Channel selection in the modulation domain for improved speech intelligibility in noise

Kamil K. Wójcicki and Philipos C. Loizou[a)]
Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas 75083-0688

(Received 2 September 2011; revised 27 January 2012; accepted 27 January 2012)

Background noise reduces the depth of the low-frequency envelope modulations known to be important for speech intelligibility. The relative strength of the target and masker envelope modulations can be quantified using a modulation signal-to-noise ratio, $(S/N)_{mod}$, measure. Such a measure can be used in noise-suppression algorithms to extract target-relevant modulations from the corrupted (target + masker) envelopes for potential improvement in speech intelligibility. In the present study, envelopes are decomposed in the modulation spectral domain into a number of channels spanning the range of 0–30 Hz. Target-dominant modulations are identified and retained in each channel based on the $(S/N)_{mod}$ selection criterion, while modulations which potentially interfere with perception of the target (i.e., those dominated by the masker) are discarded. The impact of modulation-selective processing on the speech-reception threshold for sentences in noise is assessed with normal-hearing listeners. Results indicate that the intelligibility of noise-masked speech can be improved by as much as 13 dB when preserving target-dominant modulations, present up to a modulation frequency of 18 Hz, while discarding masker-dominant modulations from the mixture envelopes.
 [http://dx.doi.org/10.1121/1.3688488]

PACS number(s): 43.66.Ba, 43.71.Rt, 43.71.Es [EB] Pages: 2904–2913

I. INTRODUCTION

The speech signal can be represented as a sum of amplitude-modulated signals in a number of narrow frequency subbands spanning the signal bandwidth (Drullman *et al.*, 1994b). The output waveforms of each subband can be described in terms of a carrier signal (fine structure) and an envelope. The temporal modulations present in the envelope convey important information involving both segmental (e.g., manner of articulation) and suprasegmental (e.g., intonation) distinctions in speech. The strength of these temporal-envelope modulations has been quantified in terms of the modulation index (Houtgast and Steeneken, 1985). Reduction in modulation depth due to, for instance, noise or reverberation has been used as a good predictor of speech intelligibility. This led to the development of the concept of modulation transfer function for intelligibility prediction in room acoustics (Houtgast and Steeneken, 1973). The modulation transfer function forms the basis for the speech transmission index (STI), an objective measure used for prediction of speech intelligibility (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985).

There is growing physiological and psychoacoustic evidence in support of modulation processing in the auditory system (e.g., Schreiner and Urbas, 1986; Bacon and Grantham, 1989; Sheft and Yost, 1990; Shamma, 1996; Ewert and Dau, 2000; Depireux *et al.*, 2001; Atlas and Shamma, 2003). Psychophysical experiments by Bacon and Grantham (1989), for example, indicated that there are channels in the auditory system which are tuned to the detection of low-frequency modulations. Follow up experiments by Ewert and Dau (2000) revealed the shapes of these modulation filters by measuring masked threshold patterns for a set of signal frequencies spanning the range from 4 to 256 Hz in the presence of $\frac{1}{2}$-octave-wide modulation maskers. The results of these psychophysical experiments were interpreted as indicating frequency selectivity in the envelope-frequency domain (i.e., modulation domain), analogous to the frequency selectivity in the acoustic-frequency domain. Experiments by Schreiner and Urbas (1986) showed that a neural representation of amplitude modulation is preserved through all levels of the mammalian auditory system, including the highest level of audition, the auditory cortex. Neurons in the auditory cortex are thought to decompose the acoustic spectrum into spectro-temporal modulation content (Mesgarani and Shamma, 2005; Schönwiesner and Zatorre, 2009), and are best driven by sounds that combine both spectral and temporal modulations (Shamma, 1996; Kowalski *et al.*, 1996; Depireux *et al.*, 2001).

There exists ample behavioral evidence in support of the contribution of low-frequency amplitude modulations to speech perception (e.g., Houtgast and Steeneken, 1985; Drullman *et al.*, 1994a,b; Elliott and Theunissen, 2009). Drullman *et al.* (1994a,b), for example, investigated the importance of low-frequency modulation frequencies for intelligibility by applying low-pass and high-pass filters to the envelopes extracted from 24 $\frac{1}{4}$-octave bands. Modulation frequencies between 4 and 16 Hz were found to contribute the most to intelligibility, with the region around 4–5 Hz being the most significant, reflecting the rate at which syllables are produced. In their studies, both target and masker modulations were present in the envelopes. The speech-reception thresholds (SRTs) obtained with the filtered stimuli, containing modulation frequencies lower than 16 Hz, were 1–6 dB higher than those obtained with the control, unfiltered,

[a)]Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu

 0001-4966/2012/131(4)/2904/10/$30.00

stimuli (the SRTs of the filtered stimuli did not differ from those of the control stimuli for modulation cutoff frequencies greater than 16 Hz). In brief, their studies did not assess speech intelligibility improvements achievable by isolating, or somehow extracting, only the target-relevant modulations from the envelopes. This is of interest, as such an approach could potentially be incorporated in noise-reduction algorithms to improve speech intelligibility.

In the present study, we consider the selection (and extraction) of target-relevant modulations from the corrupted target + masker envelopes as a means of designing algorithms that could potentially improve speech intelligibility. Finding a method for delineating target-relevant modulations from the modulations introduced by the masker is, however, not an easy task (e.g., Dubbelboer and Houtgast, 2007) and requires a perceptually meaningful selection criterion operating in the modulation domain. For signals subjected to nonlinear processing, this task becomes even more difficult due to the introduction of stochastic modulations caused by the interaction of the target and masker modulations (Dubbelboer and Houtgast, 2007). As a selection criterion in this work we consider the signal-to-noise ratio defined in the modulation spectral domain and henceforth denoted as $(S/N)_{mod}$ to distinguish it from the signal-to-noise ratio (SNR) defined in the acoustic spectral domain. Based on this modulation-selective criterion, envelopes can be constructed by retaining modulations with $(S/N)_{mod}$ greater than a prescribed threshold, while discarding modulations with $(S/N)_{mod}$ smaller than a prescribed threshold.

The importance of measures similar to $(S/N)_{mod}$ has been highlighted in a number of studies (e.g., Dubbelboer and Houtgast, 2008; Jørgensen and Dau, 2011). Dubbelboer and Houtgast (2008), for instance, have defined a similar measure for assessment of the relative strength of the signal and nonsignal modulations in the context of nonlinear envelope modification by noise-reduction algorithms. Their measure was proposed as a tool for predicting the limited effect of conventional noise-reduction algorithms, such as the spectral subtraction algorithm, on speech intelligibility. Their study was motivated by the fact that the conventional STI measure falls short of predicting the limited effects of noise-suppression on speech intelligibility (Ludvigsen, 1993). The study by Jørgensen and Dau (2011) used a metric similar to $(S/N)_{mod}$ in a model for predicting speech intelligibility in noise. Good agreement was found between the model and the intelligibility of speech in conditions involving steady noise, spectral-subtraction, and reverberation.

