An ideal quantized mask to increase intelligibility and quality of speech in noise

Eric W. Healy[a),b)] and Jordan L. Vasko[b)]
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210, USA

(Received 16 May 2018; revised 6 August 2018; accepted 20 August 2018; published online 13 September 2018)

Time-frequency (T-F) masks represent powerful tools to increase the intelligibility of speech in background noise. Translational relevance is provided by their accurate estimation based only on the signal-plus-noise mixture, using deep learning or other machine-learning techniques. In the current study, a technique is designed to capture the benefits of existing techniques. In the ideal quantized mask (IQM), speech and noise are partitioned into T-F units, and each unit receives one of N attenuations according to its signal-to-noise ratio. It was found that as few as four to eight attenuation steps (IQM_4, IQM_8) improved intelligibility over the ideal binary mask (IBM, having two attenuation steps), and equaled the intelligibility resulting from the ideal ratio mask (IRM, having a theoretically infinite number of steps). Sound-quality ratings and rankings of noisy speech processed by the IQM_4 and IQM_8 were also superior to that processed by the IBM and equaled or exceeded that processed by the IRM. It is concluded that the intelligibility and sound-quality advantages of infinite attenuation resolution can be captured by an IQM having only a very small number of steps. Further, the classification-based nature of the IQM might provide algorithmic advantages over the regression-based IRM during machine estimation. © *2018 Acoustical Society of America*.
https://doi.org/10.1121/1.5053115

[JJL] Pages: 1392–1405

I. INTRODUCTION

The perception of speech in background noise represents a challenge for a variety of listeners in a variety of settings. Normal-hearing (NH) listeners with proficiency of the language can tolerate considerable amounts of noise if conditions are otherwise ideal. But even these best listeners can struggle if the signal is acoustically deficient, as can be the case during transmission over cellular phones or other communication systems. The situation is compounded if the listener does not have complete proficiency with the language, as can be the case for non-native-language listeners, children, and other individuals. But the challenge is particularly striking for listeners with hearing loss. In fact, poor speech recognition when background noise is present is a primary auditory complaint of hearing-impaired (HI) individuals (see Moore, 2007; Dillon, 2012), and the speech-in-noise problem for these listeners represents one of our greatest challenges.

Fortunately, techniques exist to help alleviate this challenge. Time-frequency (T-F) masking represents a powerful tool for improving the intelligibility of speech in noise. In T-F masking, the speech-plus-noise mixture is divided in both time and frequency into small units, and each unit is scaled in level according to the relationship between the speech and the noise within the unit. Units with less favorable signal-to-noise ratios (SNRs) are attenuated, resulting in a signal containing T-F units largely dominated by the target speech signal.

There are two main classes of T-F masks, known as "hard" and "soft" masks. These correspond to two main T-F masking schemes, which include binary masking and ratio masking. In the ideal binary mask (IBM; Hu and Wang, 2001; Wang, 2005), each T-F unit is assigned a value of 1 if it is dominated by the target speech or 0 if it is dominated by noise. The IBM is then multiplied with the speech-plus-noise mixture, causing units dominated by the target speech to remain intact and units dominated by the noise to be discarded. In the ideal ratio mask (IRM; Srinivasan *et al.*, 2006; Narayanan and Wang, 2013; Hummerstone *et al.*, 2014; Wang *et al.*, 2014), each T-F unit is again assigned an attenuation scaling according to the speech versus noise relationship. But instead of being binary, this scaling can take any value along a continuum from 0 to 1. Units having a more favorable SNR are attenuated less and those having a less favorable SNR are attenuated more. Accordingly, the IRM is similar to the classic Wiener filter (see Loizou, 2007). As with the IBM, the speech-plus-noise mixture is multiplied with this mask to obtain an array of T-F units, each scaled according to its speech versus noise dominance.

Both masks are capable of producing vast improvements in the intelligibility of noisy speech. Brungart *et al.* (2006), Li and Loizou (2008a,b), Kim *et al.* (2009), Kjems *et al.* (2009), and Sinex (2013) all found that the IBM could produce near-perfect sentence intelligibility for NH listeners in various noises (speech-shaped noise, speech-modulated noise, 2- to 20-talker babble, and various recorded environmental sounds). Anzalone *et al.* (2006) and Wang *et al.* (2009) tested both NH and HI subjects, and found that the IBM could produce substantial speech-reception threshold improvements for sentences in different noises (speech-

[a)]Electronic mail: Healy.66@osu.edu
[b)]Also at: Center for Cognitive and Brain Sciences; The Ohio State University, Columbus, Ohio 43210, USA.

shaped noise and cafeteria noise). With regard to the IRM, Madhu *et al.* (2013) and Koning *et al.* (2015) found that it can also produce near-perfect sentence intelligibility for NH listeners in different noises (multi-talker babble and single-talker interference).

The comparison between intelligibility produced by the IBM versus that produced by the IRM is made difficult by the fact that all of the studies just cited employed sentence materials and those employing percent-correct intelligibility often observed ceiling scores at or near 100%. But Madhu *et al.* (2013) and Koning *et al.* (2015) both observed that the IRM produced ceiling intelligibility for NH subjects over a wider range of parameter values than did the IBM. In contrast, Brons *et al.* (2012) observed that the IBM led to better intelligibility than did an IRM having an attenuation of 10 dB for all units with SNR below 0 dB. Thus, the relative intelligibilities produced by the IBM versus the IRM are not clear.

What is more clear is that soft masking typically provides better subjective sound quality of speech than does hard masking. Madhu *et al.* (2013) conducted pairwise comparisons of preferred sound quality for NH subjects, and found that the ideal Weiner filter was preferred over the IBM in 88%–100% of trials.

The term "ideal" refers to the fact that the masks are created using knowledge of the pre-mixed speech and noise signals—they are oracle masks. The term also refers to the fact that the IBM produces the optimal SNR gain of all binary T-F masks under certain conditions (Li and Wang, 2009). Obviously, knowledge of the pre-mixed signals is not present in real-world settings. But translational significance for T-F masks comes from efforts to estimate them directly from the speech-plus-noise mixture, and the IBM has for many years been considered a goal of computational auditory scene analysis (Wang, 2005). Recent advances in machine learning, and particularly deep learning, have allowed both the IBM and IRM to be estimated with accuracy sufficient to produce considerable intelligibility improvements. This work has involved both NH listeners (Kim *et al.*, 2009; Healy *et al.*, 2013; Healy *et al.*, 2014; Healy *et al.*, 2015; Chen *et al.*, 2016; Healy *et al.*, 2017; Monaghan *et al.*, 2017; Bentsen *et al.*, 2018) and HI listeners (Healy *et al.*, 2013; Healy *et al.*, 2014; Healy *et al.*, 2015; Chen *et al.*, 2016; Healy *et al.*, 2017; Monaghan *et al.*, 2017) and has included a variety of background noises (speech-shaped noise, multi-talker babble, recorded environmental sounds, and single-talker interference). The intelligibility improvements have often allowed HI subjects having access to speech processed by the estimated T-F mask to equal the performance of young NH subjects without processing.

In addition to their different perceptual ramifications, the two main T-F masking schemes possess different characteristics that may be relevant for their estimation by machine-learning algorithms. Estimation of the IBM involves classification of T-F units into two categories using a single decision boundary, whereas estimation of the IRM typically involves regression and approximation of the continuous function underlying attenuation versus SNR. These represent very different learning regimes, with classification into a small number of categories potentially representing a more basic form of machine learning, underlying more elementary tasks such as object recognition (e.g., character recognition) and word/phoneme recognition. Accordingly, it has been argued that computation of a binary mask may be considerably simpler than computation of a soft mask (Wang, 2008).

But it has also been argued (e.g., Wang *et al.*, 2014) and observed (Madhu *et al.*, 2013; Koning *et al.*, 2015; Bentsen *et al.*, 2018) that binary masks are less robust to estimation errors relative to soft masks, meaning that the errors that occur during machine estimation are likely larger in magnitude in binary than in soft masks. It is easy to see that every estimation error in a binary mask is of maximum magnitude (e.g., assigning 1 to a T-F unit that should have been 0, or vice versa), and that in a soft mask, estimation errors can take any value, with an upper bound corresponding to the magnitude of the binary-mask error.

In the current study, a mask is proposed that is different from the two main T-F masking schemes. In the ideal quantized mask (IQM), the speech-noise mixture is divided into T-F units and each is assigned an attenuation based on SNR. However, this attenuation takes one of N values, where N represents a small integer value. The T-F masking conditions employed in the current study form a continuum in terms of attenuation steps, from two (IBM) to infinity (IRM). The three intermediate steps involve an IQM having 3, 4, and 8 steps (IQM_3, IQM_4, and IQM_8). The goals of the current study are to first clarify the relative intelligibilities produced by the IBM versus the IRM. Then, the intelligibility and sound quality of the IQM are established in NH subjects and compared to those resulting from the IBM and IRM. The goal is to capture the (potential) intelligibility and well-established sound-quality advantages of the IRM, and the classification nature and potential computational advantages of the IBM, in an IQM having only a very small number of attenuation steps.

