An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type

Eric W. Healy[a] and Sarah E. Yoho
Department of Speech and Hearing Science, Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA

Jitong Chen and Yuxuan Wang
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

DeLiang Wang
Department of Computer Science and Engineering, Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA

(Received 30 January 2015; revised 5 August 2015; accepted 12 August 2015; published online 29 September 2015)

Machine learning algorithms to segregate speech from background noise hold considerable promise for alleviating limitations associated with hearing impairment. One of the most important considerations for implementing these algorithms into devices such as hearing aids and cochlear implants involves their ability to generalize to conditions not employed during the training stage. A major challenge involves the generalization to novel noise segments. In the current study, sentences were segregated from multi-talker babble and from cafeteria noise using an algorithm that employs deep neural networks to estimate the ideal ratio mask. Importantly, the algorithm was trained on segments of noise and tested using entirely novel segments of the same nonstationary noise type. Substantial sentence-intelligibility benefit was observed for hearing-impaired listeners in both noise types, despite the use of unseen noise segments during the test stage. Interestingly, normal-hearing listeners displayed benefit in babble but not in cafeteria noise. This result highlights the importance of evaluating these algorithms not only in human subjects, but in members of the actual target population.
[http://dx.doi.org/10.1121/1.4929493]

[DB] Pages: 1660–1669

I. INTRODUCTION

Poor intelligibility of speech in background noise remains a major complaint of hearing-impaired (HI) listeners. The difference between quiet backgrounds, in which HI listeners can perform quite well, and noisy backgrounds, in which they typically struggle, can be quite dramatic: background noises that appear to present little challenge to normal-hearing (NH) listeners can be quite debilitating to HI listeners (see Moore, 2007; Dillon, 2012). This issue is compounded by the fact that difficulty is most pronounced in noises that fluctuate over time, as most naturally occurring backgrounds do. Whereas NH listeners demonstrate better intelligibility in fluctuating relative to non-fluctuating backgrounds, HI listeners generally benefit less from this advantageous masking release (e.g., Wilson and Carhart, 1969; Festen and Plomp, 1990; Takahashi and Bacon, 1992; ter Keurs *et al.*, 1993; Eisenberg *et al.*, 1995; Bacon *et al.*, 1998; Bernstein and Grant, 2009; Oxenham and Kreft, 2014).

In accord with the ubiquitous nature and severity of the deficit, techniques to remedy poor speech reception in noise by HI listeners have been widely pursued. One approach involves monaural (single-microphone) algorithms that segregate speech from background noise and aim to increase intelligibility for HI listeners. This may be considered an ultimate goal because it is the algorithm and not the impaired listener that is tasked with extracting intelligible speech from noise. Whereas such "speech enhancement" techniques are capable of improving acoustic signal-to-noise ratio (SNR), few are capable of producing meaningful increases in intelligibility, particularly for HI listeners (for reviews, see Loizou, 2007; Healy *et al.*, 2013).

The first demonstration of substantial intelligibility increases in HI listeners was provided by Healy *et al.* (2013). In this work, sentences were segregated from background noise using an algorithm based on binary masking, a technique in which time-frequency (T-F) units are classified based on local SNR as either dominated by speech or dominated by noise, and only units dominated by speech are retained. But unlike the ideal binary mask (IBM), defined in terms of the individual premixed speech and noise signals (Hu and Wang, 2001; Wang, 2005), the algorithm in Healy *et al.* (2013) estimated the IBM using only the speech-plus-noise mixture. The algorithm employed deep neural networks (DNNs) that were trained to classify T-F units, by receiving acoustic characteristics of the speech-plus-noise mixture and the IBM. During supervised learning, the algorithm minimized the difference between the binary mask it produced and the IBM for each sentence in noise (see Wang and Wang, 2013, for technical descriptions of the algorithm).

[a]Electronic mail: healy.66@osu.edu

 0001-4966/2015/138(3)/1660/10/$30.00

Once trained, the algorithm was employed to segregate Hearing in Noise Test sentences (HINT, Nilsson *et al.*, 1994) from speech-shaped noise and multi-talker babble, at several SNRs. It was found that algorithm-processed sentences were far more intelligible than unprocessed sentences for both NH and HI listeners. The intelligibility increases were largest for HI listeners and in the babble background, which represent the target end-user population and more typical noise backgrounds. Further, the intelligibility increases were sufficient to allow the HI listeners hearing algorithm-processed sentences to outperform the NH listeners hearing unprocessed sentences, in conditions of identical noise. Thus, having the algorithm was more advantageous in these conditions than having NH.

A subsequent study (Healy *et al.*, 2014) involved recognition of isolated consonants in order to identify the specific speech cues transmitted by the algorithm and the IBM. Consonant recognition in speech-shaped noise and babble was substantially increased by the algorithm for both NH and HI listeners, despite the lack of top-down cues associated with sentence recognition and the correspondingly increased reliance on bottom-up acoustic cues. An information-transmission analysis revealed that the speech cues transmitted by the algorithm were similar to those transmitted by the IBM, indicating that the algorithm estimated the IBM with effective accuracy.

One major hurdle that must be overcome before an algorithm such as this can have direct translational impact involves the ability to generalize to noisy conditions not used during training. Indeed, the inevitable mismatch between conditions encountered during typical use and those employed during algorithm training is a common concern in supervised learning. Some steps have been taken to deal with generalizability. In the study by Healy *et al.* (2013), sentences employed during the training of the algorithm were not employed during its operation that produced stimuli used for human-subjects testing. This same novel-sentence technique was employed by Kim *et al.* (2009) and Hu and Loizou (2010) in their demonstrations of an IBM-based algorithm employing Gaussian Mixed Model (GMM) classifiers, which produced large gains in noisy sentence intelligibility for NH listeners (Kim *et al.*, 2009) and for cochlear-implant users (Hu and Loizou, 2010).

A primary step toward generalization involves the use of novel or unseen noise segments. In the work of Kim *et al.* (2009) and Hu and Loizou (2010), the same brief noise segment was used during both algorithm training and operation. The use of the same brief noise segment has the potential to substantially increase "overfit" to the training conditions, thus limiting generalizability. Healy *et al.* (2013) and Healy *et al.* (2014) used longer-duration noise (10 s) with looping, in an effort to decrease overfit and increase the potential to generalize. However, the background noise selected for each utterance was drawn from the same overall 10-s noise segment during both algorithm training and operation. Thus, the ability to generalize to novel noise segments is likely improved relative to Kim *et al.* (2009) and Hu and Loizou (2010), but it is still limited.

