvements.

2.4.1. Stimuli

Speech stimuli were constructed using the modified AMS procedure outlined in Section 2.2. The window length of $T_w = 32$ ms was used throughout. The stimuli can be grouped into five treatment types: clean, noisy, Wang-O, Matched-O and Mismatched-O, based on the AMS settings used to construct them. The treatment types are summarised in Table 1.

There were 30 clean speech stimuli (the undistorted part of the Noizeus corpus) and corresponding 60 noisy speech stimuli (the clean stimuli corrupted by AWGN at 0 and 10 dB SNR).

There were 60 Wang-O stimuli, which were constructed by processing the noisy stimuli through the modified AMS procedure (detailed in Section 2.2) with AMS settings similar to those used by Wang and Lim (1982). The frame shift was set to $T_w/2$. The Hanning window was employed as the analysis window in both branches of the modified AMS procedure, i.e., the same analysis window was used for the estimation of the magnitude and phase spectra. The FFT analysis length was set to N, where $N(=T_wF_s)$ is the number of samples in each frame.

There were 60 Matched-O stimuli, for which stricter AMS settings were used (than those used for Wang-O stimuli construction). The frame shift was set to $T_w/8$ to minimise aliasing. The Hamming analysis window was used in both branches of the modified AMS procedure, i.e., the analysis windows were matched. The FFT analysis length was set to $L = 2N$, i.e., N zeros were appended to each frame prior to FFT analysis.

Finally, there were 600 Mismatched-O stimuli, for which the same strict AMS settings were employed as for the Matched-O stimuli, however, this time the analysis windows were mismatched. The Hamming window was used in the first analysis branch, while the Chebyshev window was used in the second branch. The dynamic range of the Chebyshev window was varied as an experimental parameter, from 5 to 50 dB in 5 dB increments, to determine its effect on objective speech quality of the Mismatched-O stimuli. Overall, 810 stimuli were constructed for the objective experiment.

2.4.2. Objective speech quality measure

Perceptual estimation of speech quality (PESQ) (ITU-T, 2001) was employed as an objective speech quality measure. PESQ prediction maps mean opinion score estimates to a range between −0.5 and 4.5, where 1.0 corresponds to *bad* and 4.5 corresponds to *distortion-less*. In our experiments, mean PESQ scores were computed over relevant treatment sub-groups.

Table 1
Treatment types for the oracle matched and mismatched analysis window experiments (case 1). Oracle-type treatments, i.e., treatments for which the clean phase spectrum is made available during AMS processing of noisy speech, have '-O' label appended as suffix.

Treatment type	Description of processing
Clean	Clean speech
Noisy	Noisy speech (clean speech degraded by AWGN)
Wang-O	Noisy speech processed through the modified AMS procedure shown in Fig. 2, noisy magnitude spectrum and oracle phase spectrum are employed, relaxed AMS settings similar to those reported by Wang and Lim (1982) are used, matched analysis windows ($w_a(n) = w_b(n)$) are employed: Hanning window is used for both $w_a(n)$ and $w_b(n)$
Matched-O	Noisy speech processed through the modified AMS procedure shown in Fig. 2, noisy magnitude spectrum and oracle phase spectrum are employed, strict AMS settings are used, matched analysis windows ($w_a(n) = w_b(n)$) are employed: Hamming window is used for both $w_a(n)$ and $w_b(n)$
Mismatched-O	Noisy speech processed through the modified AMS procedure shown in Fig. 2, noisy magnitude spectrum and oracle phase spectrum are employed, strict AMS settings are used, mismatched analysis windows ($w_a(n) \neq w_b(n)$) are employed: Hamming window is used for $w_a(n)$ and Chebyshev window used for $w_b(n)$

2.4.3. Results and discussion

Results of the objective experiment, in terms of mean PESQ scores as a function of the dynamic range of $w_b(n)$ analysis window, for 0 and 10 dB SNR AWGN are shown in Fig. 3(a) and (b), respectively. For AWGN at 0 dB SNR, the objective speech quality of Wang-O stimuli is slightly higher than the quality of noisy stimuli, while for 10 dB SNR this improvement is negligible. Thus employing clean phase spectrum provides little to no improvement in speech quality for AMS settings similar to those used in (Wang and Lim, 1982). This is consistent with the conclusions drawn by Wang and Lim (1982).

On the other hand, the use of stricter AMS settings, for the construction of the Matched-O stimuli, produces greater improvements in objective quality. This is true for both 0 and 10 dB SNR input speech. These results suggest that in order to attain full benefit (in terms of speech quality) from the clean phase spectrum, the frame shift and FFT size settings employed in the AMS procedure should be chosen carefully.

The results of Fig. 3 also show that Mismatched-O stimuli achieved higher objective speech quality scores than noisy, Wang-O and Matched-O stimuli. This was the case for all dynamic range settings investigated for the $w_b(n)$ analysis window, with the highest improvements achieved for 20–25 dB dynamic range. Thus, the use of mismatched analysis windows, such that $w_a(n)$ is the Hamming window and $w_b(n)$ is the Chebyshev window with dynamic range between 20 and 25 dB, provides highest objective speech quality improvements in this oracle case.

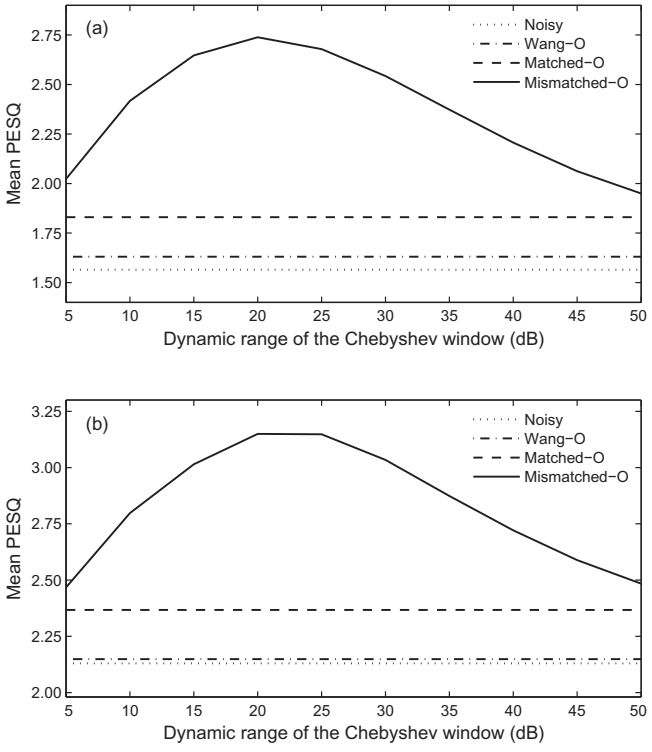

Fig. 3. Results for the objective oracle (case 1) experiment in terms of mean PESQ scores as a function of the dynamic range (dB) of the Chebyshev analysis window $w_b(n)$ for AWGN at: (a) 0 dB SNR and (b) 10 dB SNR. Note that the clean speech stimuli achieved mean PESQ of 4.50.

2.5. Subjective experiment

In the objective experiment of Section 2.4, we have determined that the dynamic range of the $w_b(n)$ Chebyshev window that produces highest objective speech quality improvements for Mismatched-O stimuli is between 20 and 25 dB. We employ this result in a formal subjective evaluation of the importance of the short-time phase spectrum for speech quality. The subjective evaluation is in a form of AB listening tests that determine subjective method

preference. The details of this subjective assessment, along with the results and discussion, are presented in the following subsections.

2.5.1. Stimuli

Stimuli for the subjective experiment were constructed using a procedure similar to the one used in the objective experiment with the following differences. In order to make the listening tests feasible, only two Noizeus sentences were included in these tests (sp10 and sp27), one belonging to a male speaker and the other to a female speaker. The dynamic range of the $w_b(n)$ Chebyshev window used during phase spectrum estimation for Mismatched-O stimuli was set to 20 dB (based on the results of the objective experiment). AWGN at 0 and 10 dB SNR was considered. The stimuli types investigated in this subjective experiment are summarised in Table 1.

2.5.2. Subjects

Twelve English speaking listeners participated in the listening tests. None of the participants reported any hearing deficits.

2.5.3. Procedure

Listening tests were conducted in a quiet room. The participants were familiarised with the task during a short practice session. The actual tests consisted of two sessions, one for each of the two SNR conditions. In each session, the participants listened to 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 if they could not prefer one stimulus over the other. Pairwise scoring was employed, with the score of +1 awarded to the preferred method and +0 to the other. For a similar preference response (i.e., the third option) each method was awarded the 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.

2.5.4. Results and discussion

Results of the subjective experiment for AWGN at 0 and 10 dB SNR, are shown in Fig. 4(a) and (b), respectively. As can be expected, clean speech stimuli achieved highest subjective preference, while noisy speech stimuli were the least preferred. The Mismatched-O stimuli attained significantly higher subjective preference than the noisy, Wang-O and Matched-O stimuli.[9] This was the case at both 0 and 10 dB SNR for AWGN. The Matched-O stimuli achieved significantly higher subjective preference than the noisy and Wang-O stimuli for both noise intensities. Wang-O stimuli achieved significantly higher preference over the noisy stimuli at 0 dB SNR, however, at 10 dB SNR the difference was not significant. Importantly, the results of the listening tests are in agreement with the results of the objective experiment detailed in Section 2.4.

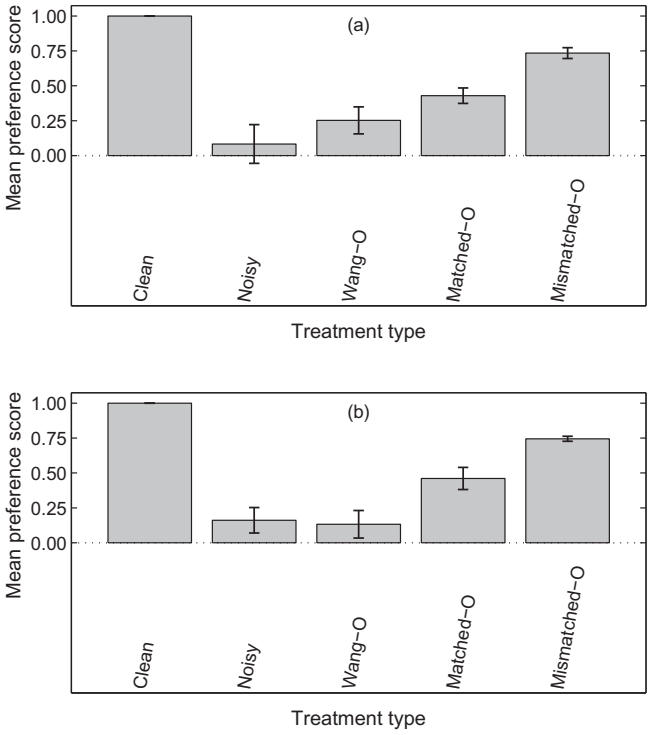

Fig. 4. Results for the subjective oracle (case 1) experiment in terms of mean preference scores for AWGN at: (a) 0 dB SNR and (b) 10 dB SNR. Chebyshev window with 20 dB dynamic range was used as $w_b(n)$ during Mismatched-O stimuli construction.

2.6. Spectrogram analysis

Fig. 5 shows spectrograms of a Noizeus sentence processed using different treatments employed in the subjective experiment detailed in Section 2.5. Spectrograms of the clean speech and noisy speech (AWGN at 10 dB SNR) are shown in Fig. 5(a) and (b), respectively. By comparing the spectrogram of the noisy speech with the spectrogram of the clean speech, it can be seen that the low-energy pitch harmonic structure has been lost due to the additive noise distortion. Some of the low-energy higher formant regions have also been lost.

Spectrogram of Wang-O stimulus shown in Fig. 5(c) is very similar to the spectrogram of noisy speech (Fig. 5(b)), with some additional pitch frequency components present due to undistorted information contained in the clean (oracle) phase spectrum.

Spectrogram of Matched-O stimulus is shown in Fig. 5(d). This spectrogram is quite similar to the spectrogram of Wang-O stimulus. However, the pitch frequency

[9] For the statistical analysis of our results we employ paired one-tailed sign test (Wackerly et al., 2007) with level of significance α set to 0.05.

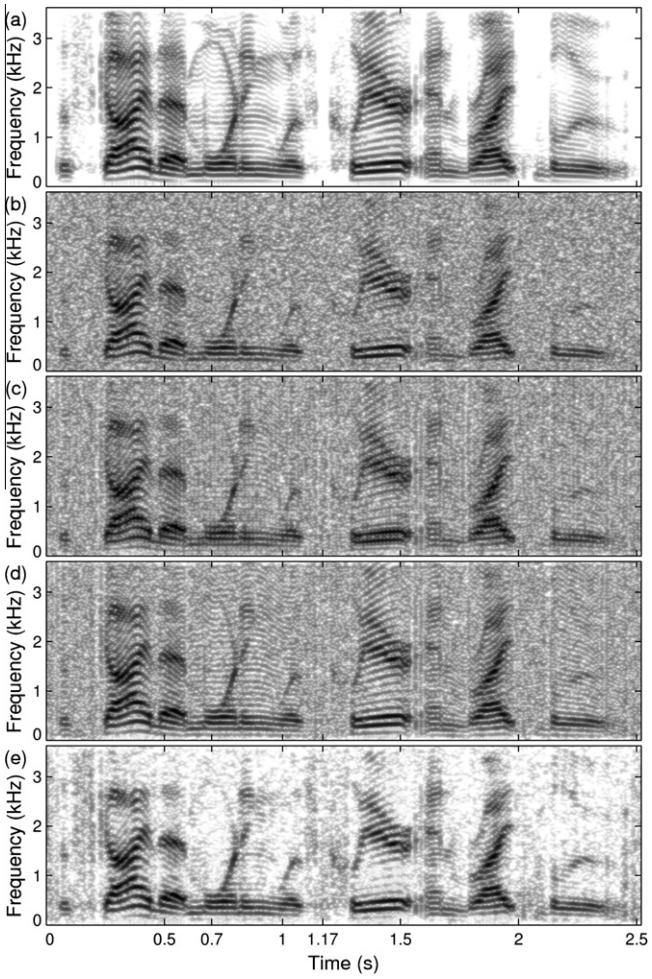

Fig. 5. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (AWGN at 10 dB SNR, PESQ: 2.07); (c) Wang-O (PESQ: 2.10); (d) Matched-O (PESQ: 2.26); and (e) Mismatched-O ($w_b(n)$ set to Chebyshev 20 dB, PESQ: 3.05).

harmonics are more pronounced for Matched-O stimulus than for Wang-O stimulus – they more closely resemble the harmonic structure present in clean speech spectrogram. Thus stricter AMS settings produce finer pitch harmonic detail when oracle phase spectrum is employed.

Finally, the spectrogram of Mismatched-O stimulus is shown in Fig. 5(e). The Mismatched-O stimulus has reduced noise compared with the other stimuli (except for the clean stimulus). Note that the noise suppression is more significant during speech presence regions. As an example of this, compare spectrogram of noisy speech (Fig. 5(b)) with the spectrogram of the Mismatched-O stimulus (Fig. 5(e)). At time 0.7 s (i.e., during speech presence) a significant reduction of noise can be observed in the spectrogram of Mismatched-O stimulus. On the other hand, the noise reduction at time 1.17 s (i.e., during speech absence) is less significant.

From Fig. 5(e) it can also be seen that the Mismatched-O stimulus does not contain the additional harmonic structure present in Wang-O and Matched-O stimuli. Instead, the harmonic structure in Mismatched-O stimulus is somewhat similar to that present before enhancement, i.e., in the noisy stimulus. In the next subsection we show that for the Mismatched-O stimuli there exists a trade-off between noise suppression and preservation of fine spectral detail, and that this trade-off can be controlled through the dynamic range of $w_b(n)$ window function.

2.6.1. Effect of dynamic range of $w_b(n)$ analysis window function on Mismatched-O stimuli

The effect of the dynamic range of $w_b(n)$ analysis window, used during construction of Mismatched-O stimuli, is demonstrated using spectrograms shown in Fig. 6. The use of mismatched analysis windows results in noise reduction, the strength of which depends on dynamic range of $w_b(n)$ Chebyshev window. Dynamic range of $w_b(n)$ close to the dynamic range of $w_a(n)$ means that very little noise is suppressed and most of the fine, low-energy, content is preserved. This is shown in Fig. 6(c), where the Hamming window was used as $w_a(n)$ and the Chebyshev window with the dynamic range of 35 dB was used as $w_b(n)$. On the other hand, dynamic range of $w_b(n)$ that is considerably lower than the dynamic range of $w_a(n)$, results in noise reduction along with some suppression of fine speech content. This is shown in Fig. 6(d), where the Chebyshev window with the dynamic range of 20 dB was used as $w_b(n)$. At the extreme where the dynamic range of $w_b(n)$ is much smaller than the dynamic range of $w_a(n)$, the high energy speech components will also begin to be suppressed. This is shown in Fig. 6(e), where the Chebyshev window with the dynamic range of 5 dB was used as $w_b(n)$. Thus the mismatch in the dynamic range of the $w_a(n)$ and $w_b(n)$ analysis windows controls the trade-off between noise suppression and preservation of fine low-energy spectral content.

How the phase spectrum computed using lower dynamic range window helps in improving the magnitude spectrum of the synthesised stimuli can be briefly explained as follows. We know that the phase and magnitude spectra are related to each other. This relationship has been used in the past to construct[10] the phase spectrum from the magnitude spectrum and vice versa (e.g., Hayes et al., 1980; Nawab et al., 1983; Yegnanarayana et al., 1987; Alsteris and Paliwal, 2007). As a result, when magnitude spectrum information is completely removed in the modification stage of the AMS procedure (i.e., by setting magnitude spectrum values to unity) and speech is synthesised using only the phase spectrum information, we see the magnitude spectrum information (such as formants and pitch frequency harmonics) appearing back in the spectrogram of the synthesised stimulus (Paliwal and Alsteris, 2005, Appendix A).[11]

[10] Under some mild conditions (Hayes et al., 1980).

[11] Note that a similar mechanism has been observed by Ghitza (2001) in the envelope and carrier decomposition framework.