II. EXPERIMENT 1. INTELLIGIBILITY RESULTING FROM VARIOUS T-F MASKS

In experiment 1, intelligibility was assessed in each of the five conditions of T-F masking. The speech materials selected were standard word lists because sentences tend to produce ceiling intelligibility values when subjected to both IBM and IRM processing. Experiment 1a involved broadband (unfiltered) word stimuli, and experiment 1b involved the same stimuli subjected to bandpass filtering in order to further avoid ceiling effects and better reveal differences across conditions. The background noise employed involved recordings from a busy cafeteria. It was selected for ecological validity and to possess variety of sound sources and types, including the babble of multiple talkers, the transient impact sound of dishes, and other environmental sounds.

A. Method

1. Subjects

A total of 20 subjects participated: 10 in experiment 1a and 10 in experiment 1b. The subjects were students at The

Ohio State University and received course credit for participating. All were native speakers of American English and had NH as defined by audiometric thresholds of 20 dB hearing level (HL) or better at octave frequencies from 250 to 8000 Hz on the day of test (ANSI, 2004, 2010). The exception was one subject with a threshold of 25 dB HL at 8000 Hz in one ear. Ages ranged from 19 to 29 yr (mean = 20.9 yr) and all were female. Care was taken to ensure that no subject had prior exposure to the speech materials employed.

2. Stimuli

The speech materials for both experiments 1a and 1b were from the Central Institute for the Deaf (CID) W-22 test (Hirsh *et al.*, 1952), drawn from an Auditec CD (St. Louis, MO). The test includes 200 phonetically balanced words in the carrier phrase, "Say the word ___." Five words were excluded (mew, two, dull, book, there), based on low frequency of occurrence or poor articulation/recording quality, to yield 195 words. The background cafeteria noise was also from an Auditec CD. It was approximately 10 min in duration and consisted of three overdubbed recordings made in a busy hospital-employee cafeteria. Noise segments having random start points and durations equal to each word in its carrier phrase were mixed with each speech utterance at an overall SNR of −10 dB. This relatively low SNR was selected to reduce ceiling intelligibility effects.

The files were down-sampled to 16 kHz for processing in MATLAB (MathWorks, Natick, MA). Preparation of the T-F masks began by dividing each speech-plus-noise mixture into a T-F representation. The cochleagram representation (Wang and Brown, 2006) was employed, which is essentially a spectrogram having attributes similar to the human cochlea. This involved first filtering into 64 gammatone bands having center frequencies ranging from 50 to 8000 Hz evenly spaced on the equivalent rectangular bandwidth scale (Glasberg and Moore, 1990). Each band was then divided into 20-ms time segments having 10 ms overlap using a Hanning window. This same T-F representation was used for all of the masks.

a. Preparation of the IBM. The IBM consists of a two-dimensional array of 1's and 0's, one value for each T-F unit. Its processing followed that employed by us previously (Healy *et al.*, 2013; Healy *et al.*, 2014). The SNR within each T-F unit was calculated based on the pre-mixed signals. If the SNR was greater than a fixed local criterion (LC) value, the unit was concluded to be target-speech dominated and it was assigned a value of 1. Inversely, if that SNR was less than or equal to LC, the unit was concluded to be noise dominated and it was assigned a value of 0. That is,

$$\mathrm{IBM}(t,f) = \begin{cases} 1, & \text{if SNR}\,(t,f) > LC \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

where $SNR(t,f)$ denotes the SNR within the T-F unit centered at time t and frequency f. LC was set to −15 dB in order to be 5 dB below the overall SNR. This relationship between LC and SNR has been considered near optimal (e.g., Brungart *et al.*, 2006; Kjems *et al.*, 2009; Vasko *et al.*, 2018), and has also been employed by us previously (Healy *et al.*, 2013; Healy *et al.*, 2014). To create the IBM-processed signals, the mask was applied to the speech-plus-noise mixture by multiplying each mixture T-F unit by the value of the IBM for that unit.

b. Preparation of the IRM. The IRM also consists of a two-dimensional array of values, one for each T-F unit, but these values are continuous between 0 and 1. It was also based on the relative energies of speech versus noise within each T-F unit, as defined by

$$\mathrm{IRM}(t,f) = \sqrt{\frac{S(t,f)}{S(t,f)+N(t,f)}} = \sqrt{\frac{\mathrm{SNR}(t,f)}{\mathrm{SNR}(t,f)+1}}, \quad (2)$$

where $S(t,f)$ is the speech energy contained within the T-F unit centered at time t and frequency f, and $N(t,f)$ is the noise energy contained within the same unit. Whereas Eq. (1) is a conditional statement and so units are irrelevant so long as they match for SNR and LC, SNR in Eq. (2) is an untransformed ratio of energies. This square-root form of the IRM has been found to be optimal (e.g., Wang *et al.*, 2014) and has been employed by us previously (Healy *et al.*, 2015; Healy *et al.*, 2017; Chen *et al.*, 2016). The mask was applied to the speech-plus-noise, again by weighting each mixture T-F unit by the value of the IRM for that unit.

c. Preparation of the IQM. IQMs were created having three, four, and eight attenuation steps (IQM_3, IQM_4, IQM_8). The SNR boundaries defining each step of the IQM and the attenuation assigned to each step of the IQM were based on the IBM and IRM functions. The SNR boundaries were centered such that the IQM_2 would equal the IBM (having an LC value 5 dB below the overall mixture SNR) once scaled for overall level. The center SNR boundaries of the IQM_4 and IQM_8 (between steps 2 and 3 in the IQM_4 and between steps 4 and 5 in the IQM_8) were also set to equal the single IBM division. The attenuation assigned to each step (the IQM value) was equal to the attenuation assigned by the IRM (the IRM value) at the lower SNR boundary for the step. The exception was that the lowest step was always assigned a value of 0, like in the IBM.

The process began with the selection of a series of points on the IRM function, according to Eqs. (3) and (4),

$$p = -\log_2\sqrt{\frac{10^{(\mathrm{LC}/10)}}{10^{(\mathrm{LC}/10)}+1}}, \quad (3)$$

$$x_n(t,f) = \left(\frac{n-1}{N}\right)^p, \quad (4)$$

where the exponent p was selected based on the LC for the IBM (−15 dB) to provide the desired relationship between the IQM center SNR boundary and the IBM boundary. $x_n(t,f)$ represents the mask gain within the T-F unit centered at time t and frequency f, N represents the total number of

steps in the IQM, and $n = 1,\ldots,N$ and represents the ordinal position of each step. These points became the SNR boundaries and attenuation values for the IQM, as in Eq. (5),

$$\mathrm{IQM}_N(t,f) = \begin{cases} x_1(t,f), & \text{if } 0 \le \mathrm{IRM}(t,f) \le x_2(t,f) \\ x_2(t,f), & \text{if } x_2(t,f) < \mathrm{IRM}(t,f) \le x_3(t,f) \\ \vdots & \\ x_N(t,f), & \text{if } x_N(t,f) < \mathrm{IRM}(t,f) \le 1. \end{cases} \quad (5)$$

As with the other two masks, the IQM was applied by multiplying the stepped mask with the speech-plus-noise mixture units. Figure 1 displays the SNR boundaries for each step of each IQM employed (top panel) and the attenuations produced by each step of each IQM employed (bottom panel). Every stimulus was scaled after processing to the same overall root-mean-square level, eliminating differences in overall level.

Whereas the IBM takes values of either 0 or 1, the IRM takes on values bounded by 0 and 1. Although the IRM is capable in theory of zeroing T-F units, it is potentially notable that this will generally not occur because the likelihood of zero signal energy within a T-F unit is nil. But like the IBM, the current IQM was designed to zero all T-F units at the lowest step [$x_1(t,f)$ always $= 0$]. This decision was made to reduce the perception of low-level noise arising from the T-F units having the least-favorable SNRs. It is also noteworthy that the current implementation of the IQM is based on existing T-F masks in order to facilitate direct comparison, and it is not yet known to what extent this particular implementation produces optimal human intelligibility and sound quality.

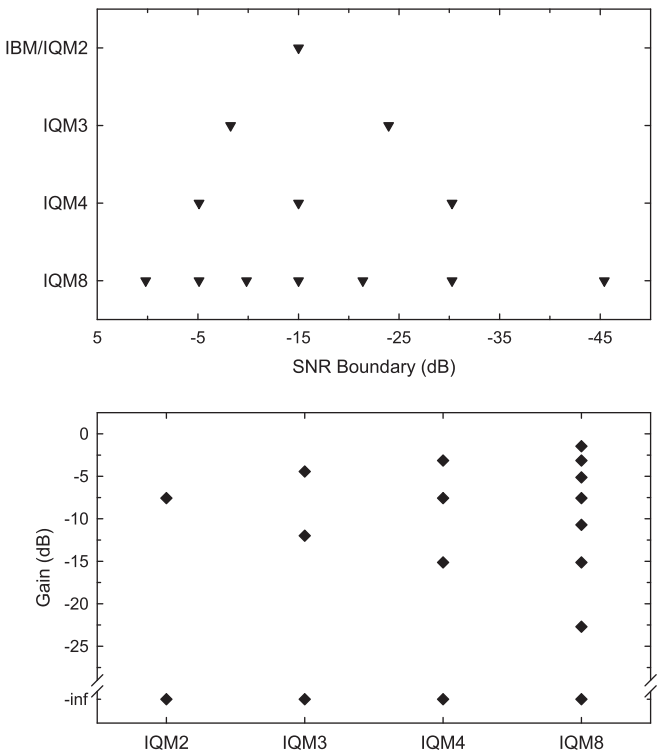

FIG. 1. The top panel displays the SNR boundaries for each step of each IQM employed, and the bottom panel displays the attenuations assigned to each step of each IQM employed.