The generalization to novel or unseen noise segments is an obvious issue for the future goal of implementation into hearing technology, including hearing aids and cochlear implants. Because it is not possible to train the algorithm on all noises that will be encountered by a user, noise employed during training will have to be different from that encountered during the operational stage. May and Dau (2014) examined the impact of a noise mismatch between training and operational (test) stages on the performance of a GMM algorithm modeled after Kim *et al.* (2009). The accuracy of the binary mask estimated by the algorithm was assessed relative to the IBM using the classification performance metric of hit rate (HIT, percentage of correctly classified speech-dominant T-F units) minus false-alarm (FA) rate (percentage of erroneously classified noise-dominant T-F units; see Kim *et al.*, 2009). It was found that the mask estimated by the algorithm was substantially reduced in accuracy when training and test noises were drawn from different segments of the same noise type. These accuracy reductions became smaller as noise durations were increased, but they remained as large as 50 percentage points in HIT-FA rate, even at training/test noise durations of 10 s each. The study of May and Dau (2014) thus highlights the importance of training and evaluating on different segments of a noise.

The focus of the current study was to investigate the ability of a new speech-segregation algorithm to generalize to unseen segments of background noise. Assessed was the ability of the DNN algorithm to generalize from training conditions involving segments of everyday, nonstationary background noises to operation/test conditions employing entirely novel segments of the same noise type. Recognition of sentences segregated by the algorithm from multi-talker babble and from cafeteria noise was assessed in NH and HI listeners to characterize this generalizability.

II. METHOD

As described below, the algorithm tested currently differs from those employed in Healy *et al.* (2013) and Healy *et al.* (2014) in several aspects. Whereas the goal of the algorithms employed by Healy *et al.* (2013) and Healy *et al.* (2014) was to estimate the IBM, the current algorithm estimates the Ideal Ratio Mask (IRM; Srinivasan *et al.*, 2006; Narayanan and Wang, 2013; Wang *et al.*, 2014). To address the challenge of unseen noise segments, the current algorithm was trained using substantially longer noises, which were further expanded using a noise-perturbation technique (Chen *et al.*, 2015).

A. Stimuli

The stimuli were sentences drawn from the IEEE corpus (IEEE, 1969). The 44.1 kHz, 16-bit files were spoken by one male talker and each sentence contained five keywords used for scoring. Although grammatically and semantically correct, these sentences are typically considered more difficult to understand than those of other corpora (e.g., HINT sentences, Nilsson *et al.*, 1994; or Central Institute for the Deaf everyday-speech sentences, Silverman and Hirsch, 1955; Davis and Silverman, 1978). Two background noises were

employed. These were 20-talker babble and cafeteria noise, each 10 min in duration, and both from an Auditec CD (St. Louis, MO, www.auditec. com). The babble included both male and female young-adult voices. The cafeteria noise consisted of three overdubbed recordings made in a hospital employee cafeteria. It therefore contained a variety of sources, including voices, transient noises from dishes, etc. SNRs were selected to obtain scores for unprocessed sentences in noise below and above 50%. These were 0 and +5 dB for the HI subjects and −2 and −5 dB for the NH subjects. All stimuli were downsampled to 16 kHz prior to processing.

The stimulus set employed during algorithm training at each SNR included 560 IEEE sentences and noise segments randomly selected from the first 8 min of each noise. The stimulus set employed to test algorithm performance included 160 IEEE sentences not used during training and noise segments randomly selected from the remaining 2 min of each noise. New random draws of noise were employed for each SNR. An unprocessed speech-in-noise condition was prepared by simply mixing each of the 160 test sentences with a randomly selected segment of babble or cafeteria noise at the appropriate SNR. The same randomly selected noise segment used for each test sentence in the algorithm-processed condition was also used for the corresponding sentence in the unprocessed condition. Thus, the only difference between these conditions was algorithm processing.

B. Algorithm description

1. Ideal ratio mask estimation

The IRM is defined as

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

where $S(t,f)$ is the speech energy contained within T-F unit (t,f) and $N(t,f)$ is the noise energy contained within the unit (t,f). Thus, in the IRM, each T-F unit is scaled down in level according to its SNR. Units having a larger (more-favorable) SNR are attenuated less, and those having a smaller (less-favorable) SNR are attenuated more, but no units are zeroed. This makes the IRM different from the IBM. In the latter, T-F units are classified as either speech dominant or noise dominant. This determination is based on the SNR of each unit relative to a local criterion SNR (LC). Speech-dominant units (SNR > LC) are retained and unaffected, whereas noise-dominant units (SNR ≤ LC) are entirely discarded. The estimated IRM has been shown to produce slightly better objective intelligibility (based on acoustic measures) but substantially better objective sound quality than the estimated IBM (Wang *et al.*, 2014).[1]

The IRM for each sentence-plus-noise mixture was estimated from the cochleagram (Wang and Brown, 2006) of the premixed speech and noise. The cochleagram is similar to the spectrogram, but as the name implies, it has additional perceptual relevance, due in part to the use of spectral filtering that mimics the shape of the auditory filters. The cochleagram had 64 gammatone frequency channels centered from 50 to 8000 Hz and equally spaced on the $\mathrm{ERB_N}$ scale (Glasberg and Moore, 1990). It employed 20-ms time frames with 10-ms frame shifts.

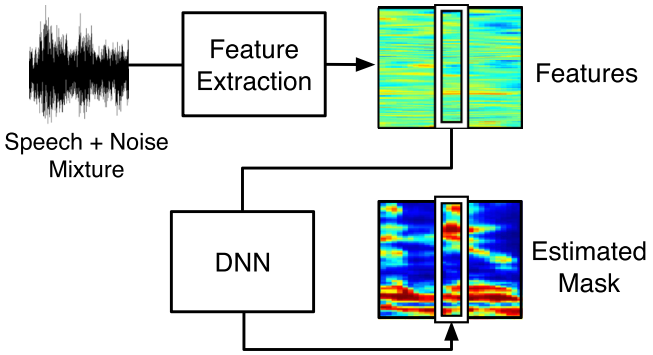

FIG. 1. (Color online) Overview of the system used to estimate the ideal ratio mask. DNN, deep neural network.