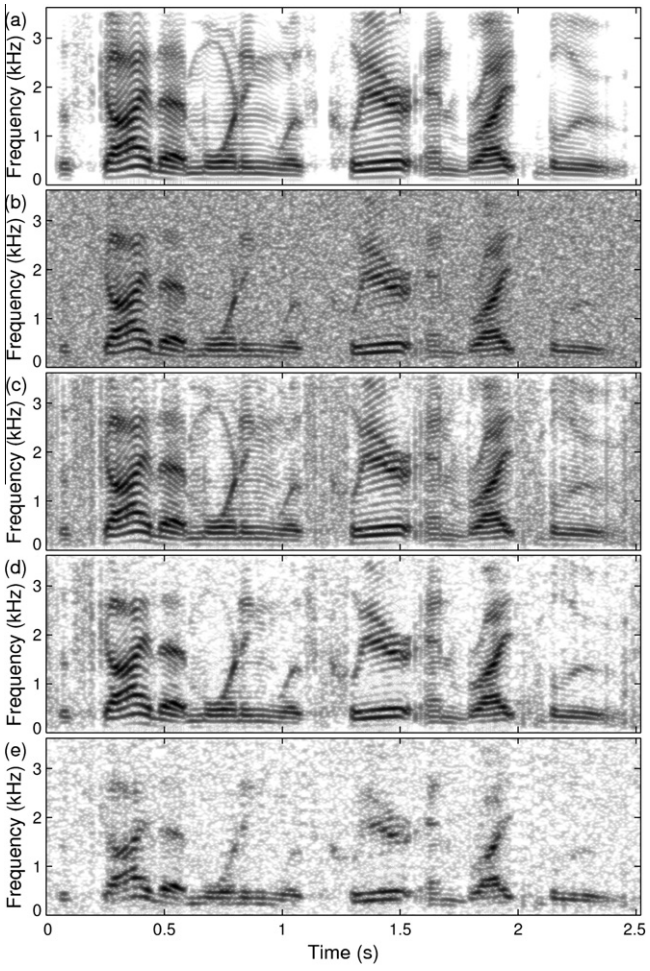

Fig. 6. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (AWGN at 10 dB SNR, PESQ: 2.07); (c) Mismatched-O ($w_b(n)$ set to Chebyshev 35 dB, PESQ: 2.76); (d) Mismatched-O ($w_b(n)$ set to Chebyshev 20 dB, PESQ: 3.05); and (e) Mismatched-O ($w_b(n)$ set to Chebyshev 5 dB, PESQ: 2.35).

We have also shown in our previous work that the relative importance of the phase spectrum can be increased with respect to the magnitude spectrum by using lower dynamic range analysis windows (Paliwal and Alsteris, 2005; Wójcicki and Paliwal, 2007). In the mismatched window experiments reported in this section, we have used the Hamming window for magnitude spectrum computation and lower dynamic range Chebyshev window for phase spectrum computation. The use of lower dynamic range window makes the phase spectrum more important than that of the matched window case (i.e., when using the Hamming window). This is reflected in the improved magnitude spectrum of the synthesised speech.

2.7. Conclusions

This section presented oracle-style experiments that investigated the importance of the short-time phase spectrum for speech quality. In these experiments, speech corrupted by AWGN was processed through modified AMS procedure, in which the undistorted phase spectrum was made available for processing and where noisy (unenhanced) magnitude spectrum was employed. The use of matched or mismatched analysis windows in the computation of magnitude and phase spectra was studied. The effect of the dynamic range of the analysis window used during phase spectrum computation was examined in a systematic manner. The experimental results show, that with the proper choice of the analysis window type and AMS settings, the short-time phase spectrum can significantly contribute towards speech quality and thus has a potential to be useful for speech enhancement. We have shown that through the use of mismatched analysis windows, significant improvements in both objective and subjective speech quality can be achieved. More specifically (in the AMS processing of noisy speech) the use of the Hamming window for magnitude spectrum computation and the use of the Chebyshev window with the dynamic range set to 20–25 dB for phase spectrum computation, produces highest speech quality improvements.

3. Case 2: non-oracle phase spectrum experiments with mismatched analysis windows

3.1. Introduction

In the oracle experiments presented in Section 2, we have determined that significant speech quality improvements are attainable if undistorted phase spectrum is available during AMS-based processing. The mismatched analysis window approach was found to work particularly well. In this section, we look at speech quality improvements that can be achieved when only the noisy spectra is available – a situation that typically arises in practice. In particular, we would like determine how well the mismatched window approach works in this non-oracle case. Note that in the experiments of the present section the noisy magnitude spectrum is again employed during AMS processing (i.e., no enhancement is performed on the magnitude spectrum).

To achieve the above, we conduct a second set of objective and subjective experiments, in which the modified AMS procedure shown in Fig. 7 is used for stimuli construction. This AMS framework is similar to the one employed in the oracle experiments detailed in Section 2.2, in that it has dual branches for estimation of the magnitude and phase spectra components, facilitating the use of matched or mismatched analysis windows. The main difference, however, is that now only the noisy speech is available for processing.

The primary goal of our objective experiment is to determine the dynamic range of the Chebyshev analysis window (used in the phase spectrum estimation) that produces highest objective speech quality scores in this non-oracle case. Based on these results, we conduct formal subjective

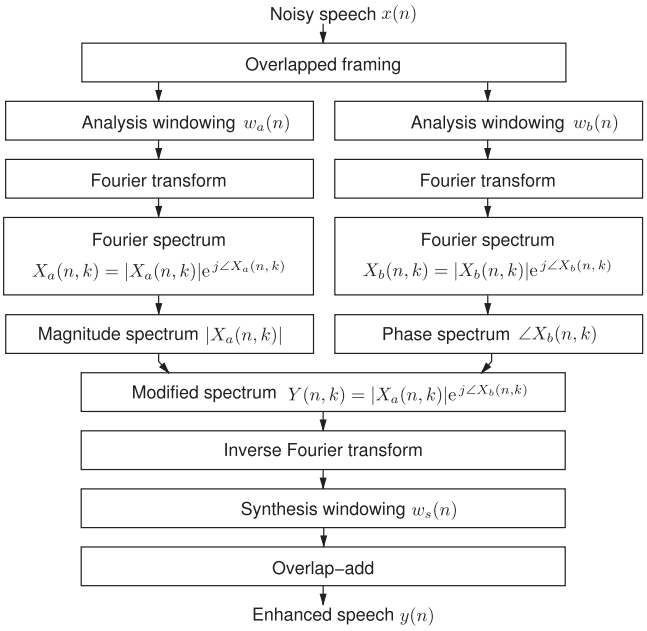

Fig. 7. Block diagram of a modified AMS procedure used for stimuli construction for the non-oracle (case 2) experiments. Here, only the degraded speech is made available for processing. This framework facilitates the use of matched ($w_a(n) = w_b(n)$) and mismatched ($w_a(n) \neq w_b(n)$) analysis windows in the estimation of the magnitude and phase spectra.

listening tests, where for construction of mismatched window stimuli we use the Chebyshev window with the optimal dynamic range determined in the objective experiment. We conclude this section with spectrogram analysis of the non-oracle stimuli.

The comparison of results of the oracle experiments, with the results of the non-oracle experiments, should tell us if there is room for further improvements in speech quality and whether further research into the short-time phase spectrum is worthwhile.

3.2. Objective experiment

Speech stimuli were constructed using the modified AMS procedure shown in Fig. 7. As is summarised in Table 2, the stimuli can be grouped into four types: clean, noisy, Wang-N and Mismatched-N, based on the AMS settings used to construct them. Only noisy speech was used during AMS processing in the construction of the latter three types.[12] The window duration of $T_w = 32$ ms was employed throughout.

Results of the objective experiment, in terms of mean PESQ scores as a function of the dynamic range of $w_b(n)$ Chebyshev window, for 0 and 10 dB input SNRs, are shown in Fig. 8(a) and (b), respectively. These results show that the objective speech quality of Wang-N stimuli is somewhat lower than the quality of noisy stimuli. This suggests that the relaxed AMS settings, employed in (Wang and Lim, 1982), cause a reduction in objective speech quality with respect to the noisy speech.

From the results shown in Fig. 8 it can been that lower dynamic range settings for the $w_b(n)$ analysis window produce Mismatched-N stimuli with higher objective speech quality than both Wang-N and noisy stimuli. The use of mismatched analysis windows, such that $w_a(n)$ is the Hamming window and $w_b(n)$ is the Chebyshev window with 10 dB dynamic range, provides highest objective speech quality improvements in this non-oracle case. As expected, these improvements are not as high as in the oracle case (detailed in Section 2), however, they are significant at a 95% confidence level.

For the construction of Mismatched-N stimuli, the empirically determined optimal dynamic range for the $w_b(n)$ Chebyshev window found in this non-oracle case (10 dB), is somewhat different from the optimal dynamic range found in the oracle case (20–25 dB). This difference can be explained as follows. Lower dynamic range settings for the $w_b(n)$ window are better suited for noisier conditions, since they induce stronger noise suppression during synthesis. The use of clean phase spectrum in the oracle case, means that there is less noise in the modified spectra and hence 20–25 dB windows work well, while in the non-oracle case, the use of noisy phase spectrum means that the resulting modified spectra contains more distortion and, hence, analysis windows with lower dynamic range (such as the 10 dB Chebyshev window function) work better. Note that if we consider AMS processing of clean speech, then the best we can do is to employ matched analysis windows ($w_a(n) = w_b(n)$).

3.3. Subjective experiment

In the objective experiment presented in the previous section, we have determined that the 10 dB setting for the dynamic range of the $w_b(n)$ Chebyshev window (used in the construction of Mismatched-N stimuli) produces highest objective speech quality scores. This result is employed in a subjective non-oracle experiment, the details of which are presented in this section.

The stimuli for the subjective experiment were constructed using the modified AMS procedure shown in Fig. 7. The dynamic range of the $w_b(n)$ Chebyshev window used during phase spectrum computation of the Mismatched-N stimuli was set to 10 dB. The stimuli types considered here are summarised in Table 2.

The subjective experiment was in the form of AB listening tests that determine subjective method preference. The listening tests consisted of two sessions. In the first session, AWGN at 0 dB was investigated, while in the second session, AWGN at 10 dB was considered. The listeners were asked to judge 24 stimuli pairs in each session.

[12] Note that AMS processing with matched analysis windows and strict AMS settings, (essentially) reproduces the input signal at the output. Thus, in this non-oracle case, noisy and Matched-N stimuli can be considered as same. For this reason we include them only once and refer to them as noisy stimuli.

Table 2
Treatment types for the non-oracle mismatched analysis window experiments (case 2). Non-oracle treatments, i.e., treatments for which only the noisy spectra are available during AMS processing of noisy speech, have '-N' label appended as suffix.

Treatment type	Description of processing
Clean	Clean speech
Noisy	Noisy speech (clean speech degraded by AWGN)
Wang-N	Noisy speech processed through the modified AMS procedure shown in Fig. 7, noisy magnitude spectrum and noisy phase spectrum are employed, relaxed AMS settings similar to those reported by Wang and Lim (1982) are used, matched analysis windows ($w_a(n) = w_b(n)$) are employed: Hanning window is used for both $w_a(n)$ and $w_b(n)$
Mismatched-N	Noisy speech processed through the modified AMS procedure shown in Fig. 7, noisy magnitude spectrum and noisy phase spectrum are employed, strict AMS settings are used, mismatched analysis windows ($w_a(n) \neq w_b(n)$) are employed: Hamming window is used for $w_a(n)$ and Chebyshev window is used for $w_b(n)$

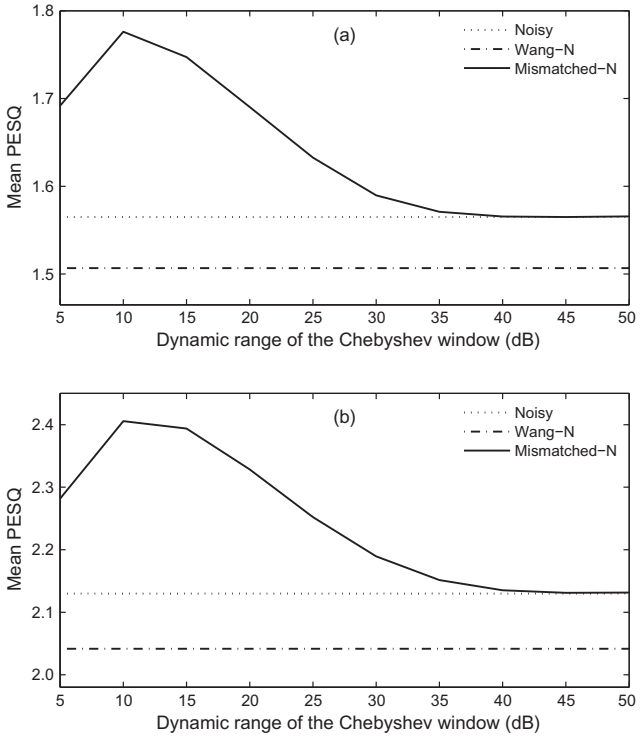

Fig. 8. Results for the objective non-oracle (case 2) experiment in terms of mean PESQ scores as a function of the dynamic range (dB) of the Chebyshev analysis window $w_b(n)$ for AWGN at: (a) 0 dB SNR and (b) 10 dB SNR. Note that the clean speech stimuli achieved mean PESQ of 4.50.

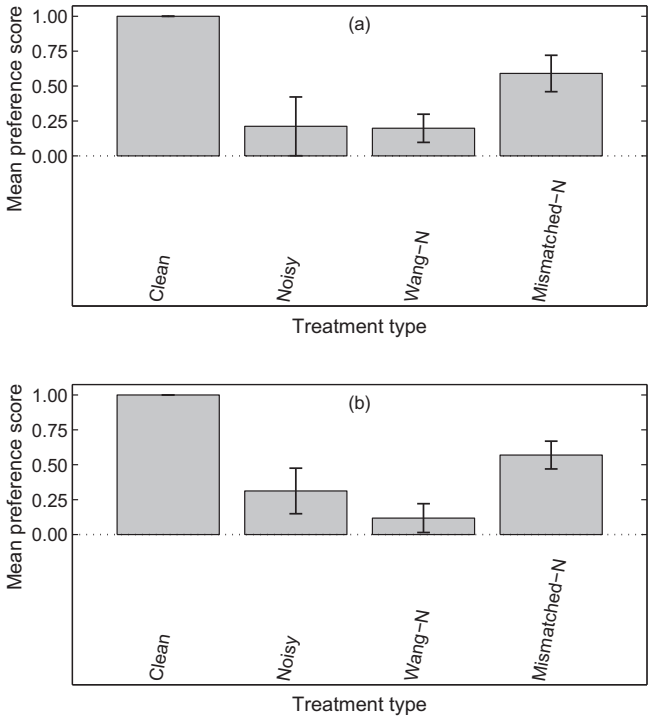

Fig. 9. Results for the subjective non-oracle (case 2) experiment in terms of mean preference scores for AWGN at: (a) 0 dB SNR and (b) 10 dB SNR. Chebyshev window with 10 dB dynamic range was used as $w_b(n)$ during Mismatched-N stimuli construction.

The results of the subjective experiment, for 0 and 10 dB SNR AWGN, are shown in Fig. 9(a) and (b), respectively. The clean speech stimuli have the highest subjective preference, while the Wang-N stimuli have the lowest preference. This reaffirms the observation made based on the earlier results of the objective experiment of Section 3.2, that the relaxed AMS settings used in the construction of Wang-N stimuli cause a reduction in speech quality with respect to the quality of noisy speech.

Mismatched-N stimuli achieved significantly higher subjective preference scores than noisy and Wang-N stimuli. This can be attributed to the noise reduction achieved through the use of mismatched analysis windows during Mismatched-N stimuli construction. Note that the above results are consistent with the results of the objective experiment presented in Section 3.2.

3.4. Spectrogram analysis

Fig. 10(a–c and e) shows spectrograms of the sp10 Noizeus sentence for stimuli types included in the subjective experiment of Section 3.3. The spectrograms of clean and noisy speech (AWGN at 10 dB SNR) are shown in

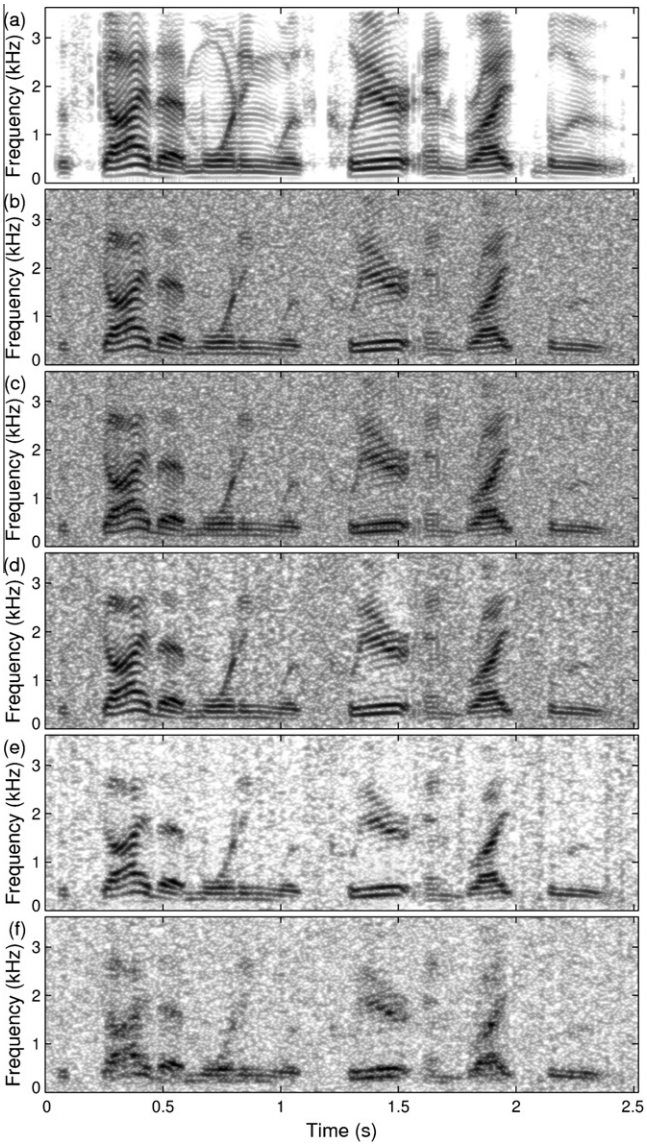

Fig. 10. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (AWGN at 10 dB SNR, PESQ: 2.07); (c) Wang-N (PESQ: 2.00); and (d) Mismatched-N ($w_b(n)$ set to Chebyshev 25 dB, PESQ: 2.15); (e) Mismatched-N ($w_b(n)$ set to Chebyshev 10 dB, PESQ: 2.27); and (f) Mismatched-N ($w_b(n)$ set to Chebyshev 2 dB, PESQ: 1.92).