The broadband stimuli processed as just described were used for experiment 1a. For experiment 1b, the same stimuli were subjected to bandpass filtering from 750 to 3000 Hz. The final processed stimuli were passed through a 2000-order finite-duration impulse response filter, resulting in steep filter slopes that exceeded 1000 dB/octave.

B. Procedure

The procedures for experiments 1a and 1b were identical. The experiment was divided into three blocks, each involving 13 words in each mask condition for a total of 39 words/condition. The order of mask conditions (IBM, IQM_3, IQM_4, IQM_8, IRM) was randomized for each block and subject, as was the word list-to-condition correspondence. The stimuli were converted to analog form using Echo Digital Audio (Santa Barbara, CA) Gina 3G digital-to-analog converters and presented diotically over Sennheiser HD 280 headphones (Wedemark, Germany). The presentation level was set to 65 dBA at each earphone at the start of each session using a flat-plate coupler and sound level meter (Larson Davis AEC 101 and 824, Depew, NY). Subjects were tested individually in a double-walled audiometric booth seated with the experimenter. The subjects were instructed to repeat each word back as best they could after hearing each and were encouraged to guess if unsure. No word was repeated for any listener. The experimenter controlled the presentation of words and recorded responses. Testing began with a brief practice in which subjects heard words from the consonant-nucleus-consonant (CNC) corpus (Lehiste and Peterson, 1959). These were also standard recordings produced by a male talker and in a carrier phrase ("Ready, ___."). Subjects heard five CNC words in each mask condition in order of decreasing number of attenuation steps (IRM, IQM_8, IQM_4, IQM_3, IBM). Feedback was provided during practice but not during formal testing.

C. Results and discussion

1. Human subjects results

The top panel of Fig. 2 displays group mean word-recognition scores for each broadband T-F mask in experiment 1a. Apparent in this panel is that all masks produced high recognition scores (above 70% correct), but that scores for the IBM were lower than those for the IQMs and the IRM, where all values exceed 90% correct. The bottom panel of Fig. 2 displays scores for the group hearing the bandpass stimuli in experiment 1b. Apparent is that scores were reduced below the ceiling and larger differences between scores emerged across the different masks, both as desired. A first notable finding is that speech recognition produced by the IBM is not equal to that produced by the IRM, despite that both produce similar ceiling scores for sentence intelligibility. Instead, mean recognition scores were better for the IRM by 36 percentage points when ceiling values were eliminated. A second primary finding is that scores were highest in the IQM_8 condition and scores for the IQM_4

approximated that for the IRM. Thus, it appears that the intelligibility benefit of the IRM can be captured with as few as four attenuation steps. Finally, it is noted that the addition of any number of attenuation steps above two produced increased speech recognition.

The scores were transformed into rationalized-arcsine units (RAUs; Studebaker, 1985) and subjected to a two-way mixed analysis of variance (ANOVA; 2 filtering groups × 5 mask conditions). The interaction between filtering and mask conditions was not significant [$F(4,72) = 0.9$, $p = 0.45$], suggesting that the pattern of performance across different mask conditions was generally consistent across experiments. As anticipated, the main effect of filtering was significant [$F(1,18) = 359.6$, $p < 0.001$], simply reflecting the desired reduction in scores associated with filtering. Most critically, the main effect of mask condition was significant [$F(4,72) = 86.3$, $p < 0.001$]. Performance across the five pooled mask conditions was examined using Holm-Sidak pairwise *post hoc* comparisons. Performance did not differ significantly among the IQM_4, IQM_8, and IRM ($p \geq 0.15$), where scores were within 4 percentage points. All other comparisons were significant, suggesting that the IBM and the IQM_3 produced lower recognition scores ($p < 0.001$). The patterns of significant main effects and pairwise comparisons were identical when the RAU data from experiments 1a and 1b were subjected to separate one-way repeated-measures ANOVAs, despite that the latter set of scores were all free of ceiling effects and therefore differed more widely across mask conditions.

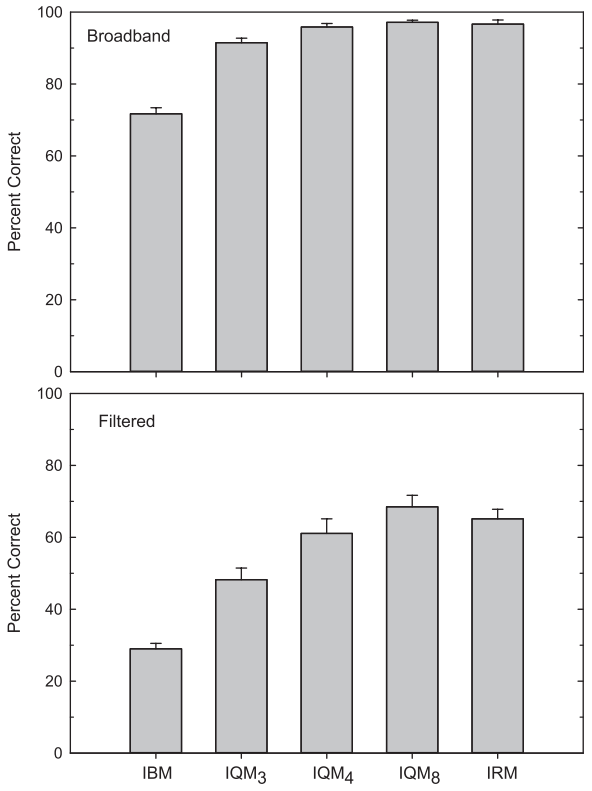

FIG. 2. Group mean W-22 word recognition (and standard errors) for NH subjects hearing speech in cafeteria noise, processed by five different T-F masks. The top panel displays scores for broadband signals and the bottom panel displays scores for a different group of subjects who heard the same signals filtered from 750 to 3000 Hz in order to avoid ceiling recognition values.

2. Acoustic intelligibility estimates

Predicted intelligibility based on the acoustic stimuli was also assessed using the standard metric, short-time objective intelligibility (STOI; Taal *et al.*, 2011). This metric reflects the correlation between the temporal amplitude envelopes of speech-plus-noise following processing and that of clean unprocessed speech. The index therefore typically ranges from 0.0 to 1.0 (although negative correlations are possible) and reflects the extent to which the envelope of the processed noisy speech reflects that of the original noise-free speech. It has been shown to be highly correlated with human speech intelligibility and is often used as an objective estimate of intelligibility.

For each mask condition, the STOI value was calculated for each of the 195 W-22 words in its carrier separately, then averaged to obtain means and variability estimates. Accordingly, standard deviations were calculated rather than standard errors because each entry in the population estimate represents a single utterance, rather than a single human subject. Figure 3 displays these STOI values for the broadband stimuli employed in experiment 1a (top panel) and the filtered stimuli employed in experiment 1b (bottom panel). Apparent is that the trend observed across conditions in Fig. 2 can also be seen in Fig. 3, except that the STOI values are somewhat similar across conditions, suggesting that they may underpredict the human speech-recognition differences observed across the five mask conditions (see Taal *et al.*, 2011, for functions mapping STOI to intelligibility). Most notable is the similarity across STOI scores observed for the experiment 1b stimuli, where ceiling effects were absent and large differences in human speech recognition were observed.

III. EXPERIMENT 2A. SOUND-QUALITY RATINGS FOR VARIOUS T-F MASKS

In this experiment, the focus was on subjective sound quality. Subjects compared utterances processed by two different T-F masks and rated which sound quality was preferred and by how much. Everyday sentences were employed in order to provide a longer duration sample to judge and a more common communication unit. Further, sentences are highly intelligible when processed by both the IBM and IRM (and so presumably by the IQM as well), removing the influence of differential intelligibility and allowing subjects to focus on sound quality. Finally, the sentence was the same across the two masks compared in each trial in order to further focus the judgment on sound quality.

