Evaluation of the importance of time-frequency contributions to speech intelligibility in noise

Chengzhu Yu,[a] Kamil K. Wójcicki, Philipos C. Loizou, and John H. L. Hansen
Department of Electrical Engineering, Erik Jonsson School of Enigneering and Computer Science, University of Texas at Dallas, Richardson, Texas 75083

Michael T. Johnson
Speech and Signal Processing Laboratory, Marquette University, 1515 West Wisconsin Avenue, Milwaukee, Wisconsin 53201-1881

(Received 13 February 2013; revised 27 December 2013; accepted 7 March 2014)

Recent studies on binary masking techniques make the assumption that each time-frequency (T-F) unit contributes an equal amount to the overall intelligibility of speech. The present study demonstrated that the importance of each T-F unit to speech intelligibility varies in accordance with speech content. Specifically, T-F units are categorized into two classes, speech-present T-F units and speech-absent T-F units. Results indicate that the importance of each speech-present T-F unit to speech intelligibility is highly related to the loudness of its target component, while the importance of each speech-absent T-F unit varies according to the loudness of its masker component. Two types of mask errors are also considered, which include miss and false alarm errors. Consistent with previous work, false alarm errors are shown to be more harmful to speech intelligibility than miss errors when the mixture signal-to-noise ratio (SNR) is below 0 dB. However, the relative importance between the two types of error is conditioned on the SNR level of the input speech signal. Based on these observations, a mask-based objective measure, the loudness weighted hit-false, is proposed for predicting speech intelligibility. The proposed objective measure shows significantly higher correlation with intelligibility compared to two existing mask-based objective measures. [http://dx.doi.org/10.1121/1.4869088]

PACS number(s): 43.71.Es, 43.71.Gv [AA] Pages: 3007–3016

I. INTRODUCTION

Understanding speech in the presence of diverse background noise is one of the most challenging tasks for listeners with hearing loss. Although development of speech enhancement techniques has produced large improvement in the quality of speech, the success in improving intelligibility has been limited (Loizou, 2007; Hu and Loizou, 2007; Loizou and Kim, 2011). Advancements in speech enhancement employing auditory masking constraints have also shown promise for improving speech quality and speech technology in noise (Nandkumar and Hansen, 1995; Hansen and Nandkumar, 1995). However, recent studies have shown that masking based on prediction of an ideal binary mask has the potential for restoring the intelligibility of speech corrupted by competing noise both for normal hearing and hearing impaired persons (Brungart *et al.*, 2006; Anzalone *et al.*, 2006).

The concept of an ideal time-frequency (T-F) binary mask was proposed by Wang (2005) as a goal for performing computational auditory scene analysis (CASA) (Bregman, 1990; Wang and Brown, 2006). Binary masking is a strategy for applying binary gains on a T-F representation. In Wang (2005), the ideal binary mask is defined by comparing the local signal-to-noise ratio (SNR) of each T-F unit against a fixed threshold (IBM-SNR). T-F units with local SNR higher than the threshold are defined as target-dominated T-F units and are retained, while others are referred to as masker-dominated T-F units and are discarded.

A number of studies have shown that application of this defined ideal binary mask approach has the potential for restoring most of the intelligibility distortion caused by background noise (Brungart *et al.*, 2006; Roman and Wang, 2006; Li and Loizou, 2008). In Brungart *et al.* (2006), an optimal SNR threshold (from −12 to 0 dB) was reported for the task of improving speech intelligibility. In Li and Loizou (2008), a wider range of optimal SNR thresholds (from −20 to 5 dB) were observed for the same task. The study observed a difference in the range of optimal thresholds, and attributed this to differences in the speech material used for experiments. A study by Kjems *et al.* (2009) showed that the optimal threshold for an ideal binary mask varies according to the SNR level of the mixture signal. This work showed that application of an ideal binary mask could bring significant improvement to mixture signals with extremely low SNRs (e.g., −60 dB) as long as the threshold is chosen to be correspondingly lower (e.g., −65 dB).

Anzalone *et al.* (2006) proposed a new definition of ideal binary mask based on the speech presence status of each T-F unit (IBM-SP). The derived ideal binary mask was also known as target binary mask (TBM) (Kjems *et al.*, 2009) as its computation relies only on the target signal. T-F units were categorized into two classes: speech-present T-F units and speech-absent T-F units, according to the status of speech activity. The status of speech activity of each T-F unit was detected by comparing its target energy to a floor

[a]Author to whom correspondence should be addressed. Electronic address: chengzhu.yu@utdallas.edu

value. The floor value of each frequency band is chosen to include a fixed percentage (e.g., 95%) of the target energy, leading to an invariable threshold with respect to SNR. Application of IBM-SP removes the portion of the signal energy localized in the speech-absent T-F units, while retaining those in speech-present T-F units. Figure 1 shows an example of applying IBM-SP on mixture signal. Despite the difference compared to the ideal binary mask defined based on SNR threshold (IBM-SNR), application of IBM-SP also indicated substantial improvement to speech intelligibility both for normal hearing and hearing impaired listeners (Anzalone *et al.*, 2006).

Several single-channel techniques have been successfully proposed to estimate the ideal mask without prior knowledge of the target signal (Kim *et al.*, 2009; Han and Wang, 2011; Seltzer *et al.*, 2004; Wang *et al.*, 2013; Kim and Hansen, 2011). In those techniques, estimation of the ideal mask has been treated as a binary classification problem, which was achieved by advanced machine learning methods. Correspondingly, several mask-based objective measures, such as hit minus false alarm rate (HIT-FA) (Kim *et al.*, 2009) and ideal binary mask ratio (IBMR) (Hummersone *et al.*, 2011) have also been developed to predict the intelligibility of binary masked speech. Mask-based objective intelligibility measures are obtained by tabulating the mismatched T-F units between the estimated binary mask and IBM. Such objective measures have two major advantages: (i) First, and perhaps the most important advantage, is that the calculation of mask-based objective measures do not require synthesized output, and is thus robust to many convolutional distortions not directly associated with the binary masking algorithm itself (Hummersone *et al.*, 2011); and (ii) second, they allow for evaluation of binary masking techniques in contrast to existing objective intelligibility measures (Goldsworthy and Greenberg, 2004; Kates and Arehart, 2005; Ma *et al.*, 2009) where binary T-F weighting effect has not been considered.

