Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 53 (2011) 327–339
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Role of modulation magnitude and phase spectrum towards speech intelligibility

Kuldip Paliwal, Belinda Schwerin *, Kamil Wójcicki

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

Received 11 June 2010; received in revised form 4 October 2010; accepted 11 October 2010
Available online 25 October 2010

Abstract

In this paper our aim is to investigate the properties of the modulation domain and more specifically, to evaluate the relative contributions of the modulation magnitude and phase spectra towards speech intelligibility. For this purpose, we extend the traditional (acoustic domain) analysis–modification–synthesis framework to include modulation domain processing. We use this framework to construct stimuli that retain only selected spectral components, for the purpose of objective and subjective intelligibility tests. We conduct three experiments. In the first, we investigate the relative contributions to intelligibility of the modulation magnitude, modulation phase, and acoustic phase spectra. In the second experiment, the effect of modulation frame duration on intelligibility for processing of the modulation magnitude spectrum is investigated. In the third experiment, the effect of modulation frame duration on intelligibility for processing of the modulation phase spectrum is investigated. Results of these experiments show that both the modulation magnitude and phase spectra are important for speech intelligibility, and that significant improvement is gained by the inclusion of acoustic phase information. They also show that smaller modulation frame durations improve intelligibility when processing the modulation magnitude spectrum, while longer frame durations improve intelligibility when processing the modulation phase spectrum.

Keywords: Analysis frame duration; Modulation frame duration; Modulation domain; Modulation magnitude spectrum; Modulation phase spectrum; Speech intelligibility; Speech transmission index (STI); Analysis–modification–synthesis (AMS)

1. Introduction

While speech is non-stationary, it can be assumed quasi-stationary, and therefore can be processed through short-time Fourier analysis. The short-time Fourier transform (STFT) of the speech signal is referred to as the acoustic spectrum, and can be expressed in terms of the short-time acoustic magnitude spectrum and the short-time acoustic phase spectrum. Thus, the signal is completely characterised by its acoustic magnitude and acoustic phase spectra.

The modulation domain has become popular as an alternative to the acoustic domain for the processing of speech signals. For a given acoustic frequency, the modulation spectrum is the STFT of the time series of the acoustic spectral magnitudes at that frequency, and can be expressed in terms of its short-time modulation magnitude spectrum and its short-time modulation phase spectrum. Therefore, a speech signal is also completely characterised by its modulation magnitude, modulation phase, and acoustic phase spectra.

Many applications of modulation domain speech processing have appeared in the literature. For example, Atlas et al. (Atlas and Vinton, 2001; Thompson and Atlas, 2003) proposed audio codecs which use the two-dimensional modulation transform to concentrate information in a small number of coefficients for better quality speech coding. Tyagi et al. (2003) applied mel-cepstrum modulation features to automatic speech recognition (ASR), to give improved performance in the presence of non-stationary noise. Kingsbury et al. (1998) applied a modulation spectrogram representation that emphasised low-frequency

* Corresponding author. Tel.: +61 0737353754.
E-mail address: belsch71@gmail.com (B. Schwerin).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.10.004

amplitude modulations to ASR for improved robustness in noisy and reverberant conditions. Kim (2004, 2005) as well as Falk and Chan (2008) used the short-time modulation magnitude spectrum to derive objective measures that characterise the quality of processed speech. The modulation magnitude spectrum has also been used for speaker recognition (Falk and Chan, 2010), and emotion recognition (Wu et al., 2009). Bandpass filtering has been applied to the time trajectories of the short-time acoustic magnitude spectrum (Falk et al., 2007; Lyons and Paliwal, 2008). Many of these studies modify or utilise only the short-time modulation magnitude spectrum while leaving the modulation phase spectrum unchanged. However, the phase spectrum is recognised to play a more important role in the modulation domain than in the acoustic domain (Greenberg, et al., 1998; Kanedera et al., 1998; Atlas et al., 2004). While the contributions of the short-time magnitude and phase spectra are very well documented in the literature for the acoustic domain (e.g., Schroeder, 1975; Oppenheim and Lim, 1981; Liu et al., 1997; Paliwal and Alsteris, 2005), this is not the case for the modulation domain. Therefore in this work, we are interested in quantifying the contribution of both modulation magnitude and phase spectra to speech intelligibility.

Typical modulation domain-based applications use modulation frame durations of around 250 ms (e.g., Greenberg and Kingsbury, 1997; Thompson and Atlas, 2003; Kim, 2005; Falk and Chan, 2008; Wu et al., 2009; Falk et al., 2010; Falk and Chan, 2010; Paliwal et al., 2010b). This is much larger than the durations typically used for acoustic-domain processing. The frames are made longer to effectively represent the time variability of speech signal spectra (Thompson and Atlas, 2003). This is justifiable since many audio signals are effectively stationary over relatively long durations. However, longer frame durations can result in the introduction of temporal smearing due to the lack of localisation of more transient signals (Thompson and Atlas, 2003; Paliwal et al., 2010b). Therefore, we are also interested in evaluating the effect of modulation frame duration on intelligibility.

In this paper, our primary aim is to evaluate the relative contributions of both the modulation magnitude and phase spectra to intelligibility. Secondly, we aim to evaluate the effect of the modulation frame duration for both modulation magnitude and phase spectra on the resulting speech intelligibility.[1] To achieve these goals, a dual analysis–modification–synthesis (AMS) framework such as proposed in (Paliwal et al., 2010b) is used. Under this framework, the short-time modulation magnitude spectrum can be investigated by discarding the modulation phase information by randomising its values. Similarly, the short-time modulation phase spectrum can be investigated by discarding the modulation magnitude information by setting its values to 1. Then by varying the modulation frame duration under this framework, we can find the frame durations which give the best speech intelligibility according to both subjective and objective testing.

The rest of this paper is organised as follows. Section 2 details the acoustic and modulation AMS-based speech processing. Section 3 describes experiments and results evaluating the contribution of the modulation magnitude and phase to intelligibility. Sections 4 and 5 describes experiments and results evaluating the effect of modulation frame duration on intelligibility for modulation magnitude and phase, respectively. Finally, conclusions are given in Section 6.

2. Analysis–modification–synthesis

One of the aims of this study is to quantify the contribution of both the modulation magnitude and phase spectra to speech intelligibility. Previous papers investigating the relative significance of the acoustic magnitude and phase spectra have made use of the short-time Fourier analysis–modification–synthesis (AMS) framework (e.g., Oppenheim and Lim, 1981; Liu et al., 1997; Paliwal and Alsteris, 2005), where AMS analysis decomposes the speech signal into the acoustic magnitude and acoustic phase spectral components. Under this framework, speech stimuli were synthesised such that only one of these spectral components (i.e., the acoustic magnitude spectrum or the acoustic phase spectrum) is retained. Intelligibility experiments were then used in the above studies to evaluate the contribution of each of these spectral components to the intelligibility of speech.

In the present work, our goal is to evaluate the contributions of magnitude and phase spectra (towards speech intelligibility) in the modulation domain. To achieve this, the acoustic AMS procedure is extended to the modulation domain, resulting in a dual AMS framework (Paliwal et al., 2010b) to which we will refer to as the modulation AMS procedure. Analysis in the above framework decomposes the speech signal into the modulation magnitude, modulation phase, and acoustic phase spectra. The relative contributions of each of these three spectral components towards speech intelligibility are then evaluated through intelligibility experiments presented in Sections 3–5. The remainder of this section describes both the acoustic and modulation AMS procedures used for the construction of stimuli, and then defines the types of stimuli constructed for experimentation using the different spectral component combinations.

2.1. Acoustic AMS procedure

Traditional acoustic-domain short-time Fourier AMS framework consists of three stages: (1) the analysis stage, where the input speech is processed using STFT analysis; (2) the modification stage, where the noisy spectrum undergoes some kind of modification; and (3) the synthesis stage,

[1] For completeness, objective speech quality results are also included in Appendix A.

where the inverse STFT is followed by overlap-add synthesis (OLA) to reconstruct the output signal.