A. Method

1. Subjects

Ten subjects who had not participated in experiment 1 were recruited from courses at The Ohio State University and received course credit for participating. All had normal

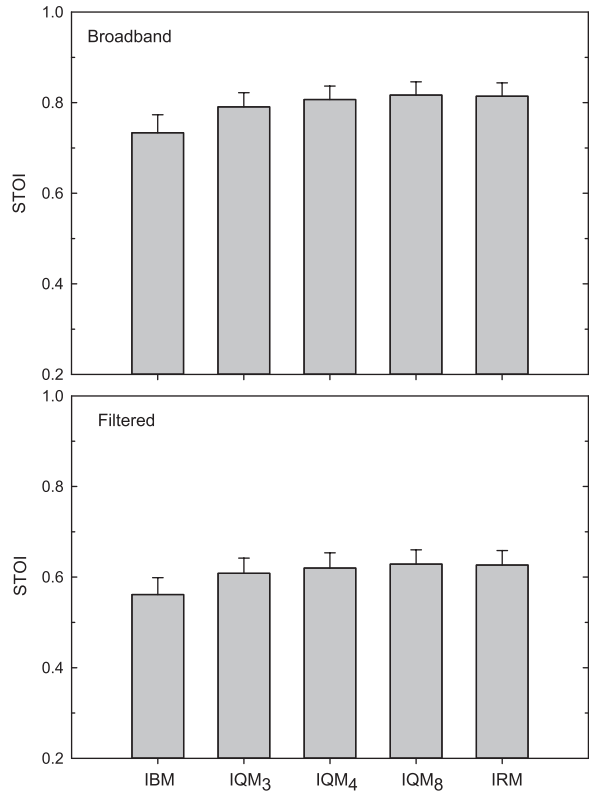

FIG. 3. STOI predictions based on the broadband acoustic stimuli employed in experiment 1a (top panel) and the filtered stimuli employed in experiment 1b (bottom panel). For each condition, STOI values were calculated for each W-22 utterance separately, and then averaged. Errors represent standard deviations.

hearing on the day of test as defined in experiment 1, ages ranged from 19 to 21 yr (average = 19.9 yr), and all were female. Care was taken to ensure that none had been previously exposed to the sentence materials employed in this experiment.

2. Stimuli

The speech stimuli employed were CID everyday American speech sentences (Silverman and Hirsch, 1955; Davis and Silverman, 1978). These 100 sentences are contextually and grammatically plausible and range in length. They were spoken by a professional male talker having a standard American English dialect and digitized at 22 kHz with 16-bit resolution. For the current experiment, sentence-length variability was reduced by selecting the 81 sentences containing 3–8 words. These were intended to provide a sound sample that was long enough to generate a clear sound-quality judgment but short enough to facilitate repetitive back-and-forth comparison. The remaining 19 sentences that were as long as 10 words or as short as 2 words were saved for practice.

The speech was mixed with the same cafeteria noise employed in experiment 1 at the same SNR of −10 dB. Each sentence was mixed with a noise segment having a different random start point in the 10-min file, two separate times, to create 162 unique mixtures. The processing of the noisy speech by the five T-F masks was identical to that employed in experiment 1a (broadband speech).

3. Procedure

The sound-quality comparison procedure was modeled after that of Madhu *et al.* (2013), Koning *et al.* (2015), and Williamson *et al.* (2015). Subjects listened to pairs of stimuli, labeled A and B, and rated their preference for one over the other based on sound quality. Each of the 5 T-F masks was compared with each of the other masks and with itself, resulting in 15 comparisons. Each comparison was made 6 times, resulting in 90 trials/subject. For each subject and trial, a sentence-plus-noise was selected randomly without replacement, and the same sentence-plus-noise was used for both masks compared within each trial. The presentation order of mask comparisons was randomized, and the assignment to position A or B was counterbalanced so that each pair appeared three times in one orientation and three times in the other.

The subjects used custom presentation software that displayed two buttons labeled A and B and a seven-point Likert-type scale (Likert, 1932) labeled from left to right, "strongly prefer A; moderately prefer A; slightly prefer A; no preference; slightly prefer B; moderately prefer B; strongly prefer B." Subjects were instructed to select how much they preferred one sentence over the other in terms of sound quality, and they were allowed to play each sentence as many times as they wished. It was suggested that they play each stimulus at least two or three times before rating. The stimuli were presented by pressing buttons A and B, and ratings were made by selecting one of the seven preferences on the scale, both using the computer mouse.

Prior to the experimental task, each subject completed practice in which each of the 15 comparisons was presented twice and the assignment to A and B was random. The practice CID sentences not used for formal testing were used for this stage. Subjects were tested while seated alone in a double-walled sound booth. As in experiment 1, stimuli were heard diotically at 65 dBA over Sennheiser HD 280 Pro headphones (Wedemark, Germany), and calibration was performed at the start of each session.

B. Results

1. Human subjects results

To quantify the sound-quality ratings, points were assigned as follows: no preference = 0; slightly prefer = 1; moderately prefer = 2; and strongly prefer = 3. Figure 4 displays the scores corresponding to each comparison averaged across subjects. Each panel corresponds to a single mask (the reference), and the columns within that panel represent the ratings for the various masks against that reference. Positive values indicate the extent to which the comparison mask was preferred over the reference, and negative values indicate the extent to which the reference was preferred. A value of 3.0 would indicate that the comparison was preferred over the reference in every trial by every subject, a value of 0.0 would indicate that no preference existed on average, and a value of −3.0 would indicate that the reference was preferred over the comparison in every trial by every subject.

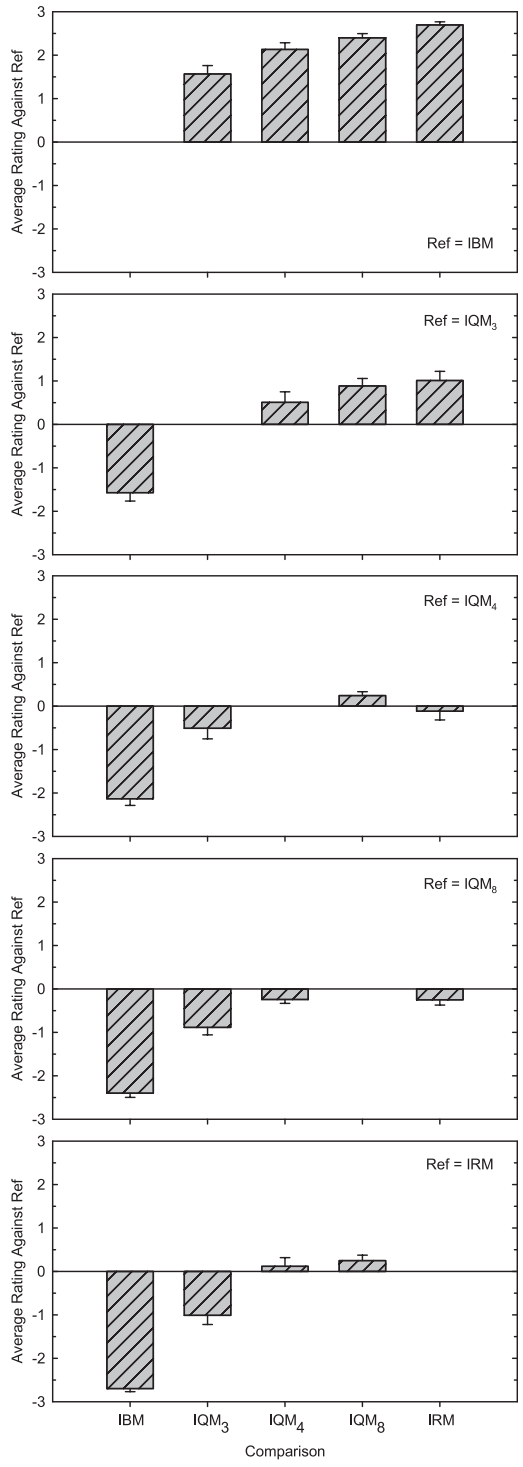

FIG. 4. Sound-quality ratings for the various T-F masks. Masks were presented in pairs and subjects rated which was preferred and by how much on a seven-point scale. "0" indicates "no preference," "1" indicates "slight preference," "2" indicates "moderate preference," and "3" indicates "strong preference." Each panel displays group mean (and standard error) ratings for each mask when compared against a given reference mask. Positive values indicate a preference for the comparison mask, and negative values indicate a preference for the reference.

It is first notable that the comparison of each mask against itself yielded a group mean rating of 0.0, corresponding to "no preference" and suggesting that the subjects were rating mask quality accurately. Second, the previously established sound-quality advantage of the IRM over the IBM is observed in these data, as the rightmost column in the top panel and the leftmost column in the bottom panel. The magnitude of this preference corresponded to "strongly prefer."

With regard to the IQM, Fig. 4 indicates that subjects preferred its sound quality over that of the IBM. This is apparent in the top panel where IQM preference values are all positive (and also in each of the IQM reference panels where the value for the IBM is negative). The magnitude of the preference was between "moderately" and "strongly prefer" for IQM_4 and IQM_8. Figure 4 also indicates that the sound quality of IQM_4 and IQM_8 matched or was slightly preferred over that of the IRM. This is apparent in the bottom panel, where the IQM_4 and IQM_8 ratings are slightly positive relative to the IRM (it can also be seen in the IQM_4 and IQM_8 reference panels where the IRM ratings are slightly negative).

Paired replicates Wilcoxon signed rank tests were conducted to compare the mean sound-quality ratings for the ten unique comparisons among T-F masks. The difference in ratings was found to be statistically significant for eight of the ten comparisons [$|W| \geq 30.0$, $p \leq 0.04$]. For each significant difference, the mask with a greater number of attenuation steps was rated as preferable over the mask having fewer steps. The sound-quality ratings did not differ significantly for the IQM_4 versus IRM [$W = 9.0$, $p = 0.55$], and for the IQM_8 versus IRM [$W = 26.0$, $p = 0.08$]. Identical analyses on medians yielded similar results (except that the intermediate comparisons between ratings for the IQM_4 versus adjacent masks IQM_3 and IQM_8 no longer differed).

