An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker

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

Masood Delfarah
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

Jordan L. Vasko[b)] and Brittney L. Carter[b)]
Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210, USA

DeLiang Wang[b)]
Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA

(Received 8 December 2016; revised 12 May 2017; accepted 14 May 2017; published online 8 June 2017)

Individuals with hearing impairment have particular difficulty perceptually segregating concurrent voices and understanding a talker in the presence of a competing voice. In contrast, individuals with normal hearing perform this task quite well. This listening situation represents a very different problem for both the human and machine listener, when compared to perceiving speech in other types of background noise. A machine learning algorithm is introduced here to address this listening situation. A deep neural network was trained to estimate the ideal ratio mask for a male target talker in the presence of a female competing talker. The monaural algorithm was found to produce sentence-intelligibility increases for hearing-impaired (HI) and normal-hearing (NH) listeners at various signal-to-noise ratios (SNRs). This benefit was largest for the HI listeners and averaged 59%-points at the least-favorable SNR, with a maximum of 87%-points. The mean intelligibility achieved by the HI listeners using the algorithm was equivalent to that of young NH listeners without processing, under conditions of identical interference. Possible reasons for the limited ability of HI listeners to perceptually segregate concurrent voices are reviewed as are possible implementation considerations for algorithms like the current one.
[http://dx.doi.org/10.1121/1.4984271]

[JFL] Pages: 4230–4239

I. INTRODUCTION

Poor speech understanding in background noise is a primary complaint of hearing-impaired (HI) and cochlear implant (CI) listeners. Work in recent years has shown that a machine-learning approach based on deep neural networks (DNNs) and time-frequency (T-F) masking can produce large increases in intelligibility for HI listeners. In this prior work, a DNN was trained to estimate the ideal binary mask (IBM) or ideal ratio mask (IRM). This is accomplished by delivering to the network during a training phase features of speech-plus-noise mixtures and the corresponding IBMs or IRMs. Once trained, the network estimates the mask using only a speech-plus-noise mixture as input.

This demonstration was provided by Healy *et al.* (2013), who had HI and normal-hearing (NH) listeners hear sentences in steady-state and modulated backgrounds prior to and following noise removal via a monaural (single-microphone) DNN-based classification algorithm. Considerable intelligibility increases were found, which were largest for the modulated background, the least-favorable signal-to-noise ratios (SNRs), and the HI listeners, particularly those displaying the poorest unprocessed performance in background noise. The average increases were sufficient to allow the HI listeners having access to the algorithm to significantly exceed the performance of young NH listeners (without the algorithm) under identical noise conditions. Subsequent work showed that the mask estimated by a simplified DNN was accurate enough to allow increases in isolated consonant recognition for both HI and NH listeners, a task that relies more heavily on the accuracy of bottom-up acoustic speech cues than does sentence recognition (Healy *et al.*, 2014). This work also showed that the estimated mask transmitted speech features (voicing, manner of articulation, and place of articulation) to listeners without any specific deficiency and in a fashion similar to that of the IBM, indicating that the algorithm was estimating the mask with effective or perceptual accuracy. This work was later extended to conditions in which the DNN was trained using one segment of a noise and tested on a novel segment of the same noise type. This condition is more challenging algorithmically than training and testing on overlapping noise segments, even when the segments are of relatively long duration (May and Dau, 2014). Increases in sentence intelligibility were again found for HI listeners in two different non-stationary noises (Healy

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

et al., 2015). Moving on to training and testing on entirely different noise types, Chen *et al.* (2016) employed large-scale training, in which the algorithm was trained on sentences mixed with 10 000 different noises of various types and tested on noises not included in this training set. Training was also conducted at a single SNR and testing conducted only at untrained SNRs. Sentence-intelligibility increases were again observed, which were largest for those HI listeners who displayed the lowest scores on unprocessed speech-in-noise. These increases were observed despite the substantial challenge associated with the highly unmatched training and test conditions and the use of subjects having milder hearing impairments than those employed in the previous studies.

These studies have employed a variety of different background noise types, from speech-shaped noise (SSN; Healy *et al.*, 2013; Healy *et al.*, 2014), to multi-talker babble containing 8–20 talkers of both genders (Healy *et al.*, 2013; Healy *et al.*, 2014; Healy *et al.*, 2015; Chen *et al.*, 2016), to recordings made in a busy hospital cafeteria (Healy *et al.*, 2015; Chen *et al.*, 2016). Thus, the backgrounds have been both steady-state and spectro-temporally complex, and in the case of cafeteria noise, have included a variety of different source types (voices, dishes, etc.). The current study extends this work to a situation in which a target talker is interfered with by a single competing talker. This "two-talker" situation is addressed currently as it represents a very different problem for both the human and machine listener.

It is known that multi-talker babble becomes acoustically similar to SSN as the number of talkers approaches infinity, and so large and similar amounts of masking are produced by these backgrounds. Conversely, babble becomes more deeply modulated and therefore less interfering when the number of interfering talkers is small (see Rosen *et al.*, 2013). Alternatively, babble containing a very small number of talkers can potentially be more interfering than babble containing a larger number of talkers due to informational masking (see Shinn-Cunningham, 2008).