For a discrete-time signal $x(n)$, the STFT is given by

$$X(n,k) = \sum_{l=-\infty}^{\infty} x(l)w(n-l)\mathrm{e}^{-j2\pi kl/N}, \quad (1)$$

where n refers to the discrete-time index, k is the index of the discrete acoustic frequency, N is the acoustic frame duration (in samples), and $w(n)$ is the acoustic analysis window function.[2] In speech processing, an acoustic frame duration of 20–40 ms is typically used (e.g., Picone, 1993; Huang et al., 2001; Loizou, 2007), with a Hamming window (of the same duration) as the analysis window function.

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

$$X(n,k) = |X(n,k)|\mathrm{e}^{j\angle X(n,k)}, \quad (2)$$

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

In the modification stage of the AMS framework, either the acoustic magnitude or the acoustic phase spectrum or both can be modified. Let $|Y(n,k)|$ denote the modified acoustic magnitude spectrum, and $\angle Y(n,k)$ denote the modified acoustic phase spectrum. Then, the modified STFT is given by

$$Y(n,k) = |Y(n,k)|\mathrm{e}^{j\angle Y(n,k)}. \quad (3)$$

Finally, the synthesis stage reconstructs the speech by applying the inverse STFT to the modified acoustic spectrum, followed by least-squares overlap-add synthesis (Quatieri, 2002). Here, the modified Hanning window (Griffin and Lim, 1984) given by

$$w_s(n) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi(n+0.5)}{N}\right), & 0 \leqslant n < N, \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

is used as the synthesis window function.

A block diagram of the acoustic AMS procedure is shown in Fig. 1.

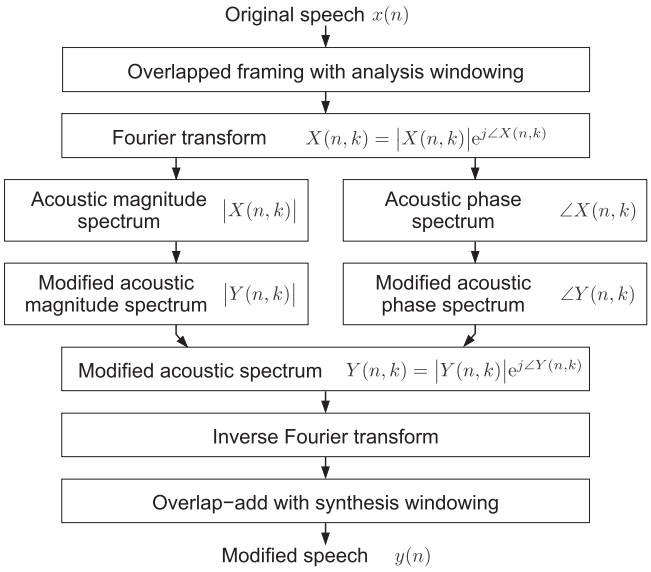

Fig. 1. Block diagram of the acoustic AMS procedure.

2.2. Modulation AMS procedure

The acoustic AMS procedure can also be extended to the modulation domain. Here, each frequency component of the acoustic magnitude spectra obtained using the AMS procedure given in Section 2.1, is processed frame-wise across time using a second AMS framework.

As mentioned earlier, we define the modulation spectrum for a given acoustic frequency as the STFT of the time series of the acoustic spectral magnitudes at that frequency. Hence, the modulation spectrum is calculated, for acoustic frequency index k, by taking its STFT as follows:

$$\mathcal{X}(\eta,k,m) = \sum_{l=-\infty}^{\infty} |X(l,k)|v(\eta-l)\mathrm{e}^{-j2\pi ml/M}, \quad (5)$$

where η is the acoustic frame number,[4] k refers to the index of the discrete-acoustic frequency, m refers to the index of the discrete modulation frequency, M is the modulation frame duration (in terms of acoustic frames), and $v(\eta)$ is the modulation analysis window function.

In polar form, the modulation spectra can be written as

$$\mathcal{X}(\eta,k,m) = |\mathcal{X}(\eta,k,m)|\mathrm{e}^{j\angle\mathcal{X}(\eta,k,m)}, \quad (6)$$

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

In the modification stage, the modulation magnitude spectrum and/or the modulation phase spectrum may be modified. Let $|\mathcal{Z}(\eta,k,m)|$ denote the modified modulation magnitude spectrum, and $\angle\mathcal{Z}(\eta,k,m)$ denote the modified modulation phase spectrum. The modified modulation spectrum is then given by

$$\mathcal{Z}(\eta,k,m) = |\mathcal{Z}(\eta,k,m)|\mathrm{e}^{j\angle\mathcal{Z}(\eta,k,m)}. \quad (7)$$

The modified acoustic magnitude spectrum $|Y(n,k)|$ can then be obtained by applying the inverse STFT to $\mathcal{Z}(\eta,k,m)$, followed by least-squares overlap-add with syn-

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

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

[4] Note that in principle, Eq. (5) could be computed for every acoustic frame, however, in practice we compute it for every modulation frame. We do not show this decimation explicitly in order to keep the mathematical notation concise.

thesis windowing (using the same window function as given in Eq. (4)). The modified acoustic spectrum $Y(n,k)$ can then be found by combining $|Y(n,k)|$ and $\angle Y(n,k)$ as given by Eq. (3).

Finally, the enhanced speech is reconstructed by taking the inverse STFT of the modified acoustic spectrum $Y(n,k)$, followed by least-squares overlap-add synthesis.

A block diagram of the modulation AMS procedure is shown in Fig. 2.

2.3. Types of acoustic and modulation domain spectral modifications considered in this study

The modulation AMS procedure described in Section 2.2 uses information contained in the modulation magnitude, modulation phase, acoustic magnitude and acoustic phase spectra to reconstruct stimuli. In the experiments of this work, we want to examine the contribution of each of these spectral components, and in particular of the modulation magnitude and phase spectra to speech intelligibility. Therefore, we construct stimuli that contain only the spectral components of interest, and remove all other spectral components. To remove acoustic or modulation magnitude spectrum information, the values of the magnitude spectrum are made unity in the corresponding modified STFT. This modified STFT is then used in the synthesis stage according to the procedure described in Section 2.2. The reconstructed signal contains no information about the short-time (acoustic or modulation) magnitude spectrum. Similarly, magnitude-only stimuli can be generated by retaining each frame's magnitude spectrum, and randomising each frame's phase spectrum values. The modified STFT then contains the magnitude spectrum and phase spectrum where phase is a random variable uniformly distributed between 0 and 2π. Note that the antisymmetry property of the phase spectrum needs to be preserved. The modified spectrum is then used for the reconstruction of stimuli, as described in Sections 2.1 and 2.2.