Figure 5 displays the percentage of trials that were preferred when masks were compared that were adjacent along the number of attenuation steps continuum. For this analysis, 50% indicates a rating of no preference. Figure 5 shows that an increase in attenuation steps from two (IBM) to three (IQM_3) caused the sound quality of the latter to be preferred in over 95% of the comparisons. The preference proportion is reduced as comparisons involve larger numbers of attenuation steps, with IQM_4 preferred more often than IQM_3 and IQM_8 preferred slightly more often than IQM_4. But that

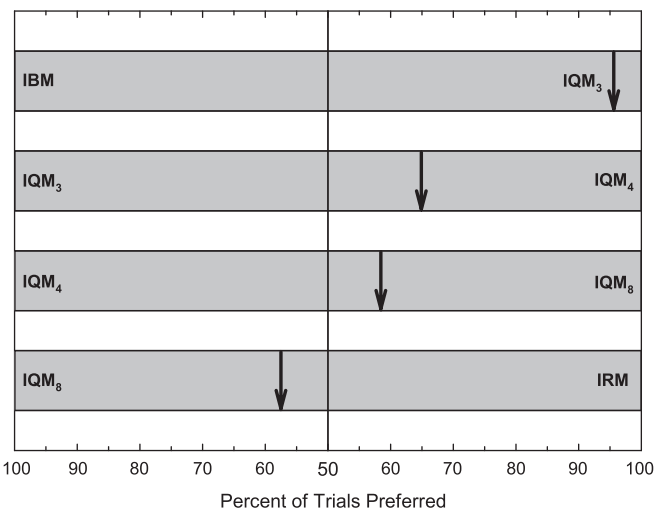

FIG. 5. Percentage of trials in which sound quality was preferred for one mask over another. The comparisons shown are for masks that are adjacent along the attenuation-step continuum. 50% reflects no preference.

trend reverses once eight attenuation steps are reached, as the sound quality of the IQM_8 was preferred slightly more often than that of the IRM.

2. Acoustic sound-quality estimates

Sound-quality estimates corresponding to the five masks were also assessed using the perceptual evaluation of speech quality (PESQ; Rix *et al.*, 2001). PESQ is a standard measure of speech sound quality based on acoustic measurement and has a scale ranging from −0.5 to 4.5. Like STOI, it reflects a comparison between speech-plus-noise following processing and clean unprocessed speech. Values were calculated for each of the CID sentence sound mixtures employed (two noises/sentence), in each of the mask conditions. Mean (and standard deviation) PESQ values are displayed in Fig. 6. Apparent is that the PESQ value increases as the number of attenuation steps exceeds two (IBM versus all IQMs), and values are highly similar for the IQM_4 and IRM. Comparisons across the scales corresponding to STOI and PESQ are difficult to make, but unlike the STOI values in Fig. 3, the PESQ values appear to display a pattern across the five mask conditions that reflects the pattern of human-subject ratings (also see the human-subject pattern in Fig. 7).

IV. EXPERIMENT 2b. CONFIRMING SIMILAR SENTENCE INTELLIGIBILITY ACROSS MASK CONDITIONS

Several steps were taken in experiment 2a to focus the judgment on subjective sound quality and control the potentially interfering influence of differential intelligibility. Those subjects participated in no intelligibility experiments, the experimenter remained outside of the sound booth to avoid exposure to another voice that could potentially influence relative sound-quality judgments, and the stimuli employed were simple sentences, which were assumed to have similar (ceiling) intelligibility across T-F mask conditions. Experiment 2b was undertaken to confirm this similarity in stimulus intelligibility.

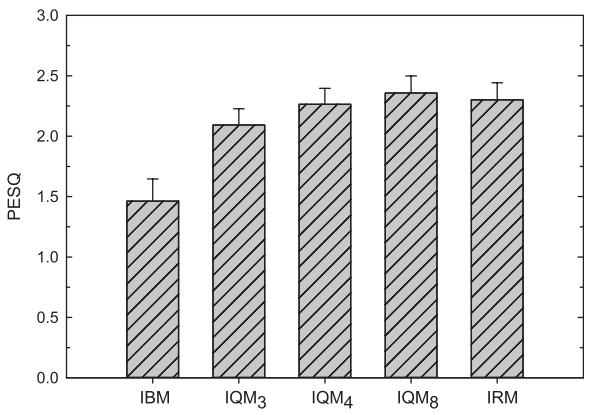

FIG. 6. PESQ estimates of sound quality based on the acoustic stimuli employed in experiment 2a. Shown are means and standard deviations for CID sentences mixed with cafeteria noise and processed by the five T-F masks.

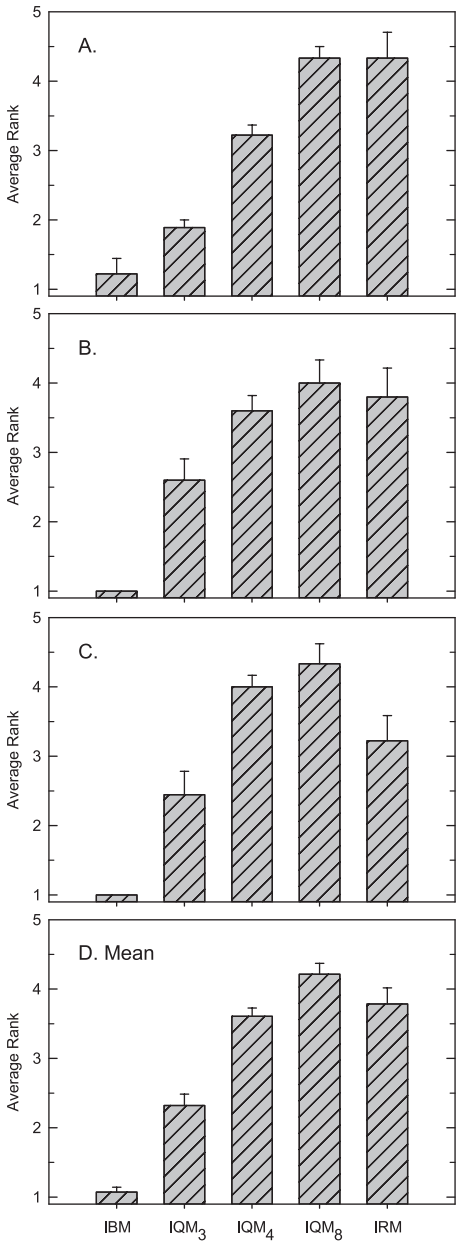

FIG. 7. Group mean subjective sound-quality rankings (and standard errors) for the five T-F masks. A ranking of "1" indicates that the sound quality was least preferred and a ranking of "5" indicates that the sound quality was most preferred. (A), (B), and (C) represent rankings produced by subjects involved in experiments 1a, 1b, and 2a, respectively. (D) displays the mean across these subgroups.

A. Method

Ten subjects were recruited from the same population as for experiments 1a, 1b, and 2a. Nine completed the current experiment after completing experiment 4 involving different speech materials. All were NH as defined in experiment 1, ages ranged from 19 to 24 yr (average = 22.0 yr), and five were female. Again, care was taken to ensure that none had been exposed to the sentence materials employed in this experiment. The stimuli were the CID sentences employed in experiment 2a, each mixed with two cafeteria-noise segments and subjected to the five T-F masks. No sentences were excluded for length in this experiment, as the inclusion of very long and very short sentences would likely only

serve to reduce intelligibility. Subjects 1–5 heard one set of mixtures (each sentence mixed with one noise segment) and subjects 6–10 heard the other set (the same sentences mixed with the other noise segment). Each subject heard 20 sentences in each of the 5 T-F mask conditions. The order of conditions was balanced such that each appeared in each serial position an equal number of times across subjects. The presentation of stimuli and collection of responses involved the same apparatus and procedures as in experiment 1. Subjects were instructed to report back each sentence after hearing it and to guess if unsure.

B. Results

The CID sentences each contain a number of scoring keywords, which generally correspond to all the content words and exclude the articles. For the current experiment, intelligibility corresponded to percentage of keywords correctly reported, with only exact matches accepted. Group-mean intelligibility was 99% for the sentences processed by the IBM and 100% for the remaining T-F masks (IQM_3, IQM_4, IQM_8, and IRM). This result confirms the high and uniform intelligibility of the stimuli employed for sound-quality judgments in experiment 2a.

V. EXPERIMENT 3. SOUND-QUALITY RANKING OF VARIOUS T-F MASKS

In this experiment, subjects ranked the T-F mask conditions in order of subjective sound quality. The same highly intelligible everyday sentence was used for each of the five masks in order to focus the judgment on quality across masks. The use of number or letter labels for the masks was avoided because they carry inherent order characteristics and, instead, each mask was assigned an arbitrary shape. Further, these shapes were arranged in a circle on the subject interface to further diminish any implication of linear ordering.

A. Method

1. Subjects

The subjects were those employed for experiments 1a, 1b, and 2a, with the exception of one subject each from experiments 1a and 1b. There were then 28 subjects, all female, aged 19 to 29 yr (average = 20.6 yr).