In the listening experiments presented in this work, we assume *a priori* knowledge of $(S/N)_{mod}$ available prior to mixing of the target and masker. This is done in order to assess the full potential of the proposed modulation channel selection scheme in terms of intelligibility improvement. Access to $(S/N)_{mod}$ is assumed within a range of modulation frequencies known to be important for speech perception (i.e., below 30 Hz). Sentence stimuli are synthesized using envelopes which are constructed by retaining modulations with $(S/N)_{mod}$ greater than a prescribed threshold, while discarding modulations with $(S/N)_{mod}$ smaller than a prescribed threshold. The modulation spectra are computed using a dual analysis-modification-synthesis framework, which allows processing in the modulation domain on relatively short intervals (256 ms). This is done for practical implementation purposes, and stands in contrast with common practices of using extremely long (in the order of minutes sometimes) speech segments from continuous discourse to compute the modulation spectra (Houtgast and Steeneken, 1985). Preliminary analysis is presented regarding the feasibility of estimating the $(S/N)_{mod}$ values directly from the mixture (target + masker) envelopes.

II. MODULATION CHANNEL SELECTION

This section gives a description of the proposed modulation channel-selection (MCS) algorithm. We begin with a brief summary of a dual analysis-modification-synthesis framework that enables processing in the short-time modulation spectral domain. Details of the MCS approach are then introduced, followed by an illustration of the MCS concept using simple synthetic stimuli.

A. Processing in the modulation domain

The proposed MCS approach uses a dual analysis-modification-synthesis framework, similar to that used by Paliwal *et al.* (2011), that allows processing in the short-time modulation spectral domain. The block diagram of the MCS processing is shown in Fig. 1. Under this framework, the speech signal is processed framewise using short-time Fourier analysis (Schafer and Rabiner, 1973). The spectrum is computed

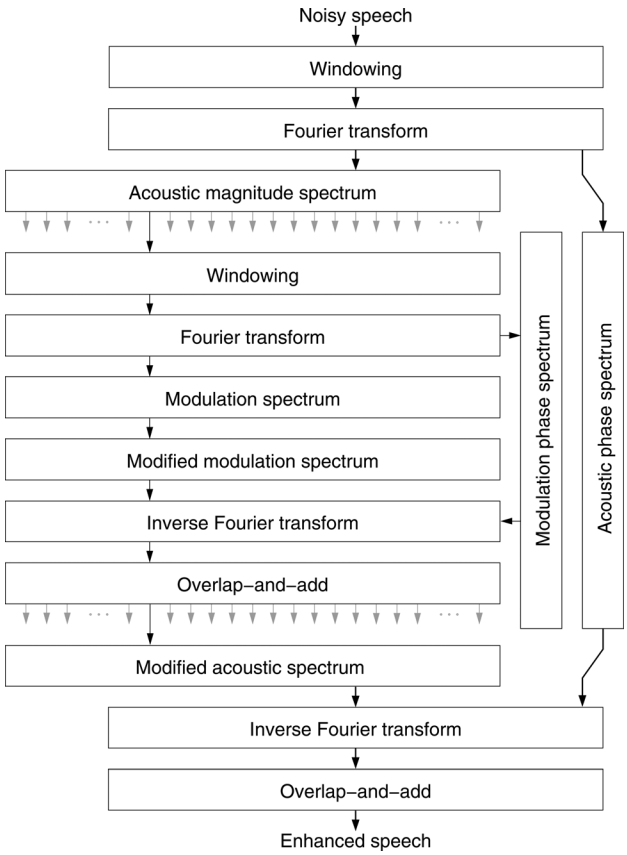

FIG. 1. Block diagram of a dual analysis-modification-synthesis approach used for processing in the short-time modulation spectral domain.

using the fast Fourier transform (FFT). The (acoustic) phase spectrum remains unmodified for synthesis, while the acoustic magnitude spectrum is further processed as follows. Time trajectories of the acoustic magnitude spectrum (at fixed acoustic frequencies) are accumulated over a finite interval of T s and subjected to a second short-time Fourier analysis to produce the modulation spectrum. At this point, the modulation spectrum can be modified, e.g., low-pass filtered or processed in some other way. In our study, modulation spectrum components (also referred to as modulation channels) satisfying a given criterion are retained while the remaining modulation components are discarded. The inverse short-time Fourier transform of the modified modulation spectrum is computed (using the unmodified modulation phase spectrum), and the overlap-and-add procedure (Griffin and Lim, 1984) is used to produce the modified trajectories of the acoustic magnitude spectrum. A subsequent inverse short-time Fourier transform of the acoustic spectrum (using the unmodified acoustic phase spectrum) is computed and the overlap-and-add procedure is finally used to synthesize the speech signal.

B. Modulation channel selection

The modulation spectra, computed using the above procedure, are modified as follows. Let us denote the modulation spectra of clean speech, noisy speech, and noise as $\mathcal{S}(f,m)$, $\mathcal{X}(f,m)$, and $\mathcal{D}(f,m)$, respectively, where f is the acoustic frequency and m is the modulation frequency. The signal-to-noise ratio in the short-time modulation spectral domain is constructed as

$$\xi(f,m)=\frac{|\mathcal{S}(f,m)|^2}{|\mathcal{D}(f,m)|^2},\quad 0\le m\le M, \tag{1}$$

where M denotes the highest modulation frequency. In the remainder of this paper, we will refer to $\xi(f,m)$ as the modulation SNR and denote it as $(\mathrm{S/N})_{\mathrm{mod}}$. It should be noted that $(\mathrm{S/N})_{\mathrm{mod}}$ was defined differently in Dubbelboer and Houtgast (2008) to account for the non-speech modulations originating from the interaction between speech and masker modulations. Our previous definition [Eq. (1)] implicitly discards the speech-masker interaction modulations and considers only the speech and masker modulations. Analysis done by Dubbelboer and Houtgast (2007) indicated that the effect of the speech-masker interaction modulations on speech intelligibility was not negligible, but it was the reduction in speech modulations that was found to be most detrimental to speech intelligibility. In the study by Jørgensen and Dau (2011), the speech-to-noise envelope power ratio, defined at the output of the modulation filter bank, did not explicitly include the interaction modulations, as those were assumed to have a negligible effect on speech intelligibility. Good agreement was obtained between the intelligibility data and the predictions of the model despite the absence of interaction modulations in the model. The reduction in speech modulations is reflected in the proposed measure by small values of $(\mathrm{S/N})_{\mathrm{mod}}$.

The relative strength of speech and noise modulations, as quantified by the $(\mathrm{S/N})_{\mathrm{mod}}$ defined in Eq. (1), is used as the channel selection criterion. More specifically, modulation channels are retained if their associated $(\mathrm{S/N})_{\mathrm{mod}}$ are sufficiently high (i.e., above a certain threshold) and are otherwise discarded (i.e., set to zero). The selection procedure is further refined by incorporating well-established findings from speech perception. Specifically, we take into account the fact that only a narrow band of modulation frequencies (2–16 Hz) contributes significantly to speech intelligibility (Houtgast and Steeneken, 1985; Drullman *et al.*, 1994a; Elliott and Theunissen, 2009). We thus select the relevant target modulation channels as follows:

$$\left|\hat{\mathcal{S}}(f,m)\right|=\begin{cases}|\chi(f,m)|, & \text{if } \xi(f,m)>\theta \text{ and } m<M_c\\ 0, & \text{otherwise}\end{cases}, \tag{2}$$

where $|\hat{\mathcal{S}}(f,m)|$ is the modified modulation magnitude-spectrum, $|\chi(f,m)|$ is the modulation magnitude-spectrum of noise-masked speech, θ is the modulation threshold, and M_c denotes the modulation cutoff frequency.