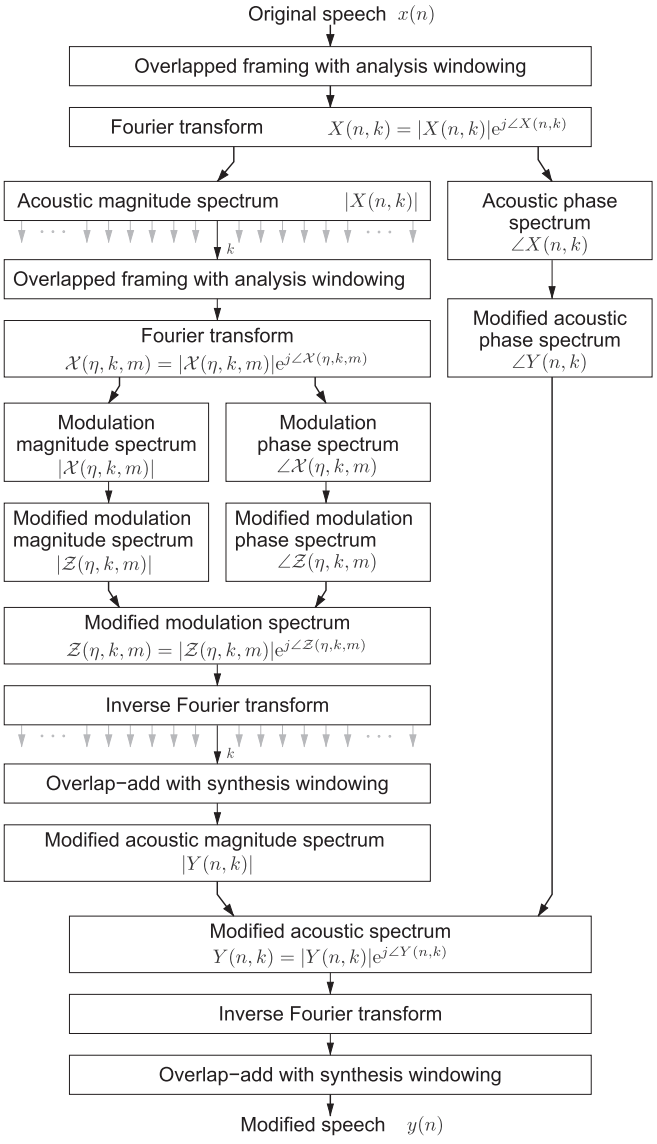

Fig. 2. Block diagram of the modulation AMS procedure.

Seven treatment types (based on types of spectral modification) were investigated in the experiments detailed in this study. These are outlined below:

- ORIG – original stimuli without modification;
- AM – stimuli generated using only the acoustic magnitude spectrum, with the acoustic phase spectrum discarded;
- AP – stimuli generated using only the acoustic phase spectrum, with the acoustic magnitude spectrum discarded;
- MM – stimuli generated using only the modulation magnitude spectrum with the modulation phase and acoustic phase spectra discarded;
- MP – stimuli generated using only the modulation phase spectrum, with the modulation magnitude and acoustic phase spectra discarded;
- MM + AP – stimuli generated using the modulation magnitude and acoustic phase spectra, with the modulation phase spectrum discarded;
- MP + AP – stimuli generated using the modulation phase and acoustic phase spectra, with the modulation magnitude spectrum discarded.

Treatment types AP and AM were constructed using the acoustic AMS procedure described in Section 2.1, and were included primarily for comparison with previous studies. Treatment types MM, MP, MM + AP, and MP + AP were constructed using the modulation AMS procedure described in Section 2.2.

3. Experiment 1: modulation spectrum intelligibility

A number of studies have investigated the significance of the acoustic magnitude and phase spectra for speech intelligibility (e.g., Schroeder, 1975; Oppenheim and Lim, 1981; Liu et al., 1997; Paliwal and Alsteris, 2005). With the

increased interest in the modulation domain for speech processing, it is therefore relevant to similarly evaluate the significance of the modulation domain magnitude and phase spectra. Therefore, in this section we evaluate the relative contributions of spectral components and their combinations to the intelligibility of speech. To achieve this, stimuli were generated to retain only selected spectral components of the modulation and acoustic spectra, as outlined in Section 2.3. Since, as previously mentioned, many modulation domain-based applications use modulation frame durations of 250 ms or more, a modulation frame duration of 256 ms was investigated here. Subjective and objective experiments were then used to evaluate the intelligibility of these stimuli.

3.1. Consonant corpus

In principle, all the vowels and consonants of the English language should be used for measuring speech intelligibility. Since this is not feasible for subjective testing, we have restricted ourselves to stop consonants in this study, as these are perhaps the most difficult sounds for human listeners to recognise. The corpus used for both the objective and subjective intelligibility tests includes six stop consonants [b, d, g, p, t, k], each placed in a vowel–consonant–vowel (VCV) context (Liu et al., 1997).[5] Four speakers were used: two male and two female. Six sentences were recorded for each speaker, giving 24 recordings in total. The recordings were made in a silent room with a SONY ECM-MS907 microphone (90° position). Each recording is around 3 s in duration, including leading and trailing silence, and sampled at 16 kHz with 16-bit precision.

3.2. Stimuli

The recordings described in Section 3.1 were processed using the AMS-based procedures detailed in Section 2. In this experiment, all acoustic domain processing used a frame duration T_{aw} of 32 ms with a 4 ms shift, and FFT analysis length of $2N$ (where $N = T_{aw}F_{as}$, and F_{as} is the acoustic domain sampling frequency). Modulation domain processing used a frame duration (T_{mw}) of 256 ms, a frame shift of 32 ms, and FFT analysis length of $2M$ (where $M = T_{mw}F_{ms}$, and F_{ms} is the modulation domain sampling frequency). In both the acoustic and modulation domains, the Hamming window was used as the analysis window function and the modified Hanning was used as the synthesis window function. Six different treatments were applied to each of the 24 recordings, including AM, AP, MM, MP, MM + AP, and MP + AP, as defined in Section 2.3. Including the original recordings, 168 stimuli files were used for these experiments. Fig. 5 shows example spectrograms for one of the recordings and each of the treatment types applied.[6,7]

3.3. Objective experiment

In this experiment, the aim was to use an objective intelligibility metric to measure the effect of inclusion or removal of different spectral components on the intelligibility of the resulting stimuli. For this purpose, the speech intelligibility index (STI) (Steeneken and Houtgast, 1980) metric was applied to each of the stimuli described in Section 3.2

3.3.1. Objective speech intelligibility metric

STI measures the extent to which slow temporal intensity envelope modulations, which are important for speech intelligibility, are preserved in degraded listening environments (Payton and Braida, 1999). For this work, a speech-based STI computation procedure was used. Here the original and processed speech signals are passed separately through a bank of seven octave band filters. Each filtered signal is squared, then low pass filtered with a 50 Hz cutoff frequency to extract the temporal intensity envelope of each signal. This envelope is then subjected to one-third octave band analysis. The components over each of the 16 one-third octave band intervals (with centres ranging from 0.5 to 16 Hz) are summed, producing 112 modulation indices. The resulting modulation spectra of the original and processed speech can then be used to calculate the modulation transfer function (MTF), and subsequently the STI.

Three different approaches were used to calculate the MTF. The first is by Houtgast and Steeneken (1985), the second is by Payton and Braida (1999), and the third is by Drullman et al. (1994). Details of the MTF and STI calculation can be found in (Goldsworthy and Greenberg, 2004). Applying the STI metric to speech stimuli returns a value between 0 and 1, where 0 indicates no intelligibility and 1 indicates maximum intelligibility. The STI metric was applied to each of the stimuli described in Section 3.2, and the average score was calculated for each treatment type.

3.3.2. Results

In the objective experiment, we have calculated the mean STI intelligibility score across the consonant corpus, using each of the three STI calculation methods, for each of the treatment types described in Section 2.3. Results of this experiment are shown in Fig. 3. Results for each of the three methods applied were found to be relatively consis-

[5] The carrier sentence used for this corpus is "hear aCa now". For example, for consonant [b] the sentence is "hear aba now".

[6] Note that all spectrograms presented in this study were generated using an acoustic frame duration of 32 ms with a 1 ms shift, and FFT length of 4096. The dynamic range is set to 60 dB. The highest peaks are shown in black, the lowest spectral valleys ($\leqslant$60 dB below the highest peaks) are shown in white, and shades of gray are used in between.

[7] The audio stimuli files are available as Supplementary materials from the Speech Communication Journal's website.

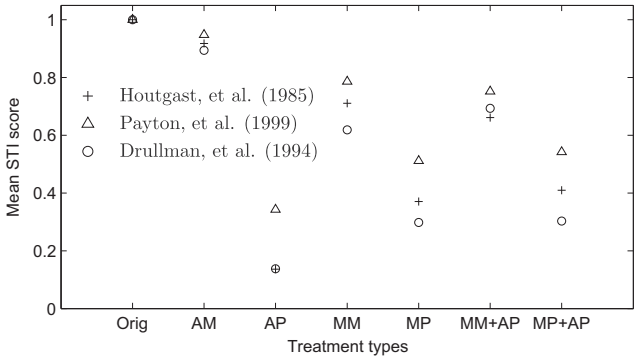

Fig. 3. Objective results in terms of mean STI scores for each of the treatments described in Section 2.3.

tent, with larger variation in results seen for types AP, MP and MP + AP, where Payton's method (Payton and Braida, 1999) attributes more importance to acoustic and modulation phase information than the other two methods. Objective results show that type AM suffers minimal loss of intelligibility with the removal of acoustic phase information. As expected, further reductions are observed for types MM and MP. Note that results also indicate type AP to have very poor intelligibility, and that little or no improvement is achieved by retaining acoustic phase information (types MM + AP and MP + AP).