2. Stimuli and procedure

This experiment was completed following the other experiment that each subject participated in at the end of that same session. The stimuli were drawn from experiment 2a and so involved the CID everyday speech sentences in cafeteria noise. The first 28 sentences used for formal testing were used in order to have a different sentence for each subject. Subjects heard that one sentence mixed with a single noise sample processed by each of the five T-F masks. The labels assigned to the masks were circle, triangle, star, diamond, and square. The correspondence between shape and mask condition was randomized for each subject, but the shapes always appeared in the same position on the screen, allowing the mask position on the screen to also be randomized for each subject. Subjects played the single sentence processed by each mask by using the computer mouse to press each of five shape-labeled buttons arranged in a circle on a computer monitor. The presentation of stimuli involved the same apparatus, presentation levels, and calibration as in experiments 1 and 2. Subjects ranked the shapes in order according to the preferred sound quality of the corresponding stimulus. They did so by ordering five paper cards, each displaying one shape, on a table in front of the computer monitor labeled "best" at one end and "worst" at the other. The subjects were instructed to play each sentence as many times as desired and they were allowed to place and move the cards as they wished. The final ordering of the cards was documented by the experimenter.

B. Results and discussion

Figure 7 displays the average rank assigned to each T-F mask by each subject subgroup, with 1 being the least preferred and 5 being the most preferred. Figures 7(A), 7(B), and 7(C) correspond to the three subject groups who performed the task at the end of experiments 1a, 1b, and 2a, respectively. Figure 7(D) displays the mean across panels. Apparent is the difference in sound-quality preference ranking for the IBM versus the IRM. The IBM value equaling 1.0 in Figs. 7(B) and 7(C) indicates that it was the least preferred of the five masks for every subject, and the value just exceeding 1.0 in Fig. 7(A) reflects that it was the least preferred by all but one subject. Also apparent is the increase in sound-quality ranking as more than two attenuation steps are introduced. On average across groups, the IQM_4 ranking approximated that for the IRM. And for each subject group, the IQM_8 ranking matched or exceeded that for the IRM.

It is not simple to predict the influence that a prior task can have on judgments of subjective sound quality. This is why the current experiment was repeated with each of the subjects in experiments 1a, 1b, and 2a. In these prior experiments, subjects heard speech that was similarly (experiment 1a), differently (experiment 1b), or equally intelligible (experiment 2a) in each condition, and they focused on intelligibility (experiments 1a and 1b) or sound quality (experiment 2a). Perhaps as a result of these differing prior experimental experiences, the patterns across Figs. 7(A)–7(C) differ somewhat. Obviously, if one had to choose which pattern was most representative of the population based on statistical variability and sampling theory, the mean would be selected [Fig. 7(D)]. But it is also likely that immediately prior conditions involving intelligibility (as is often done in research of this type) can influence subsequent judgments of subjective sound quality. It is reasonable to assume that a more understandable stimulus will become "preferred" after many intelligibility trials, potentially making it difficult to assess sound quality free of this preference bias in subsequent trials. Accordingly, it is possible to speculate that the subjects from experiment 2a whose immediate prior experience with the same processing involved only judgments of sound quality of equal-intelligibility stimuli

best represent "pure" or uninfluenced subjective sound-quality judgments [Fig. 7(C)].

VI. EXPERIMENT 4. INTELLIGIBILITY PRODUCED BY VARIOUS T-F MASKS IN A HIGHLY MODULATED BACKGROUND

In this experiment, the intelligibility resulting from each of the five T-F masks was assessed for speech in a different background noise: interference consisting of a single competing talker. The rationale is twofold. First, it was of general interest to assess the ideal quantized masking of speech in a very different background type. Whereas cafeteria noise represents a most ecologically valid masker, single-talker interference represents a masker that is one of the most acoustically different—one that is characterized by far greater spectro-temporal modulation. More specific motivation involves the possibility that the different background type may influence the IRM − IBM intelligibility difference that the IQM is attempting to bridge (see Madhu *et al.*, 2013; Koning *et al.*, 2015). The stimuli in this experiment were filtered as in experiment 1b in order to eliminate ceiling effects and maximize the ability to observe intelligibility differences across the T-F masks.

A. Method

1. Subjects

A total of ten NH subjects were recruited from the population employed for experiments 1–3. Ages ranged from 19 to 24 yr (mean = 22.0 yr), and six were female. Normal hearing was defined as in experiment 1, compensation again consisted of extra course credit, and these subjects were all entirely naive to the speech materials employed.

2. Stimuli and procedure

The stimuli were highly similar to those employed in experiment 1b in order to facilitate direct comparison. The same 195 CID W-22 word recordings were employed as target stimuli. The background consisted of standard male-talker sentence recordings from the AzBio test (Spahr *et al.*, 2012). These sentences were concatenated, and a background was selected for each target utterance by selecting a segment having a random start point and the same duration as the target speech (including carrier phrase). Each of the target utterances was mixed with background interference at both −10 and −20 dB SNR to create two sets of stimuli. The motivation for the more highly negative SNR comes in part from Madhu *et al.* (2013) and Koning *et al.* (2015), who found that the IRM − IBM intelligibility difference can be larger at more negative SNRs in modulated backgrounds. The same target speech-plus-interference pairs were employed for both SNRs in order to isolate the effect of SNR. This speech-plus-noise was subjected to the five T-F masking conditions using the same processing employed in experiment 1b, including 750–3000 Hz filtering. Also as in experiment 1, the LC was 5 dB below the input SNR.

3. Procedure

The presentation of signals and testing of subjects was accomplished using the apparatus and procedures of experiment 1. Subjects were randomly assigned to one of two groups, each hearing a different overall SNR. As in experiment 1, subjects heard 13 words in each T-F mask condition in each of 3 blocks, and condition order and word list-to-condition correspondence was randomized for each subject and block. Also as in experiment 1, practice using the same 25 CNC words preceded formal testing. In the current experiment, the first five practice words were heard unfiltered, one in each T-F mask condition, followed by four words in each filtered mask condition in order of decreasing number of attenuation steps. The SNR employed for practice was the same as that employed for formal testing.

B. Results and discussion

Figure 8 displays group mean intelligibility in each T-F mask condition. Scores for the group hearing the SNR of −10 dB are displayed in the top panel and scores for the group hearing −20 dB are in the bottom panel. The pattern of scores across T-F mask conditions is highly similar to that observed in the lower panel of Fig. 2 where the background was cafeteria noise. The IRM produced higher scores than the IBM at both SNRs, the IQM_3 showed improvements over the IBM, and scores for the IQM_4 and IQM_8 matched or exceeded those observed for the IRM. It is potentially interesting to note that scores for the IQM_8 are all highly similar (~70% correct) across the different noise types and SNRs employed across experiments involving filtering, whereas scores for the IBM and IRM appear to depend more on noise type and SNR.

Scores were subjected to RAU transform and a two-way mixed ANOVA (2 SNR groups × 5 mask conditions). Both main effects {SNR: $[F(1,8) = 29.9,\ p < 0.001]$, mask: $[F(4,32) = 44.9,\ p < 0.001]$} and the interaction $[F(4,32) = 3.8,\ p < 0.05]$ were significant. *Post hoc* Holm-Sidak pairwise comparisons among the five T-F mask conditions at −10 dB SNR indicated that the IQM and IRM scores were higher than the IBM score ($p \leq 0.02$), but that they did not differ from one another ($p \geq 0.10$). Comparisons at −20 dB SNR revealed the same pattern with the addition that scores for the IQM_8 were significantly higher than those for both the IQM_3 ($p < 0.001$) and IRM ($p < 0.05$).

VII. GENERAL DISCUSSION

It is first important to note that all of the T-F masks tested currently produced vast intelligibility improvements over unprocessed speech in noise. As part of another study, 12 NH subjects heard the same W-22 word recordings in the same cafeteria noise recording at an SNR of −8 dB, but otherwise unprocessed. This formal testing employing highly similar procedures revealed a group-mean word-recognition score of 6.0% (standard error = 1.3%). Accordingly, the SNR of −10 dB employed currently should be expected to yield an unprocessed score at or near zero. The upper panel of Fig. 2 therefore shows improvements relative to that zero

baseline. But of primary interest in the current study were the relative intelligibilities and sound qualities produced by the different masks. More specifically, it was examined whether an IQM having only a very small number of attenuation steps could match the performance of a mask having an infinite number of steps.

Speech stimuli were employed that help alleviate ceiling effects, while still possessing phonetic diversity, coarticulation, and other aspects of normal human communication. (What the W-22 word lists lack are semantic content and its associated top-down processing, which may tend to increase reliance on the bottom-up acoustic nature of the acoustic signal.) Further measures to mitigate ceiling effects included band-limiting the signal to a two-octave band in the information-rich center of the speech spectrum. Whereas sentence materials tend to produce similar scores at or near 100% correct when subjected to both IBM and IRM processing, the current study revealed substantial differences between speech-recognition accuracy resulting from these two existing techniques. The top panel of Fig. 2 shows a 25%-point advantage for the IRM over the IBM, despite that the broadband IRM score is limited by a ceiling value. In the bottom panel of Fig. 2, where ceiling effects are absent, the advantage is 36% points.