Fig. 10(a) and (b), respectively. Spectrogram of Wang-N stimulus is shown in Fig. 10(c). Compared with the spectrogram of the noisy speech, the spectrogram of Wang-N stimulus shows slightly reduced harmonic structure. This reduction could in part explain why Wang-N stimuli achieved lower objective and subjective scores than the noisy stimuli.

The spectrograms of Mismatched-N stimuli for $w_b(n)$ set to Chebyshev 25, 10 and 2 dB are shown in Fig. 10(d–f), respectively. The use of mismatched analysis windows in this non-oracle case results in noise reduction, the strength of which depends on the dynamic range of the $w_b(n)$ Chebyshev window. This is consistent with the observations made earlier in the oracle case (see the discussion in Section 2.6.1).

Out of the three dynamic range setting for $w_b(n)$ considered here, 25 dB is the closest to the dynamic range of the Hamming window used for $w_a(n)$. For this reason the spectrogram of the Mismatched-N stimulus, constructed using the 25 dB setting and shown in Fig. 10(d), shows only a very slight noise reduction and is otherwise very similar to the spectrogram of noisy speech given in the spectrogram of Fig. 10(b). This is to be expected and can be explained as follows. Speech processed through the AMS procedure with matched analysis windows, strict AMS settings and without spectral modification will appear unchanged at its output.[13] If the analysis windows are not matched but are similar, then the input speech signal will undergo little change in the AMS procedure. This is observed when comparing spectrograms of noisy speech (Fig. 10(b)) with the spectrograms of Mismatched-O stimuli constructed with $w_b(n)$ set to Chebyshev 25 dB (Fig. 10(d)). Dynamic range settings for $w_b(n)$ even closer to the dynamic range settings for $w_a(n)$ would result in an even smaller change in the reconstructed stimuli. On the other hand, as can be seen in spectrograms of Fig. 10(e and f), lower dynamic range settings for $w_b(n)$ result in much stronger noise suppression along with some loss of signal components. By comparing the three spectrograms of Mismatched-N stimuli (Fig. 10(d–f)) with the spectrogram of noisy speech of Fig. 10(b), it can be seen that a compromise between preservation of speech content and noise reduction is achieved for $w_b(n)$ set to Chebyshev 10 dB (Fig. 10(e)). This dynamic range setting for $w_b(n)$ produced highest mean PESQ scores in the objective experiment (see Section 3.2).

3.5. Conclusions

In this section, we have presented non-oracle experiments, the goal of which was to determine the usefulness of the mismatched window approach when only the noisy speech is available for processing (i.e., for stimuli construction noisy magnitude and phase spectra was employed). The results of our experiments show that the use of mismatched analysis windows in the modified AMS framework, results in noise reduction at a cost of some loss of signal components. More specifically, employing the Chebyshev window with the dynamic range set to 10 dB as the analysis window in the computation of the phase spectrum, while using the Hamming window in the computation of the magnitude spectrum, results in noise reduction and significant improvement of speech quality. Comparison of the results of the non-oracle experiments, with the results of the oracle experiments, suggests that there exists room for further improvements and that if

[13] Depending on the type of synthesis employed in the AMS procedure, perfect or least-squares reconstruction will be produced.

more accurate phase spectrum estimates can be obtained, then the speech quality can be improved much further.

4. Case 3: phase spectrum compensation experiments

4.1. Introduction

Traditional AMS-based speech enhancement techniques process only the noisy magnitude spectrum and keep noisy phase spectrum unchanged (e.g., Lim and Oppenheim, 1979; Berouti et al., 1979; Ephraim and Malah, 1984; Ephraim et al., 1985; Martin, 1994; Sim et al., 1998; Virag, 1999). Recently, a novel speech enhancement algorithm called phase spectrum compensation (PSC) has been proposed (Wójcicki et al., 2008; Stark et al., 2008). In the PSC method the phase spectrum is altered during AMS processing, while the magnitude spectrum is kept unchanged. In this section, our aim is to compare the performance of the PSC method with the performance of the matched and mismatched window techniques presented in the preceding sections.

4.2. Phase spectrum compensation procedure

The PSC method is based on the AMS framework detailed in Section 1.2. The noisy speech signal, used in the analysis stage of the AMS framework, is a real-valued signal and, therefore, its STFT is conjugate symmetric, i.e., $X(n,k) = X^*(n, L-k)$, where L is the number of samples in the frequency domain. The PSC method controls the degree to which the conjugates reinforce or cancel by altering their angular relationship. This is achieved through the use of time and frequency dependent phase spectrum compensation function, $\Lambda(n,k)$. The PSC function is computed as follows:

$$\Lambda(n,k) = \lambda\Psi(k)|\widehat{D}(n,k)|, \tag{7}$$

where λ is a real-valued empirically determined constant, $\Psi(k)$ is a time-invariant antisymmetry function given by

$$\Psi(k) = \begin{cases} 1, & \text{if } 0 < k/L < 0.5 \\ -1, & \text{if } 0.5 < k/L < 1 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

and $|\widehat{D}(n,k)|$ is an estimate of the short-time magnitude spectrum of the noise. The above formulation, facilitates for handling of time and/or frequency varying noise conditions.

Since the spectral noise estimate $|\widehat{D}(n,k)|$ is symmetric, multiplication by the scaled antisymmetry function ($\lambda\Psi(k)$) produces an antisymmetric $\Lambda(n,k)$ function. It is this antisymmetry that forms the primary basis for noise cancellation during synthesis. The $\Lambda(n,k)$ function is used to offset the complex spectrum of the noisy speech as follows:

$$X_\Lambda(n,k) = X(n,k) + \Lambda(n,k). \tag{9}$$

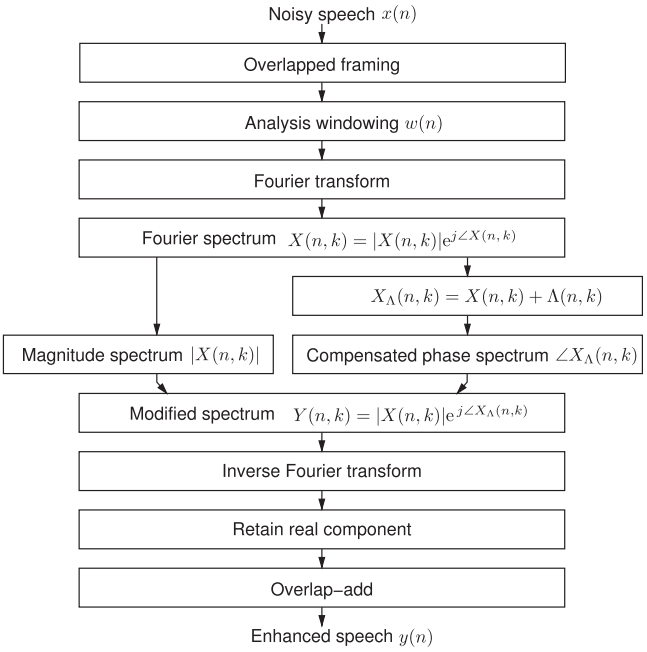

Fig. 11. Block diagram of the phase spectrum compensation procedure (Stark et al., 2008) for speech enhancement.

The compensated phase spectrum is then computed using

$$\angle X_\Lambda(n,k) = ARG[X_\Lambda(n,k)], \tag{10}$$

where ARG is the complex angle function.[14] The compensated phase spectrum is recombined with the noisy magnitude spectrum to produce a modified (complex) spectrum

$$Y(n,k) = |X(n,k)|e^{j\angle X_\Lambda(n,k)}. \tag{11}$$

In the synthesis stage, the inverse STFT is used to convert the frequency-domain frames, $Y(n,k)$, to the time-domain representation. Due to the additive offset introduced in Eq. (9), the resulting time-domain frames may be complex. In the PSC method the imaginary component is discarded. The enhanced time-domain signal, $y(n)$, is produced using overlap-add synthesis. A block diagram of the PSC method is shown in Fig. 11.

The PSC-based noise cancellation can be explained by viewing the frequency-domain representation of the speech signal as groupings of conjugate vectors. These conjugates arise naturally from the symmetry of the magnitude spectrum and anti-symmetry of the phase spectrum and are a result of applying the STFT to a real-valued signal. During the inverse STFT operation the conjugates sum together to produce a real-valued signal. By modifying the angular relationship of the conjugates, we can influence the degree to which they reinforce or cancel and, thus, the degree to which they contribute towards the reconstructed time-domain signal.

[14] Note that the compensated phase spectrum may not possess the properties of a true phase spectrum, i.e., one that is computed from a real-valued signal.

Since the PSC function is antisymmetric, applying it to the short-time Fourier spectrum offsets the conjugate vectors (and thus alters their angles) by pushing them in the opposite directions – one toward 0 radians and the other toward π radians. The further they are pushed apart, the more out of phase they become. The strength of the compensation is dependent on both the magnitude of the signal vectors and magnitude of noise vectors (since the PSC function depends on these). Low-magnitude signal components badly corrupted by high-magnitude noise components will undergo greatest cancellation. This allows for targeted suppression of noise vectors.

Let us consider two cases of angular modification for a pair of conjugate vectors depicted in Fig. 12. To make our discussion simpler, we drop both time and frequency indexes and only consider a single pair of conjugate vectors. In the first case, Fig. 12(a), the magnitudes of the conjugates, i.e., $|\vec{X}|$ and $|\vec{X}^*|$, are larger than $|\Lambda|$ (the magnitude of the PSC function). This results in limited change of the original signal. In the second case, Fig. 12(b), the vector magnitudes are smaller than $|\Lambda|$. Significant angular change occurs, as the two conjugate vectors, $\vec{X}$ and $\vec{X}^*$, are pushed toward the real-axis facing 0 and π radians, respectively. Summation produces a significant cancellation, leaving little or no real-valued component. As can be seen in Fig. 12, the strength of cancellation for a given $|\Lambda|$ is dependent on the STFT magnitude of the noisy speech, $|\vec{X}| = |\vec{X}^*|$, with larger magnitude components being less attenuated and smaller magnitude components being more attenuated. Noise frequency components are typically assumed to have much smaller magnitudes than speech signal components. This assumption is basis for many noise cancellation and noise estimation algorithms (Loizou, 2007). Using this assumption, the additive PSC function can be tuned to induce significant angular modification that results in cancellation among noise vectors, but with a limited effect on the speech signal carrying vectors.

4.3. Experiments

This section describes objective and subjective speech enhancement experiments that compare the PSC method with the short-time phase spectrum-based techniques investigated in the previous sections. The treatment types tested here are summarised in Table 3, and consist of the PSC treatment as well as the treatments considered in Section 2 oracle experiments and Section 3 non-oracle experiments.

The PSC stimuli were constructed using the modified AMS procedure shown in Fig. 11. The window duration of $T_w = 32$ ms was used throughout. The FFT length was set to $L = 2N$, i.e., each N samples long time domain frame was padded with N zeros prior to frequency analysis. The tunable parameter λ was set to 3.74 as suggested in (Stark et al., 2008).

In the objective experiment, mean PESQ scores were computed over the Noizeus corpus for each treatment type

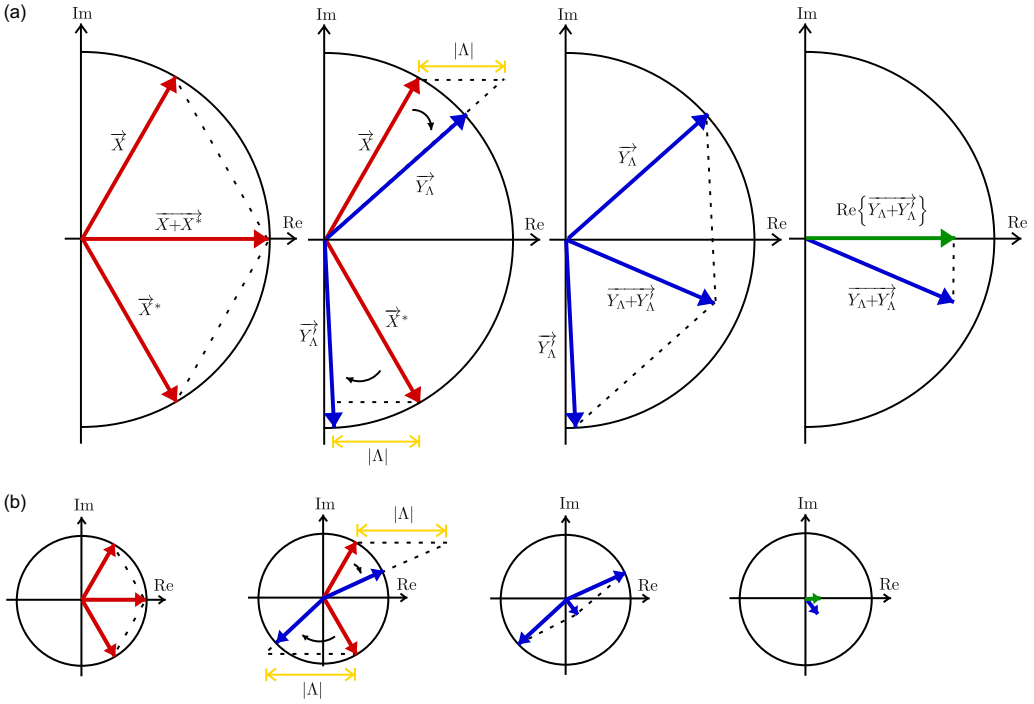

Fig. 12. Vector diagrams: modification of STFT conjugate symmetry of a single conjugate pair. Top row (a): $|\vec{X}| > |\Lambda|$. Bottom row (b): $|\vec{X}| < |\Lambda|$. Column one: conjugate vectors, $\vec{X}$ and $\vec{X}^*$, as well as their addition vector, $\overrightarrow{X+X^*}$. Column two: the real parts of the conjugate vectors are offset by $|\Lambda|$ and $-|\Lambda|$. Thus, the angles of vectors $\vec{X}$ and $\vec{X}^*$ are altered, while their magnitudes are kept unchanged to produce vectors $\vec{Y}_\Lambda$ and $\vec{Y}'_\Lambda$, respectively (see Eq. (11)). Column three: the resulting vectors are added to produce the $\overrightarrow{Y_\Lambda + Y'_\Lambda}$ vector. Column four: the imaginary part of the $\overrightarrow{Y_\Lambda + Y'_\Lambda}$ addition vector is discarded. For clarity both time and frequency indexes have been omitted in this figure.

Table 3
Treatment types for the PSC experiments (case 3). Oracle-type treatment labels have '-O' appended as suffix, while non-oracle treatment labels are suffixed with '-N'. Note that the '-N' modifier is implied for the PSC method, since PSC uses only the noisy spectra during processing, however, the '-N' suffix is not explicitly included in the PSC treatment label.

Treatment type	Description of processing
Clean	Clean speech
Noisy	Noisy speech (clean speech degraded by AWGN)
Matched-O	Noisy speech processed through the modified AMS procedure shown in Fig. 2, noisy magnitude spectrum and oracle phase spectrum are employed, strict AMS settings are used, matched analysis windows ($w_a(n) = w_b(n)$) are employed: Hamming window is used for both $w_a(n)$ and $w_b(n)$
Mismatched-N	Noisy speech processed through the modified AMS procedure shown in Fig. 7, noisy magnitude spectrum and noisy phase spectrum are employed, strict AMS settings are used, mismatched analysis windows ($w_a(n) \neq w_b(n)$) are employed: Hamming window is used for $w_a(n)$ and Chebyshev window is used for $w_b(n)$
Mismatched-O	Noisy speech processed through the modified AMS procedure shown in Fig. 2, noisy magnitude spectrum and oracle phase spectrum are employed, strict AMS settings are used, mismatched analysis windows ($w_a(n) \neq w_b(n)$) are employed: Hamming window is used for $w_a(n)$ and Chebyshev window used for $w_b(n)$
PSC	Noisy speech enhanced using the PSC procedure (Stark et al., 2008) shown in Fig. 11

at four input SNR levels: 0, 5, 10 and 15 dB (AWGN). For the subjective experiment AB listening tests were employed at two input SNR levels: 0 and 10 dB (AWGN). The remaining experimental details are similar to the ones employed in the oracle and non-oracle experiments of Sections 2 and 3, respectively.

4.4. Results and discussion

The results of the objective experiment are shown in Fig. 13. Out of the stimuli shown, Mismatched-O and noisy speech stimuli achieved highest and lowest mean PESQ scores (across the input SNR range), respectively. The PSC method was the second best, while Matched-O and Mismatched-N were comparable and were third best. The results of the subjective experiment, for 0 and 10 dB SNR, are shown in Fig. 14(a) and (b), respectively. The results of the subjective experiment are consistent with the corresponding results of the objective experiment. The following observations can be made based on the above experimental results:

- out of the non-oracle methods considered, the PSC method achieved best improvements – this can be seen by comparing results for Mismatched-N and PSC stimuli;
- the results for the best-performing oracle method, (i.e., the Mismatched-O method) are better than for any of the non-oracle methods considered in this study – this indicates that further speech quality improvements in the non-oracle scenario may be possible if more accurate phase spectrum estimates can be obtained.