3.4. Subjective experiment

While objective intelligibility tests give a quick indication of stimuli intelligibility, they are only an approximate measure. For a better indication of the intelligibility of stimuli, subjective experiments in the form of human consonant recognition tests were also conducted. The aim of this experiment was to again assess the intelligibility associated with different spectral components in the modulation-based AMS framework. For this purpose, stimuli described in Section 3.2 were used in these subjective experiments.

3.4.1. Listening test procedure

The human listening tests were conducted over a single session in a quiet room. Twelve English-speaking listeners, with normal hearing, participated in the test. Listeners were asked to identify each carrier utterance as one of the six stop consonants, and select the corresponding (labelled) option on the computer via the keyboard. A seventh option for a null response was also provided and could be selected where the participant had no idea what the consonant might have been. Stimuli audio files were played in a random order, at a comfortable listening level over closed circumaural headphones. A short practice was given at the start of the test to familiarise participants with the task. The entire test took approximately 20 min to complete.

3.4.2. Results

In the subjective experiment, we have measured consonant recognition accuracy through human listening tests. The subjective results in terms of mean consonant recognition accuracy along with standard error bars are shown with in Fig. 4. Results for type AM show that there is minimal loss of intelligibility associated with the removal of acoustic phase information from speech stimuli. Types MM and MP show a further reduction in intelligibility from the removal of modulation phase and modulation magnitude spectrum information, respectively. These results are consistent with what was observed in the objective experiments. Results of Fig. 4 also show that type MP not only has lower intelligibility scores than type MM, but its scores have a considerably greater variance than for all other types.

The subjective results also suggest that the acoustic phase spectrum contributes more significantly to intelligibility than was indicated by objective results. This is shown by the much higher intelligibility scores for type AP shown in the subjective results than in the objective results. Subjective results also show significant improvement in intelligibility for MM and MP types where acoustic phase information is also retained (types MM + AP and MP + AP). This is different to the objective results, where types MM + AP and MP + AP had mean intelligibility scores that were approximately the same (or less) as those for MM and MP types. This difference between objective and subjective results for types AP, MM + AP and MP + AP can be attributed to the way that STI is calculated. The STI metric predominantly reflects formant information, while it does not attribute importance to pitch frequency harmonics. Consequently, STI scores for type MM + AP are comparable to scores for type MM, but STI scores for type AP are worse than for all other types.

3.5. Spectrogram analysis

Spectrograms for a "*hear aba now*" utterance by a male speaker are shown in Fig. 5(a), and spectrograms for each type of treatment described in Section 2.3 are shown in Fig. 5(b)–(g). The spectrogram for type AM stimulus given

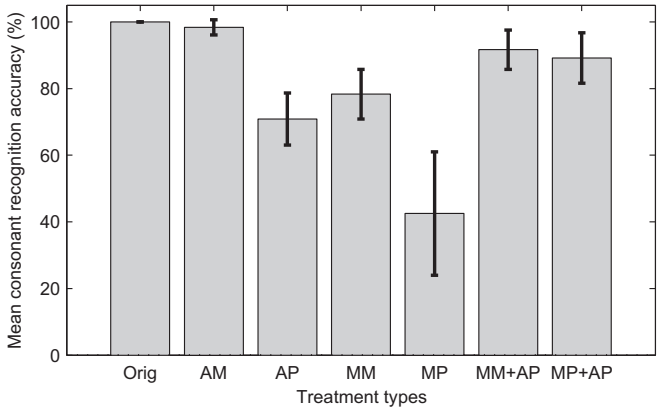

Fig. 4. Subjective intelligibility scores in terms of mean consonant recognition accuracy (%) for each of the treatments described in Section 2.3.

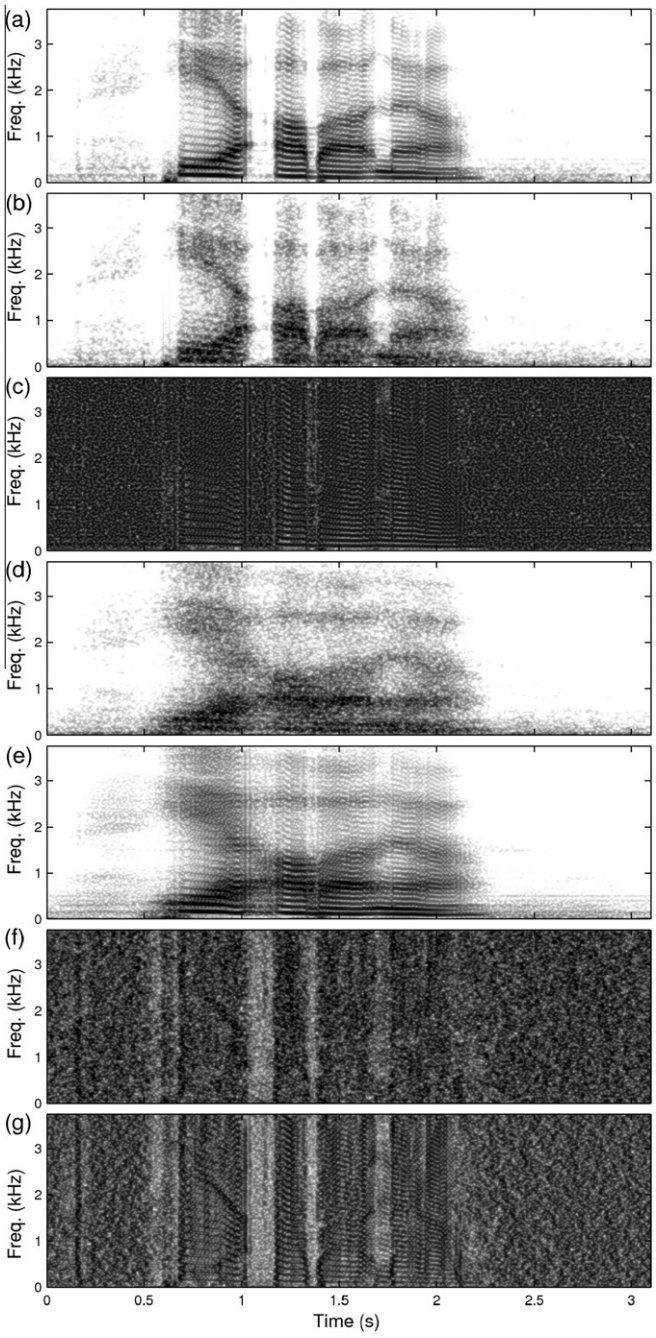

Fig. 5. Spectrograms of a "hear aba now" utterance by a male speaker. (a) Original speech. (b) and (c) Acoustic AMS processed speech using acoustic frame durations of 32 ms. Stimuli types as defined in Section 2.3 are: (b) AM, and (c) AP. (d)–(g) Modulation AMS processed speech using frame durations of 32 ms in the acoustic domain and 256 ms in the modulation domain. Stimuli types as defined in Section 2.3 are: (d) MM, (e) MM + AP, (f) MP, and (g) MP + AP.

in Fig. 5(b) shows clear formant information, with some loss of pitch frequency harmonic information. As a result, speech sounds clean and intelligible, but also has a breathy quality. On the other hand, the spectrogram for type AP stimulus in Fig. 5(c) is heavily submersed in noise without visible formant information, but with more pronounced pitch frequency harmonics than those seen in the type AM spectrogram. Type AP stimulus contains static noise, which masks speech and reduces intelligibility. The spectrograms for stimuli of type MM and MM + AP (given in Fig. 5(d) and (e), respectively) show that the modulation magnitude spectrum contains much of the formant information. The effect of temporal smearing due to the use of long modulation frame durations for processing of the modulation magnitude spectra can be clearly seen. This effect is heard as a slurring or reverberant quality. The spectrograms for stimuli of types MP and MP + AP (given in Fig. 5(f) and (g), respectively) show some formant information submersed in strong noise. The formants are more pronounced for type MP + AP than for type MP. The inclusion of the acoustic phase spectrum in the construction of MP + AP stimuli also introduces pitch frequency harmonics, as can be seen in the spectrogram of Fig. 5(g). The temporal smearing effect is not seen in the spectrograms for types MP and MP + AP. This is because the modulation phase spectrum is not affected by long window durations in the same way that the modulation magnitude spectrum is (this is further investigated in the experiments of Sections 4 and 5). The reduced intelligibility of MP and MP + AP stimuli, observed in the objective and subjective experiments, can be attributed to the presence of high intensity noise and reduced formant structure.