An overview of the IRM estimation system is shown in Fig. 1. First, a complementary feature set consisting of the amplitude modulation spectrogram (AMS), relative spectral transformed perceptual linear prediction (RASTA-PLP), mel-frequency cepstral coefficient (MFCC), and gammatone filterbank (GFB) features was extracted from each frame of the broadband speech-plus-noise mixture. These features have been employed previously in machine speech segregation, and their descriptions can be found in Chen *et al.* (2014). The GFB features were computed by passing the input signal to a 64-channel gammatone filterbank and downsampling the response signals to 100 Hz (see Chen *et al.*, 2014). Extraction of the remaining features followed the procedures described in Healy *et al.* (2013). These features of the speech-plus-noise mixture were then fed to a single DNN having four hidden layers and 1024 rectified linear units (Nair and Hinton, 2010) in each hidden layer. The DNN was trained to estimate the IRM using backpropagation and mean square error as the loss function. A minibatch size of 1024 and dropout ratio of 0.2 were used. The Adagrad algorithm (Duchi *et al.*, 2011) was used to adjust the learning rate. To incorporate context, five frames of features (two to each side of the current frame) were used to simultaneously predict five frames of the IRM. By incrementing frame-by-frame, each frame of the IRM was estimated five times and the average was taken as the final estimate for each frame (Wang *et al.*, 2014).

2. Noise perturbation for training-set expansion

Because the noises employed in the current study were nonstationary, there was potential mismatch between the noise segments used for algorithm training and those used for testing. As described in Sec. I, limited training will likely lead the DNN to overfit the training set, resulting in poor generalization. One technique to mitigate this issue involves the use of longer-duration noises during algorithm training, hence the current use of 8-min training noises. To further expand the training set, a noise perturbation technique (Chen *et al.*, 2015) was also employed here. This was accomplished by applying frequency perturbation to the spectrogram of the

noise in order to generate new noise samples. The procedure for frequency perturbation was as follows. A 161-band spectrogram of the noise was first computed by short time Fourier transform using a frame length of 20 ms and a frame shift of 10 ms. Each T-F unit of the spectrogram was then assigned a random value Δ drawn from the uniform distribution in the range of (-1000 to 1000). The value of Δ was then replaced by the average of the random values in a window centered at this T-F unit with the window spanning 201 time frames and 101 frequency bands. The energy $E(t, f)$ in the unit (t, f) was then replaced by $E(t, f+\Delta)$. Finally, the modified spectrogram was converted to a time-domain signal. This procedure results in a new noise that is acoustically and perceptually similar, but not identical, to the unperturbed noise. More details of the frequency-perturbation technique may be found in Chen *et al*. (2015).

Training sets were created using both original and perturbed noises. Each of the 560 training sentences was mixed with a different randomly selected noise segment 50 times, with half of the segments left unperturbed and the other half perturbed. This procedure resulted in a training set having 560×50 mixtures for each noise type and SNR. The stimuli employed for algorithm testing consisted of the 160 test sentences and 160 randomly selected noise segments drawn from the original unperturbed noise, for each noise type and SNR.

Figure 2 illustrates the results of the speech-segregation algorithm for a mixture of an IEEE sentence and babble noise at -5 dB SNR. The cochleagram of clean speech is shown in Fig. 2(a), and that of noisy speech in Fig. 2(b). The IRM is given in Fig. 2(c), and the estimated IRM in Fig. 2(d). Figure 2(e) shows the cochleagram of the speech utterance segregated from noise. From the figure, it is clear that the target speech is well separated from the babble noise and that the spectro-temporal structure of the speech is retained following segregation.

C. Subjects

Two groups of subjects were recruited. One group consisted of ten HI listeners representative of typical patients of the Speech-Language-Hearing Clinic at The Ohio State University. All were bilateral hearing-aid wearers having a sensorineural hearing loss. Ages ranged from 26 to 74 years (mean, 59.1 years), and seven were female. Hearing status was confirmed on day of test through otoscopy, tympanometry (ANSI, 1987) and pure-tone audiometry (ANSI, 2004, 2010). Pure-tone averages (PTAs, average of audiometric thresholds at 500, 1000, and 2000 Hz) ranged from 33 to 75 dB hearing level (HL) (average, 50.5). Hearing losses therefore ranged from mild to severe and were moderate on average. Configurations were flat or sloping. Audiograms obtained on day of test are presented in Fig. 3, along with subject number, age, and gender. Hearing-impaired subjects are numbered and plotted in order of increasing PTA.

The second group of subjects was composed of ten NH listeners. The NH subjects were recruited from undergraduate courses at The Ohio State University. Normal hearing was defined by audiometric thresholds at octave frequencies from 250 to 8000 Hz at or below 20 dB HL on day of test (ANSI, 2004, 2010). Ages ranged from 19 to 30 years (mean, 21.2 years) and all were female. All subjects received a monetary incentive or course credit for participating. As in