The above selection rule affords us the following benefits. First, the range of modulation frequencies over which estimation is to be performed in a practical implementation of the MCS algorithm is significantly reduced. This decreases the possibility of estimation errors that could potentially be detrimental to speech intelligibility, and also reduces the computational cost. Second, a properly selected modulation cutoff frequency achieves removal of non-speech modulations without degradation of speech modulations. With this approach, noise components in the modulation spectrum for $m\ge M_c$ will be eliminated regardless of the accuracy of the $(\mathrm{S/N})_{\mathrm{mod}}$ estimator. In other words, the use of modulation cutoff frequency aims at reliable removal of noise modulations above M_c, even if accurate estimates of $(\mathrm{S/N})_{\mathrm{mod}}$ in that spectral region are not available. This could be of particular interest in practical realizations of the MCS algorithm, where $(\mathrm{S/N})_{\mathrm{mod}}$ has to be estimated. The selection of modulation cutoff frequency satisfying the above considerations is investigated in the listening tests detailed in Sec. III.

C. Illustration of MCS concept using synthetic stimuli

The MCS approach described in Sec. II B can be illustrated using synthetic envelope stimuli. For this purpose, let us consider a synthetic envelope constructed as follows:

$$S_{t,f}=C+\sum_{i=1}^{N}A_i\cos(2\pi m_i t+\phi_i), \tag{3}$$

where f denotes the acoustic frequency, t is the time variable, C is a positive constant used to ensure that the envelope is non-negative, N is the number of sinusoidal components, and A_i, m_i, and ϕ_i denote the amplitude, (modulation) frequency, and phase of the ith sinusoidal component, respectively. Let us further assume an additive masker model given by

$$X_{t,f}=S_{t,f}+D_{t,f}, \tag{4}$$

where $D_{t,f}$ is the masker envelope and $X_{t,f}$ is the target + masker envelope. Note that the above model is simplistic as it assumes that there are no target-masker interaction

modulations. However, as mentioned earlier, we assume that their contribution to speech intelligibility is small (Jørgensen and Dau, 2011).

For simulation purposes, we consider envelopes at a fixed acoustic frequency f. Masker samples are drawn from a Rayleigh distribution and scaled to yield an overall envelope SNR of -5 dB. The target and target + masker envelopes are shown in Fig. 2(a). The target envelope was constructed using Eq. (3) with the following parameters: $N=4$, $A_i=[0.5, 1, 0.75, 0.5]$, $m_i=[2,4,6,8]$ Hz, $\phi_i=[\pi/8, 0, \pi/4, \pi/3]$, and $C=1.565$.

The corresponding modulation spectra of the target and target + masker envelopes are shown in Fig. 2(b). Note that these modulation spectra were normalized (for visualization purposes only) by the mean of the acoustic frequency band as per Houtgast and Steeneken (1985) to better convey modulation reduction. No modulation filterbank is applied to these spectra (as it is often done in the literature) since no such filterbank is used in MCS processing (uniform modulation bandwidth is assumed). The modulation spectra are depicted in high resolution[1] in this example, in order to effectively visualize the core idea in MCS. The modulation spectrum of the target + masker [Fig. 2(b)] shows reduced power suggesting reduction in modulation depth due to additive noise.

The target-dominant modulation channels are readily apparent from Fig. 2(b). Figure 2(c) shows the $(S/N)_{mod}$ and the modulation threshold θ. Note that the threshold is chosen such that high $(S/N)_{mod}$ regions (indicative of

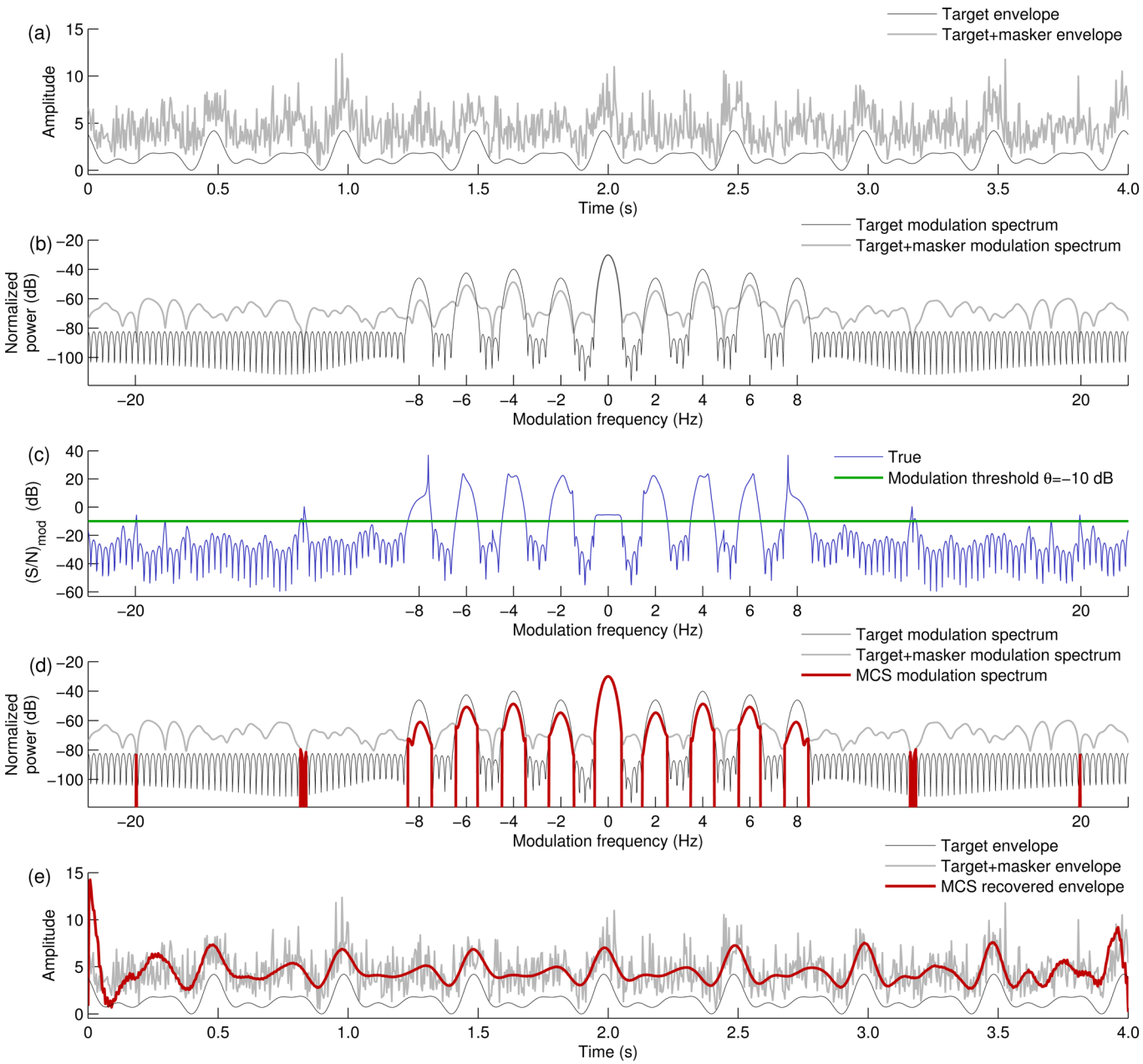

FIG. 2. (Color online) Illustration of modulation channel selection for a synthetic stimulus. (a) Acoustic envelopes: target (black line) and target + masker at -5 dB SNR (gray line); (b) modulation spectrum: target (black line) and target + masker (gray line); (c) $(S/N)_{mod}$ and modulation threshold θ (bold line); (d) modulation spectrum: target (black line), target + masker (gray line) and MCS thresholded (bold line); as well as (e) acoustic envelopes: target (black line), target + masker (gray line) and MCS recovered (bold line).