3.6. Discussion

From the results of the subjective and objective experiments, we can see that types MM and MP were both improved by including acoustic phase information. There is more variation in type MP than in any of the other types. Results support the idea that the modulation phase spectrum is more important to intelligibility than the acoustic phase spectrum, in that removal of the acoustic phase spectrum causes minimal reduction of intelligibility for type AM, while removal of modulation phase from type AM (which gives type MM), significantly reduces speech intelligibility.

4. Experiment 2: frame duration for processing of the modulation magnitude spectrum

Speech processing in the acoustic domain, typically uses acoustic frame durations between 20 and 40 ms (e.g., Picone, 1993; Huang et al., 2001; Loizou, 2007). Experiments such as those by Paliwal and Wójcicki (2008), have shown that speech containing only the acoustic magnitude spectrum is most intelligible for acoustic frame durations between 15 and 35 ms. In the modulation domain, much larger frame durations are typically used in order to effectively represent speech information. This is justifiable since the modulation spectrum of most audio signals changes relatively slowly (Thompson and Atlas, 2003). However, if the frame duration is too long, then a spectral smearing distortion is introduced. Therefore, modulation domain based algorithms generally use frame durations of around

250 ms (e.g., Greenberg and Kingsbury, 1997; Thompson and Atlas, 2003). In this section, we aim to evaluate the effect of modulation frame duration on intelligibility, in order to determine the optimum frame duration for processing of the modulation magnitude spectrum.

To achieve this, stimuli were constructed such that only the modulation magnitude spectrum was retained (type MM), with both the acoustic and modulation phase spectra removed by randomising their values. The stimuli were generated using modulation frame durations between 32 and 1024 ms. Objective and subjective intelligibility experiments were then used to determine the average intelligibility of stimuli for each duration.

4.1. Stimuli

The consonant corpus described in Section 3.1 was used for the experiments detailed in this section. Stimuli were generated using the modulation AMS procedure given in Section 2.2. Here, only the modulation magnitude spectrum was retained (type MM), with both the acoustic and modulation phase information removed by randomising their spectral values. In the acoustic domain, a frame duration of 32 ms, with a 4 ms shift, and FFT analysis length of $2N$ (where $N = T_{aw}F_{as}$) were used. In both the acoustic and modulation domains, the Hamming window was used as the analysis window function and the modified Hanning window was used as the synthesis window function (Griffin and Lim, 1984). In the modulation domain, six modulation frame durations were investigated ($T_{mw} = 32$, 64, 128, 256, 512, and 1024 ms). Here, the shift was set to one-eighth of the frame duration, with an FFT analysis length of $2M$ ($M = T_{mw}F_{ms}$).

Therefore, a total of 6 different treatments were applied to the 24 recordings of the corpus. Including the original recordings, 168 stimuli files were used for each test. Fig. 8 shows example spectrograms for each treatment applied to one of the recordings.

4.2. Objective experiment

In this section, we evaluate the intelligibility of the stimuli reconstructed from only the modulation magnitude spectrum (type MM) using the STI (Steeneken and Houtgast, 1980) intelligibility metric described in Section 3.3.1. The mean STI scores were calculated for the stimuli generated using each of the modulation frame durations considered. The mean intelligibility scores are shown in Fig. 6.

Results of the objective experiments show that for small durations, such as 32, 64, and 128 ms, the intelligibility of type MM stimuli is high. As frame duration is increased, mean intelligibility scores decrease. This trend is consistent across each of the three STI calculation methods applied. Scores returned for 32 and 64 ms show that removal of both the modulation phase and acoustic phase information causes only a small reduction in intelligibility. Objective results for type MM with a small modulation frame duration are very close to the objective results for type AM, as shown in Fig. 3.

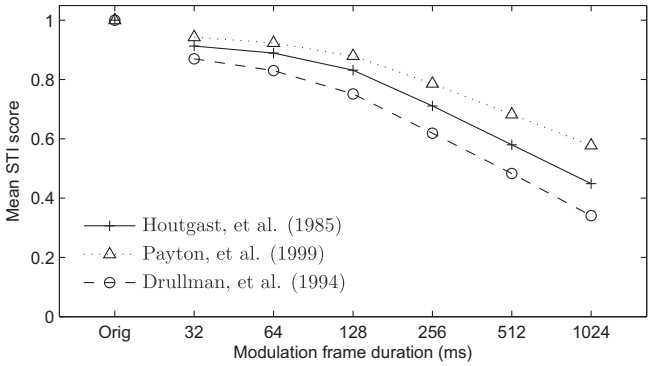

Fig. 6. Objective results in terms of mean STI scores for stimuli with treatment type MM and modulation frame durations of 32, 64, 128, 256, 512, and 1024 ms.

4.3. Subjective experiment

Subjective evaluation of the intelligibility of stimuli described in Section 4.1 was again in the form of a human listening test that measures consonant recognition performance. The test was conducted in a separate single session under the same conditions as for Experiment 1 described in Section 3.4.1. Twelve English-speaking listeners with normal hearing participated in this test.

The results of the subjective experiment, along with the standard error bars, are shown in Fig. 7. Subjective results also show that a modulation frame duration of 32 ms gives the highest intelligibility for type MM stimuli. Durations of 64, 128, and 256 ms showed moderate reductions in intelligibility compared to scores for 32 ms, while much poorer scores were recorded for larger frame durations. These results are consistent with those from objective experiments, having reduced intelligibility for increased frame durations. In particular, objective scores and subjective accuracy are approximately the same for durations 64, 128, and 256 ms. For larger durations, subjective scores

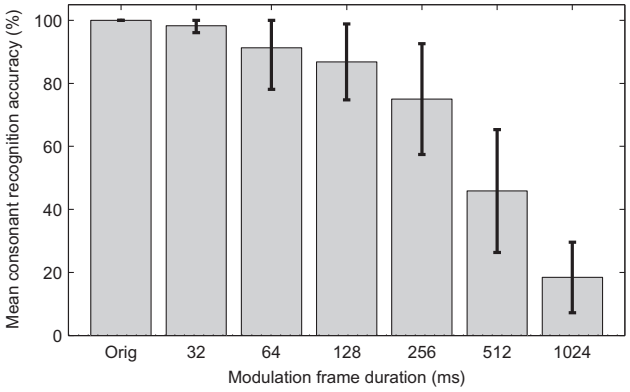

Fig. 7. Subjective results in terms of mean consonant recognition accuracy (%) for stimuli with treatment type MM and modulation frame durations of 32, 64, 128, 256, 512, and 1024 ms.

indicate intelligibility to be much poorer than predicted by the STI metrics.

4.4. Spectrogram analysis

Spectrograms of a "*hear aba now*" utterance by a male speaker are shown in Fig. 8. Fig. 8(a) shows the original signal, where formants and pitch frequency harmonics are clearly visible. For stimuli created using modulation frame durations of 32 and 64 ms (shown in Fig. 8(b) and (c), respectively), formants are relatively clear with some loss of pitch frequency harmonic information resulting in speech which sounds a little breathy, but still very intelligible.