Fig. 15 shows spectrograms of the sp10 Noizeus sentence for different stimuli types considered in the subjective experiment. Spectrograms of clean and noisy speech

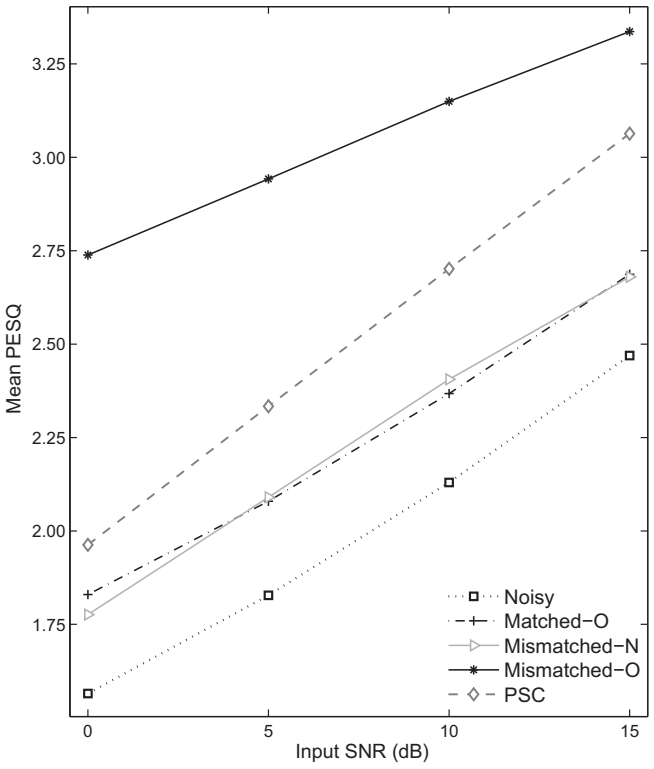

Fig. 13. Results for the objective case 3 experiment in terms of mean PESQ scores as a function of input SNR (dB) for AWGN. Chebyshev window with the dynamic range set to 10 dB and 20 dB was used as $w_b(n)$ for the construction of Mismatched-N and Mismatched-O stimuli, respectively. Note that the clean speech stimuli achieved mean PESQ of 4.50.

(AWGN at 10 dB SNR) are shown in Fig. 15(a) and (b), respectively. Spectrogram of Matched-O stimulus is shown in Fig. 15(c). Spectrograms of Mismatched-N and Mismatched-O stimuli are shown in Fig. 15(d) and (e), respectively. Spectrogram of PSC stimulus is shown in Fig. 15(f).

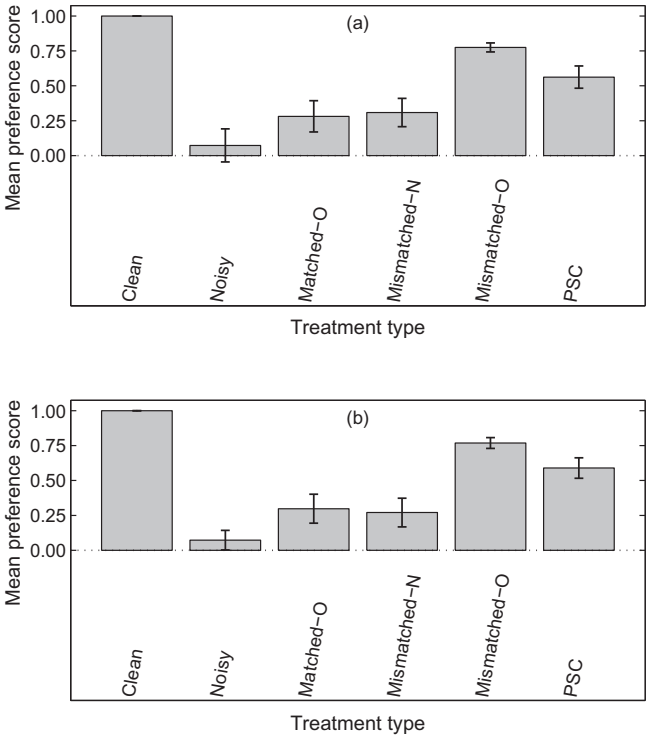

Fig. 14. Results for the subjective case 3 experiment in terms of mean preference scores for AWGN at: (a) 0 dB SNR and (b) 10 dB SNR. Chebyshev window with the dynamic range set to 10 dB and 20 dB was used as $w_b(n)$ for the construction of Mismatched-N and Mismatched-O stimuli, respectively.

By comparing the spectrogram of noisy speech (Fig. 15(b)) with the spectrogram of the PSC enhanced speech (Fig. 15(h)), it can be seen that the PSC method achieves noise reduction. This noise reduction is comparable to that achieved by the (non-oracle) Mismatched-N method (Fig. 15(d)). The Mismatched-N method, however, suffers from somewhat stronger speech signal suppression. This can be seen by comparing the spectrogram of the Mismatched-N stimulus (shown in Fig. 15(d)) with the spectrogram of the PSC stimulus (shown in Fig. 15(f)), where some of the low-energy speech signal spectral components in Fig. 15(d) are suppressed more so than in Fig. 15(f). The noise reduction achieved by the PSC method (Fig. 15(f)) is not as good as the reduction achieved by the (oracle) Mismatched-O method (Fig. 15(e)).

4.5. Conclusions

In this section we have compared a number of oracle and non-oracle AMS-based speech processing approaches that attempt to derive some benefit in terms of speech quality from the short-time phase spectrum. These include matched and mismatched analysis window techniques as well as the PSC method – which compensates the phase spectrum for additive noise distortion. In our experiments noisy (unenhanced) magnitude spectrum was employed. Out of the non-oracle methods considered in our evaluation, the PSC procedure achieved highest objective and

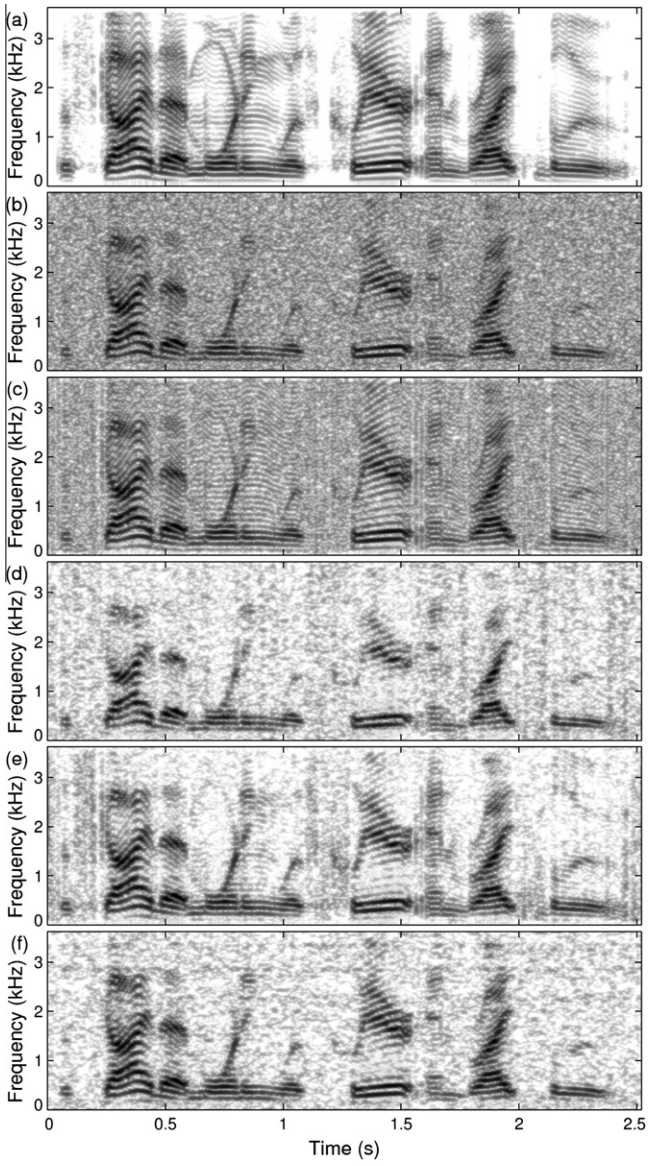

Fig. 15. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (AWGN at 10 dB SNR, PESQ: 2.07); (c) Matched-O (PESQ: 2.26); (d) Mismatched-N ($w_b(n)$ set to Chebyshev 10 dB, PESQ: 2.27); (e) Mismatched-O ($w_b(n)$ set to Chebyshev 20 dB, PESQ: 3.05); and (f) PSC (PESQ: 2.51).

subjective speech quality improvements. The best oracle method (Mismatched-O) achieved scores higher than any other method considered. This indicates that it may be possible to attain further speech quality improvements due to phase spectrum for the non-oracle approaches if more accurate phase spectrum estimates can be obtained.

5. Case 4: MMSE experiments

5.1. Introduction

In the experiments of cases 1, 2 and 3, presented in the preceding sections, we have investigated a number of approaches

Table 4
Treatment types for the MMSE experiments (case 4). Oracle-type treatment labels have '-O' appended as suffix, while non-oracle treatment labels are suffixed with '-N'. Note that, while not explicitly included, the '-N' modifier is implied for the MMSE, PSC and MMSE-PSC methods. Treatments that employ MMSE STSA estimator (Ephraim and Malah, 1984) are prefixed with the 'MMSE-' label.

Treatment type	Description of processing
Clean	Clean speech
Noisy	Noisy speech (clean speech degraded by AWGN)
MMSE	Noisy speech enhanced using the MMSE STSA method (Ephraim and Malah, 1984), reference implementation by Loizou (2007) is employed (see Fig. 17)
MMSE-Matched-O	Noisy speech processed through the modified AMS procedure shown in Fig. 16, MMSE STSA spectrum (Ephraim and Malah, 1984) and oracle phase spectrum are used, strict AMS settings are used, matched analysis windows ($w_a(n) = w_b(n)$): Hamming window is used for both $w_a(n)$ and $w_b(n)$
MMSE-Mismatched-N	Noisy speech processed through the modified AMS procedure shown in Fig. 17, MMSE STSA spectrum (Ephraim and Malah, 1984) and noisy phase spectrum are used, strict AMS settings are used, mismatched analysis windows ($w_a(n) \neq w_b(n)$): Hamming window is used for $w_a(n)$ and Chebyshev window is used for $w_b(n)$
MMSE-Mismatched-O	Noisy speech processed through the modified AMS procedure shown in Fig. 16, MMSE STSA spectrum (Ephraim and Malah, 1984) and oracle phase spectrum are used, strict AMS settings are used, mismatched analysis windows ($w_a(n) \neq w_b(n)$): Hamming window is used for $w_a(n)$ and Chebyshev window used for
PSC	Noisy speech enhanced using the PSC procedure (Stark et al., 2008) shown in Fig. 11
MMSE-PSC	Noisy speech processed through the modified AMS procedure shown in Fig. 18, MMSE STSA spectrum (Ephraim and Malah, 1984) and phase spectrum compensated for additive noise distortion using the PSC method (Stark et al., 2008) are used, strict AMS settings are used, Hamming window is employed for $w(n)$

for computation of the short-time phase spectrum within the short-time Fourier AMS framework. These approaches included matched and mismatched analysis window techniques as well as the PSC method (Stark et al., 2008). Both oracle and non-oracle scenarios were considered, in which the phase spectrum was computed from either clean or noisy speech, respectively. Our goal was to determine the potential of the above approaches for improvement of speech quality when noisy magnitude spectrum is utilised (i.e., no enhancement was performed on the magnitude spectrum).

In the present section, we investigate the potential of the above phase spectrum computation techniques for improvement of speech quality when enhanced magnitude spectrum is employed. This should tell us how useful the above approaches are in the context of current state-of-the-art speech enhancement methods, which traditionally process the magnitude spectrum, but keep the noisy phase spectrum unchanged (Loizou, 2007). Thus our objective is to determine whether it is possible to further improve the enhancement performance of the magnitude spectrum-based enhancement by also processing the phase spectrum. For this purpose, in the experiments reported in this section, the minimum mean square error (MMSE) short-time spectral amplitude (STSA) estimator by Ephraim and Malah (1984) is used to estimate the magnitude spectrum of clean speech from noisy observations. The MMSE[15] estimate of clean magnitude spectrum is then combined with complex exponential of the phase spectrum computed using techniques investigated in Sections 2–4. Reconstructed stimuli are then employed in a series of objective and subjective speech enhancement experiments.

The remainder of this study is organised as follows. Section 5.2 provides details of the methods used for stimuli construction for speech enhancement experiments of subsequent sections. Section 5.3 presents a series of objective experiments used firstly to determine suitable settings for the methods under investigation, and then to objectively compare the performance of these methods for enhancement of speech corrupted by AWGN under various SNR conditions. Section 5.4 provides details of the subjective experiment along with a discussion of its results. Spectrogram analyses of the stimuli considered in this study are presented in Section 5.6. Concluding remarks are given in Section 5.7.

5.2. Methods

This section provides details of procedures used to apply treatment types outlined in Table 4 for stimuli construction for objective and subjective experiments presented in Sections 5.3 and 5.4, respectively. The treatment types prefixed with the 'MMSE-' label employ the MMSE STSA estimator (Ephraim and Malah, 1984) for estimation of the clean speech magnitude spectrum from noisy speech. We will refer to the resulting MMSE STSA spectral estimate as the MMSE magnitude spectrum or MMSE STSA spectrum.

[15] In this work, the MMSE modifier implies MMSE STSA estimate (or estimator) by Ephraim and Malah (1984), unless otherwise stated.

Throughout this work, for the computation of the MMSE magnitude spectrum, we use a reference implementation by Loizou (2007). We set the frame duration to $T_w = 32$ ms and the frame shift to $T_s = T_w/4$. We keep all other reference settings given in (Loizou, 2007) unchanged, including the $\alpha = 0.98$ parameter used in the decision-directed approach for *a priori* SNR estimation. The Hamming window function is used for the magnitude spectrum estimation throughout, while for the phase spectrum, either the Hamming or Chebyshev window function is employed, depending on the treatment type. The FFT size is set to $L = 2N$, where N is frame duration in samples. Block diagrams of the procedures used for the construction of stimuli types summarised in Table 4 are shown in Figs. 16–18, with the block diagram of the PSC method (Stark et al., 2008) given in Fig. 11. In the following subsections, we detail method specific settings for each of the above procedures.

5.2.1. MMSE STSA spectrum and oracle phase spectrum with matched or mismatched analysis windows

For construction of MMSE-Matched-O and MMSE-Mismatched-O stimuli, we employ the modified AMS procedure shown in Fig. 16. Here, the MMSE magnitude spectrum is computed from noise corrupted speech signal, while the phase spectrum is computed from the corresponding clean speech signal.

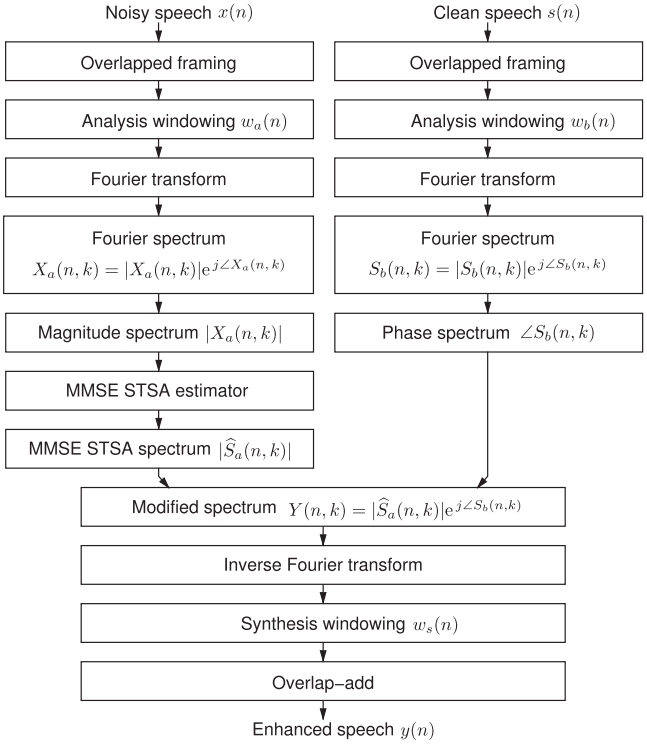

Fig. 16. Block diagram of a modified AMS procedure used for oracle stimuli construction (namely, for MMSE-Matched-O and MMSE-Mismatched-O stimuli) for the MMSE experiments (case 4). Both noisy speech and clean speech are used during processing. The phase spectrum is estimated from clean speech, while the MMSE STSA estimator (Ephraim and Malah, 1984) is used for the magnitude spectrum estimation from noisy speech. This framework facilitates the use of matched ($w_a(n) = w_b(n)$) or mismatched ($w_a(n) \neq w_b(n)$) analysis windows in the estimation of the magnitude and phase spectra.

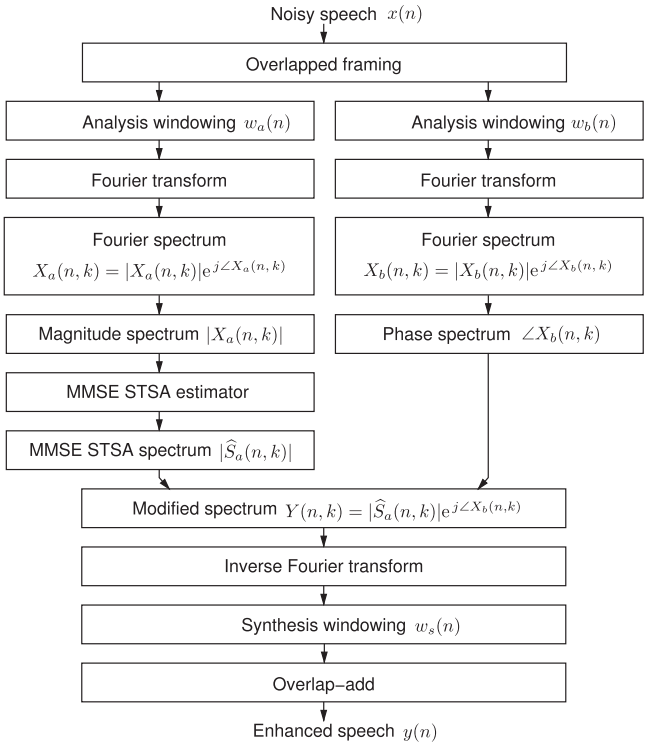

Fig. 17. Block diagram of a modified AMS procedure used for non-oracle stimuli construction (specifically, for MMSE and MMSE-Mismatched-N stimuli) for the MMSE experiments (case 4). Here, only the degraded speech is made available for processing. The MMSE STSA estimator (Ephraim and Malah, 1984) is used for the magnitude spectrum estimation. This framework facilitates the use of matched ($w_a(n) = w_b(n)$) or mismatched ($w_a(n) \neq w_b(n)$) analysis windows in the estimation of the magnitude and phase spectra.