Figure 2 also shows that the IQM is capable of matching or numerically exceeding the speech-recognition performance of the IRM with as few as 4–8 steps. This is apparent in both broadband (top panel) and filtered conditions (bottom panel). These data in the ecologically valid cafeteria noise are replicated for a single interfering talker in Fig. 8. This latter background represents a more special case, and in many ways a limiting case, because it represents one of our most highly modulated naturally occurring backgrounds. The same pattern of results was found, with the IQM_8 significantly outperforming the IRM at the less favorable SNR employed.

Sound quality was also assessed in the current study because it is an important consideration for a variety of applications. Most notably, HI listeners are highly sensitive to speech sound quality, and unnatural quality would serve to exacerbate the already low satisfaction and compliance with hearing-aid use (see Knudsen *et al.*, 2010). The sound-quality advantage of the ideal soft mask over the ideal hard mask is well established (see Madhu *et al.*, 2013, for comparison of ideal masks and Wang *et al.*, 2014, for comparison of algorithm-estimated masks). In the current study, subjects performed numerical ratings of sound quality and also ranked the various masks according to sound quality. It was confirmed that the sound-quality advantage of the IRM is considerable (it was "strongly preferred" over the IBM). And it was found that this advantage can be entirely captured by an IQM having as few as 4–8 steps. In fact, the numerical ratings, the proportion of preferences, the objective PESQ measure, and the rankings were all better for the IQM_8 than the IRM.

One question that may arise is how a quantized mask can numerically or significantly outperform a mask with infinite resolution. One possible explanation involves the decision to zero units at the lowest SNR step in the current IQM implementation. This makes the current IQM like the IBM in which units are zeroed and replaced with silence (noise floor of the system), but unlike the IRM in which units are attenuated but rarely zeroed. The decision may have served to improve the sound quality, and possibly even the intelligibility, of the IQM so that it could exceed that of the IRM. This implementation represents one example of how manipulation of attenuation scalars may serve to improve sound quality and/or intelligibility. Another example is provided by the work of Anzalone *et al.* (2006), who employed an IBM having scalars of 0.2 and 1 (corresponding to an attenuation difference of less than 20 dB), rather than the traditional 0 and 1, in order to avoid "musical noise."

The current work is also reminiscent of that from Loizou *et al.* (2000), who quantized the temporal amplitude envelopes of speech into 2–32 steps. It was found that speech recognition asymptoted at roughly eight steps for cochlear-implant users and for NH listeners hearing a vocoder simulation, suggesting that high amplitude resolution is not necessary. But the apparent similarity is only on the surface. Loizou *et al.* (2000) examined the amount of amplitude-envelope resolution required to understand speech in quiet when that speech was restricted to a small number of spectral channels. Although the number-of-steps results are consistent with those found currently, there is little reason to believe that results involving envelope detail for low spectral resolution speech in quiet would apply to the current investigation of T-F unit attenuation resolution required to most

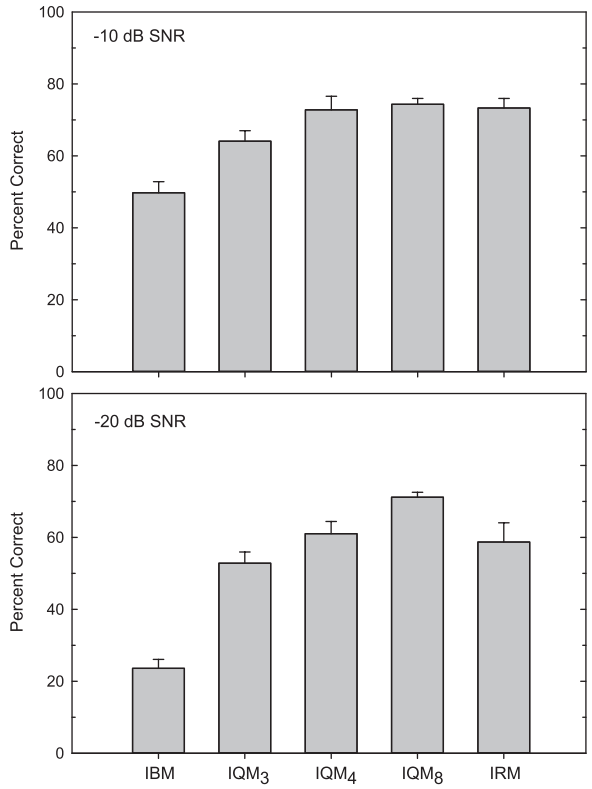

FIG. 8. Group mean W-22 word recognition (and standard errors) for NH subjects hearing speech in a single-talker background at SNRs of −10 dB (top panel) and −20 dB (bottom panel) after processing by five different T-F masks. As in the bottom panel of Fig. 2, stimuli were filtered from 750 to 3000 Hz in order to avoid ceiling recognition values.

effectively improve the intelligibility and quality of normal spectral-resolution speech in background noise.

Factors remaining to be considered include the particular mapping of IQM attenuation to SNR and the number of steps in the IQM. As indicated earlier, the current IQM mapping was based on current masks for comparison. Performance improvements may be possible with different mappings. With regard to the number of steps, the lower range was investigated in the current study. Because computational simplicity is a goal, maximum performance at a minimum number of steps is desirable. It is possible that an IQM having more than eight steps might yield even better performance, but the improvement ceiling may be limited given that the IQM_4 and IQM_8 can match or exceed the performance of the IRM. On the other hand, increasing the IQM step number above 4–8 might possibly yield better performance if the IQM is estimated, rather than calculated in ideal fashion.

The masks tested currently were all ideal. But their estimation based on only the speech-plus-noise mixture is possible using techniques of deep learning, as described in Sec. I. Two main considerations for the estimation of the masks involved in the current study include the fundamental learning scheme involved and the impact of estimation errors. As described in Sec. I, IBM estimation involves classification in which the machine learns to divide T-F units into N categories, where $N = 2$. The machine is essentially learning the single boundary. The scheme is fundamentally different during IRM estimation in which the machine learns to assign attenuation to each T-F unit. This latter scheme involves learning of the regression function underlying the continuous relationship between attenuation and SNR characteristics (based on features delivered to the algorithm). The algorithmic advantages of IQM estimation are yet to be established. But the nature of the IQM dictates that its estimation involves classification, and so corresponding advantages may be anticipated. This is especially true because the number of classification boundary values can be very low ($IQM_4 = 3$ boundaries and $IQM_8 = 7$ boundaries).

Another issue surrounding the machine estimation of these masks involves estimation errors. Healy *et al.* (2014) examined the perceptual consequences of estimated versus ideal hard masks and found that estimation by a trained deep neural network (DNN) resulted in a mask that was not deficient in any one category of speech cues but instead delivered each cue with a similar degree of imperfect accuracy. As described in Sec. I, Madhu *et al.* (2013), Wang *et al.* (2014), Koning *et al.* (2015), and Bentsen *et al.* (2018) all suggested or found that soft masks are more robust to estimation errors relative to binary masks. The explanation was that errors in the IBM are necessarily vast in magnitude (i.e., discard units that should be retained or retain units that should be discarded). But in the soft mask, errors can be large or very slight in attenuation magnitude, and so are smaller on average. The IQM likely captures this advantage as well. IQM error magnitude necessarily has a lower bound determined by the magnitude of the attenuation difference across adjacent steps. But an error of one classification step will always be smaller in the IQM (having > 2 steps) than in the IBM.

The comparisons described in the current study between the performance of the IBM versus that of the IRM largely involve ideal versions of the masks. These comparisons allow inherent properties of the masks to be assessed. But comparisons have also been made between machine-estimated versions of these masks. Wang *et al.* (2014) and Bentsen *et al.* (2018) estimated both masks using the same DNN. Wang *et al.* (2014) found that the IRM outperformed the IBM on objective measures of sound quality, and Bentsen *et al.* (2018) found that the IRM outperformed the IBM in terms of human speech intelligibility. These results are important, but multifaceted. At least two main factors are involved in the estimated IBM versus IRM comparison—those associated with inherent properties of the masks and those associated with their estimation via machine learning. The current results suggest that at least some of the IRM advantage is attributable to fundamental aspects of the mask itself. But it is also possible that some of the observed differences between estimated masks reflect different estimation-error magnitudes or other factors associated with machine estimation. Additional work is required to establish the extent to which fundamental attributes of these masks interact with estimation aspects.

VIII. CONCLUSIONS

A. Intelligibility

(1) When ceiling effects are removed, the IBM and IRM produce different intelligibilities of speech in noise, with the IRM being superior.
(2) Intelligibility benefit is observed when more than two attenuation steps are introduced to the T-F mask. Accordingly, all IQMs displayed higher intelligibility than the IBM.
(3) The intelligibility benefit of the IRM can be entirely captured with an IQM having as few as 4–8 attenuation steps (IQM_4, IQM_8). The IQM_8 numerically or significantly exceeded the intelligibility of the IRM in every condition.
(4) Conclusions (1)–(3) hold for the ecologically valid cafeteria noise (experiments 1a, 1b) and the far more highly modulated single-talker interference (experiment 4), and also across different SNRs, suggesting that the results are not restricted to a single set of conditions.
(5) The acoustic analysis STOI did not predict the magnitude of intelligibility differences observed currently across the T-F masks tested.