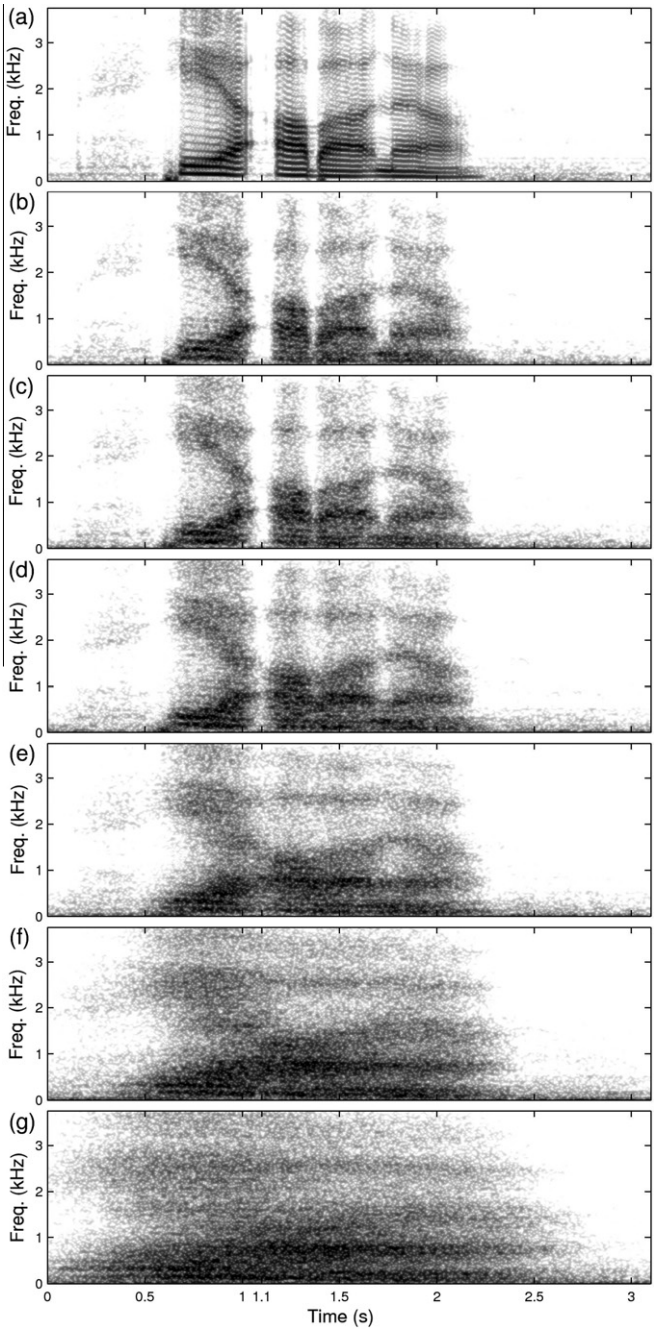

Fig. 8. Spectrograms of a "hear aba now" utterance, by a male speaker: (a) Original speech (passed through AMS procedure with no spectral modification). (b)–(g) Processed speech – MM stimuli for the following modulation frame durations: (b) 32 ms; (c) 64 ms; (d) 128 ms; (e) 256 ms; (f) 512 ms and (g) 1024 ms.

As frame duration is increased, a spectral smearing distortion due to a lack of localisation of speech information becomes noticeable. In the spectrograms of type MM stimuli for durations of 128 and 256 ms (shown in Fig. 8(d) and (e), respectively), this spectral smearing can be easily seen in the silent region at 1.1 s, where energies from earlier frames have spread into the low energy silence region. This spectral smearing gives the speech a reverberant quality. Again, the reduction in harmonic structure makes the speech sound breathy. However, because formants are still defined, the speech is still intelligible.

The spectrograms of stimuli of type MM for durations of 512 and 1024 ms are shown in Fig. 8(f) and (g), respectively. As can be seen, there is extensive smearing of spectral energies with formants difficult to distinguish. Listening to stimuli, speech has accentuated slurring making intelligibility poor.

4.5. Discussion

While the frame duration generally used in modulation domain processing is around 250 ms, the above results suggest that smaller frame durations, such as 32 or 64 ms, may improve the intelligibility of stimuli based on the modulation magnitude spectrum. They also suggest that intelligibility, for stimuli retaining only the modulation magnitude spectrum and using a modulation frame duration of 32 ms, is quite close to that obtained by retaining the whole acoustic magnitude spectrum. These results are consistent with results of similar intelligibility experiments in the acoustic domain by Liu et al. (1997) as well as Paliwal and Alsteris (2005), where smaller frame durations gave higher intelligibility for stimuli retaining only the acoustic magnitude spectrum, with intelligibility decreasing for increasing frame durations.

5. Experiment 3: frame duration for processing of the modulation phase spectrum

In the acoustic domain, there has been some debate as to the contribution of acoustic phase spectrum to intelligibility (e.g., Schroeder, 1975; Oppenheim and Lim, 1981; Wang and Lim, 1982; Liu et al., 1997; Paliwal and Alsteris, 2005; Wójcicki and Paliwal, 2007). For instance, in speech enhancement the acoustic phase spectrum is considered unimportant at high SNRs (Loizou, 2007). On the other hand, the modulation phase spectrum is considered to be more important than the acoustic phase spectrum (e.g., Greenberg, et al., 1998; Kanedera et al., 1998; Atlas et al., 2004). In this experiment we would like to further evaluate the contribution of the modulation phase spectrum to intelligibility, as modulation frame duration is increased. For this purpose, stimuli are generated to retain

only the modulation phase spectrum information for modulation frame durations ranging between 32 and 1024 ms.

5.1. Stimuli

The consonant corpus described in Section 3.1 was again used for experiments presented in this section. The stimuli were generated using the modulation AMS procedure detailed in Section 2.2. Here, only the modulation phase spectrum was retained (type MP), with acoustic phase and modulation magnitude information removed. In the acoustic domain, a frame duration of 32 ms, a frame shift of 4 ms, and FFT analysis length of $2N$ (where $N = T_{aw}F_{as}$) was used. In both the acoustic and modulation domains, the Hamming window was used as the analysis window function and the modified Hanning was used as the synthesis window function. In the modulation domain, six modulation frame durations were investigated ($T_{mw} = 32$, 64, 128, 256, 512, and 1024 ms). Here, the frame shift was set to one-eighth of the frame length, with an FFT analysis length of $2M$ (where $M = T_{mw}F_{ms}$). A total of 6 different treatments were applied to the 24 recordings of the corpus. Including the original recordings, 168 stimuli files were used for the tests. Fig. 11 shows example spectrograms for one of the recordings and each of the treatments applied.

5.2. Objective experiment

In the objective experiment, we evaluate the intelligibility of stimuli constructed using only the modulation phase spectrum information (type MP) using the STI intelligibility metric described in Section 3.3.1. The mean STI score was calculated for stimuli generated using each of the modulation frame durations investigated. These mean intelligibility scores are shown in Fig. 9.

Results of the objective experiments show that intelligibility increases as frame duration increases. For small frame durations, intelligibility was only around 20% (using the Houtgast et al. STI calculation method),[8] while for high frame durations the intelligibility was around 59%. These results are relatively consistent for each of the STI methods applied.

5.3. Subjective experiment

Human listening tests measuring consonant recognition performance were used to subjectively evaluate the intelligibility of stimuli described in Section 5.1. The test was conducted in a separate session under the same conditions as for Experiment 1 (Section 3.4.1). Twelve English-speaking listeners participated in the test.

[8] Please note that figures giving objective results show intelligibility scores for three objective speech-based STI metrics. However, our in-text discussions refer (for brevity) to the STI results for the (Houtgast and Steeneken, 1985) method only.