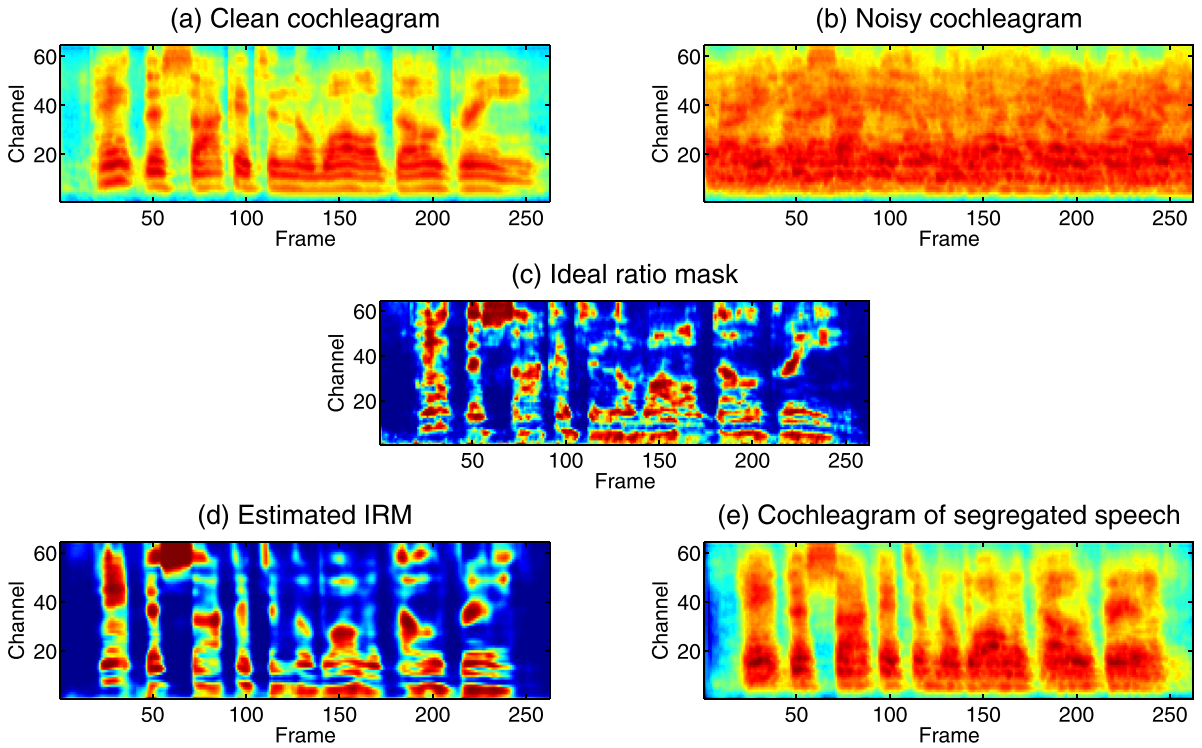

FIG. 2. (Color online) Segregation of an IEEE sentence ("Paint the sockets in the wall dull green") from babble noise at -5 dB SNR; (a) the cochleagram of the utterance in quiet; (b) the cochleagram of the speech-plus-noise mixture; (c) the IRM for this mixture; (d) the IRM for this mixture estimated by the algorithm; and (e) the cochleagram of the utterance after applying the estimated IRM to segregate the speech from noise.

our previous work on this topic (Healy *et al.*, 2013; Healy *et al.*, 2014), age matching between HI and NH subjects was not performed because the goal was to assess the abilities of typical (older) HI listeners relative to the gold-standard performance of young NH listeners.

D. Procedure

A total of eight conditions were employed (2 noise types $\times$ 2 SNRs $\times$ 2 processing conditions). Subjects heard 20 sentences in each condition for a total of 160 sentences. Sentence list-to-condition correspondence was pseudo-randomized for each subject. Noise type and SNR were blocked to allow unprocessed and algorithm conditions to appear juxtaposed in presentation order for each noise type and SNR. Half the listeners heard unprocessed prior to algorithm for each noise type and SNR and the other half heard the opposite order. Half of the subjects heard the babble conditions followed by the cafeteria-noise conditions and the other half heard the opposite order. No subjects had prior exposure to the sentence materials, and no sentence was repeated in any condition for any listener.

The total RMS level of each stimulus in each condition was equated for playback. Presentation level was 65 dBA for NH listeners and 65 dBA plus frequency-specific gains as prescribed by the NAL-R hearing-aid fitting formula (Byrne and Dillon, 1986) for each individual HI listener. The NAL-R fitting procedure employed in Healy *et al.* (2014) was also employed here. The only exception was that a RANE DEQ 60L digital equalizer (Mukilteo, WA) was used currently to shape the stimuli, rather than digital filtering in MATLAB. The signals were transformed to analog form using Echo Digital Audio Gina 3G digital-to-analog converters (Santa Barbara, CA), shaped using the RANE equalizer, routed to a Mackie 1202-VLZ mixer (Woodinville, WA) to adjust gain, and presented diotically over Sennheiser HD 280 Pro headphones (Wedemark, Germany). Hearing-impaired listeners were tested with hearing aids removed, and presentation levels were calibrated using a sound-level meter and flat-plate headphone coupler (Larson Davis models 824 and AEC 101; Depew, NY).

The subjects were seated with the experimenter in a double-walled audiometric booth. Prior to testing, a familiarization was performed during which listeners heard five IEEE sentences in quiet followed by five sentences each in the algorithm and then unprocessed conditions in either babble or cafeteria noise, corresponding to whichever noise the subject was to receive first. This familiarization was repeated half way through the experiment using the other noise type, prior to switching noise types. After presentation of the initial sentences in quiet, the HI subjects were asked if the stimuli were at a comfortable level. Two of the ten HI subjects indicated that the stimuli sounded loud, and so the experimenter reduced the presentation level by 5 dB. These individuals then judged the stimuli to be comfortable. The overall presentation level did not exceed 96 dBA for any subject. The experimenter instructed the listener to repeat back as much of each sentence as possible and controlled the presentation of each sentence.

III. RESULTS AND DISCUSSION

A. Objective acoustic measures of intelligibility

Before presenting the human-subjects results, we provide predicted intelligibility scores, in part to facilitate comparisons with the obtained listener scores and to give an objective benchmark for future segregation studies. The short-time objective intelligibly (STOI) (Taal *et al.*, 2011) score based on the mean of all 160 test sentences was employed to evaluate the speech processed by the algorithm. STOI compares the envelopes of speech segregated from noise and clean speech. First, the effect of noise perturbation was examined. It was found that the use of noise perturbation during algorithm training improved the STOI score for cafeteria noise by approximately 0.02 at negative SNRs and 0.01 at 0 dB SNR, while it decreased STOI for babble noise

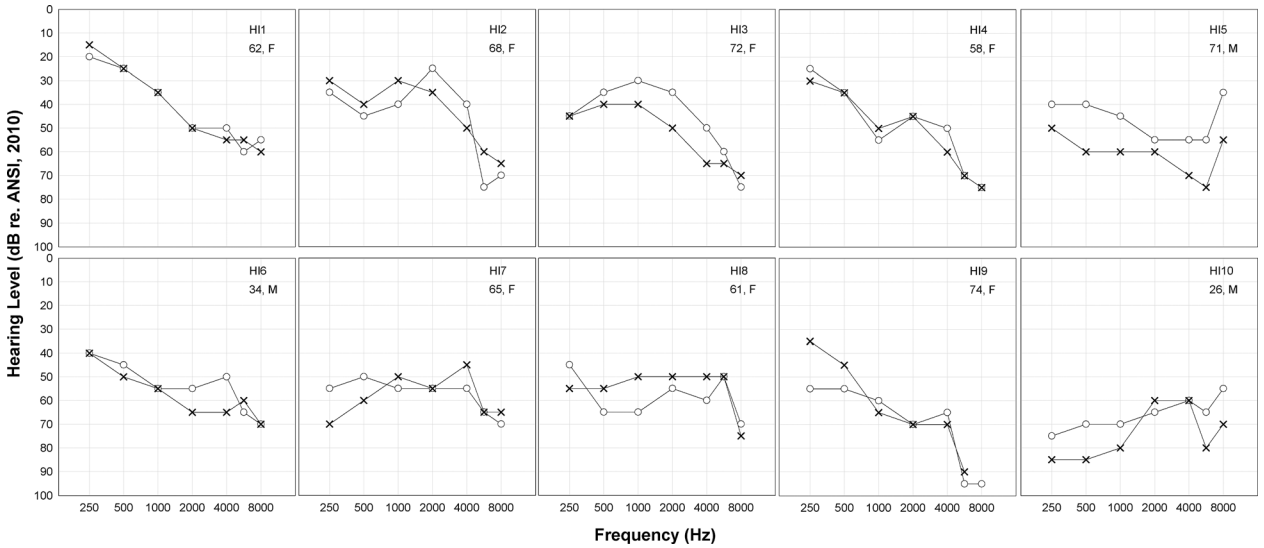

FIG. 3. Pure-tone air-conduction audiometric thresholds for the listeners with sensorineural hearing impairment. Right ears are represented by circles and left ears are represented by X's. Also displayed is subject number, listener age in years, and gender.