target-dominated modulations) fall above the threshold, while low $(S/N)_{mod}$ regions (indicative of predominantly non-target modulations) fall below the threshold. In the MCS approach, only modulation channels falling above the threshold are retained, while the remaining channels are set to zero.

The modulation spectra of the clean, noisy, and MCS-processed signals are shown in Fig. 2(d). As can be seen, the frequency location of the peaks in the MCS modulation spectrum matches those present in the target envelope. That is, the MCS processing captured (detected) the modulation frequencies present in the target envelope. The masker-dominated channels have been mostly removed. Finally, Fig. 2(e) compares the MCS recovered envelope against that of the target envelope and target + masker envelope. It can be clearly seen that MCS restores the temporal envelope of the target signal.

A few points should be noted about the example in Fig. 2. First, for this example, all modulation frequencies were considered in MCS processing, i.e., $M_c = M$ was used in Eq. (2). Due to the finite data record and the stochastic nature of the masker, as well as due to the choice of the modulation threshold, a few masker-dominated channels (e.g., at 13 and 20 Hz) were also selected [Fig. 2(d)]. Second, the boundary effects seen in the MCS recovered envelopes [Fig. 2(e)] do not present an issue in practice, since tapered analysis windows are used in the dual analysis-modification-synthesis procedure described in Sec. II A. Last, the zeroth (DC) frequency component of the modulation spectrum was preserved in this example and, hence, the target and MCS recovered envelopes shown in Fig. 2(e) are offset by a constant.

III. IMPACT OF MODULATION CHANNEL SELECTION ON SPEECH RECOGNITION IN NOISE

The objectives of the listening tests presented in this section are twofold. The first and primary goal is to determine the upper bound of performance, in terms of speech intelligibility, attainable by the MCS approach detailed in Sec. II. To achieve this, we consider an ideal scenario, where the $(S/N)_{mod}$ is assumed to be known. The above assumption is necessary in order to truly assess the full potential of the MCS method and its efficacy for future use in non-ideal scenarios. The secondary goal is to experimentally determine the lowest modulation cutoff frequency which can be used within the MCS framework without observing a significant reduction in speech intelligibility. To determine this, we systematically vary the modulation cutoff frequency and assess the intelligibility of the MCS processed stimuli.

A. Subjects

Ten normal-hearing subjects (five males and five females) participated in the listening tests. The subjects were recruited from the University of Texas at Dallas community. All were native speakers of American English, and were paid for their participation.

B. Speech materials

Sentence materials taken from the IEEE–Harvard corpus (IEEE Subcommittee, 1969) were used. The sentences, uttered by a male speaker, were recorded in a sound booth (Acoustic Systems, Inc., Austin, Texas). The original recordings were sampled at 25 kHz and are available from Loizou (2007). In our study, we consider speech corrupted by multi-talker babble, recorded in a canteen occupied by approximately one hundred speakers. This masker was taken from the NOISEX-92 noise database (Varga and Steeneken, 1993). The original masker recordings were sampled at 20 kHz. Both target and masker stimuli were downsampled to 16 kHz for the purpose of our experiments. The target utterances were mixed with the babble masker during the testing procedure at a desired SNR level. More specifically, the SNR was adjusted by keeping the level of the target fixed and varying the level of the masker.

C. Types of stimuli

Conditions included corrupted (unprocessed) stimuli and stimuli processed using the MCS dual analysis-modification-synthesis framework described in Sec. II A. For the MCS stimuli construction, speech was segmented into 32 ms duration frames using a Hanning window with 75% overlap between frames. The envelopes were thus sampled at 125 Hz ($M = 62.5$ Hz). For the second transform (i.e., for envelope processing), frames of $T = 256$ ms were used, corresponding to 32 frames of acoustic magnitude spectra. Note that the 256 ms modulation frame duration was selected in order to obtain sufficiently good resolution near 4 Hz in the modulation spectrum. Poorer resolution would be obtained for shorter durations, while longer durations would introduce more smearing in the acoustic spectrum. Hanning windowing and 75% overlap between segments was used for envelope processing. Frames in both acoustic and modulation domains were padded with zeros to double length prior to FFT computation, resulting in an acoustic spectrum composed of 1024 bins, and a modulation spectrum composed of 64 bins. The FFT bin spacing of the acoustic spectrum was 15.62 Hz, while the bin spacing of the modulation spectrum was 1.95 Hz. Hence, the bandwidth of each modulation channel was 1.95 Hz. Uniform modulation frequency spacing was used throughout this work. This stands in contrast to the $\frac{1}{3}$-octave bandwidth filterbanks often applied to the modulation spectrum (Houtgast and Steeneken, 1985). Uniform modulation filterbanks were used in the present study since these filterbanks facilitate easy synthesis of the processed stimuli via the dual-AMS framework (Fig. 1). A 75% frame overlap with a Hanning window was used for the overlap-and-add procedures. Two modulation thresholds were investigated, namely $\theta = -5$ dB and $\theta = -10$ dB. Also, nine different settings for the low-pass modulation cutoff frequency were considered: $M_c = 2$, 4, 6, 8, 10, 12, 14, 18, and 30 Hz.

For comparative purposes, stimuli processed using the ideal channel selection (ICS) algorithm operating in the acoustic, rather than the modulation, domain were also included. The ICS implementation was similar to that used by Li and Loizou (2008). The (acoustic) frame duration (32 ms with 75% frame overlap) was set the same as in the MCS implementation. Two local SNR thresholds, −5 and −10 dB, were considered.

In summary, there were 21 conditions consisting of: one noisy (unprocessed) condition, 18 MCS conditions (2 modulation thresholds × 9 modulation cutoff frequencies), and two ICS conditions (two local SNR thresholds).