B. Sound quality

(1) Sound-quality ratings improve when more than two attenuation steps are applied to the T-F mask. Accordingly, all of the IQMs were rated more favorably than the IBM. The IQM_3 was preferred in over 95% of the comparisons to the IBM, and the IQM_4 and IQM_8 were "moderately" to "strongly" preferred over the IBM.

(2) The sound-quality advantage of the IRM over the IBM can be entirely captured by an IQM having as few as 4–8 attenuation steps (IQM_4, IQM_8). The numerical sound-quality ratings for the IQM_4 and IQM_8 were slightly above that for the IRM.
(3) The ranking of T-F masks from least to most preferred based on subjective sound quality also revealed that the sound-quality advantage of the IRM over the IBM can be entirely captured with as few as 4–8 attenuation steps. On average, the sound-quality rankings for the IQM_4 and IQM_8 approximated or exceeded that for the IRM.
(4) The acoustic analysis PESQ predicted the pattern of sound-quality ratings and rankings across T-F masks observed currently.
(5) It is suggested that prior exposure to conditions involving an intelligibility task can influence subsequent judgments of subjective sound quality for stimuli processed in a similar fashion. But one can speculate that this influence is diminished if the prior task involves (i) the intelligibility of stimuli all having highly similar or the same intelligibilities, or (ii) only sound-quality judgments of speech stimuli having highly similar or the same intelligibilities.

C. Computational aspects

(1) Estimation of typical T-F masks using deep learning or other means involves either classification into a small number of categories or regression, with the former potentially representing a more basic form of machine learning. IQM estimation involves classification into a small number of categories, making it like the IBM and unlike the regression-based IRM. This characteristic may allow the IQM to possess computational advantages over the IRM.
(2) Soft masks (e.g., IRM) are typically more robust to estimation errors relative to hard masks (e.g., IBM), because errors are smaller in average magnitude. The IQM may also possess this advantage over the IBM.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication Disorders (Grant No. R01 DC015521) and The Ohio State University Center for Cognitive and Brain Sciences. J.L.V. was supported in part by a fellowship from The Ohio State University Graduate School. We thank DeLiang Wang for helpful comments and for sharing the code used to generate the cochleagram representations. We also thank Victoria Sevich and Masood Delfarah for assistance with the manuscript preparation.

ANSI (**2004**). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (Acoustical Society of America, New York).
ANSI (**2010**). S3.6, *American National Standard Specification for Audiometers* (Acoustical Society of America, New York).
Anzalone, M. C., Calandruccio, L., Doherty, K. A., and Carney, L. H. (**2006**). "Determination of the potential benefit of time-frequency gain manipulation," Ear Hear. **27**, 480–492.
Bentsen, T., May, T., Kressner, A. A., and Dau, T. (**2018**). "The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility," PLoS One **13**(5), e0196924.
Brons, I., Houben, R., and Dreschler, W. A. (**2012**). "Perceptual effects of noise reduction by time-frequency masking of noisy speech," J. Acoust. Soc. Am. **132**, 2690–2699.
Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. L. (**2006**). "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," J. Acoust. Soc. Am. **120**, 4007–4018.
Chen, J., Wang, Y., Yoho, S. E., Wang, D. L., and Healy, E. W. (**2016**). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," J. Acoust. Soc. Am. **139**, 2604–2612.
Davis, H., and Silverman, S. R. (**1978**). *Hearing and Deafness*, 4th ed. (Holt, Rinehart, and Winston, New York), pp. 492–495.
Dillon, H. (**2012**). *Hearing Aids*, 2nd ed. (Boomerang, Turramurra, Australia), p. 232.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Healy, E. W., Delfarah, M., Carter, B. L., Vasko, J. L., and Wang, D. L. (**2017**). "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," J. Acoust. Soc. Am. **141**, 4230–4239.
Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. L. (**2015**). "An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type," J. Acoust. Soc. Am. **138**, 1660–1669.
Healy, E. W., Yoho, S. E., Wang, Y., Apoux, F., and Wang, D. L. (**2014**). "Speech-cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **136**, 3325–3336.
Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029–3038.
Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., and Benson, R. W. (**1952**). "Development of materials for speech audiometry," J. Speech Hear. Disord. **17**, 321–337.
Hu, G., and Wang, D. L. (**2001**). "Speech segregation based on pitch tracking and amplitude modulation," in *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 79–82.
Hummersone, C., Stokes, T., and Brooks, T. (**2014**). "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*, edited by G. R. Naik and W. Wang (Springer, Berlin), pp. 349–368.
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. L. (**2009**). "Role of mask pattern in intelligibility of ideal binary-masked noisy speech," J. Acoust. Soc. Am. **126**, 1415–1426.
Knudsen, L. V., Öberg, M., Nielsen, C., Naylor, G., and Kramer, S. E. (**2010**). "Factors influencing help seeking, hearing aid uptake, hearing aid use and satisfaction with hearing aids: A review of the literature," Trends Amplif. **14**, 127–154.
Koning, R., Madhu, N., and Wouters, J. (**2015**). "Ideal time-frequency masking algorithms lead to different speech intelligibility and quality in normal-hearing and cochlear implant listeners," IEEE Trans. Biomed. Eng. **62**, 331–341.
Lehiste, I., and Peterson, G. E. (**1959**). "Linguistic considerations in the study of speech intelligibility," J. Acoust. Soc. Am. **31**, 280–286.
Li, N., and Loizou, P. C. (**2008a**). "Effect of spectral resolution on the intelligibility of ideal binary masked speech," J. Acoust. Soc. Am. **123**, EL59–EL64.
Li, N., and Loizou, P. C. (**2008b**). "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," J. Acoust. Soc. Am. **123**, 1673–1682.
Li, Y., and Wang, D. L. (**2009**). "On the optimality of ideal binary time-frequency masks," Speech Commun. **51**, 230–239.
Likert, R. (**1932**). "A technique for the measurement of attitudes.," Arch. Psychol. **22**(140), 5–55.

Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL), Chaps. 5–8.
Loizou, P. C., Dorman, M., Poroy, O., and Spahr, T. (**2000**). "Speech recognition by normal-hearing and cochlear implant listeners as a function of intensity resolution," J. Acoust. Soc. Am. **108**, 2377–2387.
Madhu, N., Spriet, A., Jansen, S., Koning, R., and Wouters, J. (**2013**). "The potential for speech intelligibility improvement using the ideal binary mask and the ideal Wiener filter in single channel noise reduction systems: Application to auditory prostheses," IEEE Trans. Audio Speech, Lang. Process **21**, 63–72.
Monaghan, J. J. M., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C. M., and Bleeck, S. (**2017**). "Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners," J. Acoust. Soc. Am. **141**, 1985–1998.
Moore, B. C. J. (**2007**). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*, 2nd ed. (Wiley, Chichester, UK), pp. 45–91.
Narayanan, A., and Wang, D. L. (**2013**). "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7092–7096.
Rix, A., Beerends, J., Hollier, M., and Hekstra, A. (**2001**). "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 749–752.
Silverman, S. R., and Hirsh, I. J. (**1955**). "Problems related to the use of speech in clinical audiometry," Ann. Otol. Rhinol. Laryngol. **64**, 1234–1245.
Sinex, D. G. (**2013**). "Recognition of speech in noise after application of time-frequency masks: Dependence on frequency and threshold parameters," J. Acoust. Soc. Am. **133**, 2390–2396.
Spahr, A. J., Dorman, M. F., Litvak, L. M., Van Wie, S., Gifford, R. H., Loizou, P. C., Loiselle, L. M., Oakes, T., and Cook, S. (**2012**). "Development and validation of the AzBio sentence lists," Ear Hear. **33**, 112–117.
Srinivasan, S., Roman, N., and Wang, D. L. (**2006**). "Binary and ratio time-frequency masks for robust speech recognition," Speech Commun. **48**, 1486–1501.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Taal, C. H., Hendriks, R. C., 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.
Vasko, J. L., Healy, E. W., and Wang, D. L. (**2018**). "The optimal noise-rejection threshold for normal and impaired hearing," J. Acoust. Soc. Am. **143**, 1940.
Wang, D. L. (**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, Norwell MA), pp. 181–197.
Wang, D. L. (**2008**). "Time-frequency masking for speech separation and its potential for hearing aid design," Trends Amplif. **12**, 332–353.
Wang, D. L., and Brown, G., eds. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (Wiley-IEEE, Hoboken, NJ), pp. 1–44.
Wang, D. L., Kjems, U., Pedersen, M., Boldt, J., and Lunner, T. (**2009**). "Speech intelligibility in background noise with ideal binary time-frequency masking," J. Acoust. Soc. Am. **125**, 2336–2347.
Wang, Y., Narayanan, A., and Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE/ACM Trans. Audio Speech Lang. Process. **22**, 1849–1858.
Williamson, D. S., Wang, Y., and Wang, D. L. (**2015**). "Estimating nonnegative matrix model activations with deep neural networks to increase perceptual speech quality," J. Acoust. Soc. Am. **138**, 1399–1407.