Previous studies on binary masking algorithms as well as objective measures for predicting the intelligibility of binary masked speech have often placed a constant weight on all T-F units. However, a number of studies have shown that each T-F unit has a different perceptual effect depending on its intensity (Zhang *et al.*, 2001; Yu *et al.*, 2013) as well as different acoustic correlates (Li and Allen, 2009, 2011). To the best of our knowledge, no studies have yet assessed the relative contribution of individual T-F units to speech intelligibility in the context of speech separation in noise. Since accurate estimation of ideal binary gains for all T-F units is unattainable, it is of interest to see if certain T-F units are more important to speech intelligibility than others, and should therefore be further emphasized for an overall measure of intelligibility. The study of Anzalone *et al.* (2006) on IBM-SP demonstrated that speech-present T-F units have differential contributions toward speech intelligibility in noise compared to speech-absent T-F units. It could be further expected that the positive contribution of speech-present T-F units comes from characteristics of the underlying target component, while the negative contribution of speech-absent T-F units is caused by the characteristics of the masker component.

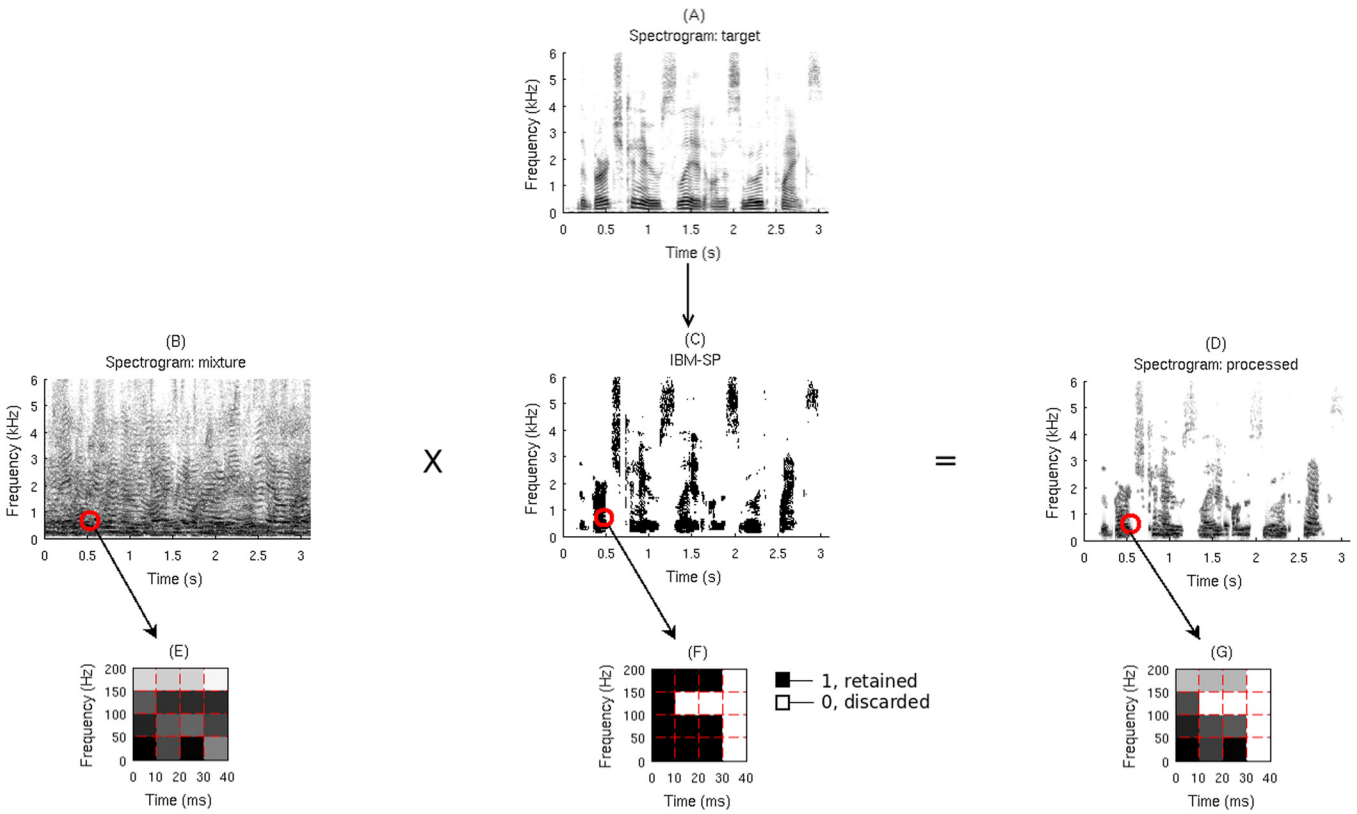

FIG. 1. (Color online) Illustration of applying IBM-SP on mixture signal. (A) Spectrogram of target speech utterance. (B) Spectrogram of mixture signal corrupted at −5 dB SNR with babble noise. (C) IBM-SP derived from target speech utterance, where 1 is indicated by black and 0 by white. (D) Spectrogram of resynthesized speech after applying IBM-SP on mixture signal. (E), (F), (G) T-F representations of 16 T-F units taken from mixture signal, IBM-SP, and processed signal.

In the present study we hypothesize that the positive contribution of each speech-present T-F unit is related to the degree of loudness of its target component, while the negative contribution of each speech-absent T-F unit is varied according to the loudness of its masker component. In other words, speech-present T-F units with a louder target signal are expected to contribute more toward speech intelligibility than those with quieter target components. Similarly, speech-absent T-F units with louder masker signals are expected to degrade speech intelligibility more than those with quieter maskers. The above hypotheses will be assessed in Experiment 1. In Experiment 2, we will evaluate the importance of two different types of mask errors, miss and false alarm errors (Li and Loizou, 2008), on speech intelligibility. Miss errors occur when T-F units originally marked as ones in ideal binary mask are flipped to zeros, while false alarm errors occur when T-F units originally marked as zeros in ideal binary mask are flipped to ones. A previous study by Li and Loizou (2008) showed that false alarm errors are more harmful to speech intelligibility than miss errors for a mixture SNR of −5 dB. In the present work, we extend that study by varying the mixture SNR level to determine if the relative importance between the two mask errors varies according to the mixture SNR level. Finally, based on the results from Experiments 1 and 2, we will propose an objective intelligibility metric, the loudness weighted hit-false (LWHF) score, as an improvement over conventional mask-based objective measures.

II. EXPERIMENT 1: IMPORTANCE OF T-F UNITS ACCORDING TO THE LOUDNESS OF TARGET OR MASKER CONTENT

A. Methods

1. Subjects

Eight normal-hearing listeners participated in this experiment, and subjects were paid for their participation. Listeners were all native speakers of American English and were undergraduate students from the University of Texas at Dallas. Subjects age ranged from 18 to 30 yr with five of them being males and others are females.