D. Procedure

The stimuli were presented in a sound booth over closed circumaural headphones (Sennheiser HD428, Wennebostel, Germany). Each subject was familiarized with the task during a short practice session. During the practice session, the participants were allowed to adjust the stimulus level to a comfortable listening level. The adjusted level was then used throughout the listening test, which involved an adaptive method designed for measuring the SRT (Plomp, 1986). The SRT was measured using a simple up-down procedure that determines the SNR at which average sentence intelligibility reaches 50% correct. The order of the conditions was randomized across listeners, with each condition assigned a randomly selected sentence list from a pool of lists not previously selected for a given subject. During the test, the participants were asked to repeat the words they heard (if any). For each condition, sentence stimuli were presented starting at a very low SNR. The SNR was then progressively increased (in steps of 2 dB) until the listener was able to correctly reproduce more than half of the words. Subsequent trials employed the following adaptive up-down method (Levitt, 1971) for SRT computation. In the first trial, the SNR was decreased by 2 dB. Then, depending on whether or not the listener recognized more than half of the words correctly, the SNR was either decreased or increased by 2 dB, respectively. Word recognition was assessed by the experimenter over ten trials. The result of the tenth trial was used to determine the SNR level for the eleventh trial (the eleventh trial was not actually conducted). The SRT score for a given condition was computed as an average over the SNRs from the last eight trials (i.e., trials 4 through to 11). No feedback was given to the listeners. Each listener completed the testing in less than two hours, including breaks.

E. Results

Mean SRT values as a function of low-pass modulation cutoff frequency are shown in Fig. 3. The results are given for two modulation thresholds, namely $\theta = -5$ dB and $\theta = -10$ dB. Mean SRT values for unprocessed (control) stimuli are also included for comparison. Two-way analysis of variance with modulation thresholds and cutoff frequencies as within-subject factors revealed a significant effect [$F(1,9) = 9.1$, $p = 0.014$] of modulation threshold θ, a significant effect [$F(8,72) = 80.3$, $p < 0.01$] of modulation cutoff frequency M_c, and a non-significant interaction [$F(8,72) = 2.0$, $p = 0.054$] between modulation threshold and modulation cutoff frequency.

Post hoc tests (Tukey HSD) showed that the SRT values obtained with $M_c \geq 8$ Hz (and threshold $\theta = -10$ dB) did not differ significantly. That is, the SRT value obtained with $M_c = 8$ Hz was not found to be significantly different ($p = 0.22$) from the SRT score obtained with $M_c = 30$ Hz. With the threshold set to $\theta = -5$ dB, the SRT values obtained with $M_c \geq 10$ Hz did not differ significantly. Overall, better performance was obtained when a lower modulation threshold ($\theta = -10$ dB) was used, particularly in low modulation cutoff frequencies, $M_c < 6$ Hz.

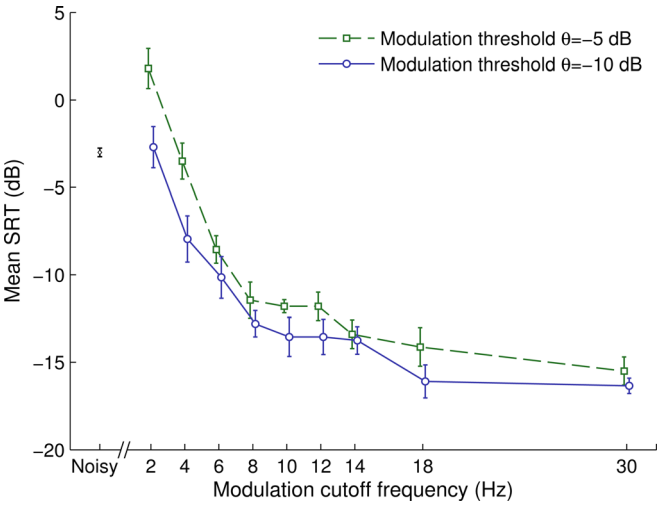

FIG. 3. (Color online) Mean SRT values as a function of modulation cutoff frequency M_c, for modulation thresholds of $\theta = -5$ dB and $\theta = -10$ dB. Error bars indicate standard errors of the mean.

SRT performance of MCS processed stimuli ($\theta = -10$ dB) improved significantly ($p = 0.015$) from -3 dB (unprocessed) to -7.9 dB when the modulation cutoff frequency was set to 4 Hz. Hence, preserving the low-frequency target-dominant modulations seems to be sufficient in terms of observing significant benefits in intelligibility with MCS processing. Substantially larger improvements in intelligibility were noted when the modulation cutoff frequency was set higher than 4 Hz.

The corresponding acoustic domain ICS processed stimuli achieved mean SRTs of -26.95 dB (s.e. = 0.72 dB) and -31.25 dB (s.e. = 0.63 dB) for local SNR criteria of -5 and -10 dB, respectively, where s.e. denotes standard error of the mean. These scores are comparable to those reported in other studies (e.g., Kjems *et al.*, 2009). In contrast, the lowest (best) score obtained with the MCS processed stimuli was -17 dB. Nonetheless, the overall improvement in intelligibility that can be obtained via MCS processing relative to that of the unprocessed stimuli is quite substantial and amounts to 13 dB.

IV. DISCUSSION

The outcomes of the present study indicate that the $(S/N)_{mod}$ is an effective criterion that can be used for improving speech intelligibility in noise. More specifically, the $(S/N)_{mod}$ can be used as a criterion for discerning between target-dominated modulations and masker-dominated modulations. Retaining the target-dominated modulations within a narrow band of modulation frequencies (0–8 Hz), while discarding the masker-dominant modulations was found to significantly improve speech intelligibility in noise (see Fig. 3). The finding regarding the importance of preserving low-frequency envelope modulations is consistent with previous studies of modulation spectrum filtering (e.g., Drullman *et al.*, 1994a,b; Arai *et al.*,

1996). In the study by Drullman *et al.* (1994a), for instance, the SRT value obtained in speech-shaped noise with modulation cutoff frequency of 16 Hz was not found to be significantly different from the SRT value of the control (unprocessed) stimuli.

Unlike previous studies that assessed speech intelligibility in situations where both masker- and target-dominated modulations were present in the low frequencies (e.g., Drullman *et al.*, 1994a), the present study aimed to isolate the contributions of masker and target modulations. The data from the present study suggest that by discarding the masker-dominated modulations, we can design a signal processing algorithm that can improve speech intelligibility in noise. Such an algorithm might offer advantages over existing noise-reduction algorithms, which generally offer no benefit in terms of intelligibility (e.g., Hu and Loizou, 2007a; Bentler *et al.*, 2008).

This point can be illustrated by considering the spectrograms shown in Fig. 4, along with their corresponding envelopes (at acoustic frequency $f = 500$ Hz) shown in Fig. 5. Plots of the target and masker corrupted stimuli (for babble masker at -5 dB SNR) are shown in panels (a) and (b), respectively, while plots of the MCS-processed stimuli for modulation cutoff frequencies of 14 Hz and 2 Hz are shown in panels (c) and (d), respectively. As shown in Fig. 4(b), aspects of the spectrum conveying important phonetic cues (e.g., formants, harmonics, etc.) are completely masked by babble at low SNR levels. This presents a considerable challenge for conventional noise-reduction algorithms in terms of being able to recover the heavily masked target from the mixture. In contrast, by considering the envelope trajectories, it is relatively easier to identify and track the target-dominant modulations. As shown in Fig. 5(b), while the modulation depth is greatly reduced due to additive noise, the target modulations are readily apparent (see, for example, envelope segments around $t = 0.24$ s and $t = 2.33$ s in Fig. 5). By retaining modulation components within the range of 0–14 Hz and with $(S/N)_{mod} > -10$ dB, we can recover the target envelope [see Fig. 5(c)]. The corresponding spectrogram of the MCS processed stimulus is shown in Fig. 4(c). The formants are recovered to some extent (e.g., see formants F1/F2/F3 at $t = 2.33$ s) and the vowel/consonant boundaries are more evident. A small degree of temporal smearing is also present. In contrast, use of the more aggressive modulation filtering, i.e., with modulation cutoff frequency set to 2 Hz, results in stronger temporal smoothing and significant modulation depth reduction. This is demonstrated by the spectrogram of Fig. 4(d) and MCS recovered envelope shown in Fig. 5(d).