In the two-talker situation, NH listeners are quite good at separating concurrent voices and attending to a target talker to understand what is being said. Thus, the interference produced by a single competing talker is far less than that produced by 2- to 8-talker babble (e.g., Miller, 1947; Brungart, 2001; Brungart *et al.*, 2006; Rosen *et al.*, 2013; Kidd *et al.*, 2016). In contrast, HI listeners perform this task more poorly. For example, when presented with concurrent vowels, HI listeners often report hearing only one vowel (Arehart *et al.*, 2005). They identify the constituent vowels with less accuracy than do NH listeners and tend to benefit less from differences in fundamental frequency (Arehart *et al.*, 1997; 2005; Summers and Leek, 1998). Related work shows that HI listeners produce similar recognition performance when target speech is masked by a single competing talker or by steady-state noise, whereas NH listeners perform substantially better in the presence of a single competing talker (e.g., Carhart and Tillman, 1970; Festen and Plomp, 1990).

One possible reason for this difficulty that HI listeners have understanding speech in the presence of a competing talker involves the fact that broadened auditory tuning smears the speech spectra, making peaks attributable to each voice less pronounced and the benefit from masker modulations smaller (e.g., ter Keurs *et al.*, 1993). Further, broad tuning causes a larger number of harmonic components to fall within single auditory filters, reducing the resolvability of harmonics and saliency of voice pitches (e.g., Culling and Darwin, 1993).

More recently, it has been suggested that temporal fine structure (TFS) plays a role in the ability to perceptually segregate concurrent sounds (e.g., Qin and Oxenham, 2003; Hopkins and Moore, 2009; Apoux and Healy, 2011, 2013; Lunner *et al.*, 2012; Apoux *et al.*, 2013; Jackson and Moore, 2013). Apoux and Healy (2013) examined sentence intelligibility in the presence of a single competing talker when the speech and background envelopes were mixed and then imposed on (a) the target-speech TFS or (b) the competing-talker TFS. It was found that these conditions produced equivalent target-sentence intelligibility, indicating that the presence of the speech TFS was not beneficial and suggesting that this TFS did not supply speech information. In a third condition (c), the envelope from the target speech and that from the competing talker were each imposed on its own TFS, and then mixed. This dual-TFS condition produced far better target intelligibility, suggesting that the presence of two TFS streams was important for segregation of the target from the background and the resulting intelligibility of the target.

These findings suggesting that the TFS is important for the perceptual segregation of concurrent sounds indicate that HI and CI listeners face particular challenges. It has been argued that HI listeners have reduced access to speech TFS (Buss *et al.*, 2004; Lorenzi *et al.*, 2006; Hopkins *et al.*, 2008; Ardoint *et al.*, 2010; Hopkins and Moore, 2011). Accordingly, a deficit in the processing of this cue would limit the ability to segregate a target voice from a competing voice and understand what is said. The difficulty faced by CI listeners is likely even greater. These individuals generally receive no speech TFS (devices having specialty TFS coding excluded) and instead hear the envelopes of incoming sounds imposed on a single pulse-train carrier. Therefore, when more than one speech source is present, the complex ensemble modulation resulting from mixed envelopes is imposed on a single TFS, which may indicate to the auditory system that only one sound is present.

The two-talker situation also represents a different challenge to speech segregation algorithms than does the speech-plus-noise situation. When the interference is non-speech noise, classification as embodied in IBM estimation provides a natural framework as speech and non-speech noise belong to two distinct classes of signals. Is the binary or ratio masking framework applicable to two-talker separation? The answer is yes, as shown in several recent studies in supervised speech separation. Huang *et al.* (2014, 2015) trained DNNs and recurrent neural networks (RNNs) to estimate the IBM and the IRM for separating two talkers. Their results show that supervised T-F masking outperforms nonnegative matrix factorization methods in two-talker separation. At about the same time, Tu *et al.* (2014) and Du *et al.* (2014) trained DNNs to map from the spectrogram of two-talker mixtures to those of individual talkers. Zhang and Wang

(2016) recently trained multiple DNNs to form a deep ensemble to perform two-talker separation. Their study also contrasts T-F masking and spectral mapping approaches in two-talker conditions. However, to our knowledge, none of the two-talker separation algorithms were tested on human listeners.

Whatever the cause of the inability of HI and CI listeners to segregate concurrent voices and understand one talker in the presence of another, an effective algorithm would render these issues moot by performing this segregation task for the listener. In the current study, a monaural DNN-based IRM estimation algorithm is introduced to increase the intelligibility of a male voice in the presence of a competing female talker. Its performance is then assessed at different SNRs in both HI and NH listeners.

II. METHOD

A. Subjects

Two groups of listeners participated, and particular care was taken to ensure that no subject in either group had any prior exposure to any of the sentences employed. The HI group was composed of ten individuals who had sensorineural hearing impairment and wore bilateral hearing aids. They were recruited from The Ohio State University Speech-Language-Hearing Clinic and were selected to represent typical patients. Accordingly, hearing losses ranged from mild to moderately severe and were moderate on average. Configurations were flat to sloping. Ages ranged from 60 to 74 years of age (mean = 69.7 years of age), and five listeners were female. Hearing status was confirmed on day of test using otoscopy, tympanometry (ANSI, 1987), and pure-tone audiometry (ANSI, 2004, 2010). Otoscopy was unremarkable and middle-ear pressures were within normal limits for all subjects. Pure-tone averages (PTAs) for each subject, based on audiometric thresholds at 500, 1000, and 2000 Hz and averaged across ears, ranged from 35 to 67 dB hearing level (HL), with a mean of 48.0. These subjects were numbered in order of increasing PTA so that higher subject number corresponded to greater mid-frequency hearing loss. Figure 1 displays these audiograms.