2. Stimuli

Speech sentences were taken from the IEEE database (1969) (IEEE, 1969). The sentences were produced by a male speaker in a soundproof booth at a sampling rate of 25 kHz then down sampled to 12 kHz. Details of the recording process and setup can be found in Loizou (2007). A multi-talker babble noise source from AUDITEC CD (St. Louis, MO) was used as the masker to corrupt the sentences at −5 dB SNR. The babble noise was produced by recording 20 young adults reading different passages simultaneously.

3. Signal processing

Signals (target, masker, and mixture) were first segmented in time using a Hamming window of 20 ms duration with 50% overlap between frames. A fast Fourier transform (FFT) was then applied to each frame, followed by magnitude-squared power spectrum computation. The derived T-F representations were composed of T-F units having equal area, the length of which is 20 ms along time and the width is 50 Hz along frequencies. The T-F analyzed signals were pre-emphasized by an equal-loudness curve, to simulate the perceptual sensitivity of the human ear to the intensity of sound at different frequency locations (Hermansky, 1990). The correlation between intensity and the perceptual loudness of sound was then modeled using a power law compression (Fastl and Zwicker, 2007).

IBM-SP is computed similar as described in Anzalone *et al.* (2006). In order to detect speech activity in a given T-F unit, the local energy (magnitude-squared power spectrum of T-F unit) of the target signal at the given T-F unit was compared to a floor value. For each sentence, this floor level was chosen separately within each frequency band to retain 95% of the total target loudness of that individual frequency band. Speech-present T-F units were assigned a value of 1, while speech-absent T-F units were assigned a value of 0.

The speech-present T-F units were categorized into four groups L1, L2, L3, and L4, according to increasing target loudness. L1 consisted of speech-present T-F units having target loudness in the lowest level, while L4 consisted of speech-present T-F units having target loudness in the highest level. Each group was chosen to include 40% of speech-present T-F units, so that there exists a 20% overlap between T-F units belonging to adjacent groups (i.e., some T-F units are shared between adjacent groups). Similarly, speech-absent T-F units were also categorized into four groups T1, T2, T3, and T4, but according to the masker loudness rather than the target loudness.

To compare the importance of speech-present T-F units belonging to different loudness groups, a new binary mask was calculated using the IBM-SP approach to introduce mask errors on all speech-present T-F units of a given loudness group, thereby changing all 1's (speech present) in that group to 0's (speech not present). We repeated this process separately for each of the four groups L1, L2, L3, and L4, to create four new binary masks. No errors were introduced to the speech-absent T-F units. Since each group contains the same number of T-F units, these binary masks have the same error rates, but with errors localized on speech-present T-F units with varying target loudness.

Similarly, the importance of speech-absent T-F units belonging to different loudness groups, T1, T2, T3, and T4, was also compared by constructing new binary masks based on the IBM-SP, by changing all 0's in that group to 1's. As before, we repeated this process for each of the four speech-absent groups, and derived four new binary masks with no errors introduced to the speech-present T-F units. The new binary masks have the same error rates, but with errors localized on speech-absent T-F units with varying degree of masker loudness.

The new derived binary masks were applied to mixture signals to produce stimuli. Figure 2 shows an example of the resulting stimuli spectrograms.

4. Procedure

The listening experiments were conducted in a soundproof booth (Acoustic Systems, Inc.) with a personal computer (PC) connected to a Tucker-Davis system. Stimuli were