In this work, a dual analysis-modification-synthesis framework was used to compute the modulation spectra. These spectra were computed every 64 ms based on 256 ms long segments. This makes the proposed MCS algorithm potentially amenable to real-time implementation, subject to acceptable latency and computational complexity constraints.

While in this study, access to true values of $(S/N)_{mod}$ was assumed, in practice these values need to be estimated from the mixture envelopes. In order to demonstrate the

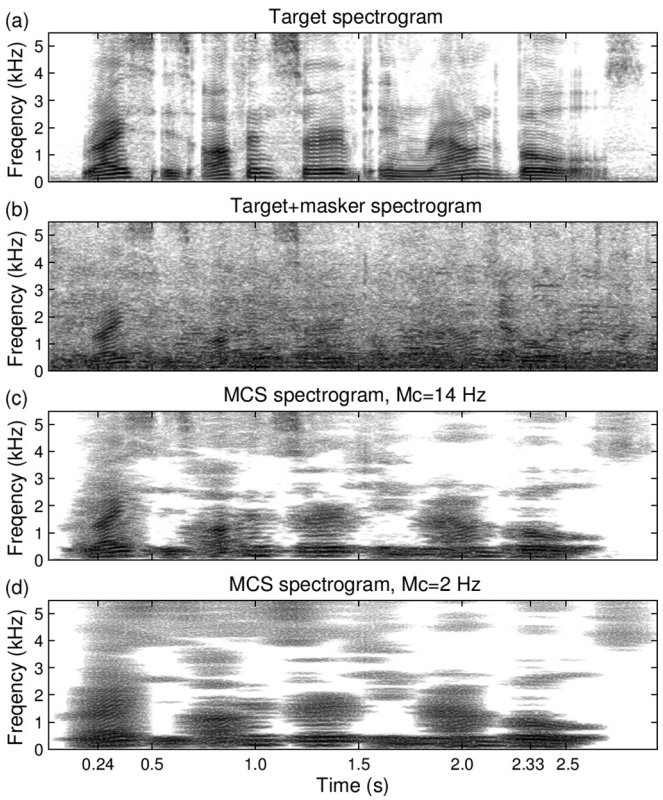

FIG. 4. Wideband spectrograms of the (a) target stimulus; (b) target + masker stimulus (babble masker at -5 dB SNR); (c) MCS−processed stimulus: $\theta = -10$ dB, $M_c = 14$ Hz; and (d) MCS-processed stimulus: $\theta = -10$ dB, $M_c = 2$ Hz.

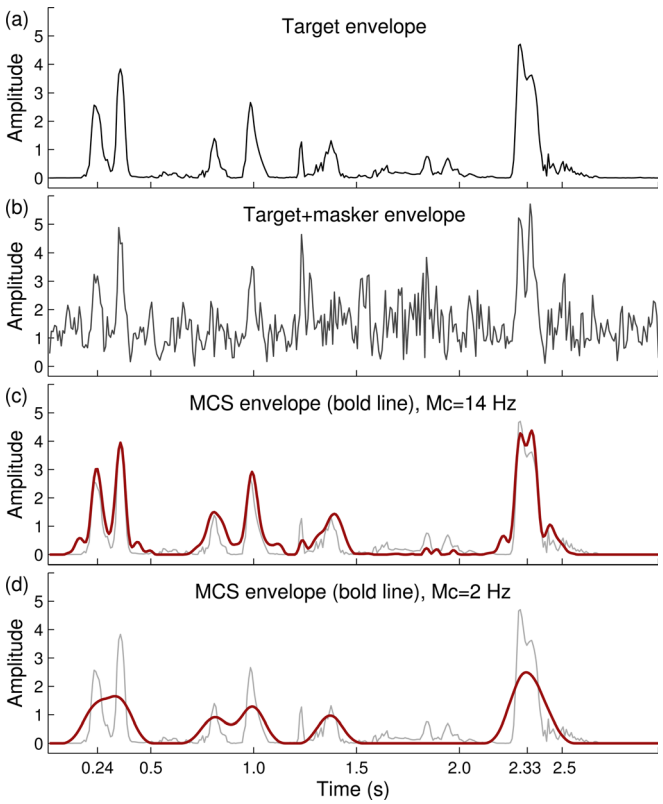

FIG. 5. (Color online) Stimuli envelopes at acoustic frequency $f = 500$ Hz. (a) Target envelope; (b) target + masker envelope (babble masker at -5 dB SNR); (c) MCS-derived envelope (bold line): $\theta = -10$ dB, $M_c = 14$ Hz; and (d) MCS-derived envelope (bold line): $\theta = -10$ dB, $M_c = 2$ Hz.

feasibility of this task, some preliminary experiments were conducted. Specifically, we considered estimating $(S/N)_{mod}$ from the mixture envelopes as

$$\hat{\xi}(f,m) = \frac{|\mathcal{Y}(f,m)|^2}{\left|\hat{\mathcal{D}}(f,m)\right|^2}, \quad 0 \le m \le M, \tag{5}$$

where $|\mathcal{Y}(f,m)|$ is an estimate of clean modulation spectrum, computed using the spectral subtraction method applied in the modulation domain (Paliwal *et al.*, 2010), and $\left|\hat{\mathcal{D}}(f,m)\right|$ is an estimate of modulation spectrum of noise, computed from the leading silent (speech-absent) portion of the noise masked stimulus. Note that the equation of the above estimate of $(S/N)_{mod}$, i.e., $\hat{\xi}(f,m)$, is similar to that used by Jørgensen and Dau (2011) for computing the envelope SNR (SNR_{env}) in the modulation domain. The main difference is that in the above equation the modulation spectrum of the masker (i.e., $\left|\hat{\mathcal{D}}(f,m)\right|$) is estimated from the mixture envelopes, whereas in the study by Jørgensen and Dau (2011) the modulation spectrum of the masker was assumed to be available prior to mixing.