The NH group was composed of ten listeners having audiometric thresholds of 20 dB HL or better at octave frequencies from 250 to 8000 Hz (ANSI, 2010). The exceptions were two listeners who had thresholds of 25/30 dB HL at 250 Hz in one ear. The NH subjects were recruited from undergraduate courses at The Ohio State University and received course credit for participating. Ages ranged from 19 to 23 years of age (mean = 19.8 years of age), and all were female. Young listeners with NH were selected for the current task to represent an upper bound for human performance.

B. Stimuli

The stimuli were drawn from the Institute of Electrical and Electronics Engineers (IEEE) corpus Revised List of Phonetically Balanced Sentences (Harvard Sentences; IEEE, 1969). This set is composed of 720 grammatically and semantically correct sentences each having 5 scoring keywords. The sentences were recorded at 44.1 kHz with 16-bit resolution and were down-sampled to 16 kHz for processing and presentation. The stimuli consisted of target sentences mixed with interfering sentences. Each target sentence was mixed with only one interfering sentence, and there was no overlap between the set of target sentences and the set of interfering sentences. All target sentences were spoken by the same male talker (average fundamental frequency = 132 Hz, standard deviation = 41 Hz), and all interfering sentences were spoken by the same female talker (average fundamental frequency = 209 Hz, standard deviation = 42 Hz).

Target and interfering sentences were paired such that each was approximately equal in duration. The paired sentences had durations within 0.015 s on average and no difference between members of a pair exceeded 0.113 s. Target

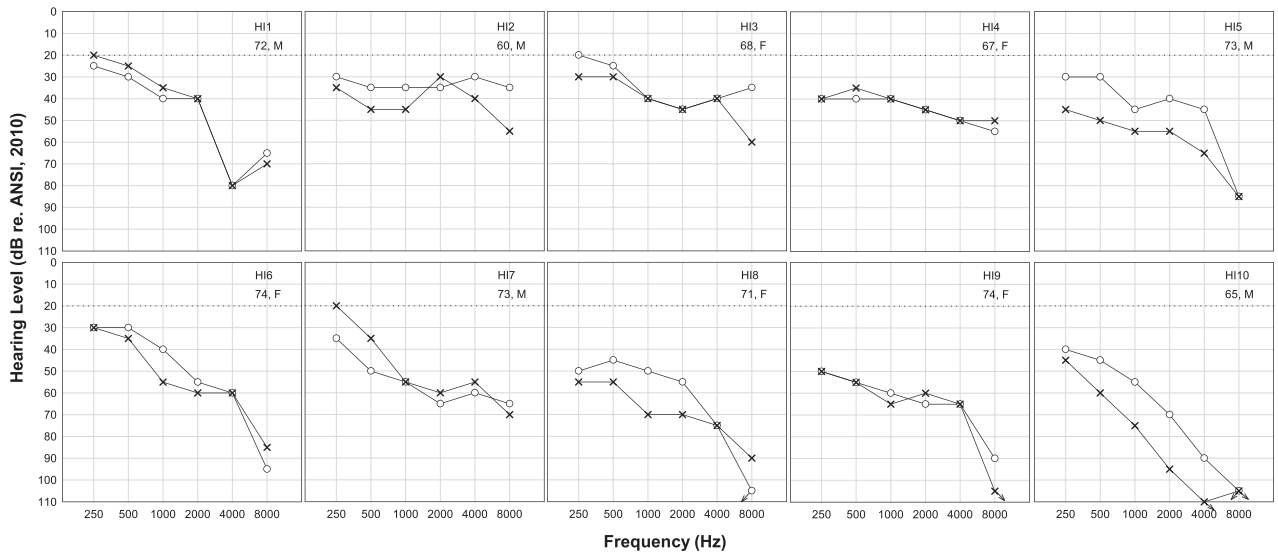

FIG. 1. Pure-tone air-conduction audiometric thresholds for the listeners with sensorineural hearing impairment. Right ears are represented by circles and left ears are represented by X's. Arrows indicate thresholds exceeding audiometer limits. Also displayed is subject number, listener age in years, and gender. The NH limit of 20 dB HL is represented by the horizontal dotted line in each panel.

signals tended to be slightly longer in duration. Mixture durations ranged from 1.781 s to 3.349 s (mean = 2.628 s, standard deviation = 0.302 s). Target and interfering sentence files were mixed so that their onsets aligned, in order to eliminate any precursor or preceding fringe containing only one voice, which could serve to facilitate segregation and inflate intelligibility. Because the materials were sentences having no fixed structure or format, the alignment of vowels and consonants across sentences was essentially random.

Stimuli employed for testing human subjects consisted of 120 target sentences mixed with 120 interfering sentences. These 240 sentences were not used during algorithm training, in order to determine how well the DNN generalized to new, unseen sentences. Each subject heard sentences mixed at three SNRs. SNRs for the HI listeners were −3, −6, and −9 dB. Those for the NH listeners were −6, −9, and −12 dB. These SNRs were selected to produce a variety of unprocessed intelligibility scores and to reduce floor and ceiling effects.

The stimuli employed for algorithm training were also drawn from the IEEE corpus (also 44.1 kHz and 16-bit, down-sampled to 16 kHz). The training set included 600 sentences spoken by the male target talker and 600 sentences spoken by the female competing talker. Two thousand mixtures were generated at each of 8 SNRs ranging from −15 dB to 6 dB in steps of 3 dB by randomly mixing one target sentence with one interfering sentence. In each mixture, the interfering sentence was either truncated or looped and repeated so that it matched the target-sentence duration. Further, the interfering-sentence start point was selected randomly for each mixture in order to produce many mixtures out of the same two sentences. Finally, the training set was down-sampled by a factor of 10 by randomly discarding 90% of all mixture time frames, in order retain diversity in the training set but reduce the amount of time required to train the DNN. Five percent of the remaining frames were set aside for the purposes of cross validation. This yielded ∼421 000, 20-ms time frames of training data. The entire training process took about 10 h on a GPU server.