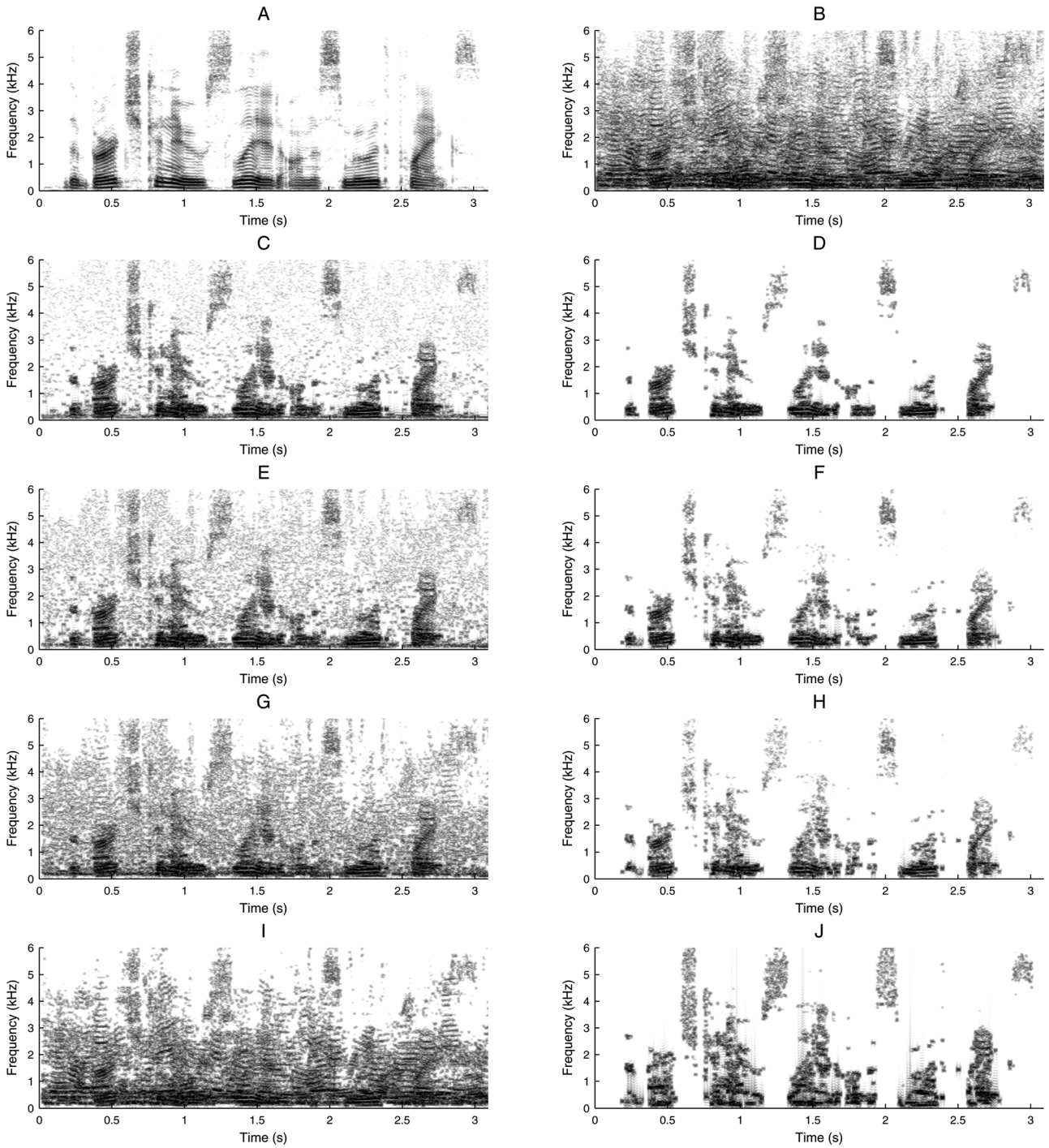

FIG. 2. (A) Spectrogram of clean speech utterance. (B) Spectrogram of noisy speech corrupted at −5 dB SNR with babble noise. (C), (E), (G), (I) Spectrograms of sentences synthesized from four new binary masks derived by masking speech-absent T-F units belonging to T1, T2, T3, and T4, respectively, from top to bottom. (D), (F), (H), (J) Spectrograms of sentences synthesized from new binary masks derived by masking speech-present T-F units belonging to L1, L2, L3, and L4, respectively, from top to bottom.

played to the listeners monaurally through Sennheiser HD 250 circumaural headphones at a comfortable listening level. The speech was present to the subjects on both ears. To become familiar with the test procedure, each subject listened to a set of noisy sentences before the actual test. During the test, subjects were asked to write down the words they recognized. Subjects participated in a total of eight conditions (four speech-present conditions and four speech-absent conditions). Each condition used two IEEE sentence lists of non-repeated sentences (i.e., 20 sentences). The order of test conditions was randomly selected for each subject. The duration of the test lasted 1.5 h and subjects were given 5 min breaks every half hour. The intelligibility score was computed with dividing the numbers of the correctly recognized words by the total counts words contained in 20 sentences.

B. Results and discussion

Results of the experiments are shown in Fig. 3. Two-way analysis of variance (ANOVA) with repeated

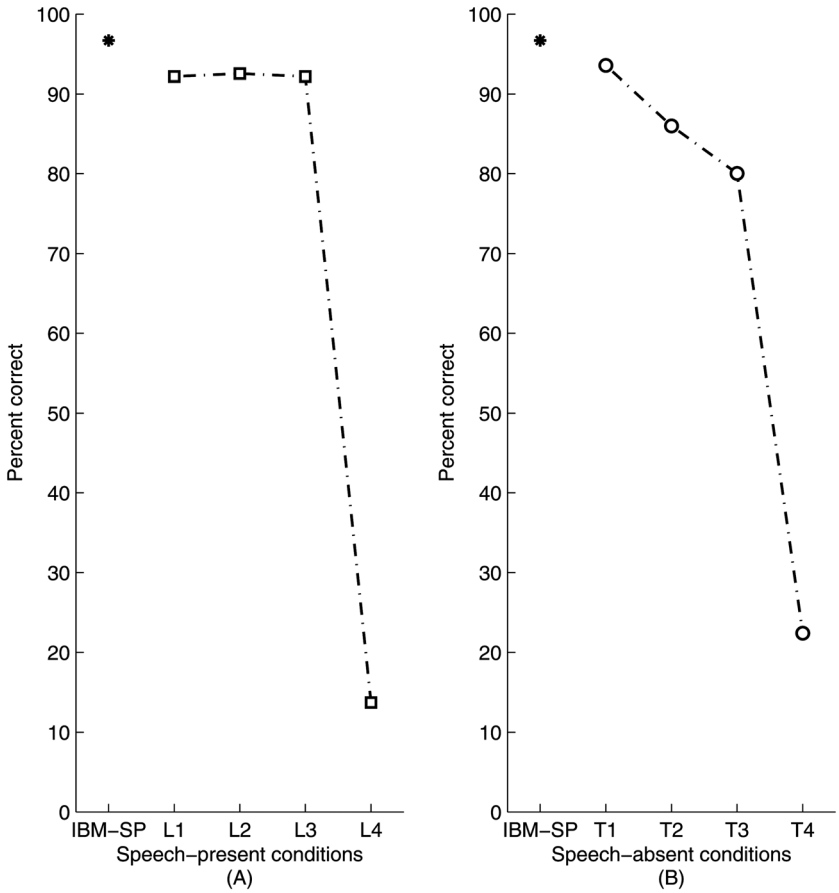

FIG. 3. Performance (percent of correctly recognized words) as a function of loudness groups with artificial masking errors and SNR of −5 dB SNR. (A) Left panel corresponds to the treatments for which errors were introduced only to the speech-present T-F units, while (B) the right panel corresponds to stimuli for which errors were introduced only to the speech-absent T-F units. In both plots, IBM-SP indicates the condition where no mask errors were introduced.

measures indicates a significant effect of loudness level of T-F units ($F[3, 56] = 886.3$, $p < 0.0001$) and a significant effect of interaction between loudness level and error types ($F[3, 56] = 14.36$, $p < 0.0001$). The left panel shows the intelligibility when mask errors were introduced only to speech-present T-F units. Consistent with the previous study by Anzalone *et al.* (2006), performance of speech synthesized from IBM-SP was very high (near 98%). Corresponding to IBM-SP, only a minor degradation (5%) in performance was observed when errors were introduced to the speech-present T-F units belonging to L1, L2, or L3 levels. This suggests that speech-present T-F units having lower target loudness have a reduced contribution toward total speech intelligibility. Alternatively, performance drops significantly (by nearly 80%) when the same number of errors are introduced into the speech-present T-F units belonging to L4 level. This result indicates that speech-present T-F units belonging to the highest loudness group of L4 are critically important to overall speech intelligibility. This suggests that contribution of each speech-present T-F unit to speech intelligibility is highly related to the loudness of its target content.

A similar tendency is observed when mask errors were introduced to speech-absent T-F units. A gradual drop in performance is observed as the location of the mask errors shifts from T1 to T3. A dramatic degradation in performance occurs when mask errors are introduced to T-F units belonging to T4. This illustrates the fact that the importance of speech-absent T-F units varies in accordance with the loudness of its masker content.

III. EXPERIMENT 2: EFFECT OF TYPES OF ERROR ACCORDING TO MIXTURE SNR LEVEL

In the previous study by Li and Loizou (2008), it was reported that false alarm errors are more harmful to speech intelligibility than miss errors. That study was performed using only a single input SNR level (−5 dB). In the current experiment, we extend the previous study Li and Loizou (2008) to evaluate the relative importance of the two types of mask errors by varying the mixture SNR levels from −15 to 5 dB, in increments of 2, 3, or 5 dB. The aim of this experiment is to determine if the relative importance between these error types varies according to the level of input SNR.