TABLE I. Short-time objective intelligibly (STOI) values for speech embedded in (unprocessed), and segregated from (processed), babble and cafeteria noise, at the SNRs indicated. Also shown are hit minus false-alarm (HIT − FA) and false alarm (FA) rates for the IRM estimated by the algorithm.

	Babble			Cafeteria Noise		
	Unprocessed STOI	Processed STOI	HIT − FA (FA)	Unprocessed STOI	Processed STOI	HIT − FA (FA)
5 dB	0.783	0.908	79% (12%)	0.761	0.889	84% (7%)
0 dB	0.668	0.856	78% (9%)	0.651	0.820	78% (6%)
−2 dB	0.616	0.830	76% (9%)	0.604	0.778	75% (6%)
−5 dB	0.548	0.783	72% (8%)	0.544	0.697	66% (8%)

by around 0.02 at negative SNRs and 0.01 at 0 dB SNR. Noise perturbation did not affect STOI scores at 5 dB SNR for either cafeteria or babble noise. The greater positive impact of perturbation on cafeteria noise is likely due to the fact that cafeteria noise contains strong transient components, making it more nonstationary than multi-talker babble and therefore more suitable for noise perturbation. Despite the differential impact of noise perturbation on processing of cafeteria noise versus babble, we decided to use perturbation (i.e., the same system) for both noise types to avoid the complexity of using two different systems. The STOI scores for the 160 test sentences are shown in Table I. The current algorithm substantially improves STOI scores compared with unprocessed mixtures, with increases ranging from 0.13 to 0.24 proportion points depending on SNR and noise type.

The HIT-FA rate was also computed to facilitate comparisons with binary masking algorithms (e.g., Kim *et al.*, 2009; Healy *et al.*, 2013). To compute the HIT-FA rate, the estimated IRM was converted to the estimated IBM using an LC set to be 5 dB lower than the overall mixture SNR. HIT-FA rates for the 160 test sentences are also shown in Table I. These HIT-FA values are broadly consistent with STOI scores and indicate that the system was estimating the mask with reasonable accuracy.

B. Intelligibility by human listeners

Figure 4 shows intelligibility based on percentage of keywords reported by each HI and NH listener, in each condition. Hearing-impaired listeners are represented by filled symbols and NH listeners are represented by open symbols. Scores on unprocessed speech in noise are represented by circles and scores on algorithm-processed speech are represented by triangles. Algorithm benefit for each listener is therefore represented by the height of the line connecting these symbols. As in Fig. 3, HI subjects are numbered and plotted in order of increasing PTA.

In the babble background, all HI and NH subjects received some benefit at the less favorable SNR, and all but one listener in each group received some benefit at the more favorable SNR. Benefit at the less favorable babble SNR ranged as high as 67 percentage points (HI3) and was 45 points or greater for 6 of the 10 HI listeners. Benefit at the more favorable babble SNR ranged as high as 49 percentage points (HI3) and was 30 points or greater for 7 of the 10 HI listeners. Like their HI counterparts, the NH listeners also displayed substantial benefits in babble.

In the cafeteria-noise background, all but one HI listener received some benefit at the less favorable SNR and all but two HI listeners received some benefit at the more favorable SNR. Benefit at the less favorable cafeteria-noise SNR ranged as high as 43 percentage points (HI6) and was 30 points or greater for 6 of the 10 HI listeners. Benefit at the more favorable cafeteria-noise SNR also ranged as high as 43 percentage points (HI3) and was 20 points or greater for 5 of the 10 HI listeners. In contrast to these substantial benefits observed for HI listeners, and in contrast to what was observed in babble, the listeners with NH generally did not receive benefit from algorithm processing in the cafeteria-noise background.

Group-mean intelligibility in each condition is displayed in Fig. 5. The average benefit from algorithm processing in babble was 27.8 and 44.4 percentage points for the HI listeners (at 5 and 0 dB SNR) and 21.5 and 26.8 percentage points for the NH listeners (at −2 and −5 dB SNR). A series of planned comparisons (paired, uncorrected t tests) between unprocessed and processed scores in each panel of Fig. 5 indicated that algorithm processing produced significant increases in intelligibility for both HI and NH listeners at all babble SNRs [$t(9) \geq 4.8$, $p < 0.001$].

The average benefit from algorithm processing in cafeteria noise was 18.2 and 26.9 percentage points for the HI listeners (at 5 and 0 dB SNR). Planned comparisons indicated that algorithm processing produced significant increases in intelligibility for the HI listeners at both cafeteria-noise SNRs [$t(9) \geq 3.5$, $p \leq 0.007$]. In contrast, algorithm processing resulted in numerical decreases in average intelligibility scores for the NH listeners. Benefit was −3.0 and −1.5 percentage points for the NH listeners (at −2 and −5 dB SNR). Planned comparisons indicated that scores in unprocessed and processed conditions were statistically equivalent for the NH listeners at both cafeteria-noise SNRs [$t(9) \leq 1.0$, $p \geq 0.35$].