We should point out that the MCS approach presented in the present study differs in the following ways from the work reported by Paliwal *et al.* (2010). In the MCS approach, a binary decision is made as to whether a given channel in the modulation spectrum is target-dominated or masker-dominated. This binary decision is then used to either retain or discard the energy present in that modulation channel. That is, the MCS approach aims to keep the target-dominated channels unaltered (undisturbed), while removing the masker-dominated channels. In contrast, in the modulation spectral subtraction approach (Paliwal *et al.*, 2010), estimates of the masker modulation spectrum are subtracted from all components of the modulation spectrum, regardless of whether they are target- or masker-dominated. That is, the subtraction

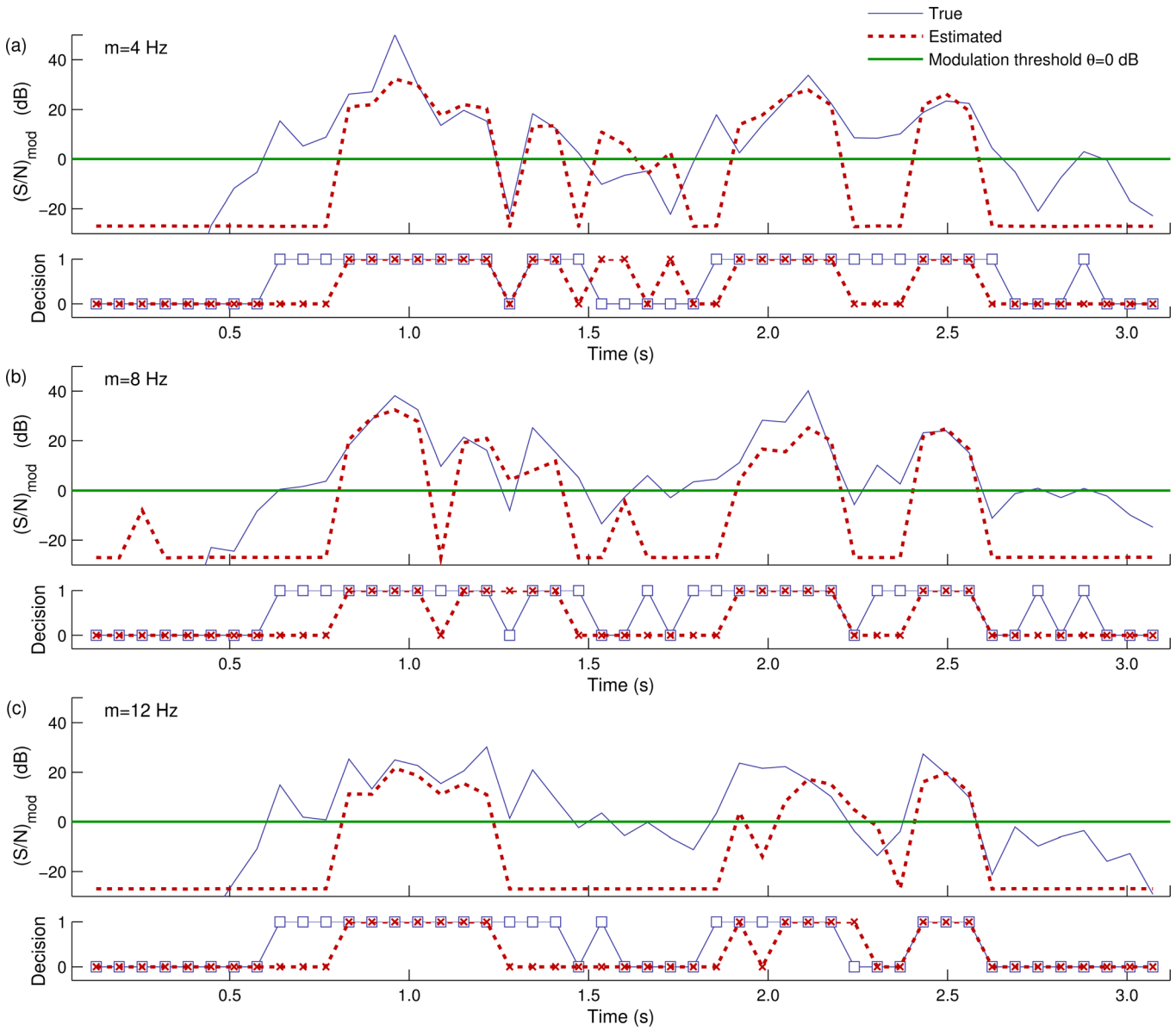

FIG. 6. (Color online) Plots of the true $(S/N)_{mod}$ values (thin lines) and estimated [as per Eq. (5)] $(S/N)_{mod}$ values (dotted lines) at acoustic frequency $f = 1500$ Hz and for the following modulation cutoff frequencies: (a) $m = 4$ Hz, (b) $m = 8$ Hz, and (c) $m = 12$ Hz. The modulation threshold $\theta = 0$ dB is also shown (bold line). Smaller sub-panels show the binary decisions made by comparing the $(S/N)_{mod}$ values against the threshold $\theta = 0$ dB.

operation alters both target-dominated and masker-dominated components.

Examples of $(S/N)_{mod}$ estimates, computed using Eq. (5) for acoustic frequency $f = 1500$ Hz and for modulation frequencies of 4, 8, and 12 Hz, are shown in panels (a), (b) and (c) of Fig. 6, respectively. The true (ideal) and estimated modulation-channel (binary) decisions (i.e., decisions made regarding as to whether a specific modulation channel is selected or discarded) are also shown in the bottom panels. An IEEE sentence with 500 ms of leading silence was used in this example, and speech was corrupted by a steady masker (speech-shaped noise) at 5 dB SNR. As can be seen from Fig. 6, the $(S/N)_{mod}$ estimates follow—for the most part—the true $(S/N)_{mod}$ values. The $(S/N)_{mod}$ estimates are typically more accurate at the high $(S/N)_{mod}$ regions than at the low $(S/N)_{mod}$ regions. It should be noted, however, that in practice the $(S/N)_{mod}$ estimates need not be very accurate as long as they fall in the right region (either smaller or larger than the prescribed threshold). In other words, the modulation-selection algorithm will be effective as long as the channel selection decisions remain consistent with the ideal (true) decisions (shown as squares in Fig. 6).

To objectively evaluate the above approach in terms of accuracy of binary decisions, hit rate (HIT), false alarm rate (FA), HIT-FA, and percent agreement (PA) were computed using the first 10 sentences of the NOIZEUS corpus (Hu and Loizou, 2007b).[2] The hit and false alarm rates were calculated by comparing the estimated decisions against the true decisions (made assuming access to the signals prior to mixing). More specifically, the hit rate was computed as the probability of a correct decision for target-dominated channels, while the false alarm rate was computed as the probability of an incorrect decision for masker-dominated channels (Hu and Loizou, 2008). The hit and false alarm rates were then used to compute the HIT–FA metric, which in the recent work of Kim *et al.* (2009), operating in the acoustic spectral domain, has been shown to correlate modestly high with speech intelligibility. Finally, the PA measure was calculated as the probability of making correct decisions irrespective of the error type, i.e., for all target-dominated and masker-dominated channels. The results of the objective evaluation are shown in Table I. Overall, the percent agreement was high (>84%) and the HIT and FA rates compared favorably to those obtained with other noise reduction algorithms operating in the acoustic spectrum domain (Hu and Loizou, 2008). While this preliminary work and the above results demonstrate potential feasibility of $(S/N)_{mod}$ estimation for practical applications, further—more exhaustive— research in this direction is warranted, particularly for tackling non-stationary maskers.

TABLE I. Objective evaluation of non-ideal MCS processing in terms of hit rate (HIT), false alarm rate (FA), HIT-FA and percentage agreement (PA) for different modulation cutoff frequencies, M_c.