A. Methods

1. Subjects and material

The same eight subjects who participated in Experiment 1 also participated in the present experiment. Also, the procedure and speech material used in this experiment is the same as those of Experiment 1.

2. Signal processing

The IBM-SP was computed as in Experiment 1 and used as the benchmark for introducing artificial mask errors. To compare the relative importance between the two types of mask errors, new binary masks were created by introducing a fixed percentage of miss errors and false alarm errors into the IBM-SP separately. Specifically, for assessing the effect of miss errors, a fixed percentage of speech-present T-F units

originally labeled as 1 in IBM-SP were flipped to 0, while no mask errors were created on speech-absent T-F units. Similarly, for assessing the effect of false alarm errors, a fixed percentage of speech-absent T-F units originally labeled as 0 in IBM-SP were flipped to 1, while no mask errors were created on speech-present T-F units. Stimuli were created from the new binary masks containing the fixed rate of miss or false alarm errors. This procedure was repeated for a range of mixture SNRs. More specifically, the relative importance of the two types of errors was evaluated for the following mixture SNRs: −15, −10, −7, −5, 0. Only two fixed error levels, 25% and 50%, were examined in this experiment due to the limited number of sentences available in the IEEE corpus.

3. Procedure

The procedure was the same as in Experiment 1. Subjects participated in 24 conditions (2 fixed rates ×6 SNR levels ×2 types of errors). As in Experiment 1, two lists of non-repeated sentences were used for each condition. Because of the large number of conditions used for testing, listening tests were separated into two independent sessions on different days. Each session duration took approximately 1.5−2 h with 5 min breaks every 40 min of testing. The test conditions were assigned to subjects in a randomized order.

B. Results and discussion

Figure 4 indicates the performance when a fixed level of miss errors or false alarm errors was introduced to the input signal at various mixture SNRs. Three-way analysis of variance (ANOVA) with repeated measures indicates a significant effect of input SNR ($F[5, 149] = 269.3$, $p < 0.0001$), a significant effect of mask error type ($F[1, 149] = 1552.0$, $p < 0.0001$), a significant effect of mask error level, ($F[1, 149] = 115.1$, $p < 0.0001$), a significant interaction between input SNR and mask error type ($F[5, 149] = 170.7$, $p < 0.0001$), a significant interaction between input SNR and mask error level ($F[5, 149] = 8.67$, $p < 0.0001$), and a significant interaction between mask error type and mask error level ($F[1, 149] = 63.2$, $p < 0.0001$).

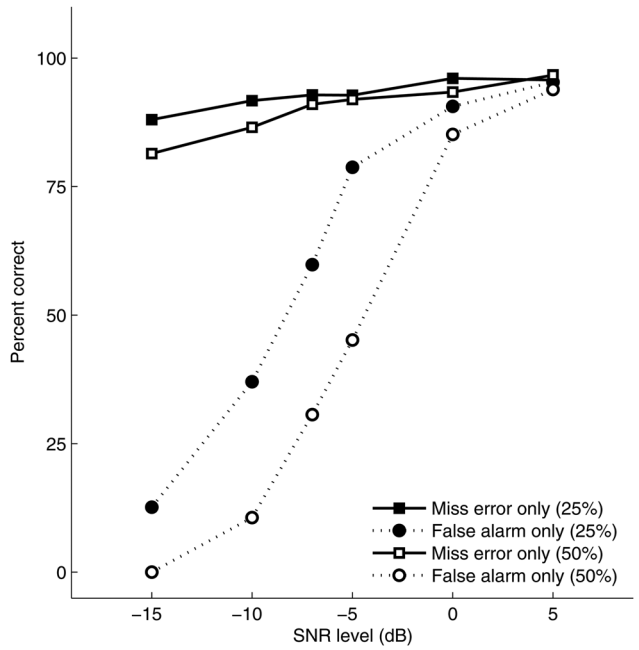

FIG. 4. Speech intelligibility scores of the speech stimuli synthesized from binary masks that included artificial classification errors (miss or false alarm errors) as a function of mixture SNR. Two fixed error levels (25% and 50%) are considered. The abscissa axis indicates the global SNR level of the mixture signals, while the ordinate axis indicates the subjective intelligibility scores.

Results show that the performance due to fixed rate of miss errors does not change as much across various mixtures of SNRs. For miss errors at both 25% and 50% error levels, a gradual degradation in performance was observed as the mixture SNR decreased from 5 to −15 dB. The net decrease in percent correct between 5 and −15 dB was 7% for the 25% error level, and slightly higher, 14%, for the 50% error level. In stark contrast, the impact of introducing the fixed rate of false alarm errors varied considerably with mixture SNR level, with the percent correct decreasing significantly as the mixture SNR decreases from 5 to −15 dB. For false alarm errors at the 25% level, the percent correct dropped dramatically, from 95% at 5 dB SNR to only 12% at −15 dB SNR. Similarly, for false alarm errors at the 50% error level, the percent correct dropped dramatically from 93% obtained at 5 dB to about 0% at −15 dB SNR.

Similar to the findings of the study by Li and Loizou (2008), the impact of introducing false alarm errors is much greater than introducing the same level of miss errors, when the input SNR is equal to or less than 0 dB. Moreover, the difference in the performance between the two error types varied with the mixture SNRs. As the input SNR is decreased from 5 to −15 dB, the difference increased from near 0% to over 80%. This suggests that in the context of speech intelligibility, the importance of limiting false alarm errors increases as the background SNR level decreases.

IV. THE LOUDNESS WEIGHTED HIT-FA

Mask-based objective speech intelligibility measures such as hit rate minus false alarm rate (HIT-FA) and ideal binary mask ratio (IBMR) have been proposed and frequently used as measures to evaluate binary masking techniques. These mask-based objective intelligibility measures are often obtained by comparing the estimated binary mask against the IBM. Since the calculation of mask-based objective measures does not require resynthesized output, these are robust against convolutional distortions associated with acoustic resynthesis. While these measures have been shown to have a modestly high correlation with subjective scores, the contribution of all T-F units is equally weighted. However, in Experiments 1 and 2, we have demonstrated that the importance of each T-F unit toward speech intelligibility varies significantly. In this section, we propose a new mask-based objective intelligibility measure, LWHF, to predict the intelligibility of binary masked speech.