Matched analysis windows, i.e., $w_a(n) = w_b(n)$, are used for construction of MMSE-Matched-O stimuli. The Hamming window is employed for this purpose. Note that the MMSE-Matched-O treatment is essentially the MMSE STSA method proposed by Ephraim and Malah (1984), however, here the oracle (clean) phase spectrum is used instead of the noisy phase spectrum. Our aim is to determine how much benefit can be attained in terms of speech quality in the context of traditional magnitude spectrum-based speech enhancement methods, when accurate estimates of the phase spectrum are available. Mismatched analysis windows, i.e., $w_a(n) \neq w_b(n)$, are employed during construction of MMSE-Mismatched-O stimuli. The Hamming window is used as $w_a(n)$ analysis window for computation of the magnitude spectrum from noisy speech, while the Chebyshev window is employed as $w_b(n)$ analysis window for computation of the phase spectrum from clean speech. Our goal is to determine if the mismatched window approach can provide us with further speech quality improvements when enhanced magnitude spectrum is used along with the oracle phase spectrum. This should tell us, how useful the mismatched window approach can be (in

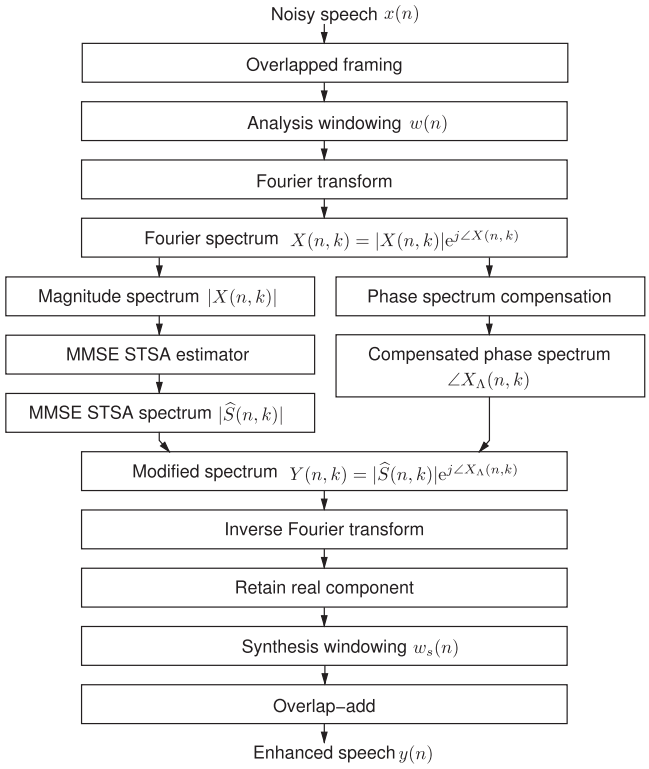

Fig. 18. Block diagram of the proposed MMSE-PSC method for speech enhancement.

the context of traditional magnitude spectrum-based enhancement methods) if accurate phase spectrum estimates can be obtained.

5.2.2. MMSE STSA spectrum and non-oracle phase spectrum with matched or mismatched analysis windows

For construction of MMSE and MMSE-Mismatched-N stimuli, we employ the modified AMS procedure shown in Fig. 17. Here, both the MMSE magnitude spectrum and the phase spectrum are computed from noisy speech.

MMSE stimuli are constructed using the MMSE STSA method (Ephraim and Malah, 1984). For this purpose, a reference implementation by Loizou (2007) is employed. As stated previously, the reference settings used in (Loizou, 2007) are kept unchanged, with the exception of frame duration and frame shift (see Section 5.2). Matched analysis windows are used, where the Hamming window is employed for both $w_a(n)$ and $w_b(n)$.

Mismatched analysis windows are used during construction of MMSE-Mismatched-N stimuli. The Hamming window is used as $w_a(n)$ analysis window for computation of the magnitude spectrum, while the Chebyshev window is employed as $w_b(n)$ analysis window for computation of the phase spectrum. Our aim is to determine if the mismatched window approach can provide us with speech quality improvements when enhanced magnitude spectrum is used along with noisy phase spectrum.

5.2.3. Noisy magnitude spectrum and compensated phase spectrum – the PSC method

For the construction of the PSC stimuli, the PSC procedure (Stark et al., 2008) detailed in Section 4.2 and shown in Fig. 11 is used.

5.2.4. MMSE STSA spectrum and compensated phase spectrum – the MMSE-PSC method (proposed)

The modified AMS procedure shown in Fig. 18 is used for construction of MMSE-PSC stimuli. Here, only the noisy speech is available for processing. The Hamming window function is employed for both $w_a(n)$ and $w_b(n)$ analysis windows. MMSE STSA estimate (Ephraim and Malah, 1984) of the clean magnitude spectrum is combined with the complex exponential of the phase spectrum compensated for additive noise distortion using the PSC method (Stark et al., 2008).

5.3. Objective experiments

In the objective experiments, the enhancement performance of the methods listed in Table 4 was investigated using the PESQ measure (ITU-T, 2001). The procedures detailed in Section 5.2 were employed for stimuli construction. The experiments were performed over the Noizeus corpus for AWGN at various noise intensities, and comprised of three parts – the details of which are presented in the following subsections.

5.3.1. Part 1 – Effect of dynamic range of $w_b(n)$ analysis window function on objective quality of MMSE-Mismatched-N and MMSE-Mismatched-O stimuli

In the first part, our goal was to examine the effect of dynamic range of the Chebyshev analysis window function on the objective quality of the MMSE-Mismatched-N and MMSE-Mismatched-O stimuli. The MMSE and MMSE-Matched-O methods were also included in this comparison. AWGN case was considered at two SNR levels, 0 and 10 dB. The results of this evaluation are shown in Fig. 19. It can be seen that the use of the oracle phase spectrum improves the objective quality of the MMSE enhanced speech for both matched and mismatched analysis window approaches (i.e., MMSE-Matched-O and MMSE-Mismatched-O treatments, respectively). The improvements were more significant, however, for the mismatched window method, where 20–40 dB dynamic range of the Chebyshev $w_b(n)$ analysis window function worked well, with highest improvements achieved for the 30 dB setting. The above results show that there exists a good potential for significant improvement of enhancement performance for traditional magnitude spectrum-based methods, if accurate phase spectrum estimates can be obtained and especially when suitable mismatched analysis windows are employed. Note that while accurate phase spectrum estimation from noise corrupted speech signal is by no means an easy task (Wang and Lim, 1982; Ephraim and Malah, 1984), our hope is that the results presented in

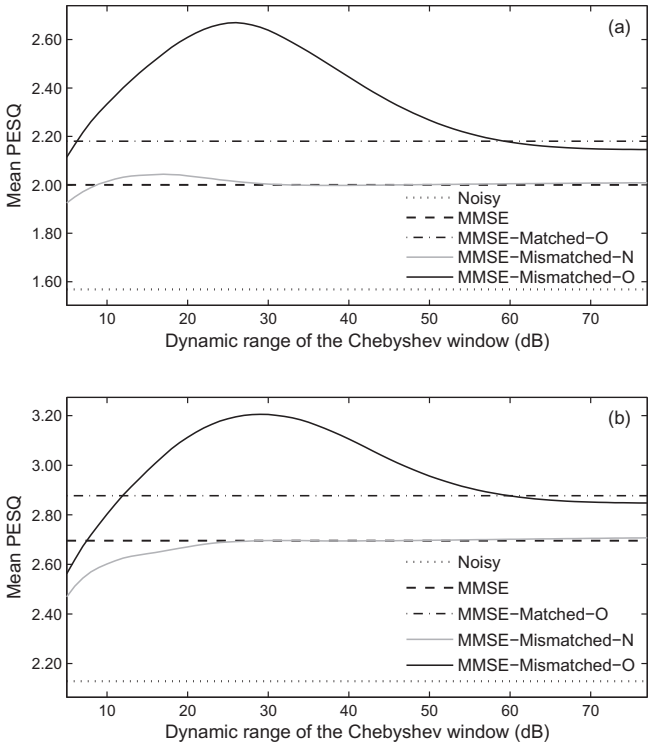

Fig. 19. Results for the objective MMSE experiment (case 4) in terms of mean PESQ scores as a function of the dynamic range (dB) of the Chebyshev analysis window $w_b(n)$ for AWGN at: (a) 0 dB SNR and (b) 10 dB SNR. Note that the clean speech stimuli achieved mean PESQ of 4.50.

this study will encourage further research into phase spectrum processing for speech enhancement.

The results of Fig. 19 indicate that the mismatched window approach used in conjunction with the MMSE magnitude spectrum (i.e., for MMSE-Mismatched-N stimuli construction) does not afford an improvement in objective speech quality. Instead, the enhancement performance of the MMSE-Mismatched-N method matches the baseline performance of the MMSE STSA method (Ephraim and Malah, 1984) when the dynamic ranges of $w_a(n)$ and $w_b(n)$ analysis windows are quite similar. This is to be expected, since in the MMSE STSA method (Ephraim and Malah, 1984) same analysis window is used during computation of the magnitude and phase spectra. The PESQ measure also indicates that the speech quality of the MMSE-Mismatched-N stimuli reduces at very low dynamic range settings for the $w_b(n)$ analysis window. Interestingly, for Mismatched-N stimuli (where noisy magnitude spectrum was employed – i.e., no enhancement was performed on the magnitude spectrum – and where mismatched analysis windows were used) some objective speech quality improvements were observed (see the objective results of the non-oracle case 2 experiment presented in Section 3.2). This can be explained as follows. When the noisy magnitude spectrum is employed, the mismatched window approach can be used to improve speech quality, since there exists much room for improvement. On the other hand, when the MMSE STSA estimate is employed instead of the noisy magnitude spectrum, there is much less room for further improvements. Consequently, the mismatched window approach is less effective when enhanced magnitude coefficients are employed along with the noisy phase spectrum. Recall that the mismatched window approach significantly improved noisy speech quality for the oracle MMSE-Mismatched-O stimuli. This seems to suggest that in order for magnitude spectrum-based enhancement methods to benefit from mismatched window approach better estimates of the phase spectrum are needed. Note, however, that during informal listening tests we have found that the 25 dB setting for the $w_b(n)$ Chebyshev analysis window produced minor noise reduction for MMSE-Mismatched-N stimuli as compared with MMSE stimuli (see also the spectrogram analyses presented in Section 5.6). Consequently, this setting is included in the subsequent objective and subjective experiments of Sections 5.3.3 and 5.4, respectively.

5.3.2. Part 2 – Effect of λ parameter on objective quality of PSC and MMSE-PSC stimuli

In the second part, our aim was to investigate the effect of λ parameter[16] on objective quality of PSC and MMSE-PSC stimuli. For comparison, the MMSE and MMSE-Matched-O methods were also included. AWGN case was investigated at 0 and 10 dB SNR. The result of this evaluation are shown in Fig. 20. In terms of mean PESQ scores, the MMSE method rated slightly higher at 0 dB SNR than the PSC method with λ set to 3.74. At 10 dB SNR, the MMSE and PSC methods were comparable for $3.5 \leqslant \lambda \leqslant 7$. The performance of the MMSE method was surpassed by the MMSE-PSC method, with highest objective improvements achieved for $\lambda \approx 2.6$. At this λ setting, however, the residual noise was still noticeable for the AWGN case. Using informal listening tests, $\lambda = 3.74$ was selected for use in experiments of later sections. Note that the effect of different values of the λ parameter on the PSC and MMSE-PSC stimuli is discussed in much detail as part of spectrogram analysis presented in Section 5.6.2.

The MMSE-Matched-O method achieved highest improvements in this second part of the objective evaluation – again highlighting the premise that if accurate phase spectrum estimates can be obtained, then the performance of the existing magnitude spectrum-based enhancement methods can be significantly improved.

5.3.3. Part 3 – Comparative evaluation for AWGN

In the third part, we objectively evaluate all treatment types listed in Table 4 (and detailed in Section 5.2) over a range of SNRs for AWGN. Dynamic range settings for

[16] Recall that the λ parameter in the PSC method (Stark et al., 2008) controls the overall strength of synthesis-based cancellation. Thus, larger values of λ will lead to stronger noise suppression. However, for very large λ values the speech signal components will also be attenuated. Refer to Section 4.2 for further details of the PSC method.

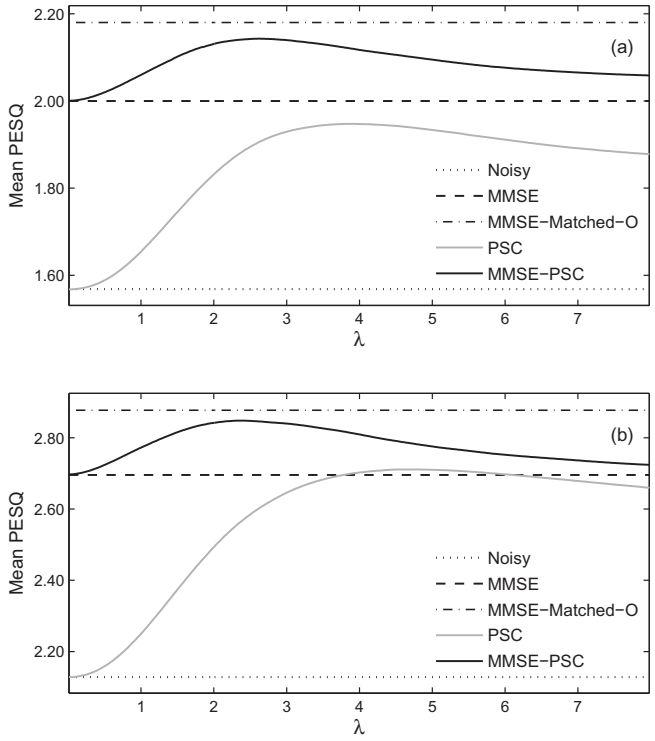

Fig. 20. Results for the objective MMSE experiment (case 4) in terms of mean PESQ scores as a function of λ parameter for AWGN at: (a) 0 dB SNR and (b) 10 dB SNR. Note that the clean speech stimuli achieved mean PESQ of 4.50.

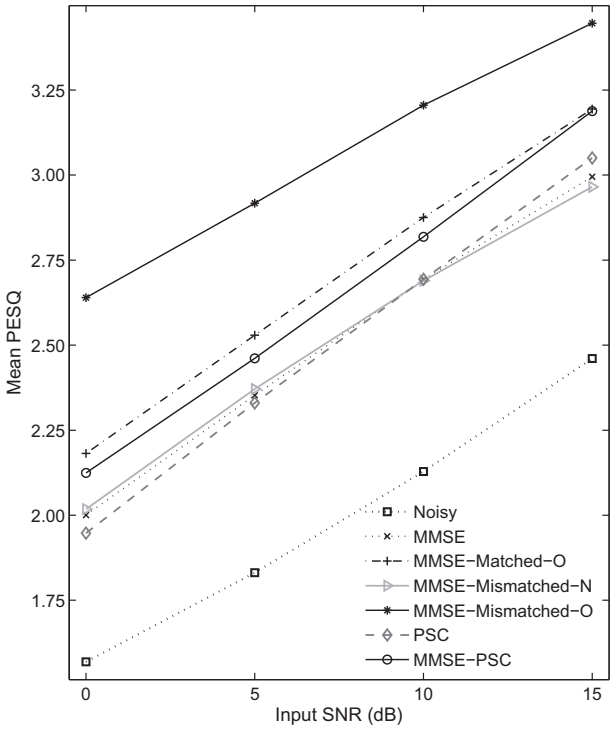

Fig. 21. Results for the objective MMSE experiment (case 4) in terms of mean PESQ scores as a function of input SNR. Chebyshev window function, with the dynamic ranges set to 25 and 30 dB was used as $w_b(n)$ for the construction of Mismatched-N and Mismatched-O stimuli, respectively. For PSC and MMSE-PSC methods, λ was set to 3.74. Note that the clean speech stimuli achieved mean PESQ of 4.50.

the $w_b(n)$ analysis window (for mismatched window approaches) as well as the λ parameter setting (for PSC-based methods) arrived at in the previous parts of this objective evaluation were employed. More specifically, $w_b(n)$ was set to Chebyshev 25 dB for the MMSE-Mismatched-N treatment, and to 30 dB for the MMSE-Mismatched-O treatment. The λ parameter was set to 3.74 for both, the PSC and MMSE-PSC methods.[17] The results of this evaluation are shown in Fig. 21. All methods significantly improved noisy speech quality. The MMSE-Mismatched-O method achieved consistently higher improvements than all other methods considered in this evaluation. The next-best approach was the MMSE-Matched-O method, closely followed by the MMSE-PSC method, which was the best performing non-oracle approach. The MMSE, MMSE-Mismatched-N and PSC methods achieved comparable objective scores.

The objective results for the MMSE-Mismatched-O method, demonstrate that processing of the short-time phase spectrum in addition to processing of the short-time magnitude spectrum can significantly improve speech quality – well beyond the improvements attainable by processing the magnitude spectrum alone (provided that accurate phase spectrum estimates can be obtained). While the mismatched analysis window approach does not significantly improve objective speech quality for the non-oracle stimuli (i.e., for MMSE-Mismatched-N stimuli), the encouraging oracle results (i.e., for MMSE-Matched-O and MMSE-Mismatched-O stimuli) should provide the motivation needed for further research into better methods for phase spectrum estimation. Moreover, the MMSE-PSC method achieved objective scores higher than the MMSE, MMSE-Mismatched-N and PSC methods, suggesting that the PSC processing has a good potential for combination with existing magnitude spectrum-based speech enhancement methods.

5.4. Subjective experiment

The aim of the subjective experiment was to compare the treatment types listed in Table 4 (and detailed in Section 5.2) using human speech perception tests for AWGN at two SNR levels, 0 and 10 dB. Same settings were used for stimuli construction as in part 3 of the objective experiment presented in Section 5.3.3. The test format detailed in Section 2.5 was employed.

The results of the subjective experiment for AWGN at 0 and 10 dB SNR, are shown in Fig. 22(a) and (b),

[17] Note that while $\lambda = 3.74$ was found to work well for MMSE-PSC method in the case of AWGN, for coloured noises a lower λ setting (e.g., $\lambda = 2.6$) was found to work better in general. However, in an effort to keep settings as consistent as possible throughout this work, $\lambda = 3.74$ was used for both AWGN presented here, as well as for coloured noises presented in Appendix A.