M_c (Hz)	HIT (%)	FA (%)	HIT-FA (%)	PA (%)
30	50.79	4.99	45.80	85.13
62.5	46.90	6.49	40.41	84.36

V. CONCLUSIONS

Motivated by psychoacoustic evidence of frequency selectivity in the modulation domain (e.g., Bacon and Grantham, 1989; Ewert and Dau, 2000), this paper introduced the concept of channel selection in the modulation spectral domain as a potential means of improving speech intelligibility. The proposed approach allows for selective retention or removal of modulation channels from mixture (target + masker) envelopes over short intervals (256 ms). Specifically, target-dominated modulations which are important for speech intelligibility were identified [based on the $(S/N)_{mod}$ criterion] and retained, while masker-dominated modulations which are potentially detrimental to speech perception were discarded. The selection of modulation channels was based on the $(S/N)_{mod}$ criterion over a narrow range of modulation frequencies relevant for speech intelligibility. Our study considered an ideal scenario where $(S/N)_{mod}$ was assumed to be known. This allowed us to determine the upper bound of performance that can be attained via MCS processing and its potential for future implementation in noise reduction algorithms. The main conclusions of the present study are summarized below:

(1) The criterion $(S/N)_{mod}$, which quantifies the relative strength of speech and noise modulations, is an effective selection criterion appropriate for modulation-domain processing.
(2) Modulation channel selection based on $(S/N)_{mod}$ over a narrow range of modulation frequencies is an effective approach for improving speech intelligibility in noise. The proposed approach can yield large improvements in intelligibility—up to 13 dB improvement in SNR (see Fig. 3).
(3) Modulation channel selection can be performed effectively (i.e., without significant degradation of speech intelligibility) over a narrow range of modulation frequencies (0–10 Hz). Specifically, it was observed that removal of modulation components above 10 Hz does not significantly reduce intelligibility of MCS-processed speech.
(4) Results of preliminary experiments (see Fig. 6) suggest that estimation of $(S/N)_{mod}$, for MCS application in non-ideal scenarios, is feasible—at least for stationary type maskers.

ACKNOWLEDGMENTS

This research was supported by Grant No. R01 DC010494 from the National Institute of Deafness and Other Communication Disorders (NIDCD), National Institutes of Health (NIH). The authors would like to thank the two anonymous reviewers for their helpful comments.

[1]This was achieved through the use of a long data record, 4 s in duration. The envelope was sampled at 256 Hz. Prior to the FFT computation, the synthetic stimuli were padded with zeros resulting in FFT bin spacing of 0.0156 Hz in the modulation spectral domain.

[2]The original NOIZEUS recordings (sampled at 25 kHz) were down-sampled to 16 kHz to match the (acoustic) sampling frequency used throughout this work.

Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (**1996**). "Intelligibility of speech with filtered time trajectories of spectral envelopes", in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, October 3–6, Philadelphia, PA, pp. 2490–2493.

Atlas, L., and Shamma, S. A. (**2003**). "Joint acoustic and modulation frequency," EURASIP J. Appl. Signal Process. **2003**, 668–675.
Bacon, S. P., and Grantham, D. W. (**1989**). "Modulation masking: Effects of modulation frequency, depth, and phase," J. Acoust. Soc. Am. **85**, 2575–2580.
Bentler, R., Wu, Y., Kettel, J., and Hurtig, R. (**2008**). "Digital noise reduction: Outcomes from laboratory and field studies," Int. J. Audiology **47**, 447–460.
Depireux, D. A., Simon, J. Z., Klein, D. J., and Shamma, S. A. (**2001**). "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," J. Neurophysiol. **85**, 1220–1234.
Drullman, R., Festen, J. M., and Plomp, R. (**1994a**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.
Drullman, R., Festen, J. M., and Plomp, R. (**1994b**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670–2680.
Dubbelboer, F., and Houtgast, T. (**2007**). "A detailed study on the effects of noise on speech intelligibility," J. Acoust. Soc. Am. **122**, 2865–2871.
Dubbelboer, F., and Houtgast, T. (**2008**). "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," J. Acoust. Soc. Am. **124**, 3937–3946.
Elliott, T., and Theunissen, F. (**2009**). "The modulation transfer function for speech intelligibility," PLos Comput. Biol. **5**, 1–14.
Ewert, S. D., and Dau, T. (**2000**). "Characterizing frequency selectivity for envelope fluctuations," J. Acoust. Soc. Am. **108**, 1181–1196.
Griffin, D., and Lim, J. (**1984**). "Signal estimation from modified short-time Fourier transform," IEEE Trans. Acoust., Speech, Signal Process. **32**, 236–243.
Houtgast, T., and Steeneken, H. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica **28**, 66–73.
Houtgast, T., and Steeneken, H. J. M. (**1985**). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069–1077.
Hu, Y., and Loizou, P. C. (**2008**). "Techniques for estimating the ideal binary mask," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, September 14–17, 2008, Seattle, WA.
Hu, Y., and Loizou, P. C. (**2007a**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.
Hu, Y., and Loizou, P. C. (**2007b**). "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun. **49**, 588–601.
IEEE Subcommittee (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust., September 14–17, 2008, Seattle, WA, **AU-17**, 225–246.
Jørgensen, S., and Dau, T. (**2011**). "Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing," J. Acoust. Soc. Am. **130**, 1475–1487.
Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.
Kjems, U., Boldt, J. B., Pedersen, M. S., Lunner, T., and Wang, D. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.
Kowalski, N., Depireux, D., and Shamma, S. (**1996**). "Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses to moving ripple spectra," J. Neurophysiol. **76**, 3503–3523.
Levitt, H. (**1971**). "Transformed up-down methods in psychoacoustics," J. Acoust. Soc. Am. **49**, 467–477.
Li, N., and Loizou, P. C. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.
Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (Taylor and Francis, Boca Raton, FL), pp. 589–599.
Ludvigsen, C. (**1993**). "Evaluation of a noise reduction method-comparison between observed scores and scores predicted from STI," Scand. Audiol. Suppl. **38**, 50–55.
Mesgarani, N., and Shamma, S. (**2005**). "Speech enhancement based on filtering the spectrotemporal modulations," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, March 18–23, 2005, Philadelphia, PA, Vol. 1, 1105–1108.
Paliwal, K., Schwerin, B., and Wójcicki, K. (**2011**). "Role of modulation magnitude and phase spectrum towards speech intelligibility," Speech Commun. **53**, 327–339.
Paliwal, K., Wójcicki, K., and Schwerin, B. (**2010**). "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," Speech Commun. **52**, 450–475.
Plomp, R. (**1986**). "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," J. Speech Hear. Res. **29**, 146–154.
Schafer, R., and Rabiner, L. (**1973**). "Design and simulation of a speech analysis-synthesis system based on short-time Fourier analysis," IEEE Trans. Audio Electroacoust. **21**, 165–174.
Schönwiesner, M., and Zatorre, R. J. (**2009**). "Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI," Proc. Natl. Acad. Sci. U.S.A. **106**, 14611–14616.
Schreiner, C. E., and Urbas, J. V. (**1986**). "Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF)." Hear. Res. **21**, 227–241.
Shamma, S. (**1996**). "Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method," Network Comput. Neural Syst. **7**, 439–476.
Sheft, S., and Yost, W. A. (**1990**). "Temporal integration in amplitude modulation detection," J. Acoust. Soc. Am. **88**, 796–805.
Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Varga, A., and Steeneken, H. (**1993**). "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun. **12**, 247–251.