C. Algorithm description

The current study employed a DNN to estimate the IRM from two-talker mixtures. To generate the IRM, the target-talker signal $s(t)$ and the interfering-talker signal $n(t)$ were divided into 20-ms time frames with 10-ms overlap, and then a Hamming window was applied to each signal. Each frame was transformed into 161 frequency bins via a short-time Fourier transform. Accordingly, each frequency bin corresponded to a bandwidth of ∼50 Hz. Magnitude spectrograms of $s(t)$ and $n(t)$ at time frame m and frequency bin c are denoted as $S(m,c)$ and $N(m,c)$. The IRM for the target talker is defined in the current study as

$$\mathrm{IRM}(m,c) = \frac{S^2(m,c)}{S^2(m,c)+N^2(m,c)}.$$

In this case, the IRM amounts to the Wiener filter, which is the optimal estimator of the target signal in the power spectrum (Wang *et al.*, 2014). The IRM for the interfering talker, 1 − IRM, was also included in the training target since this constraint acts as a regularizer during training (Huang *et al.*, 2015).

An overview of the separation process is given in Fig. 2. The process starts with the extraction of acoustic features from a speech mixture; these features were normalized to zero mean and unit variance in every dimension. After training, extracted features were fed into the trained DNN to obtain an estimate of the target IRM, denoted as RM. Estimated target magnitude $\tilde{S}(m,c)$ is calculated as

$$\tilde{S}(m,c) = \sqrt{\mathrm{RM}(m,c)} \times Y(m,c),$$

where $Y(m,c)$ is the magnitude response of the mixture signal. Mixture phase and the estimated target magnitude were used to generate the separated target speech in the time domain. Specifically, 15-dimensional (15-D) amplitude modulation spectrogram (AMS), 13-dimensional (13-D) relative spectral transformed perceptual linear prediction (RASTA-PLP), 31-dimensional (31-D) mel-frequency cepstral coefficients (MFCCs), 64-dimensional (64-D) gammatone frequency features (GF), and 31-D power-normalized cepstral coefficients (PNCCs) were extracted from each time frame and used as a feature set for DNN training. This set combines two complementary feature sets: AMS, RASTA-PLP, and MFCC revealed from Wang *et al.* (2013), and GF and PNCC from Delfarah and Wang (2016). It was found that this set of five features produced better results than an individual feature set. Detailed descriptions of these features

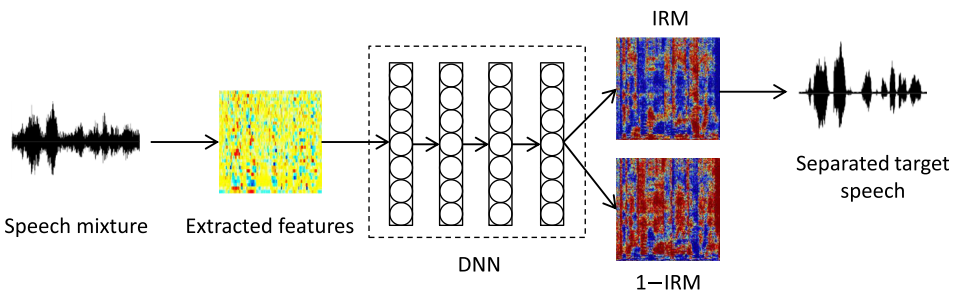

FIG. 2. (Color online) Diagram of the proposed DNN-based speech separation framework. A two-talker mixture first undergoes feature extraction. A DNN is trained using these features to estimate the IRM for the male target talker as well as the IRM for the female interfering talker. The estimated IRM for the target talker is pointwise multiplied with the magnitude spectrogram of the two-talker mixture, which results in the estimated magnitude spectrogram of the target speech. Finally, an overlap-add method is used to resynthesize the target speech signal from the estimated magnitude spectrogram and the mixture phase.

are given in these earlier studies (see also Chen *et al.*, 2014). The trained DNN can estimate one time frame of the IRM from the corresponding frame of features. However, in order to provide temporal context, a feature window of 13 feature frames was used to simultaneously predict 3 frames of the training target, all centered at the current frame (Wang *et al.*, 2014). This process was shifted frame-by-frame, allowing each time frame of the IRM to be predicted three times, and the three estimates (one for each shift) were then averaged to produce the final estimate of one IRM frame.

The DNN had an input layer with 13×154 units, 4 hidden layers of 2048 units each, and an output layer with 6×161 units. Rectified linear units (Nair and Hinton, 2010) were used for all hidden layers, and the output layer used sigmoid units. The dropout rate of 0.2 was used for regularization purposes. The training algorithm was adaptive gradient descent, and it was run for 100 epochs. These DNN parameters and choices were made on the basis of previous studies (see, e.g., Chen *et al.*, 2016) and pilot experimentation. Note that separation results were not very sensitive to these parameter values.