IV. GENERAL DISCUSSION

The current study demonstrates that an algorithm designed to estimate the IRM using a trained DNN can successfully generalize to novel segments of the same type of nonstationary noise to produce substantial improvements in intelligibility in HI listeners. For these listeners, substantial benefit was observed in both babble and cafeteria noises. Benefit was largest at the least favorable SNRs and in the babble background. Benefit also tended to be greatest for the HI listeners who performed most poorly on unprocessed

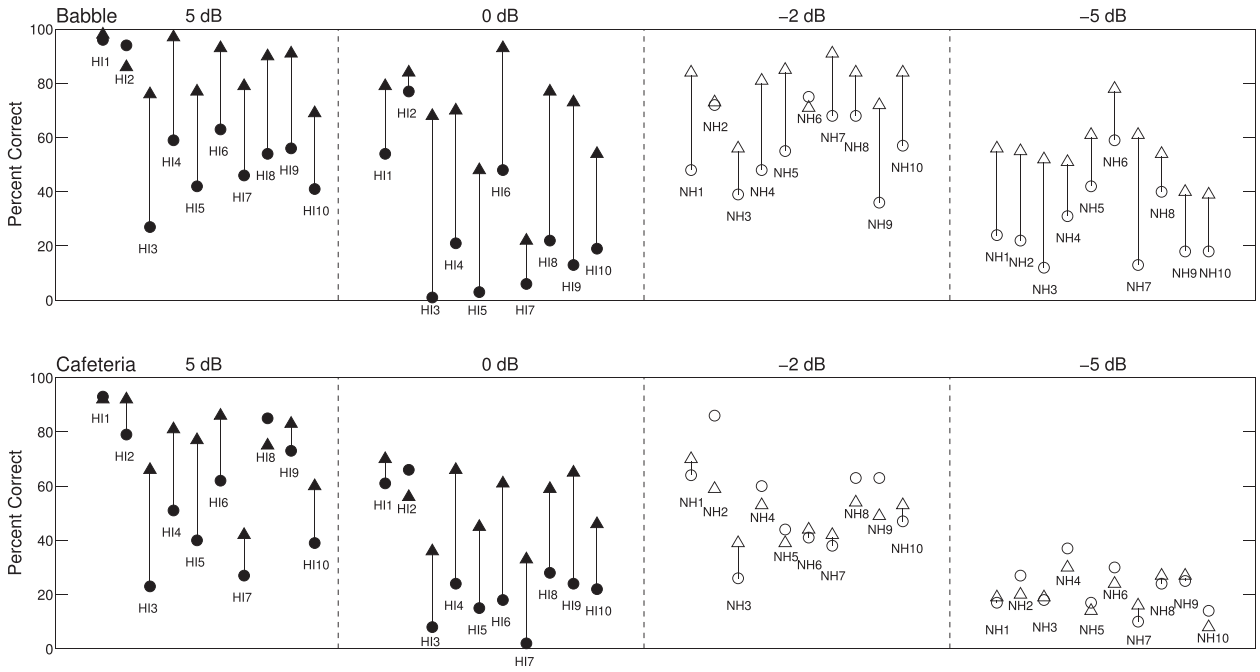

FIG. 4. Intelligibility of IEEE sentences based on percentage of keywords reported. The top panels represent scores in, or segregated from, babble noise, and the bottom panels represent scores in, or segregated from, cafeteria noise, all at the SNRs indicated. Individual HI listeners are represented by filled symbols and individual NH listeners are represented by open symbols. Scores on unprocessed speech in noise are represented by circles and scores on algorithm-processed speech are represented by triangles. Algorithm benefit is represented by the height of the line connecting these symbols.

speech in background noise. In particular, HI1 and HI2 displayed less benefit than others, primarily as a result of their high scores in the unprocessed conditions (see Fig. 4). Thus, the algorithm appears to operate most effectively for those listeners who need it most and under conditions of high-level noise where these listeners perform most poorly.

For an algorithm to have translational significance for HI listeners, it must produce gains in intelligibility under conditions of high-level noise, but not degrade intelligibility in conditions of lower-level noise where intelligibility is relatively good. To assess this, SNRs were selected in the current study to produce intelligibility of unprocessed speech in noise both below and above 50%. Figures 4 and 5 show that algorithm processing still provides performance benefit relative to unprocessed conditions when SNRs were more favorable and sentence-intelligibility scores were above 50%.

The successful generalization to unseen segments of the same noise type was undoubtedly related to the magnitude

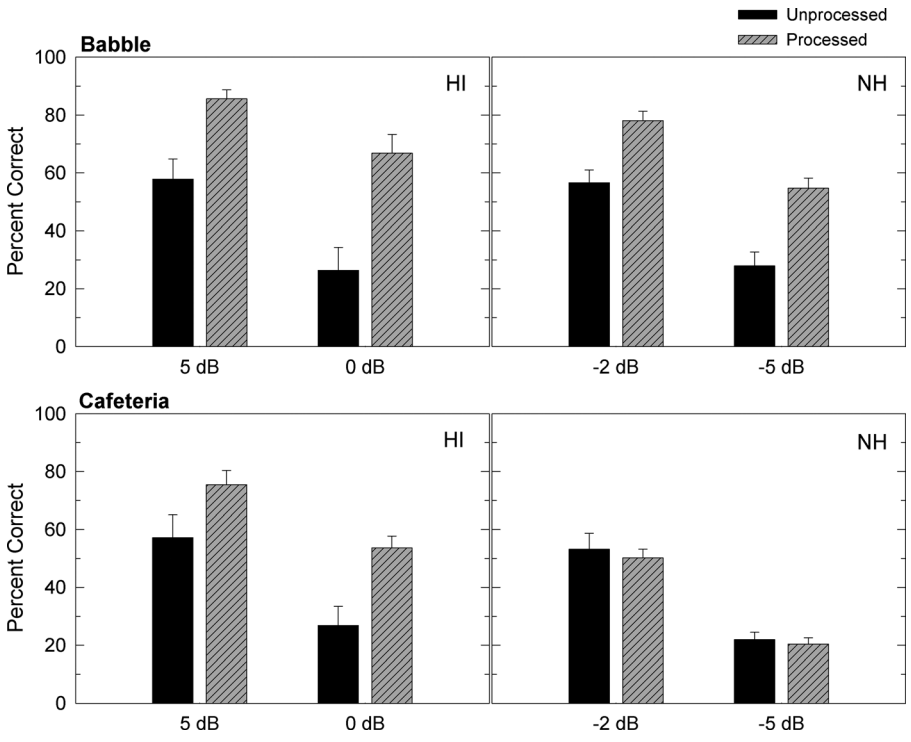

FIG. 5. Group-mean intelligibility scores and standard errors for HI and NH listeners hearing unprocessed IEEE sentences in noise and sentences following algorithm processing. The top panels show scores for a babble background, and the bottom panels show scores for a cafeteria-noise background, both at the SNRs indicated.

of training. First, to mitigate overfit to the nonstationary noise, relatively long-duration training noises were employed. Second, perturbation (Chen *et al.*, 2015) was employed to expand the set of noises employed during training. The modest modifications to the acoustic content of the training-noise segments allowed the training set to better capture the variations that can potentially exist in unseen segments of the nonstationary backgrounds.