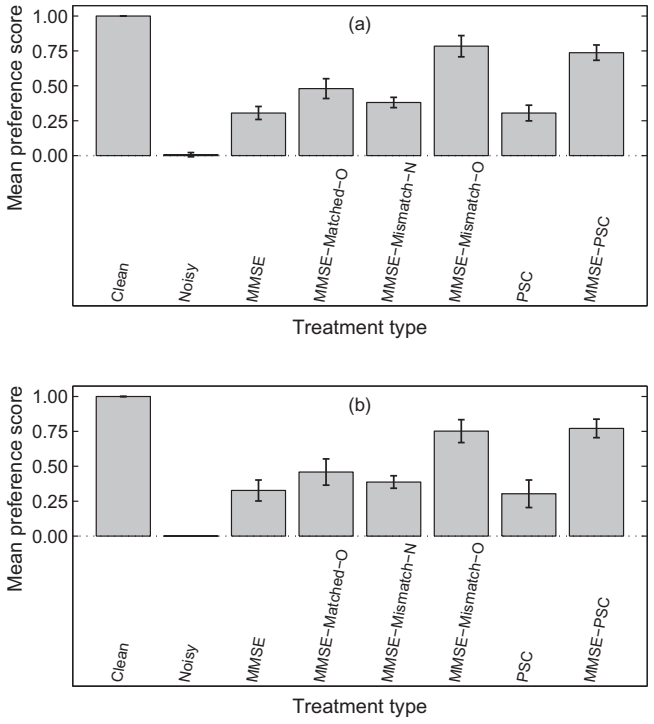

Fig. 22. Results for the subjective MMSE experiment (case 4) in terms of mean preference scores for AWGN at: (a) 0 dB SNR and (b) 10 dB SNR. Chebyshev window function, with the dynamic ranges set to 25 and 30 dB was used as $w_b(n)$ for the construction of Mismatched-N and Mismatched-O stimuli, respectively. For PSC and MMSE-PSC methods, λ was set to 3.74.

respectively. Clean speech stimuli achieved significantly higher preference than all other stimuli types. All treatments achieved significantly higher preference than noisy speech. Listener preference for the MMSE, MMSE-Mismatched-N and PSC treatments did not differ significantly. Out of the MMSE-based methods, the preference for the MMSE-Mismatched-O and MMSE-PSC approaches did not differ significantly, but was significantly higher than for the MMSE, MMSE-Matched-O, MMSE-Mismatched-N and PSC methods. Worth noting is that the subjective results across the two noise intensities are consistent.

The subjective results for clean, noisy, MMSE, MMSE-Mismatched-N and PSC stimuli are in agreement with the objective results presented in Section 5.3.3. On the other hand, while the objective results indicate that the MMSE-PSC method is: only slightly better than the baseline MMSE approach, not quite as good as the MMSE-Matched-O method and far worse than the MMSE-Mismatched-O approach, the subjective measure indicates that the MMSE-PSC method is significantly better than both the MMSE baseline and the MMSE-Matched-O method, while being comparable to the MMSE-Mismatched-O approach. This is a very significant result as it suggests that processing phase spectrum alongside magnitude spectrum can provide a significant benefit for speech enhancement.

The difference between objective and subjective results can be explained as follows. The MMSE-Matched-O method benefits from extra information contributed by the oracle phase spectrum. As a result, MMSE-Matched-O stimuli contain additional spectral details, such as fine pitch frequency harmonic structure and formant details.[18] These additional spectral details are not present in the (non-oracle) MMSE-PSC stimuli. However, the MMSE-PSC method results in a significantly better noise reduction, than the MMSE-Matched-O method. Note that both MMSE-PSC and MMSE-Matched-O approaches employ the MMSE magnitude spectrum, hence, both methods *inherit* noise reduction due to the MMSE processing. The PSC component of the MMSE-PSC method, however, reduces the noise even further, while the inclusion of the oracle phase spectrum in the MMSE-Matched-O method, does not reduce the noise beyond the reduction already achieved due to the MMSE processing (i.e., the noise reduction of the MMSE-Matched-O and MMSE methods is comparable, while the noise reduction for the MMSE-PSC method is significantly better than for the MMSE method). The PESQ metric seems to favour the extra spectral details over noise suppression, while the listeners preferred the reduced noise of MMSE-PSC stimuli over the contribution of increased spectral detail of MMSE-Matched-O stimuli.

The results of the subjective experiment support the observations made based on the objective results of Section 5.3.3. While the mismatched analysis window approach works very well when oracle phase spectrum is employed, the quality of the non-oracle stimuli is not significantly improved over the MMSE baseline. This is indicated by the subjective performance of the MMSE-Mismatched-O and MMSE-Mismatched-N methods, respectively. This again suggests that better phase spectrum estimates have a potential to produce better results in the non-oracle scenarios.

The PSC approach has proven an effective counterpart to the MMSE processing, by significantly reducing the intensity of the residual noise. This is demonstrated by the subjective results for the MMSE-PSC method.

5.5. A note regarding optimality of noisy phase spectrum

In their work, Ephraim and Malah (1984) derived an optimal MMSE estimator of the complex exponential of the phase under the constraint that its modulus is unity. They showed that this estimator is equal to the complex exponential of the noisy phase. They also derived an unconstrained estimator. Though both the estimators are optimal, the former one was preferred because of its speech enhancement performance.

While many speech enhancement methods derive the enhanced signal spectrum by optimising some mathematically tractable error criterion (e.g., Ephraim and Malah, 1984; Ephraim et al., 1985; Sim et al., 1998; Loizou, 2007), they need not necessarily be optimal from the point

[18] For a detailed comparison of spectral characteristics of stimuli constructed using the methods investigated in this study the reader is refered to the spectrograms and discussions of Section 5.6.

of view of human speech perception. The human ear, on the other hand, can be considered as the ultimate judge of speech quality. For this reason, in the present work we have employed listening tests involving human subjects to evaluate the performance of the investigated methods. Using these tests, we have shown that it is possible to significantly improve enhancement performance of the MMSE STSA method by also processing the phase spectrum.

5.6. Spectrogram analyses

Spectrograms of a Noizeus sentence, processed using different treatment types investigated in the subjective experiment of Section 5.4, are shown in Fig. 23. Spectrograms of clean and noisy speech (AWGN at 10 dB SNR) are shown in Fig. 23(a) and (b), respectively. Spectrogram of a MMSE stimulus (i.e., noisy speech enhanced using the MMSE STSA method by Ephraim and Malah (1984)) is shown in Fig. 23(c). With respect to noisy speech (Fig. 23(b)) the MMSE stimulus (Fig. 23(c)) has significantly reduced noise. Some residual noise does remain, however, its nature is non-musical. Spectrogram of a MMSE-Matched-O stimulus (i.e., noisy speech enhanced using the MMSE STSA method (Ephraim and Malah, 1984), where the clean (oracle) phase spectrum was employed) is shown in Fig. 23(d). By comparing the MMSE stimulus (Fig. 23(c)) with the MMSE-Matched-O stimulus (Fig. 23(d)), it can be seen that the oracle phase spectrum contributes additional spectral details, such as fine harmonic structure and lower energy higher formant details – some of which were destroyed by noise and thus are absent in the spectrogram of noisy speech (Fig. 23(b)).

Spectrogram of a MMSE-Mismatched-N stimulus in Fig. 23(e) shows a minor reduction in noise during speech presence, as compared to the spectrogram of MMSE stimulus shown in Fig. 23(c). On the other hand, the spectrogram of MMSE-Mismatched-O stimulus shown in Fig. 23(f), has significantly reduced noise along with significantly augmented spectral features as compared to both, the spectrogram of noisy speech (Fig. 23(b)) and the spectrogram of MMSE stimulus (Fig. 23(c)). The spectrogram of a PSC stimulus in Fig. 23(g) shows a significant noise reduction, as compared to the spectrogram of noisy speech in Fig. 23(b), without an adverse effect on speech signal components. Finally, the spectrogram of a MMSE-PSC stimulus given in Fig. 23(h) (and produced by combining the MMSE STSA method (Ephraim and Malah, 1984) with the PSC method (Stark et al., 2008)) shows best noise reduction compared to the other techniques investigated in this study. Importantly, the improved noise reduction of the MMSE-PSC method was achieved without a substantial loss of spectral detail, i.e., the formant and pitch frequency harmonic structures of the MMSE-PSC enhanced speech do not differ significantly with those present in the MMSE enhanced speech. This can be seen by comparing

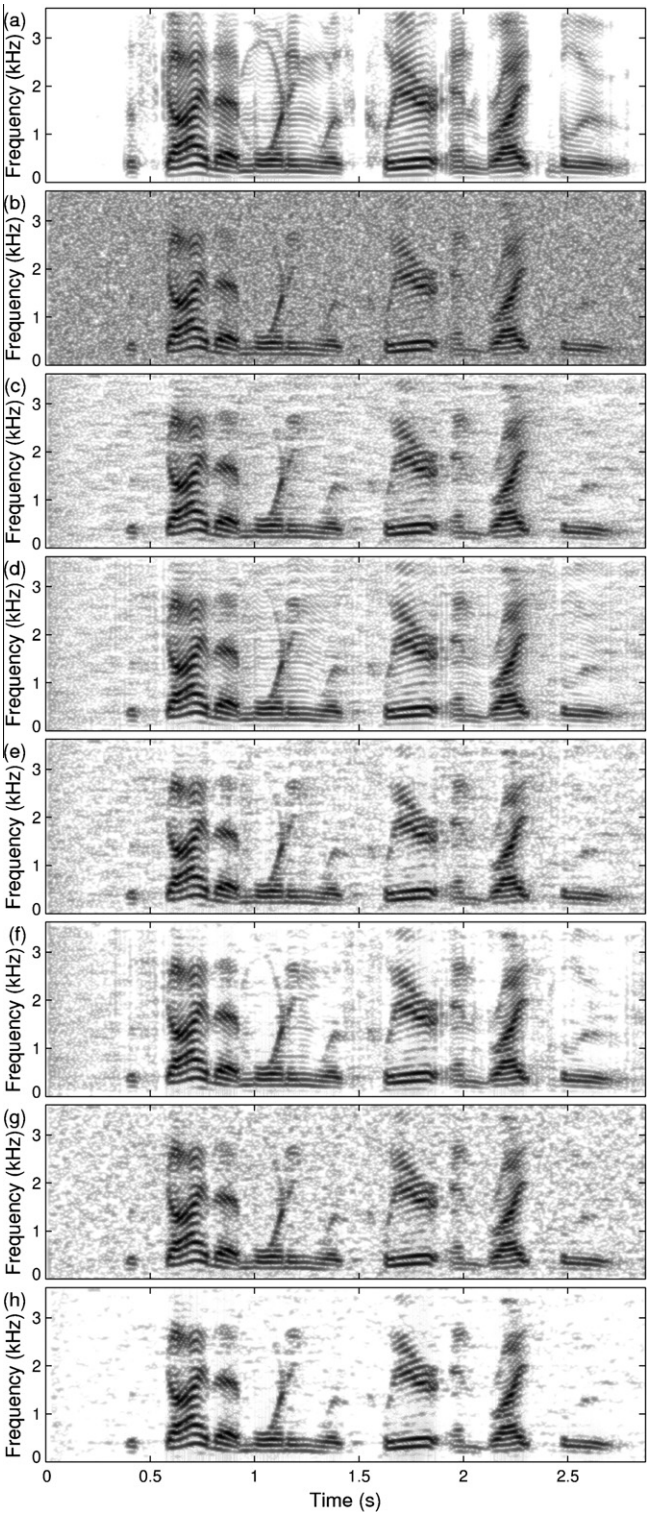

Fig. 23. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (AWGN at 10 dB SNR, PESQ: 2.09); (c) MMSE (PESQ: 2.61); (d) MMSE-Matched-O (PESQ: 2.74); (e) MMSE-Mismatched-N ($w_b(n)$ set to Chebyshev 25 dB, PESQ: 2.58); (f) MMSE-Mismatched-O ($w_b(n)$ set to Chebyshev 30 dB, PESQ: 3.12); (g) PSC (PESQ: 2.56); and (h) MMSE-PSC (PESQ: 2.65).

the spectrogram of Fig. 23(h) with the spectrogram of Fig. 23(c) and was also confirmed using informal listening tests. This is an important result, as it demonstrates that the performance of the magnitude spectrum-based enhancement methods can be significantly improved by also processing the phase spectrum.

5.6.1. Effect of dynamic range of $w_b(n)$ analysis window function on MMSE-Mismatched-N and MMSE-Mismatched-O stimuli

Spectrograms of Fig. 24 demonstrate the effect of dynamic range of $w_b(n)$ analysis window function on MMSE-Mismatched-N and MMSE-Mismatched-O stimuli. Spectrograms of clean and noisy speech (AWGN at 10 dB SNR) are shown in Fig. 24(a) and (b), respectively. Spectrograms of MMSE-Mismatched-N stimuli for $w_b(n)$ set to Chebyshev 5, 25 and 45 dB are shown in Fig. 24(c–e), respectively. For the 5 dB setting, the spectrogram of MMSE-Mismatched-N stimulus (Fig. 24(c)) shows well suppressed noise, however, the speech signal components are badly distorted – with much of the fine spectral detail significantly deteriorated with respect to that observable in the spectrogram of noisy speech (Fig. 24(b)). For the 25 dB setting, the spectrogram of MMSE-Mismatched-N stimulus (Fig. 24(d)) shows some noise reduction – slightly more so during speech presence than absence – with important spectral details considerably less affected than for the 5 dB setting. For the 45 dB setting, the spectrogram of MMSE-Mismatched-N stimulus (Fig. 24(e)) is essentially no different from the spectrogram of the MMSE stimulus shown in Fig. 23(c). This is because the analysis windows are approximately matched (i.e., $w_a(n) \approx w_b(n)$) and, hence, the MMSE-Mismatched-N stimulus is essentially noisy speech enhanced using the MMSE STSA method (Ephraim and Malah, 1984).

Spectrograms of MMSE-Mismatched-O stimuli for $w_b(n)$ set to Chebyshev 10, 30 and 50 dB are shown in Fig. 24(f–h), respectively. For the 10 dB setting, the spectrogram of MMSE-Mismatched-O stimulus (Fig. 24(f)) shows well suppressed noise. While some residual noise and signal distortion are still present, these are notably lower than those observed for MMSE-Mismatched-N stimulus of Fig. 24(c). For the 30 dB setting, the spectrogram of MMSE-Mismatched-O stimulus (Fig. 24(g)) shows much reduced noise throughout the spectrum, with particularly good suppression observable during speech presence. Formant and pitch frequency harmonic details not present in the noisy speech (Fig. 24(b)) are also visible due to the use of the oracle phase spectrum during MMSE-Mismatched-O stimulus construction. For the 50 dB setting, the spectrogram of MMSE-Mismatched-O stimulus (Fig. 24(h)) shows a noise reduction comparable to that of MMSE stimulus shown in Fig. 23(c). This is because the analysis windows are now approximately matched (i.e., $w_a(n) \approx w_b(n)$) and, hence, the MMSE-Mismatched-O stimulus essentially is the noisy speech enhanced using the MMSE STSA method (Ephraim and Malah, 1984)

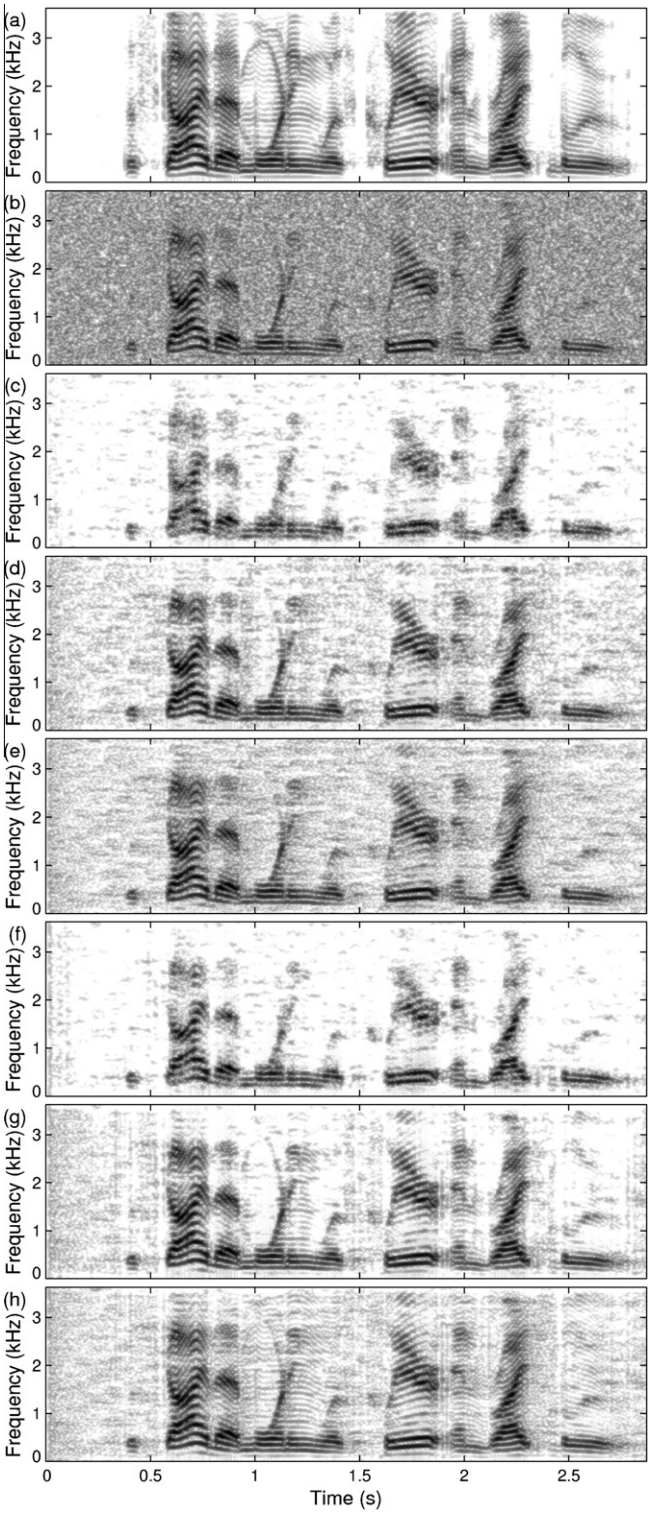

Fig. 24. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (AWGN at 10 dB SNR, PESQ: 2.09); (c) MMSE-Mismatched-N ($w_b(n)$ set to Chebyshev 5 dB, PESQ: 2.41); (d) MMSE-Mismatched-N ($w_b(n)$ set to Chebyshev 25 dB, PESQ: 2.58); (e) MMSE-Mismatched-N ($w_b(n)$ set to Chebyshev 45 dB, PESQ: 2.61); (f) MMSE-Mismatched-O ($w_b(n)$ set to Chebyshev 10 dB, PESQ: 2.67); (g) MMSE-Mismatched-O ($w_b(n)$ set to Chebyshev 30 dB, PESQ: 3.12); and (h) MMSE-Mismatched-O ($w_b(n)$ set to Chebyshev 50 dB, PESQ: 2.82).

with, however, the clean phase spectrum employed during AMS processing instead of the noisy one.