Compared to the earlier DNN based two-talker separation studies discussed in Sec. I, the current algorithm is distinct mainly in two aspects. First, although related, our training target differs from those used in these studies. Second, the current algorithm utilizes a larger and more diverse set of acoustic features, and benefits from larger contextual windows both in the input and the output layers of the DNN. In Huang *et al.* (2014, 2015), magnitude spectrogram and log-mel spectrogram features were used as input features with a context window of at most seven frames. Du *et al.* (2014) and Tu *et al.* (2014) used log-power spectrogram features and a context window of seven frames. Zhang and Wang (2016) also used magnitude spectrogram features but exploited contextual information using an ensemble of DNNs with different window lengths. However, all of these earlier studies predicted only one frame of the ideal mask, not three as in the current study. As a result, the current scores measured on objective metrics (presented in Sec. III B) are better.

D. Procedure

Each subject heard a total of 6 conditions (3 SNRs $\times$ 2 processing conditions), with 20 target sentences per condition. Conditions were blocked first for SNR, with the three SNRs presented in a new random order for each subject. The two processing conditions (unprocessed, algorithm processed) were heard juxtaposed within each SNR. These two conditions were randomized and balanced such that each was heard first during half of the blocks across all subjects. Condition orders were first determined for the HI subjects, then used for the NH subjects by pairing each with a randomly selected HI subject.

Stimuli were converted to analog form using an Echo Digital Audio Gina 3 G digital-to-analog converter (Santa Barbara, CA), amplified using a Mackie 1202-VLZ mixer (Woodinville, WA), and presented diotically over Sennheiser HD 280 Pro headphones (Wedemark, Germany). Hearing-impaired listeners were tested with hearing aids removed. The overall root-mean-square (RMS) level of each stimulus was set to 65 dBA in each ear for the NH subjects using a sound-level meter and flat-plate coupler (Larson Davis models 824 and AEC 101, Depew, NY). This same 65 dBA level was used for the HI subjects, with the addition of frequency-specific gains as prescribed by the NAL-R hearing-aid fitting formula (Byrne and Dillon, 1986). These gains were customized for each participant's specific hearing loss and were implemented using a RANE DEQ 60 L digital equalizer (Mukilteo, WA), as described in Healy *et al.* (2015). The total RMS level following NAL-R amplification did not exceed 99 dBA for any participant.

A brief familiarization prior to the start of the experiment involved 15 sentences spoken by the male target talker and 15 by the female interfering talker, all drawn from the pool used for algorithm training. The first five sentences were presented three times each, first in quiet (spoken by the target talker), then following algorithm processing, and finally as an unprocessed mixture with a competing sentence. This was followed by five algorithm-processed sentences, then five sentences in unprocessed mixtures. The middle SNR employed for each listener group was employed for familiarization (-6 dB for HI and -9 dB for NH). During this familiarization, the HI subjects were asked if the sound level was uncomfortably loud. No subject reported the level to be uncomfortable. Following the familiarization, subjects heard the six blocks of experimental conditions. Subjects completed the experiment while seated with the experimenter in a double-walled sound booth. They were instructed to attend to the male voice, to repeat back each sentence as best they could, and to guess if unsure of the content of the sentence. No sentence was repeated for any listener. The experimenter controlled the presentation of each stimulus and scored keywords correctly reported.

III. RESULTS AND DISCUSSION

Figure 3 illustrates the operation of the current separation algorithm at the lowest SNR of -12 dB. Spectrograms for the target signal, the interfering signal, and the mixture are given in Figs. 3(a), 3(b), and 3(c), respectively. Figure 3(d) shows the IRM, and its estimate is given in Fig. 3(f). These are aligned vertically for ease of comparison. Figure 3(e) shows the spectrogram of the separated speech. As shown in Fig. 3(e), the weaker target signal [Fig. 3(a)] is largely recovered from the mixture [Fig. 3(c)].

A. Human performance

In addition to scoring male-voice target-sentence intelligibility, the experimenter monitored the content of the interfering female-voice sentences to ensure that the subjects were not inadvertently reporting the wrong voice. Inadvertent reports did not occur, and on the rare occasion that a word from the competing sentence was reported ($\sim$1% of the unprocessed HI trails), the subject indicated knowing that it was from the wrong voice.

Figure 4 shows sentence intelligibility based on keywords correct for each subject in each condition. Scores for the individual HI listeners at each SNR are displayed in the