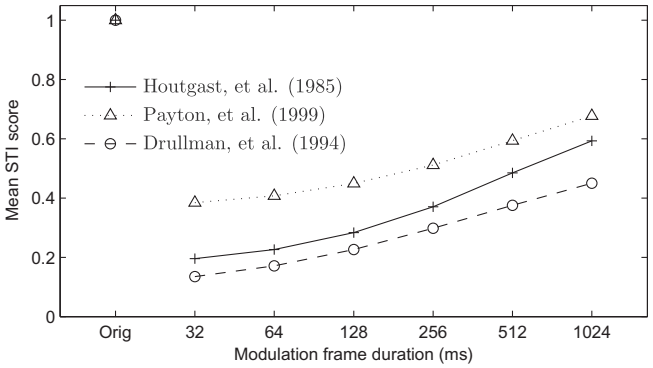

Fig. 9. Objective results in terms of mean STI scores for stimuli with treatment type MP and modulation frame durations of 32, 64, 128, 256, 512, and 1024 ms.

The results of the subjective experiment, along with standard error bars, are shown in Fig. 10. Consistent with the objective results, the subjective speech intelligibility is shown to increase for longer modulation frame durations, where stimuli are generated using only the modulation phase spectrum. As can be seen, much longer modulation analysis frame durations are required for reasonable intelligibility compared to the modulation magnitude spectrum. For small frame durations (32 and 64 ms), intelligibility is negligible, while for large frame durations (1024 ms), intelligibility is around 86%, which is close to the intelligibility of type AM stimuli. These results also show that intelligibility, as a function of modulation frame duration, varies much more than indicated by the objective metrics, with subjective results ranging from 0% to 86% compared to objective results ranging from 20% to 59%.

5.4. Spectrogram analysis

Spectrograms for a "*hear aba now*" utterance by a male speaker are shown in Fig. 11. Fig. 11(a) shows the original signal, while the stimuli where only the modulation phase spectrum is retained (i.e., type MP) for modulation frame durations of 32, 64, 128, 256, 512, and 1024 ms are shown

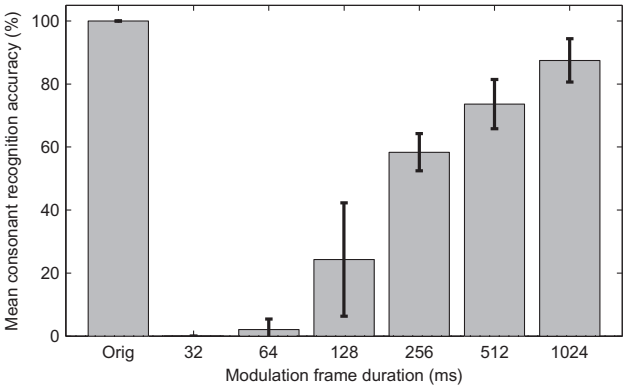

Fig. 10. Subjective results in terms of mean consonant recognition accuracy (%) for stimuli with treatment type MP and modulation frame durations of 32, 64, 128, 256, 512, and 1024 ms.

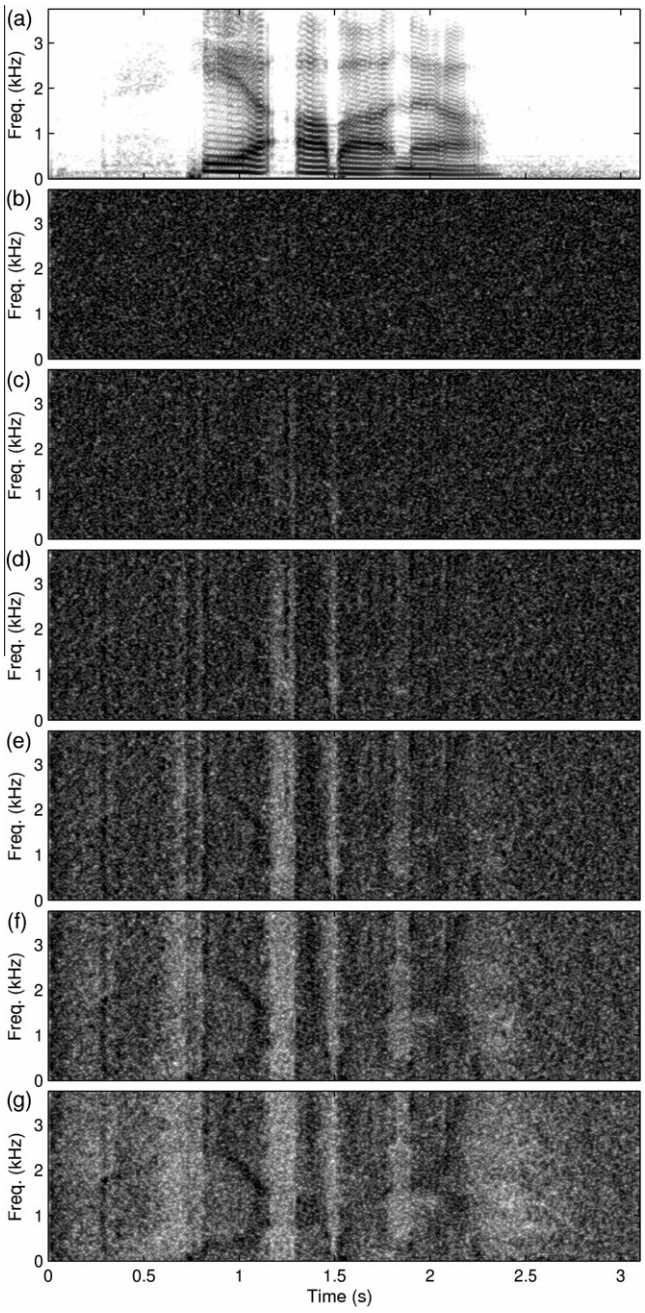

Fig. 11. Spectrograms of utterance "hear aba now", by a male speaker. (a) Original speech (passed through AMS procedure with no spectral modification). (b)–(g) Processed speech – MP stimuli for the following modulation frame durations: (b) 32 ms; (c) 64 ms; (d) 128 ms; (e) 256 ms; (f) 512 ms and (g) 1024 ms.

in Fig. 11(b)–(g), respectively. For 32–128 ms frame durations (Fig. 11(b)–(d)), the spectrograms are submersed in noise with almost no observable formant information. Informal listening tests indicate that these stimuli sound predominantly like static or white noise. Breathy sounding speech can be heard for stimuli generated using 128 ms, but it is heavily submersed in noise. For 256 ms frame durations (Fig. 11(e)), the spectrogram begins to show formant information, with background noise of slightly lower intensity. Listening to stimuli, the sentence can now be heard and understood, but speech sounds breathy due to lack of pitch frequency harmonic information. For 512 and 1024 ms frame durations (Fig. 11(f) and (g)), background noise is further reduced and formant information is clearer. Listening to stimuli, the background noise is quieter (though more metallic in nature), and speech is more intelligible. Thus larger frame durations result in improved intelligibility because there is less background noise swamping the formant information in the spectrum.

5.5. Discussion

The above results can be explained as follows. The results of both the objective and subjective experiments show that there is an increase in intelligibility for an increase in modulation frame duration for stimuli generated using only the modulation phase spectrum (type MP). Results show that 256 ms is the minimum reasonable frame duration for intelligibility, but that intelligibility improves if the frame duration is further increased. Spectrograms also show that the modulation phase spectrum is not susceptible to the effects of localisation (i.e., spectral smearing) like the modulation magnitude spectrum is.

Results shown in this section are consistent with results of similar experiments in the acoustic domain where intelligibility was shown to increase for increasing acoustic frame durations (Paliwal and Alsteris, 2005). However, here, intelligibility is much lower for smaller durations than observed for the acoustic phase spectrum.

6. Discussion and conclusion

In this paper, we firstly considered a modulation frame duration of 256 ms, as is commonly used in applications based on the modulation magnitude spectrum. We investigated the relative contribution of the modulation magnitude and phase spectra towards speech intelligibility. The main conclusions from this investigation are as follows. For the above frame duration, it was observed that the intelligibility of stimuli constructed from only the modulation magnitude or phase spectra is significantly lower than the intelligibility of the acoustic magnitude spectrum. Notably, the intelligibility of stimuli generated from either the modulation magnitude or modulation phase spectra was shown to be considerably improved by also retaining the acoustic phase spectrum.