In addition to expanded training and the use of unseen noise segments, another major difference between the current algorithm and those described previously (Healy *et al.*, 2013; Healy *et al.*, 2014) involves estimation of the IRM, rather than the IBM. In the IBM, the binary classification and discard of noise dominant T-F units leaves a signal having "holes" in its spectro-temporal pattern. In the IRM, all T-F units are retained, and each is scaled in level according to its SNR. As mentioned earlier, this ratio-masking strategy leads to significantly better objective speech quality with no loss of (even slightly better) predicted speech intelligibility (Wang *et al.*, 2014).

The amount of intelligibility improvement observed in this study appears to be somewhat lower than that observed in Healy *et al.* (2013). Speech corpora and noises used are both different across these two studies, making direct comparisons difficult. Nonetheless, for the 8-talker babble at −5 dB, for example, the improvement for NH listeners in Healy *et al.* (2013) was 35 percentage points, compared to 27 percentage points in the current study with a 20-talker babble. On the other hand, the test conditions employed currently, i.e., on segments of nonstationary noises unseen during training, are far more difficult to handle than the matched, briefer-noise conditions tested in Healy *et al.* (2013).

To assess the ability to generalize to an entirely new noise, the algorithm trained on the cafeteria noise was tested using a recording of noise from a French restaurant. This new noise recording was obtained from Sound Ideas (2015) and was created by recording the actual ambient sounds within a medium-sized restaurant in Paris. As with the cafeteria noise employed in the formal study, a 2-min segment of the French-restaurant noise was employed during testing. Although similar in a general sense, the two separate recordings represented a substantial mismatch between training and test conditions. Three of the original HI listeners (HI3, HI4, and HI6) returned for additional testing. They heard unprocessed speech in French-restaurant noise at the same SNRs employed previously (0 and 5 dB), and corresponding conditions in which the algorithm was used to extract the speech from noise, for a total of four conditions. Each heard a different set of 18 IEEE sentences in each condition, sentences that were not heard previously, and the conditions were presented in random order for each subject. All other test conditions were identical to those employed in the formal experiments. It was found that substantial intelligibility improvements remained when the new noise was tested. All subjects displayed benefit at both SNRs. At 5 dB SNR, this benefit ranged from 9 to 29 percentage points and averaged 16 percentage points (group mean percent correct = 62.6 unprocessed versus 78.5 algorithm processed). At 0 dB SNR, benefit ranged from 12 to 28 percentage points and averaged 21 percentage points (group mean percent correct = 27.0 unprocessed versus 48.2 algorithm processed). We stress that these increases were obtained with no retraining performed on the new noise at all.

One result of potential interest involves the algorithm benefit displayed by the HI versus the NH subjects in cafeteria noise. Whereas the HI subjects displayed substantial improvements in intelligibility at both SNRs, the NH subjects did not. One way to interpret this result is to recognize that the algorithm benefit we observe (Healy *et al.*, 2013; Healy *et al.*, 2014) is typically larger for HI than for NH listeners. As more challenging conditions are introduced (such as the current novel-segment cafeteria-noise conditions) and benefit is made smaller for the HI subjects, it is eventually reduced to zero for the NH subjects. The generally smaller benefit displayed by NH listeners is likely related to their remarkable ability to extract speech from background noise in challenging conditions—an ability not shared by their HI counterparts. It is the high scores achieved by NH listeners in unprocessed conditions that cause their benefit to be generally reduced. Of course, objective measures of intelligibility (e.g., STOI or HIT-FA rate) are based on acoustic analyses of the signal and do not reflect differences across human-subject types. These results underscore the importance of testing not only in human subjects, but in subjects who represent the actual target end-user population.

The human-subject results in Fig. 5 and STOI scores in Table I afford an opportunity to assess the accuracy of intelligibility prediction. STOI has been shown to be more accurate than many alternative metrics, such as the classic speech intelligibility index, and has become a standard objective speech intelligibility metric (Taal *et al.*, 2011; Yu *et al.*, 2014) for NH listeners. For the IEEE corpus with the same male speaker employed here, Taal *et al.* (2011) provide parameter values of a logistic transfer function that maps STOI scores to percent-correct numbers. After such mapping, STOI predicts, for NH listeners, improvement of 25 and 51 percentage points for babble noise (at −2 and −5 dB, respectively), and 28 and 47 percentage points for cafeteria noise (at −2 and −5 dB, respectively). Comparing these predicted recognition values and the recognition scores obtained currently shows that the STOI numbers are far off (see Fig. 6). In general, the predicted gains are much larger than the actual ones. The best match occurs for babble noise at −2 dB with a 3 percentage-point difference in terms of predicted gain, but even in this case STOI overestimates the human-subject performance for the unprocessed noisy speech. The worst case appears for the cafeteria noise at −5 dB, where STOI predicts a large improvement of 47 percentage points even though there is actually none. This assessment shows the challenge of predicting human speech intelligibility despite a considerable amount of recent work on this topic (e.g., Yu *et al.*, 2014; Kates and Arehart, 2014); see Valentini-Botinhao *et al.* (2011) for a related assessment in the context of modified speech. While we still consider the STOI metric to be a useful reference, its overestimation of intelligibility gain should be kept in mind when interpreting STOI scores. As mentioned earlier, there is no substitute

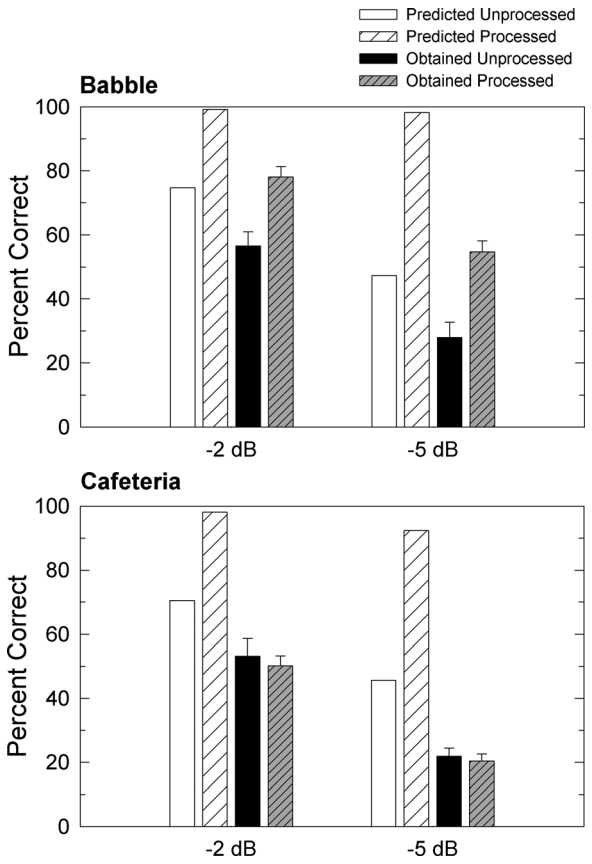

FIG. 6. Comparison of STOI-predicted versus obtained NH percent-correct IEEE sentence intelligibilities, for unprocessed and algorithm-processed speech, in the noise types and at the SNRs indicated.

for conducting actual listening tests on human subjects representative of the desired target population.