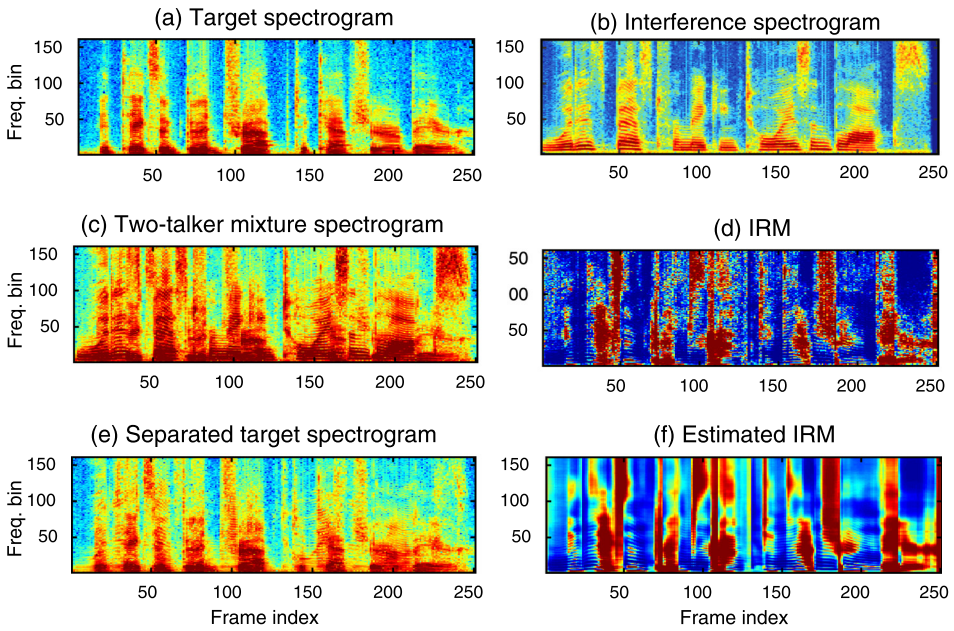

FIG. 3. (Color online) Illustration of separating an IEEE sentence uttered by a male talker ("It takes a good trap to capture a bear.") from another IEEE sentence uttered by a female talker ("Wood is best for making toys and blocks."), mixed at −12 dB SNR. (a) Spectrogram of the target utterance. (b) Spectrogram of the interfering utterance. (c) Spectrogram of the two-talker mixture. (d) IRM for the mixture. (e) Estimated target spectrogram. (f) Estimated IRM for the mixture. "Freq." indicates frequency.

top panels,[1] and those for the individual NH listeners are displayed in the bottom panels. Unprocessed scores are represented by circles and algorithm-processed scores are represented by triangles. As in Fig. 1, HI subjects are numbered and plotted in order of increasing PTA. Apparent is that performance in unprocessed conditions is very different between HI and NH listeners. Whereas the NH listeners could understand a large majority of keywords in most conditions, the HI listeners performed far more poorly. Understandably, these unprocessed scores for the HI listeners generally decrease from left to right in each panel, as degree of hearing loss increases. At the two least-favorable SNRs (−6 and −9 dB), roughly half of the HI listeners were able to report no more than 5% of the component keywords within sentences in the presence of the competing female talker.

Also apparent is that algorithm processing increased scores for the HI listeners substantially. Intelligibility for those HI listeners just described as having unprocessed scores of 5% or below, increased to average over 70% at those same two SNRs. Algorithm-processed scores were over 80% correct on average across all HI listeners and

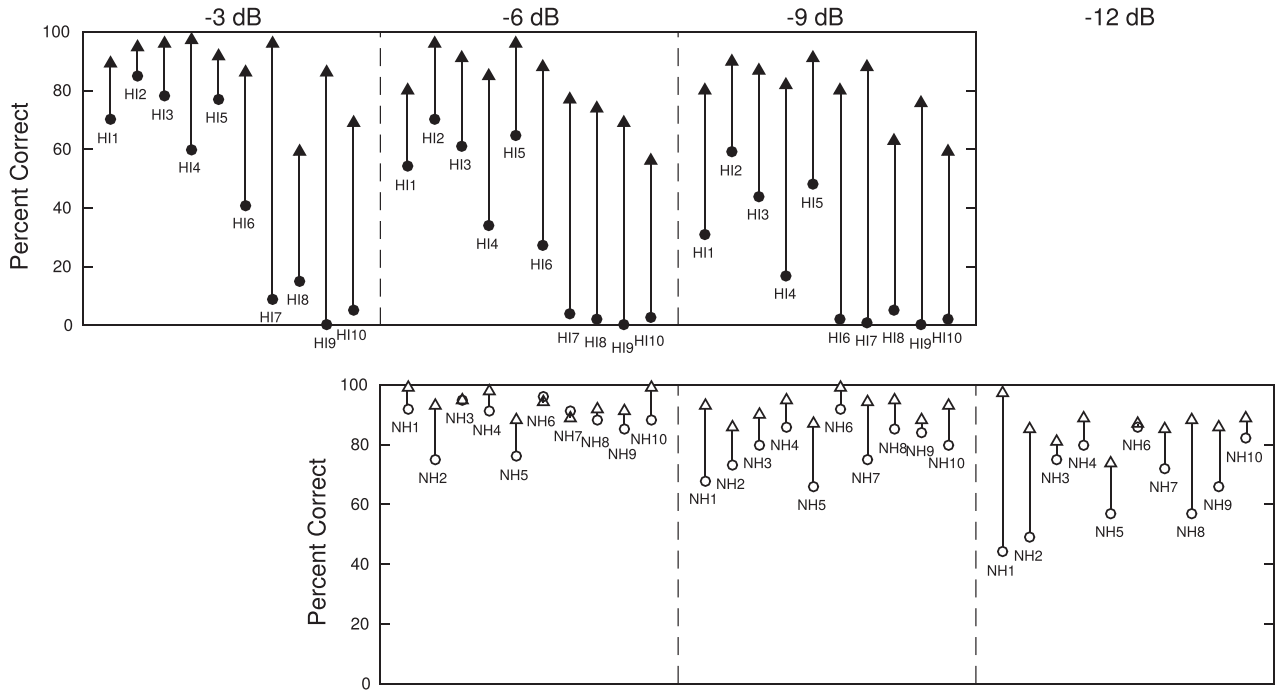

FIG. 4. Intelligibility of IEEE sentences based on percentage of keywords reported correctly. Circles represent scores in the presence of a competing talker, and triangles represent scores following algorithm processing of this mixture. Algorithm benefit is therefore represented by the height of the line connecting these symbols. Individual HI listeners are represented by filled symbols in the top panels and individual NH listeners are represented by open symbols in the bottom panels. The four SNRs employed are labeled at the top of the figure.