From Fig. 24(g–h) it can be seen that the oracle knowledge of the short-time phase spectrum for mismatched window approach can significantly contribute towards important spectral details, with respect to those visible in noisy speech of Fig. 24(b). On the other hand, the spectrograms of Fig. 24(c and d) show that the benefits of mismatched window processing, when only noisy spectra is available (i.e., for non-oracle phase spectrum) are modest at best – suggesting the need for better phase spectrum estimation algorithms.

5.6.2. Effect of λ parameter on PSC and MMSE-PSC stimuli

In the PSC method (Stark et al., 2008) reviewed in Section 4.2, λ is a tunable parameter that governs the overall strength of phase spectrum compensation. Thus, larger λ values lead to stronger noise suppression. Note that in (Stark et al., 2008) $\lambda = 3.74$ was suggested. Spectrograms of Fig. 25 demonstrate the effect of λ parameter on both, PSC and MMSE-PSC stimuli. Spectrograms of clean speech and noisy speech (for AWGN at 10 dB SNR) are shown in Fig. 25(a) and (b), respectively.

Spectrograms of PSC stimuli for λ set to 1, 3.74 and 7, are shown in Fig. 25(c–e), respectively. For the $\lambda = 1$ setting, the spectrogram of PSC stimulus (Fig. 25(c)) shows almost no noise reduction and is very similar to the spectrogram of noisy speech (Fig. 25(b)). This is because, very little compensation takes pace for the $\lambda = 1$ setting. For the $\lambda = 3.74$ setting, the intensity of the noise is considerably reduced without an adverse effect on speech signal components, but with some residual noise still remaining. This can be seen by comparing spectrogram of PSC stimulus in Fig. 25(d) with the spectrogram of noisy speech in Fig. 25(b). Further reduction in noise is possible by increasing the strength of compensation even more, at an expense of very minor loss of spectral detail. This is shown in the spectrogram of PSC stimulus in Fig. 25(e) for λ set to 7.[19]

The spectrograms of MMSE-PSC stimuli for λ set to 1, 3.74 and 7 are shown in Fig. 25(f–h), respectively. For all three λ settings considered here, significant noise reduction was achieved with respect to noisy speech (Fig. 25(b)). For the latter two λ settings, the noise reduction was also considerable with respect to speech enhanced using the MMSE STSA method (Ephraim and Malah, 1984) (Fig. 23(c)). For the $\lambda = 1$ setting, the spectrogram of MMSE-PSC stimulus (Fig. 25(f)) shows some noise reduction, however, much of the noise still remains. For the $\lambda = 3.74$ setting (Fig. 25(g)) the noise is mostly removed with little loss of spectral detail. While for $\lambda = 7$ setting (Fig. 25(h)) the noise is all but gone, at an expense of some suppression of signal components.

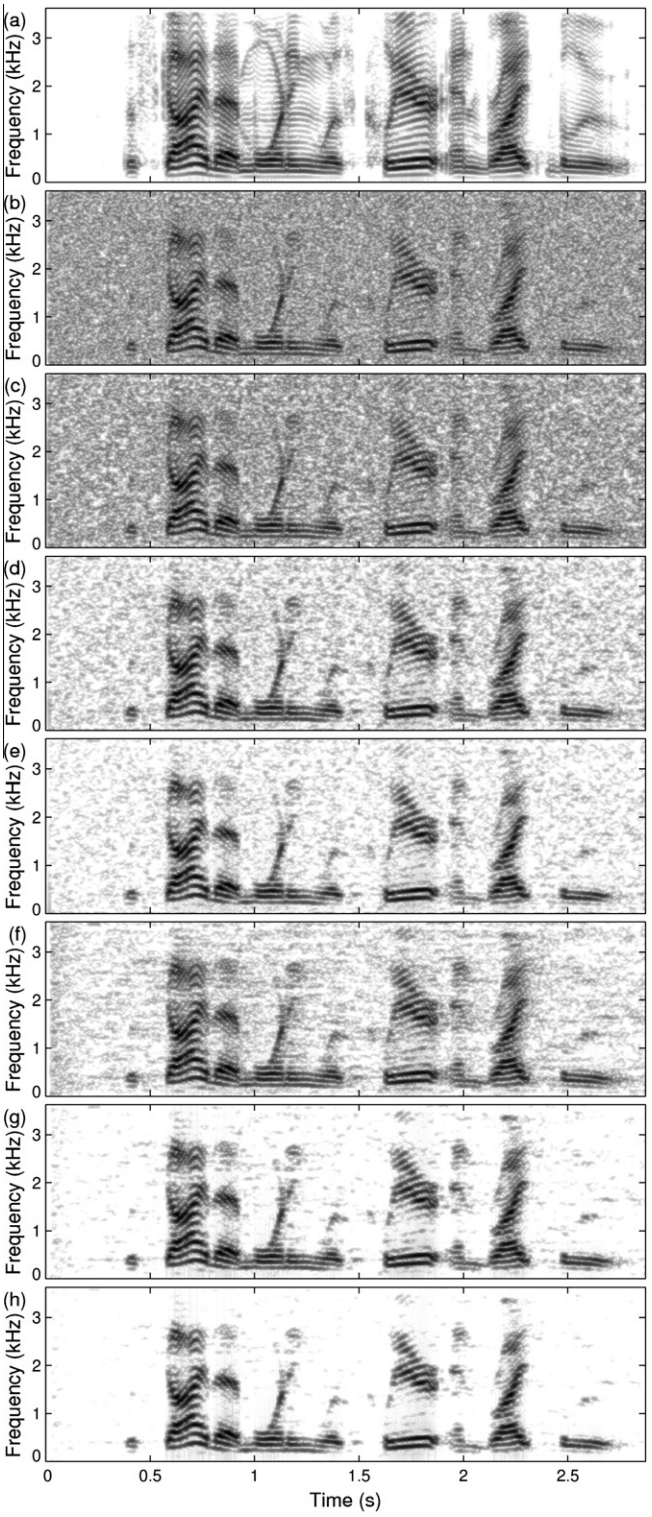

Fig. 25. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (AWGN at 10 dB SNR, PESQ: 2.09); (c) PSC ($\lambda = 1.00$, PESQ: 2.20); (d) PSC ($\lambda = 3.74$, PESQ: 2.56); (e) PSC ($\lambda = 7.00$, PESQ: 2.49); (f) MMSE-PSC ($\lambda = 1.00$, PESQ: 2.67); (g) MMSE-PSC ($\lambda = 3.74$, PESQ: 2.65); and (h) MMSE-PSC ($\lambda = 7.00$, PESQ: 2.55).

[19] Note that while the $\lambda = 7$ setting results in significantly better noise reduction (than the $\lambda = 3.74$ setting), for consistency with previous work, $\lambda = 3.74$ is employed in the enhancement experiments of Sections 4 and 5.

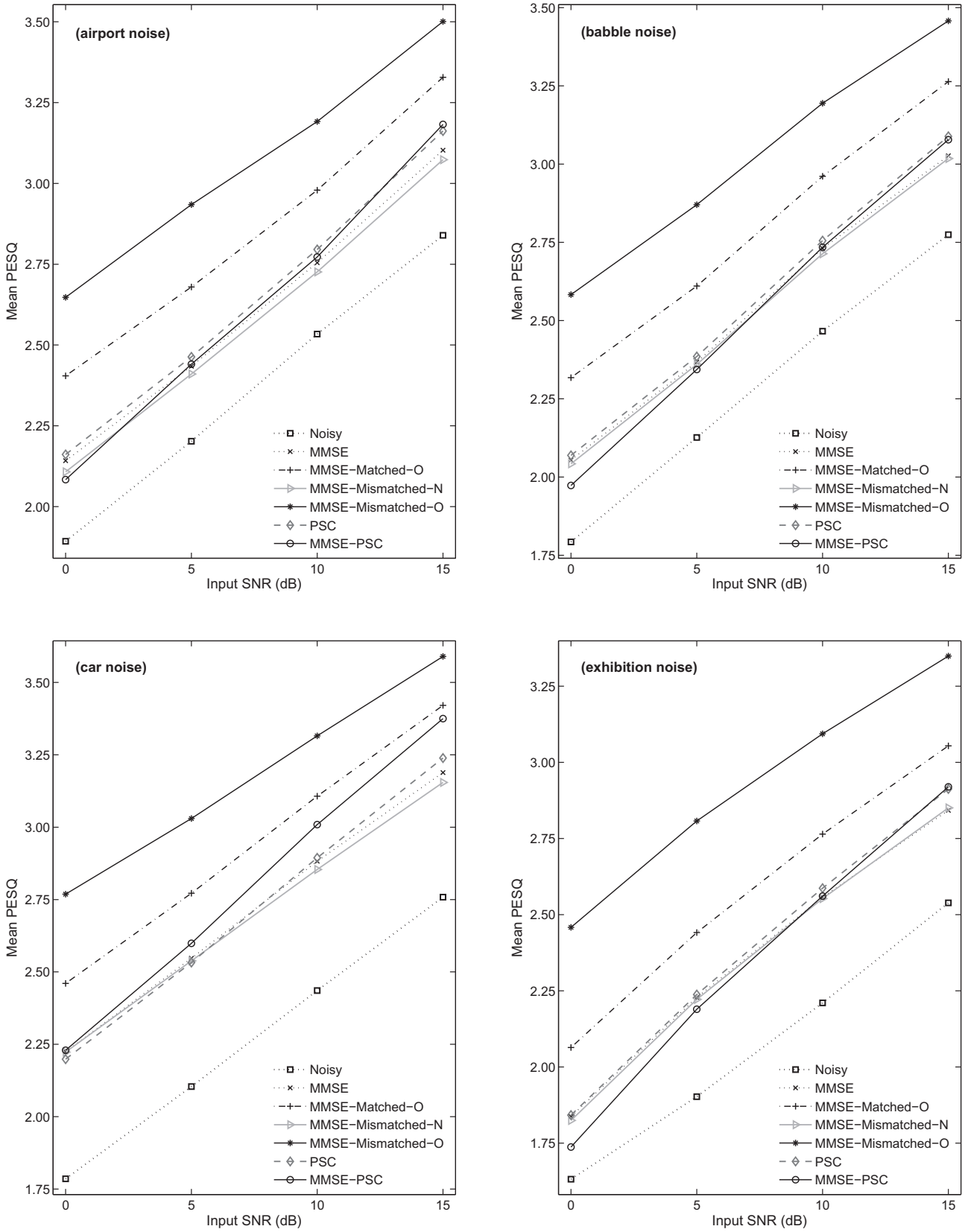

Fig. 26. Coloured noise results for the objective MMSE experiment (case 4) detailed in Section 4.3. The results are in terms of mean PESQ scores as a function of input SNR (dB) for airport, babble, car and exhibition noises over the Noizeus corpus.

5.7. Conclusions

In this section, we have investigated the usefulness of a number of approaches to phase spectrum computation when the magnitude spectrum is estimated using the MMSE STSA method (Ephraim and Malah, 1984). Traditional AMS-based speech enhancement techniques process only the magnitude spectrum, and leave the noisy phase spectrum unchanged. Our goal here was to determine if speech quality could be further improved by also processing the noisy phase spectrum. For this purpose objective and subjective speech enhancement experiments were conducted, in which both oracle and non-oracle phase spectrum knowledge was considered. Matched and mismatched analysis window approaches were included, along with the PSC method and a MMSE STSA–PSC combination. The results of our experiments show that it is possible to further improve quality of enhanced speech by processing both, the short-time phase and magnitude spectra. Notably, very significant improvements were observed for the oracle treatments, and in particular when mismatched analysis windows were employed. Smaller improvements were attained for the non-oracle treatments, with the MMSE fusion with the PSC method working particularly well. Our results suggest that better phase spectrum estimates have a potential for significant improvement of enhanced speech quality, and that further research into phase spectrum processing may well be worthwhile in the context of speech enhancement.

6. Summary

In this study, we have investigated short-time Fourier AMS-based speech enhancement approaches, that attempt to derive some benefit – in terms of speech quality – from the short-time phase spectrum. Throughout our experiments, the frame duration of 32 ms was employed. Both oracle and non-oracle scenarios were considered. In the oracle experiments the undistorted phase spectrum was made available during AMS-based processing of noisy speech. The objective was to determine the maximum speech quality improvements attainable due to availability of undistorted phase spectrum. In the non-oracle experiments only the noisy spectra were available for processing. The objective was to determine improvements attainable in a more realistic scenario. The effect of matched and mismatched analysis windows on speech quality of processed stimuli was systematically studied. The PSC method was also investigated. The following two scenarios were also considered. First, in which the noisy magnitude spectrum was used; and second, in which the MMSE STSA estimate of the clean magnitude spectrum was computed from noisy speech. Both objective and subjective speech quality measurements were employed in our investigations.

The results of our experiments, where the noisy magnitude spectrum was employed, show that significant improvements of speech quality are possible especially when clean phase spectrum is known. For the approaches considered in this study, only modest improvements were

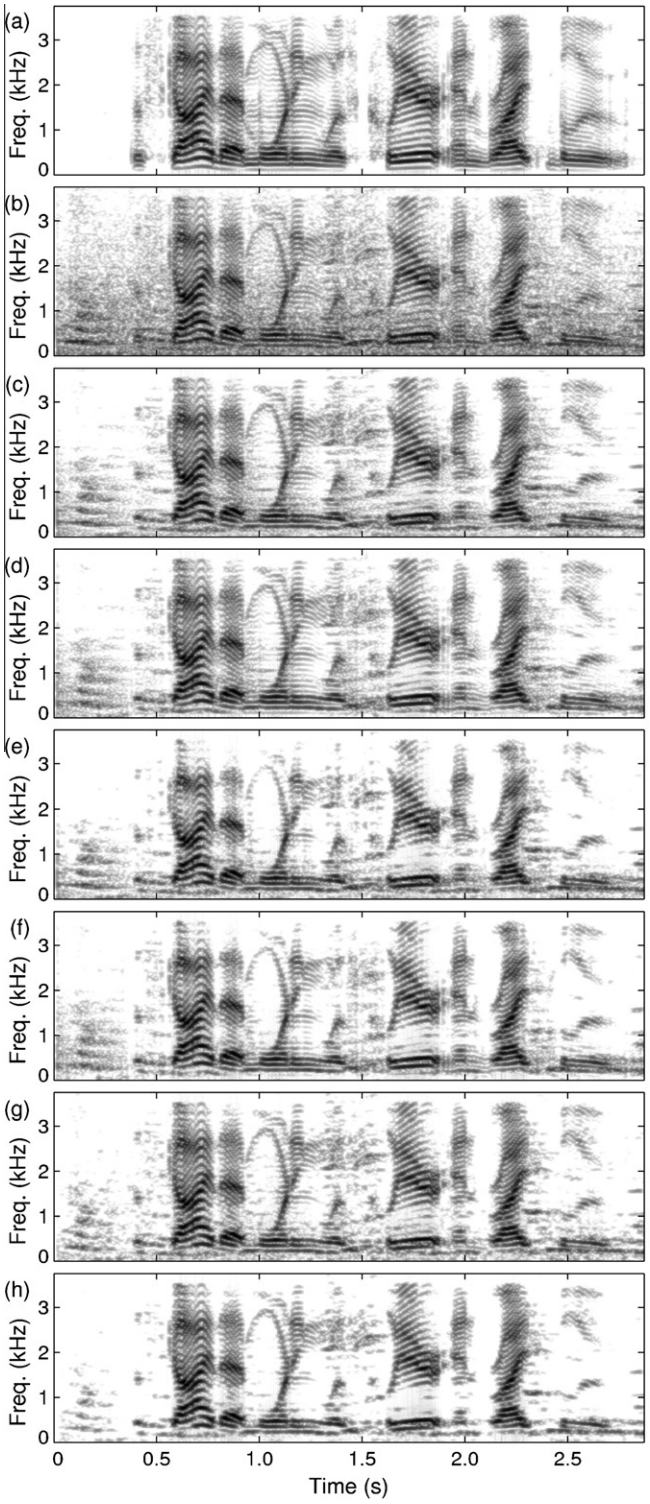

Fig. 27. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (airport noise at 10 dB SNR, PESQ: 2.55); (c) MMSE (PESQ: 2.77); (d) MMSE-Matched-O (PESQ: 2.95); (e) MMSE-Mismatched-N (PESQ: 2.71); (f) MMSE-Mismatched-O (PESQ: 3.12); (g) PSC (PESQ: 2.84); and (h) MMSE-PSC (PESQ: 2.65).