A. Loudness spectrogram computation

Let $y(n) = x(n) + d(n)$ be the mixture signal, with $x(n)$ denoting the target signal and $d(n)$ denoting the masker signal. Signals [$y(n)$, $x(n)$, and $d(n)$] are first segmented in time

using a Hamming window (20 ms) with 50% overlap between frames. A fast Fourier transform (FFT) is then applied to each frame. T-F analyzed signals [$Y(t, f)$, $X(t, f)$, and $D(t, f)$] are pre-emphasized by an equal-loudness curve, simulating the perceptual sensitivity of the human ear to the intensity of sound at different frequency locations (Hermansky, 1990),

$$\overline{Y(t,f)} = Y(t,f)E(f), \tag{1}$$

$$\overline{X(t,f)} = X(t,f)E(f), \tag{2}$$

$$\overline{D(t,f)} = D(t,f)E(f), \tag{3}$$

$E(f)$ is an approximation of the equal loudness contour which simulates the sensitivity of human hearing at 40 dB level. It is valid up to 5000 Hz and is given by

$$\mathrm{E(f)} = \frac{[(f^2 + 56.8 \times 10^6)f^4]}{[(f^2 + 6.3 \times 10^6)^2 \times (f^2 + 0.38 \times 10^9)}. \tag{4}$$

After multiplying by the equal-loudness contour, the loudness spectrogram is calculated by applying a power law compression amplitude compression (Fastl and Zwicker, 2007):

$$L_Y(t,f) = [\overline{Y(t,f)}]^{0.23}, \tag{5}$$

$$L_X(t,f) = [\overline{X(t,f)}]^{0.23}, \tag{6}$$

$$L_D(t,f) = [\overline{D(t,f)}]^{0.23}, \tag{7}$$

where $L_Y(t, f)$, $L_X(t, f)$, and $L_D(t, f)$ indicate the loudness spectrogram of the mixture, target, and masker signals, respectively.

B. Loudness weighted miss error rate

The loudness weighted miss error rate (R_1) of the binary masked speech is defined as follows:

$$R_1 = \frac{\sum \mu(t,f) \times \mathrm{MISS}(t,f)}{\sum \mu(t,f) \times \mathrm{SP}(t,f)}, \tag{8}$$

where MISS(t, f) is the binary indication of miss error of each T-F unit, SP(t, f) is the binary indication of speech-present T-F units, and $\mu(t, f)$ is the weight value associated with each speech-present T-F unit. Since miss errors occur only in speech-present T-F units, $\mu(t, f)$ is related to the loudness of the local target component. Thus, we define $\mu(t, f)$ as follows:

$$\mu(t,f) = g[L_X(t,f)], \tag{9}$$

where $g(\ )$ is a sigmoid function for mapping each target-present T-F unit to the perceptual weight according to its target loudness,

$$g(x) = \frac{1}{1 + \exp\left(\frac{-(x - \alpha_1)}{\beta_1}\right)}. \tag{10}$$

C. Loudness weighted false alarm error rate

The loudness weighted false alarm error rate (R_2) of the binary masked speech is defined as follows:

$$R_2 = \frac{\sum \nu(t,f) \times \mathrm{FA}(t,f)}{\sum \mu(t,f) \times \mathrm{SP}(t,f)}, \tag{11}$$

where FA(t, f) is the binary indication of the false alarm error of each T-F unit, and $\nu(t, f)$ is the weight value associated with each false alarm error. Since false alarm errors occur only in speech-absent T-F units, $\nu(t, f)$ is related to the loudness of the local masker component. Thus, we define $\nu(t, f)$ as follows:

$$\nu(t,f) = h[L_d(t,f)], \tag{12}$$

where $h(\ )$ is a sigmoid function used for mapping each speech-absent T-F units to the perceptual weight according to its masker loudness,

$$h(x) = \frac{1}{1 + \exp\left(\frac{-(x - \alpha_2)}{\beta_2}\right)}. \tag{13}$$

D. Proposed objective intelligibility measure

Since miss and false alarm errors contribute different effects on speech intelligibility, the loudness weighted miss error rate and loudness weighted false alarm error rate need to be further weighted. According to the previous study by Li and Loizou (2008), the distortion of miss error rate on speech intelligibility is nonlinear, while the distortion of the false alarm rate on speech intelligibility is approximately linear. Thus, the final LWHF is defined as follows:

$$\mathrm{LWHF} = 1 - G(R_1) - \gamma \times R_2, \tag{14}$$

where $G(\ )$ is a sigmoid function used for approximating the influence of miss error on speech intelligibility,

$$G(x) = \frac{1}{1 + \exp\left(\frac{-(x - \alpha_3)}{\beta_3}\right)} \tag{15}$$

and γ is the weight associated with the false alarm error.

E. Evaluation of LWHF

In order to evaluate the proposed objective intelligibility measure, we compare it against two other existing mask-based objective measures, HIT-FA and IBMR, based on the stimuli produced in Experiment 1. Table I denotes the values of free parameters used to compute the LWHF. The parameters shown here were jointly optimized to maximize the correlation with subjective intelligibility scores. Results are shown in Fig. 5. It is clear from Fig. 5 that existing

TABLE I. Used values of free parameters.