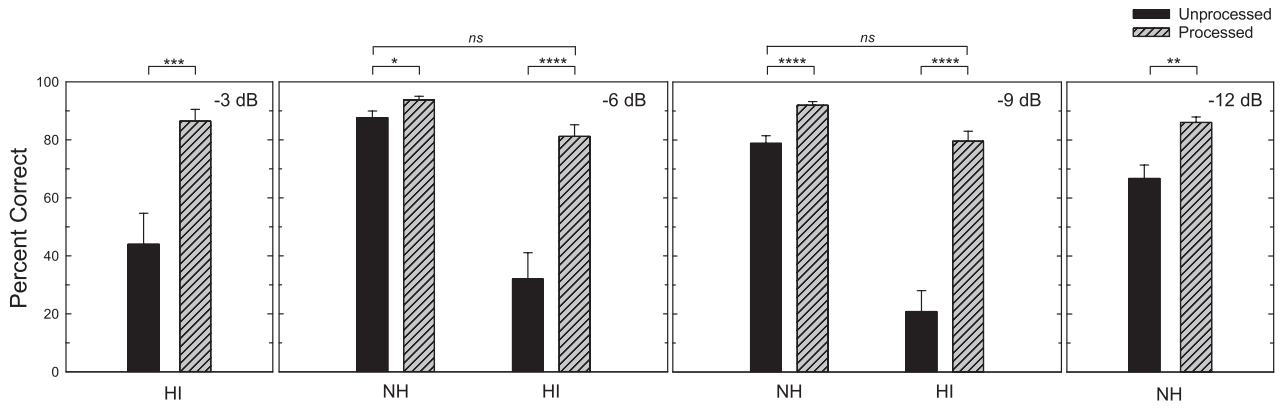

FIG. 5. Group-mean intelligibility scores and standard errors for HI and NH listeners hearing unprocessed IEEE sentences in the presence of a competing talker and sentences following algorithm processing, at the SNRs indicated. Statistical significance is indicated as follows: $*p < 0.05$, $**p < 0.005$, $***p < 0.001$, $****p < 0.00005$, ns = nonsignificant.

SNRs. Algorithm benefit was as large as 87%-points (HI 7, −3 dB and −9 dB SNR). Mean algorithm benefit exceeded 65%-points at all three SNRs for the half of the HI group having the greatest degree of hearing loss (HI 6–10). As expected, the presence of higher unprocessed scores for the better-hearing half of the HI group and the NH listeners led to smaller algorithm benefit. However, the half of the HI group having the mildest hearing loss (HI 1–5) displayed algorithm benefits of 20, 33, and 46%-points at −3, −6, and −9 dB SNR, respectively, despite unprocessed scores of 74, 57, and 40% correct. The small minority of NH scores that were around 50% correct—low enough to allow benefit to be easily observed–increased to average over 85% correct.

Figure 5 displays group mean performance in each condition for each listener group. Mean algorithm benefit was 42.5, 49.2, and 58.7%-points for the HI listeners and 6.1, 13.1, and 19.3%-points for the NH listeners, at the three SNRs heard by each group. Planned comparisons (uncorrected paired *t*-tests) were performed on RAU-transformed scores (Studebaker, 1985). Comparisons between unprocessed and algorithm-processed scores indicated that significant algorithm benefit was observed for the HI group at each SNR [$t(9) \geq 4.8$, $p < 0.001$] and for the NH group at each SNR [$t(9) \geq 3.1$, $p < 0.05$]. Further planned comparisons between the unprocessed scores for NH listeners and the processed scores for HI listeners at the SNRs common to both listener groups indicated that scores were not statistically different at −6 or −9 dB SNR [$t(18) \leq 1.3$, $p \geq 0.22$].

B. Objective measures of intelligibility

As mentioned in Sec. I, previous computational studies on two-talker separation did not use human speech intelligibility as the evaluation metric, but rather used objective measures such as SNR. To facilitate replication and future comparison, the current separation performance was also evaluated using two objective metrics: short-time objective intelligibility (STOI; Taal *et al.*, 2011) and output SNR. STOI is a standard objective measure for intelligibility with values typically ranging from 0 to 1, roughly corresponding to proportion correct. Another benefit of providing STOI results, along with human speech intelligibility results, is that such results should be valuable for improving computational models of speech intelligibility (see, e.g., Kressner *et al.*, 2016). The average objective scores for all test mixtures are shown in Table I. The results in the table show that the current DNN-based separation algorithm led to large and consistent improvements in SNR and STOI. The amount of improvement increased as the input SNR decreased. At the input SNR of −12 dB, the algorithm produced a very large SNR gain of 17.1 dB, and improved the STOI score by 40%-points.

IV. GENERAL DISCUSSION

A. Human performance

The current conditions highlight the difficulty that HI individuals have in understanding speech in the presence of a single competing talker, and represent one of the situations in which HI and NH listeners perform most differently. This vast performance difference can be seen in the unprocessed scores of Fig. 4, where the listeners with large amounts of hearing loss (HI 6–10) performed far more poorly than their milder-loss HI or NH counterparts. This variability across HI subjects on unprocessed conditions is also reflected by the error bars displayed in Fig. 5.

In contrast, the performance provided by the current monaural algorithm can be seen in the high scores of most all listeners, even at the highly unfavorable SNRs employed. The algorithm produced significant intelligibility increases for all listener groups at all SNRs, but improvements were largest for the HI listeners having the most hearing loss and correspondingly lowest unprocessed scores. Also important,

TABLE I. Average separation performance measured in SNR and short-time objective intelligibility (STOI) across all test mixtures at the different input SNRs.