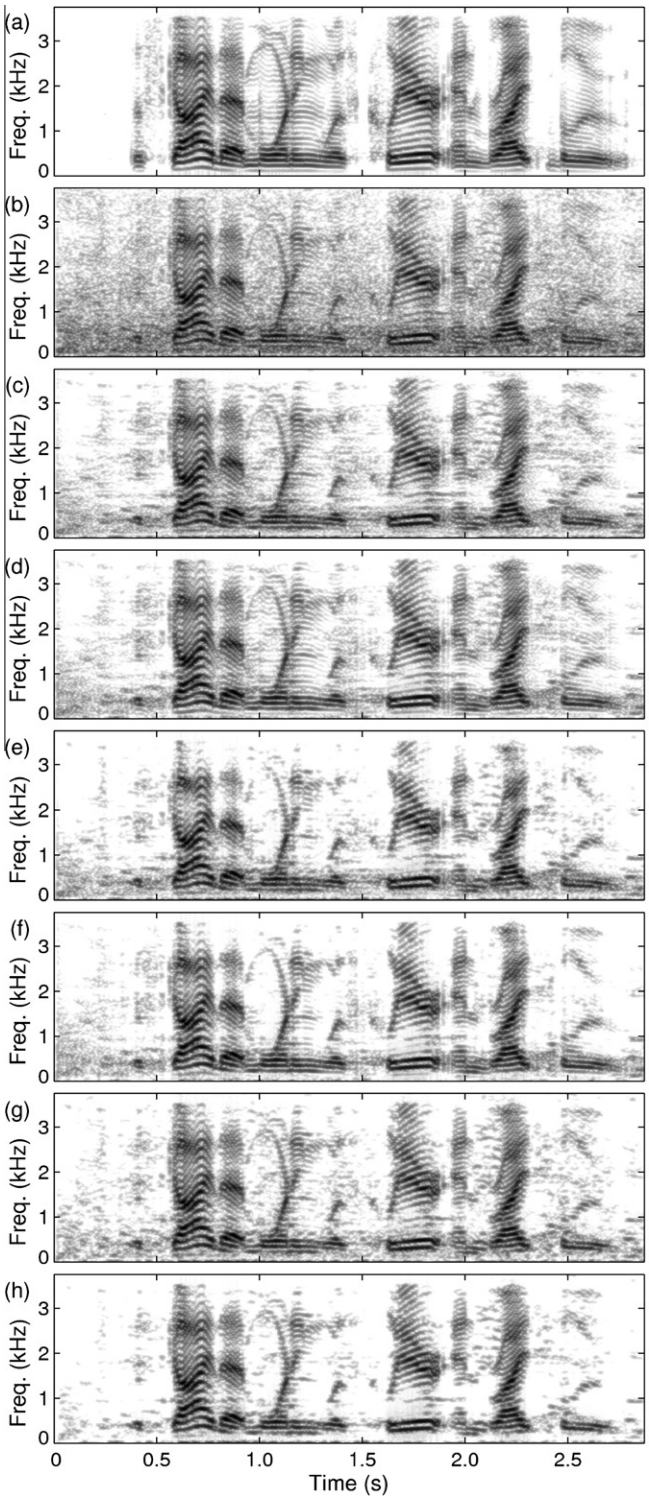

Fig. 28. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (babble noise at 10 dB SNR, PESQ: 2.56); (c) MMSE (PESQ: 2.87); (d) MMSE-Matched-O (PESQ: 3.14); (e) MMSE-Mismatched-N (PESQ: 2.89); (f) MMSE-Mismatched-O (PESQ: 3.37); (g) PSC (PESQ: 2.93); and (h) MMSE-PSC (PESQ: 2.85).

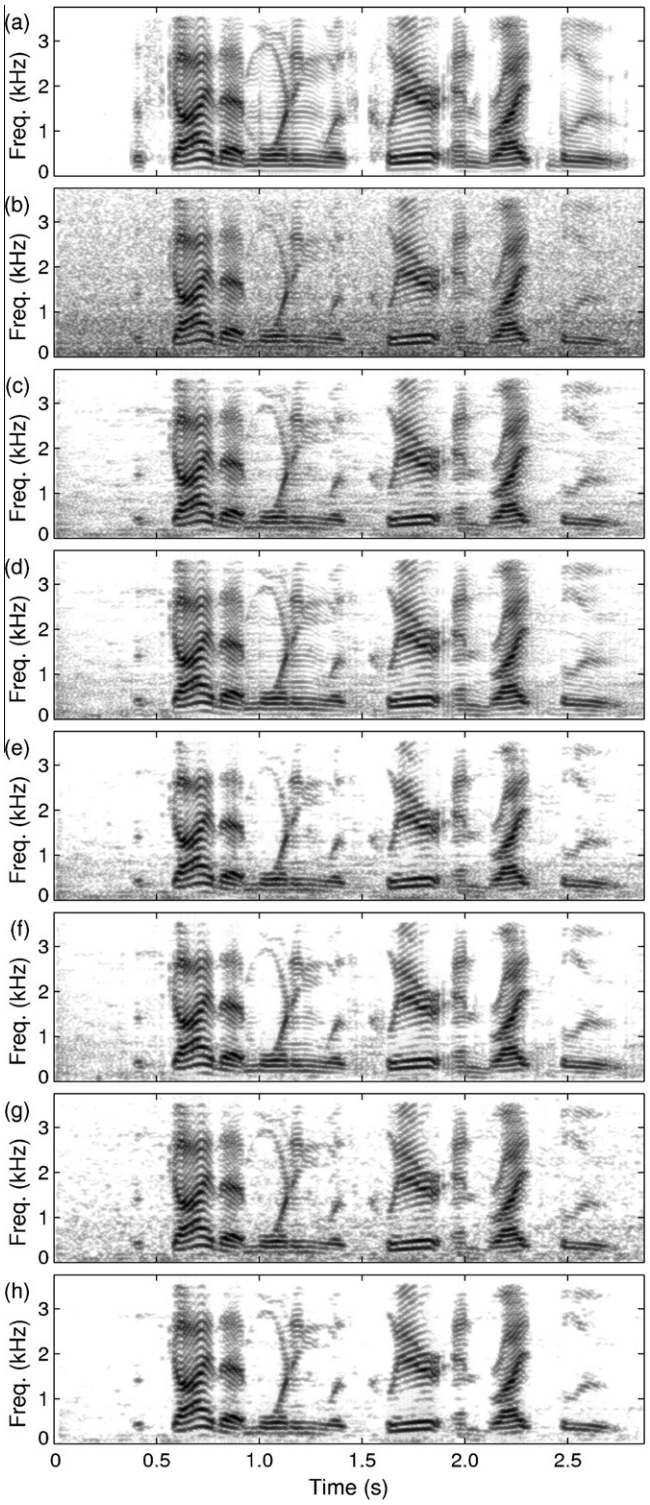

Fig. 29. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (car noise at 10 dB SNR, PESQ: 2.46); (c) MMSE (PESQ: 2.94); (d) MMSE-Matched-O (PESQ: 3.13); (e) MMSE-Mismatched-N (PESQ: 2.93); (f) MMSE-Mismatched-O (PESQ: 3.30); (g) PSC (PESQ: 2.89); and (h) MMSE-PSC (PESQ: 3.03).

attained when only the noisy spectrum was available for processing. This suggests that better phase spectrum estimation algorithms are needed in order to attain full benefit from the short-time phase spectrum. While this is by no

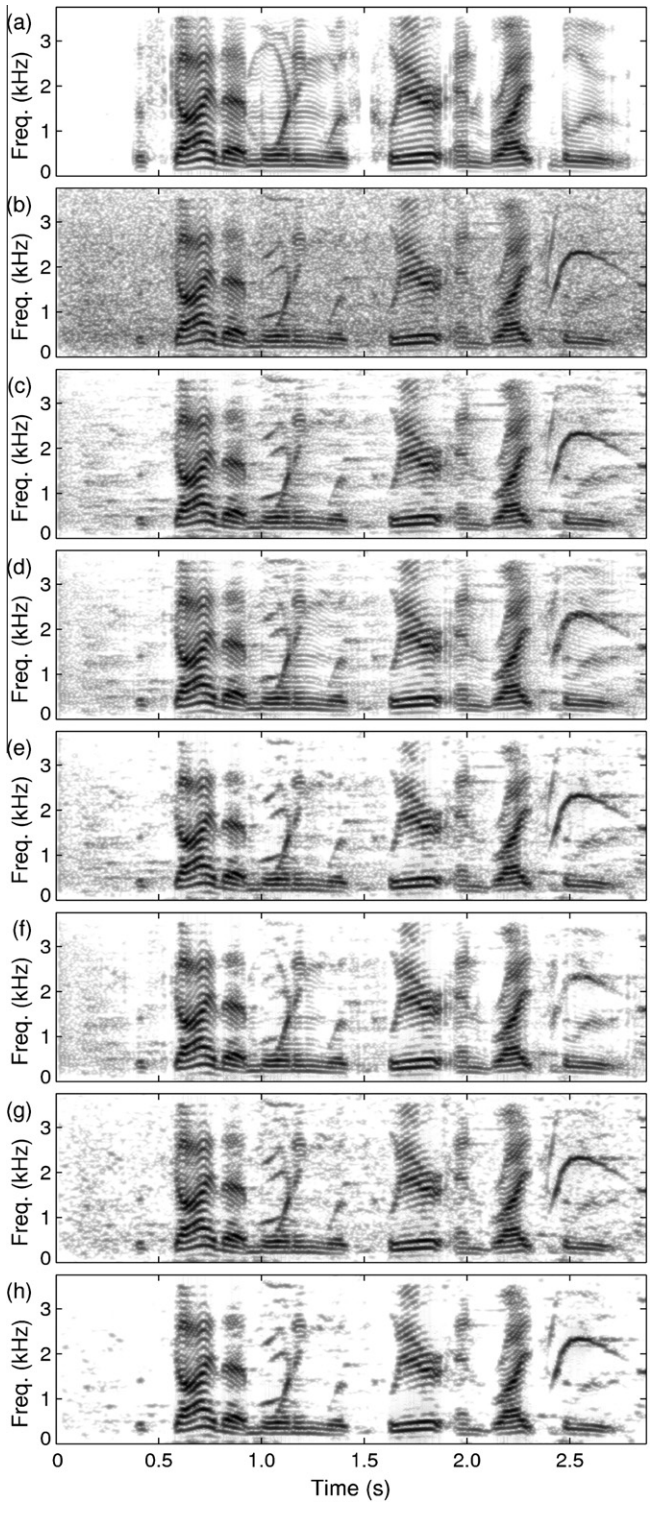

Fig. 30. Spectrograms of the sp10 Noizeus utterance, "The sky that morning was clear and bright blue", by a male speaker. Spectrograms for the following stimuli types are shown: (a) clean speech (PESQ: 4.50); (b) noisy speech (exhibition noise at 10 dB SNR, PESQ: 2.19); (c) MMSE (PESQ: 2.38); (d) MMSE-Matched-O (PESQ: 2.53); (e) MMSE-Mismatched-N (PESQ: 2.32); (f) MMSE-Mismatched-O (PESQ: 3.00); (g) PSC (PESQ: 2.39); and (h) MMSE-PSC (PESQ: 2.31).

means an easy task, we hope that our results show the potential for further improvements and that our work will encourage further research in this direction. The results of our experiments, where the MMSE STSA estimate of the clean magnitude spectrum was computed from noisy speech and where mismatched analysis windows were employed, show that significant speech quality improvements can be attained when the clean phase spectrum is known (i.e., when accurate phase spectrum estimates are available), with only modest improvements achieved for the non-oracle phase spectrum. This again highlights the need for research into better phase spectrum estimation algorithms. Importantly, our results do show that the potential for further improvements due to phase spectrum definitely exists. For the non-oracle methods, the combination of the MMSE STSA method with the PSC method, was shown to work particularly well and achieved high subjective preference scores.

Appendix A. Coloured noise results for the MMSE experiments (case 4)

This appendix, includes some additional results for various coloured noises, including airport, babble, car and exhibition, for the case 4 experiments. Mean PESQ scores are shown in Fig. 26. The MMSE-Mismatched-O stimuli achieved consistently highest objective scores, while the noisy stimuli achieved lowest scores. The MMSE-Matched-O was consistently second best method, with the remaining methods (MMSE, MMSE-Mismatched-N, PSC, MMSE-PSC) – in general – achieving relatively comparable objective sores. Example spectrograms for the various noise types are shown in Figs. 27–30.

References

Alsteris, L., Paliwal, K., 2004. Importance of window shape for phase-only reconstruction of speech. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process, (ICASSP), vol. 1. Montreal, QC, Canada, pp. 573–576.

Alsteris, L., Paliwal, K., 2005. Evaluation of the modified group delay feature for isolated word recognition. In: Proc. Int. Sympos. Signal Process. and Applications (ISSPA). Sydney, NSW, Australia.

Alsteris, L., Paliwal, K., 2006. Further intelligibility results from human listening tests using the short-time phase spectrum. Speech Commun. 48 (6), 727–736.

Alsteris, L., Paliwal, K., 2007. Iterative reconstruction of speech from short-time Fourier transform phase and magnitude spectra. Comput. Speech Lang. 21 (1), 174–186.

Berouti, M., Schwartz, R., Makhoul, J., 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.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Process. ASSP-28 (4), 357–366.

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.

Ephraim, Y., Trees, H.V., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3, 251–266.

Ghitza, O., 2001. On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception. J. Acoust. Soc. Am. 110 (3), 1628–1640.

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

Harris, F., 1978. On the use of windows for harmonic analysis with the discrete Fourier transform. Proc. IEEE 66 (1), 51–83.

Hayes, M., Lim, J., Oppenheim, A., 1980. Signal reconstruction from phase or magnitude. IEEE Trans. Acoust. Speech Signal Process. 28 (6), 672–680.

Hayes, M.H., 1996. Statistical Digital Signal Processing and Modeling. John Wiley and Sons, Inc., New York, NY, USA.

Hermansky, H., 1990. Perceptual linear predictive (PLP) analysis of speech. J. Acoust. Soc. Am. 87 (4), 1738–1752.

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

ITU-T, 2001. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs (ITU-T Recommendation P.862). Series P: Telephone Transmission Quality, Telephone Installations, Local Line Networks. Telecommunication Standardization Sector of International Telecommunication Union. Geneva, Switzerland.

Kim, D., 2003. Perceptual phase quantization of speech. IEEE Trans. Speech Audio Process. 11 (4), 355–364.

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

Liu, L., He, J., Palm, G., 1997. Effects of phase on the perception of intervocalic stop consonants. Speech Commun. 22 (4), 403–417.

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

Loveimi, E., Ahadi, S., 2010. Objective evaluation of magnitude and phase only spectrum-based reconstruction of the speech signal. In: Proc. Int. Sympos. Commun. Control Signal Process (ISCCSP). Limassol, Cyprus, pp. 1–4.

Lu, Y., Loizou, P., 2008. A geometric approach to spectral subtraction. Speech Commun. 50 (6), 453–466.

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

McAulay, R., Quatieri, T.F., 1995. Sinusoidal coding. In: Kleijn, W., Paliwal, K. (Eds.), Principles of Speech Coding. Elsevier Science, New York, NY, USA, pp. 121–173, Chapter 4.

Nakagawa, S., Asakawa, K., Wang, L., 2007. Speaker recognition by combining MFCC and phase information. In: Proc. ISCA Conf. Int. Speech Commun. Assoc. (INTERSPEECH). Antwerp, Belgium, pp. 2005–2008.

Nawab, S., Quatieri, T., Lim, J., 1983. Signal reconstruction from short-time Fourier transform magnitude. IEEE Trans. Acoust. Speech Signal Process. 31 (4), 986–998.

Oppenheim, A.V., Lim, J.S., 1981. The importance of phase in signals. Proc. IEEE 69 (5), 529–541.

Oppenheim, A.V., Lim, J.S., Kopec, G., Pohlig, S.C., 1979. Phase in speech and pictures. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process (ICASSP), vol. 4. Washington, DC, USA, pp. 632–637.

Paliwal, K., Feb 2003. Usefulness of phase in speech processing. In: Proc. IPSJ Spoken Language Processing Workshop. Gifu, Japan, pp. 1–6.

Paliwal, K., Alsteris, L., 2003. Usefulness of phase spectrum in human speech perception. In: Proc. ISCA European Conf. Speech Commun. and Technology (EUROSPEECH). Geneva, Switzerland, pp. 2117–2120.

Paliwal, K., Alsteris, L., 2005. On the usefulness of STFT phase spectrum in human listening tests. Speech Commun. 45 (2), 153–170.

Paliwal, K., Basu, A., 1987. A speech enhancement method based on Kalman filtering. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process (ICASSP), vol. 12. pp. 177–180.

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

Pobloth, H., Kleijn, W., 1999. On phase perception in speech. In: Proc. IEEE Int. Conf. Acoustics, Speech, Signal Process (ICASSP). Phoenix, Arizona, USA, pp. 29–32.

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

Reddy, N., Swamy, M., 1985. Derivative of phase spectrum of truncated autoregressive signals. IEEE Trans. Circuits Syst. 32 (6), 616–618.

Schlüter, R., Ney, H., 2001. Using phase spectrum information for improved speech recognition performance. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process (ICASSP), vol. 1. IEEE Comput. Soc., Los Alamitos, CA, USA, pp. 133–136.

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

Shi, G., Shanechi, M., Aarabi, P., 2006. On the importance of phase in human speech recognition. IEEE Trans. Audio Speech Lang. Process. 14 (5), 1867–1874.

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.

Skoglund, J., Bastiaan Kleijn, W., Hedelin, P., 1997. Audibility of pitch-synchronously modulated noise. In: Proc. IEEE Workshop on Speech Coding for Telecommunications Proceeding. Pocono Manor, PA , USA, pp. 51–52.

Stark, A., Wójcicki, K., Lyons, J., Paliwal, K., 2008. Noise-driven short-time phase spectrum compensation procedure for speech enhancement. In: Proc. ISCA Conf. Int. Speech Commun. Assoc. (INTERSPEECH). Brisbane, QLD, Australia, pp. 549–552.

Vary, P., 1985. Noise suppression by spectral magnitude estimation – mechanism and theoretical limits. Signal Process. 8 (4), 387–400.

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.

Wackerly, D., Mendenhall, W., Scheaffer, R.L., 2007. Mathematical Statistics with Applications. Duxbury Press, Pacific Grove, CA, USA.

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

Wang, L., Ohtsuka, S., Nakagawa, S., 2009. High improvement of speaker identification and verification by combining MFCC and phase information. In: Proc. IEEE Int. Conf. Acoustics Speech and Signal Process. (ICASSP). Taipei, Taiwan, pp. 4529–4532.

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

Wójcicki, K., Milacic, M., Stark, A., Lyons, J., Paliwal, K., 2008. Exploiting conjugate symmetry of the short-time Fourier spectrum for speech enhancement. IEEE Signal Process. Lett. 15, 461–464.

Wójcicki, K., Paliwal, K., 2007. Importance of the dynamic range of an analysis window function for phase-only and magnitude-only reconstruction of speech. In: Proc. IEEE Int. Conf. Acoustics Speech and Signal Process (ICASSP), vol. IV. Honolulu, HI, USA, pp. 729–733.

Wójcicki, K., Paliwal, K., 2008. On the relative importance of the short-time magnitude and phase spectra towards speaker dependent information. In: Proc. ISCA Tutorial and Research Workshop (ITRW). Aalborg, Denmark.

Yegnanarayana, B., Fathima, S., Murthy, H., 1987. Reconstruction from Fourier transform phase with applications to speech analysis. In: Proc. IEEE Int. Conf. Acoustics Speech and Signal Process. (ICASSP), vol. 12. Dallas, Texas, USA, pp. 301–304.