α_1	β_1	α_2	β_2	α_3	β_3	γ
0.9	0.12	1.15	0.1	1.0	0.14	0.1

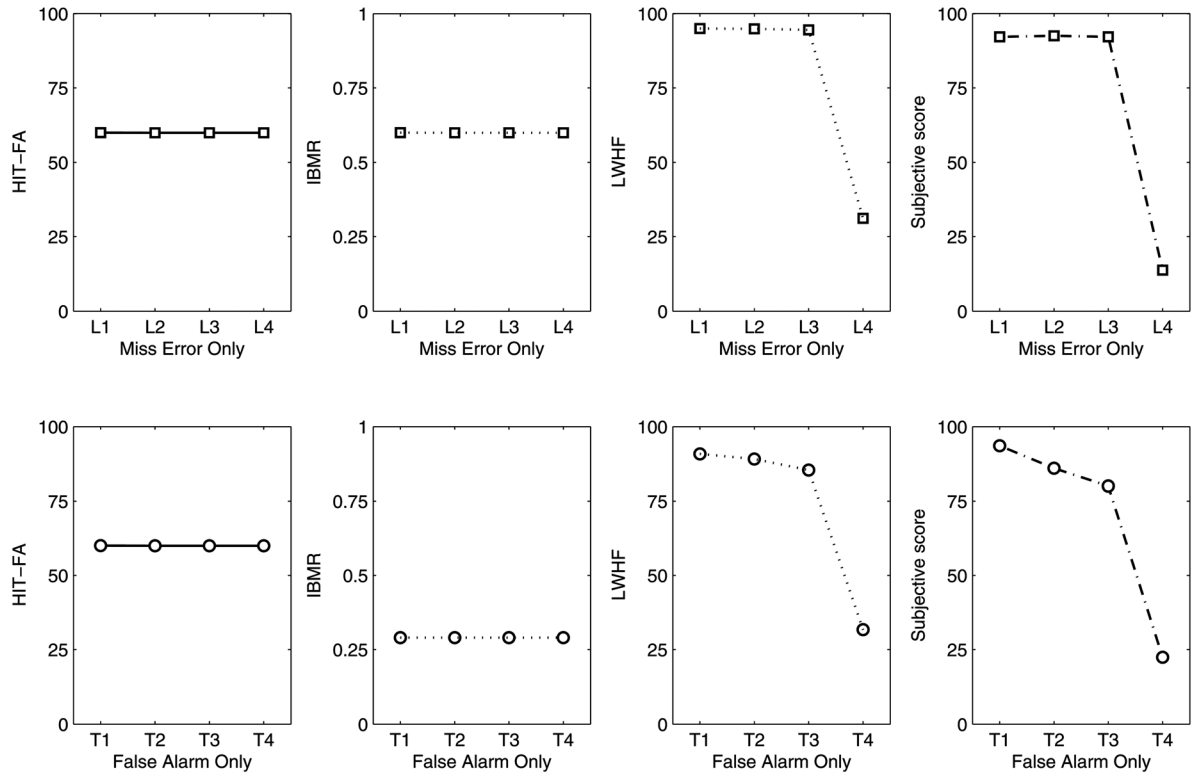

FIG. 5. Comparison of the proposed objective speech intelligibility measure (LWHF) with two existing mask-based objective intelligibility measures, hit rate minus false alarm rate (HIT-FA) and ideal binary mask ratio (IBMR), on speech stimuli from Experiment 1. Subjective performance is used as reference for comparison.

mask-based objective measures, HIT-FA and IBMR, do not provide consistent prediction on stimuli created from binary masks having asymmetric mask errors, as in those from Experiment 1. This is due to the fact that HIT-FA and IBMR assume that each T-F unit provides an equal contribution to speech intelligibility. Alternatively, the proposed mask-based objective measure (LWHF) is consistent with subjective listening scores.

In addition, we compare the proposed measure with two well known objective measures: short-time objective intelligibility measure (STOI) (Taal *et al.*, 2011) and IBM-modulated SNR (Hu and Wang, 2004). In STOI, the intelligibility of speech is estimated by computing the correlation between T-F representations of clean and binary masked speech in a short term basis (386 ms) and the IBM-modulated SNR is computed by using speech resynthesized from ideal binary mask as ground truth. In contrast to HIT-FA and IBMR, the intensity of each T-F unit has certain effects on the computation of STOI and IBM-modulated SNR. However, none of those metrics have included perceptually motivated T-F mapping as in our proposed metric. Previous studies (Taal *et al.*, 2011; Hu and Wang, 2004) have shown that both STOI and IBM modulated SNR have high correlation with speech intelligibility. However, the evaluations were mostly based on the stimuli resynthesized from binary masks having symmetric mask errors. In this study, we compare those metrics with proposed objective measure on speech stimuli generated from binary masks having highly asymmetric mask errors as in Experiments 1 and 2. The results are shown in Fig. 6. The listening scores

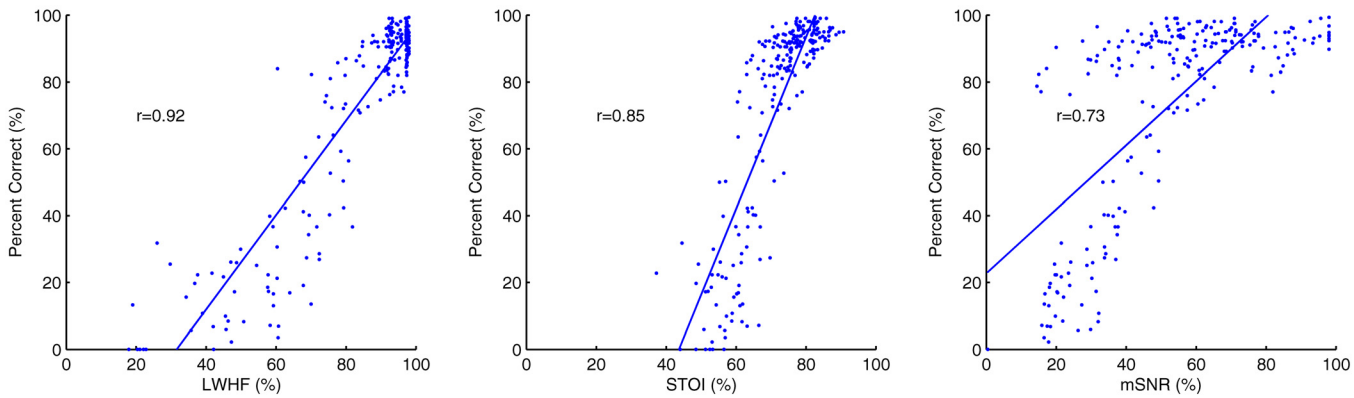

FIG. 6. (Color online) Scatter plots showing the correlation between subjective listening scores of the stimuli obtained from Experiments 1 and 2 and three objective measures: proposed mask-based objective intelligibility metric (LWHF), STOI, and IBM modulated SNR (mSNR).

in each condition of individual listener is the average of 20 sentences of that condition. The results show that proposed objective measure has a higher correlation with intelligibility than the other two objective measures. The above results suggest that the LWHF measure has the potential to achieve higher correlation with subjective intelligibility scores than existing mask-based objective measures as well as more general objective measures, in particular for the conditions when mask errors are not symmetrically distributed with respect to loudness as well as error types.

V. DISCUSSION AND CONCLUSIONS

The present study has assessed the contribution of individual T-F units to speech intelligibility in the context of IBM. Results from Experiment 1 indicate, consistent with our hypothesis, that there is a strong correlation between target or masker loudness in a specific T-F cell and how it contributes to overall intelligibility. That is, mask errors localized in the speech-present T-F units having higher target loudness create more intelligibility distortion than those with lower target loudness. Similarly, mask errors in speech-absent T-F units having higher masker loudness produce more intelligibility degradation than those having lower masker loudness. The result also shows that speech-present T-F units having low target loudness, and speech-absent T-F units having low masker loudness have minimal influence on speech intelligibility.

From Experiment 2, we extended the findings of the study by Li and Loizou (2008) by comparing the influence of false alarm errors and miss alarm errors across a range of mixture SNRs. The result from Experiment 2 confirmed that false alarm errors are more harmful to speech intelligibility than miss errors when mixture SNR is lower than 0 dB. The new finding from Experiment 2 is that the relative importance between the two types of errors varies significantly according to the mixture SNR. As the SNR of the input signal decreases, the false alarm errors become more harmful to speech intelligibility, while the effect of miss errors does not vary significantly. This result agrees with the findings from Experiment 1 that the importance of miss errors (located in speech-present T-F units) is highly related to its target loudness, and relatively unaffected by the background noise level. Alternatively, the importance of false alarm errors (located in speech-absent T-F units) is highly related to its masker content, which increases as the mixture SNR is reduced.