Although the preliminary results on an entirely new restaurant noise described earlier are encouraging, we point out that generalization to completely different noisy backgrounds remains to be addressed. Another limitation is that no room reverberation is considered in this study. Although we think that large-scale training with a variety of background interference, including typical room reverberation, is a promising approach (Wang and Wang, 2013), its viability has to be verified in future research. Furthermore, the input to the DNN consists of five frames of features, including two future frames. The inclusion of future frames in the input cannot be done in real-time processing, which is required for hearing aid and cochlear implant applications.

V. CONCLUSIONS

A trained DNN algorithm that estimates the IRM produced substantial improvements in sentence intelligibility for HI listeners in two types of nonstationary noise, despite training on one segment of noise and testing on different segments of the same noise type. Improvements were largest for those HI listeners who performed most poorly in background noise, in the multi-talker babble background, and at the least-favorable SNRs. This ability of the algorithm to generalize to novel segments of the same noise type likely resulted from the large training set, which consisted of a long-duration training noise and noise perturbation (Chen *et al.*, 2015) to further expand the training set. Of potential interest is the fact that, for one of the noise types, algorithm processing substantially improved intelligibility for HI listeners, but not for NH listeners. This result underscores the importance of testing in human subjects representative of actual target end users. The current results are promising for the translational significance of algorithms such as this, as the ability to generalize to novel noise backgrounds is an inescapable requirement for actual devices such as hearing aids and cochlear implants.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication Disorders (R01 DC008594 to E.W.H. and R01 DC012048 to D.L.W.) and the Air Force Office of Scientific Research (FA9550-12-1-0130 to D.L.W.). We gratefully acknowledge computing resources from the Ohio Supercomputer Center and thank Jordan Vasko for manuscript-preparation assistance.

[1]Whereas the ideal binary or ratio mask is an oracle mask, the estimated IBM or IRM is computed from only the speech + noise mixture.

ANSI (**1987**). S3.39 (R2012), *American National Standard Specifications for Instruments to Measure Aural Acoustic Impedance and Admittance (Aural Acoustic Immittance)* (Acoustical Society of America, New York).
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).
Bacon, S. P., Opie, J. M., and Montoya, D. Y. (**1998**). "The effects of hearing loss and noise masking on the masking release for speech in temporally complex backgrounds," J. Speech Lang. Hear. Res. **41**, 549–563.
Bernstein, J. G. W., and Grant, K. W. (**2009**). "Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **125**, 3358–3372.
Byrne, D., and Dillon, H. (**1986**). "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear Hear. **7**, 257–265.
Chen, J., Wang, Y., and Wang, D. L. (**2014**). "A feature study for classification-based speech separation at low signal-to-noise ratios," IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**, 1993–2002.
Chen, J., Wang, Y., and Wang, D. L. (**2015**). "Noise perturbation improves supervised speech separation," in *Proceedings of LVA/ICA*, pp. 83–90.
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), 232 pp.
Duchi, J., Hazan, E., and Singer, Y. (**2011**). "Adaptive subgradient methods for online learning and stochastic optimization," J. Machine Learning Res. **12**, 2121–2159.
Eisenberg, L. S., Dirks, D. D., and Bell, T. S. (**1995**). "Speech recognition in amplitude-modulated noise of listeners with normal and listeners with impaired hearing," J. Speech Hear. Res. **38**, 222–233.
Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.
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., 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.
Hu, Y., and Loizou, P. C. (**2010**). "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," J. Acoust. Soc. Am. **127**, 3689–3695.
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.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Kates, J. M., and Arehart, K. H. (**2014**). "The hearing-aid speech perception index (HASPI)," Speech Comm. **65**, 75–93.
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.
Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL), Chaps. 5–8.
May, T., and Dau, T. (**2014**). "Requirements for the evaluation of computational speech segregation systems," J. Acoust. Soc. Am. **136**, EL398–EL404.
Moore, B. C. J. (**2007**). *Cochlear Hearing Loss*, 2nd ed. (Wiley, Chichester, UK), pp. 201–232.
Nair, V., and Hinton, G. E. (**2010**). "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of ICML*, pp. 807–814.
Narayanan, A., and Wang, D. L. (**2013**). "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proceedings of ICASSP*, pp. 7092–7096.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
Oxenham, A. J., and Kreft, H. A. (**2014**). "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," Trends Hear. **18**, 1–14.
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.
Sound Ideas (**2015**). "Sound effects library," www.sound-ideas.com (Last viewed April 16, 2015).
Srinivasan, S., Roman, N., and Wang, D. L. (**2006**). "Binary and ratio time-frequency masks for robust speech recognition," Speech Comm. **48**, 1486–1501.
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.
Takahashi, G. A., and Bacon, S. P. (**1992**). "Modulation detection, modulation masking, and speech understanding in noise in the elderly," J. Speech Hear. Res. **35**, 1410–1421.
ter Keurs, M., Festen, J. M., and Plomp, R. (**1993**). "Effects of spectral envelope smearing on speech reception. II," J. Acoust. Soc. Am. **93**, 1547–1552.
Valentini-Botinhao, C., Yamagishi, J., and King, S. (**2011**). "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?," in *Proceedings of INTERSPEECH 2011*, pp. 1837–1840.
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, Norwell, MA), pp. 181–197.
Wang, D. L., and Brown, G. J., Eds. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (John Wiley, Hoboken, NJ), pp. 1–44.
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.
Wang, Y., and Wang, D. L. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio. Speech Lang. Process. **21**, 1381–1390.
Wilson, R. H., and Carhart, R. (**1969**). "Influence of pulsed masking on the threshold for spondees," J. Acoust. Soc. Am. **46**, 998–1010.
Yu, C., Wojcicki, K. K., Loizou, P. C., Hansen, J. H. L., and Johnson, M. T. (**2014**). "Evaluation of the importance of time-frequency contributions to speech intelligibility in noise," J. Acoust. Soc. Am. **135**, 3007–3016.