Secondly, we investigated the effect of the modulation frame duration on intelligibility for both the modulation magnitude and phase spectra. Results showed that speech reconstructed from only the short-time modulation phase spectrum has highest intelligibility when long modulation frame durations ($>$256 ms) are used, and that for small durations ($\leqslant$64 ms) the modulation phase spectrum can be considered relatively unimportant for intelligibility. On the other hand, speech reconstructed from only the short-time modulation magnitude spectrum is most intelligible when small modulation frame durations ($\leqslant$64 ms) are used,

with the intelligibility due to modulation magnitude spectrum decreasing with increasing modulation frame durations. These conclusions were supported by objective and subjective intelligibility experiments, as well as spectrogram analysis and informal listening tests. The decrease in intelligibility with increasing frame duration for the stimuli constructed from only the modulation magnitude spectrum, and the increase in intelligibility for stimuli constructed from only the modulation phase spectrum, is consistent with the results of similar intelligibility experiments in the acoustic domain (Liu et al., 1997; Paliwal and Alsteris, 2005).

Thus, the main conclusions from the research presented in this work are two-fold. First, for applications based on the short-time modulation magnitude spectrum, short modulation frame durations are more suitable. Second, for applications based on the short-time modulation phase spectrum, long modulation frame durations are more suited. Contrary to these findings, many applications which process the modulation magnitude spectrum use modulation frame durations of 250 ms or more (e.g., Greenberg and Kingsbury, 1997; Thompson and Atlas, 2003; Kim, 2005; Falk and Chan, 2008; Wu et al., 2009; Falk et al., 2010; Falk and Chan, 2010). Therefore an implication of this work is the potential for improved performance of some of these modulation magnitude spectrum based applications by use of much shorter modulation frame durations (such as 32 ms). Example applications which may benefit from use of shorter modulation frame durations include speech and speaker recognition, objective intelligibility metrics as well as speech enhancement algorithms. These will be investigated in future work.

It should also be noted that for applications that use the modulation spectrum (i.e., both the magnitude and phase spectra), the choice of optimal frame duration will depend on other considerations. For example, delta-cepstrum and delta–delta-cepstrum are used in automatic speech recognition with modulation frame durations of around 90 and 250 ms, respectively (Hanson and Applebaum, 1993). Similarly, in speech enhancement, we have used modulation frame durations of 250 ms for modulation domain spectral subtraction method (Paliwal et al., 2010b) and 32 ms for modulation domain MMSE magnitude estimation method (Paliwal et al., 2010a).

Appendix A. Objective quality evaluation

Speech quality is a measure which quantifies how nice speech sounds and includes attributes such as intelligibility, naturalness, roughness of noise, etc. In the main body of this paper we have solely concentrated on the intelligibility attribute of speech quality. More specifically our research focused on the objective and subjective assessment of speech intelligibility of the modulation magnitude and phase spectra at different modulation frame durations. However, in many speech processing applications, the overall quality of speech is also important. Therefore in this appendix, for the interested reader, we provide objective speech quality results for the modulation magnitude and phase spectra as a function of modulation frame duration. Two metrics commonly used for objective assessment of speech quality are considered, namely the perceptual evaluation of speech quality (PESQ) (Rix et al., 2001), and the segmental SNR (Quackenbush et al., 1988). Mean scores for the PESQ and segmental SNR metrics, computed over the Noizeus corpus (Hu and Loizou, 2007), are shown in Figs. 12 and 13, respectively.

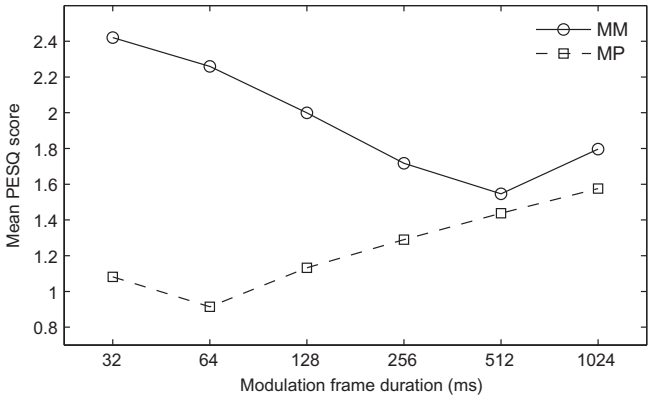

Fig. 12. Objective results in terms of mean PESQ score for stimuli with treatment types MM and MP, for modulation frame durations of 32, 64, 128, 256, 512, and 1024 ms.

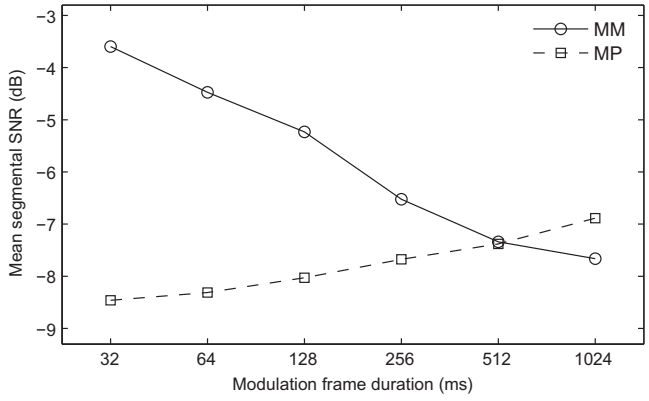

Fig. 13. Objective results in terms of mean segmental SNR (dB) for stimuli with treatment types MM and MP, for modulation frame durations of 32, 64, 128, 256, 512, and 1024 ms.

In general, both measures suggest that the overall quality of the MM stimuli improves with decreasing modulation frame duration, while for the MP stimuli this trend is reversed. For the most part, these indicative trends are consistent with those observed for intelligibility results given in Sections 4 and 5.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.specom.2010.10.004.

References

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

Atlas, L., Vinton, M., 2001. Modulation frequency and efficient audio coding. In: Proc. SPIE Internat. Soc. Opt. Eng., Vol. 4474, pp. 1–8.

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

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

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

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

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

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

Greenberg, S., Arai, T., Silipo, R., 1998. Speech intelligibility derived from exceedingly sparse spectral information. In: Proc. Internat. Conf. Spoken Lang. Process. (ICSLP), Vol. 6, Sydney, Australia, pp. 2803–2806.

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

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

Hanson, B., Applebaum, T., 1993. Subband or cepstral domain filtering for recognition of lombard and channel-distorted speech. In: Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP), Vol. 2, Minneapolis, MN, USA, pp. 79–82.

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

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

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

Kanedera, N., Hermansky, H., Arai, T., 1998. Desired characteristics of modulation spectrum for robust automatic speech recognition. In: Proc. IEEE Internat. Conf. Acoust. Speech Signal Process. (ICASSP), Seattle, USA, pp. 613–616.

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

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

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

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

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

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

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

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

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

Paliwal, K., Schwerin, B., Wójcicki, K., 2010a. Speech enhancement using minimum mean-square error short-time spectral modulation magnitude estimator. Technical report, SPL-10-1, Griffith University, Brisbane, Australia.

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

Payton, K., Braida, L., 1999. A method to determine the speech transmission index from speech waveforms. J. Acoust. Soc. Amer. 106 (6), 3637–3648.

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

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

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

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

Schroeder, M., 1975. Models of hearing. Proc. IEEE 63 (9), 1332–1350.

Steeneken, H., Houtgast, T., 1980. A physical method for measuring speech-transmission quality. J. Acoust. Soc. Amer. 67 (1), 318–326.

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

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

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

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 Internat. Conf. Acoust. Speech Signal Process. (ICASSP), Vol. 4, Honolulu, Hawaii, USA, pp. 729–732.

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