Drawing from the findings from Experiments 1 and 2, we proposed a new mask-based objective intelligibility metric, the LWHF score, to incorporate T-F variation into the prediction of speech intelligibility. The new LWHF showed a high correlation ($r = 0.92$) on stimuli synthesized from binary masks having asymmetric mask errors where existing mask-based objective metrics such as HIT-FA and IBMR could not provide consistent scores with subjective listening scores. By comparing with two recently proposed objective measures namely STOI and IBM modulated SNR, we confirm that proposed mask-based objective measure has the potential of achieving higher correlation with subjective intelligibility scores than existing objective measures.

Although a relatively higher number of free parameters were used for LWHF, the high correlation results are attributed to the modeling of the perceptual effect of each T-F unit as well as different types of mask errors on overall intelligibility, rather than an extensive fine tuning process of the free parameters. In order to explore the robustness of the objective intelligibility measure, the seven parameter settings from Table I were modified to assess changes in correlation with listeners' results. Each parameter was modified sequentially (i.e., increased and decreased) by 10%, and a regression analysis performed to determine any change in the overall correlation coefficient. The results showed that the correlation coefficient varied by less than 2% absolute, when any individual parameter setting was varied by 10%. This confirms the robustness of the measure, and the fact that once reasonable parameter settings are determined, the proposed intelligibility measure is effective.

The above results have important implications for speech separation algorithms that are based on the estimation of ideal binary mask as well as the objective evaluations of those algorithms. For those algorithms to achieve intelligibility gains, it is important to focus on the accurate classification of those speech-present T-F units having high target loudness, and speech-absent T-F units having high masker loudness. It is also necessary to assign appropriate emphasis to false alarm errors in accordance with the SNR level of the background noise. In this work, we have developed a new mask-based objective intelligibility measure specifically for binary masked speech based on the findings from two experiments. However, it is important to note that these findings may also be beneficial for other general purpose objective intelligibility measures.

In this study, we have therefore shown that the importance of each T-F unit is highly related to its loudness content. However, this does not necessarily infer that loudness content is the only factor to be considered. Future studies could investigate the effects of other attributes of T-F units, such as frequency content location or prosodic/f_0 structure, on overall speech intelligibility.

Although the proposed objective measure has shown promising result for predicting the intelligibility of binary masked speech, a further validation on other datasets as well as noisy conditions need to be performed in our future work. In the mean time, the proposed intelligibility model could not predict the binary masked speech generated directly from IBM modulated noise (Wang *et al.*, 2008) where intelligible speech was produced without any speech-present T-F units.

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).

Anzalone, M., Calandruccio, L., Doherty, K., and Carney, L. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.

Bregman, A. (**1990**). *Auditory Scene Analysis* (MIT Press, Cambridge, MA), pp. 1–570.
Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.
Fastl, H., and Zwicker, E. (**2007**). *Psychoacoustics: Facts and Models* (Springer, New York), Vol. 22, pp. 220–233.
Goldsworthy, R. L., and Greenberg, J. E. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.
Han, K. and Wang, D. (**2011**). "An SVM based classification approach to speech separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Prague, Czech Republic, pp. 4632–4635.
Hansen, J. H. L., and Nandkumar, S. (**1995**). "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process," J. Acoust. Soci. Am. **97**(6), 3833–3849.
Hermansky, H. (**1990**). "Perceptual linear predictive (PLP) analysis of speech," J. Acoust. Soc. Am. **87**, 1738–1752.
Hu, G., and Wang, D. (**2004**). "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Networks **15**, 1135–1150.
Hu, Y., and Loizou, P. C. (**2007**). "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun. **49**, 588–601.
Hummersone, C., Mason, R., and Brookes, T. (**2011**). "Ideal binary mask ratio: a novel metric for assessing binary-mask-based sound source separation algorithms," IEEE Trans. Audio, Speech, Lang. Process. **19**, 2039–2045.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **AU-17**, 225–246.
Kates, J. M., and Arehart, K. H. (**2005**). "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. J. **117**, 2224–2237.
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.
Kim, W., and Hansen, J. (**2011**). "A novel mask estimation method employing posterior-based representative mean estimate for missing-feature speech recognition," IEEE Trans. Audio, Speech, Lang. Process. **19**, 1434–1443.
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.
Li, F., and Allen, J. (**2009**). "Additivity law of frequency integration for consonant identification in white noise," J. Acoust. Soc. Am. **126**, 347–353.
Li, F., and Allen, J. (**2011**). "Manipulation of consonants in natural speech," IEEE Trans. Audio, Speech, Lang. Process. **19**, 496–504.
Li, N., and Loizou, P. (**2008**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.
Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (Taylor and Francis, Boca Raton, FL), pp. 589–599.
Loizou, P., and Kim, G. (**2011**). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," IEEE Trans. Audio, Speech, Lang. Process. **19**, 47–56.
Ma, J., Hu, Y., and Loizou, P. C. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**, 3387–3405.
Nandkumar, S., and Hansen, J. H. L. (**1995**). "Dual-channel iterative speech enhancement with constraints based on an auditory spectrum," IEEE Trans. Speech Audio Process. **3**(1), 22–34.
Roman, N., and Wang, D. (**2006**). "Pitch-based monaural segregation of reverberant speech," J. Acoust. Soc. Am. **120**, 458–469.
Seltzer, M., Raj, B., and Stern, R. (**2004**). "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," Speech Commun. **43**, 379–393.
Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio, Speech, Lang. Process. **19**, 2125–2136.
Wang, D. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Dordrecht, Netherlands), pp. 181–197.
Wang, D., and Brown, G., eds. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley, Hoboken, NJ/IEEE, New York), pp. 1–395.
Wang, D., Kjems, U., Pedersen, M. S., Boldt, J. B., and Lunner, T. (**2008**). "Speech perception of noise with binary gains," J. Acoust. Soc. Am. **124**, 2303–2307.
Wang, Y., Han, K., and Wang, D. (**2013**). "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio, Speech, Lang. Process. **21**, 270–279.
Yu, C., Wójcicki, K. K., Loizou, P. C., and Hansen, J. H. (**2013**). "A new mask-based objective measure for predicting the intelligibility of binary masked speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Vancouver, Canada, pp. 7030-7033.
Zhang, X., Heinz, M., Bruce, I., and Carney, L. (**2001**). "A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression," J. Acoust. Soc. Am. **109**, 648–670.