Input SNR (dB)	Output SNR (dB)	Unprocessed STOI	Processed STOI
−12.00	5.10	0.435	0.835
−9.00	6.09	0.499	0.866
−6.00	7.11	0.569	0.891
−3.00	8.20	0.642	0.911

the current DNN algorithm produced no significant decreases in intelligibility. Accordingly, the algorithm appears to work for those in need, but does not hinder performance under conditions where it is unnecessary. One way to assess algorithm performance is to compare the intelligibility it affords HI listeners relative to the unprocessed performance of young NH listeners, under identical noise conditions. The current algorithm allowed the HI listeners to come on average within 6.5%-points of the NH listeners at −6 dB SNR, and to exceed the performance of NH listeners by 1%-point at −9 dB SNR (see Fig. 5).

Possible reasons for the inability of HI listeners to segregate concurrent voices and understand a target talker in the presence of a competing talker were addressed in Sec. I. Subject reports support the notion that HI subjects are unable to segregate. They all reported knowing which voice they were supposed to report, but the data show that they were largely unable to do so. Another issue surrounds listener age. The current HI subjects were selected to represent typical HI listeners and so they are older (aged $\geq$ 60 years of age). There is evidence that the use of TFS cues, a potentially important cue for segregation, declines with age (for review, see Moore, 2016). But whatever the cause for the inability to segregate voices, and the resulting poor performance of typical HI individuals, the current algorithm renders these limitations moot by performing the segregation task that the listener cannot and allowing substantial increases in intelligibility.

B. Translational potential

The translational potential of algorithms such as the current one requires two main considerations: (1) The ability to generalize to untrained situations and (2) the computational demands of the algorithm.

1. Generalization

A first consideration involves the ability of an algorithm to generalize to conditions not encountered during training. This has been a focus of the series of works by Healy, Wang, and colleagues. Thus far, our algorithms have been shown to generalize to untrained sentences (Healy *et al.*, 2013; Healy *et al.*, 2015; Chen *et al.*, 2016), untrained segments of the same noise type (Healy *et al.*, 2015), entirely novel noise types (Chen *et al.*, 2016), and untrained SNRs (Chen *et al.*, 2016). Like the current algorithm, this work has thus far involved the same target talker for training and testing. There are at least two approaches to talker generalization in the two-talker situation. In a first approach, the algorithm is trained using target speech from a frequent communication partner, such as a spouse, whereas interfering speech comes from a variety of talkers. Such an approach, called target-dependent training, has been described in supervised speech separation literature (Du *et al.*, 2014; Zhang and Wang, 2016). In a second approach, which is likely more appropriate for general noise reduction, the algorithm is trained in a talker-independent way and thus generalizes to untrained talkers. This can be potentially accomplished through large-scale training using multiple talkers. Recent work has also demonstrated that a RNN represents an effective approach for talker-independent segregation (Chen and Wang, 2016).

2. Computational aspects

A second consideration involves processing delays and load. With regard to algorithm training, processing delays are largely inconsequential, because training is completed prior to algorithm operation. For the same reason, processing load associated with training, as well as training duration, is inconsequential. Large-scale training, while costly, represents an effective method to achieve generalizability (Chen *et al.*, 2016). In contrast, the processing delays and load encountered during operation are critical considerations for translation. It is important to note that the works by Healy *et al.*, have targeted effectiveness. But the implementation considerations are also clear.

Processing delays during operation of an algorithm implemented for telecommunications are less of a concern because brief delays are not problematic. But an algorithm optimized for a hearing aid or CI will need to limit group delays to roughly 20–30 ms, corresponding to values that do not disrupt various aspects of aided speech perception, including auditory-visual synchrony (see Stone and Moore, 1999). In terms of algorithm design, this affects the time windows from which features are extracted from a noisy input and fed to the trained DNN. One way to limit processing delay is to employ past time frames and the current time frame, and to limit the use of future time frames to within the tolerable delay values. Our analysis with the aforementioned RNN (Chen and Wang, 2016) indicates that it benefits little beyond the first few future frames.

Processing load encountered during operation and algorithm complexity are again less of a concern in a telecommunication application, where the processing can be handled by centralized equipment or by the users' phone. But here again, hearing technology is seemingly different. The processing power of the behind-the-ear processor is extremely limited, and it is tempting to view this as the platform constraining operational complexity. One approach would be to reduce algorithm complexity (e.g., number of layers and units/layer) in order to run on the limited hardware platform. But this would likely negatively affect algorithm performance.

A likely superior solution exists. This solution involves hearing aid or CI processing by a smartphone-type device that is carried by the listener and transmits bi-directionally to and from an earpiece worn by the listener. This solution has the advantages of both the vastly superior processing power and battery life of the smartphone processor, and the extremely small size of a wireless microphone/speaker earpiece. This technology already exists. In fact, current smartphones offer tremendous computing power and bi-directional transmission to and from wireless earpieces.

V. CONCLUSIONS

The current work demonstrates the particular difficulty that HI listeners have in understanding speech in the presence of a competing talker—a task that is handled

effectively by NH listeners. Possible reasons for this difficulty are addressed, and an algorithm based on a trained DNN is introduced to deal with this situation. The DNN was trained using a novel set of features extracted from the speech signals. The IRMs estimated for two concurrent voices were used to segregate a male voice from the mixture of a male and female voice. Increases in sentence intelligibility were observed following algorithm processing at a variety of SNRs. These were largest for the HI listeners, with an average of 59%-points at the least-favorable SNR and a maximum of 87%-points. The increases afforded by the current algorithm allowed the HI listeners to perform equivalently to young NH listeners (without the algorithm) in conditions of identical background interference. To our knowledge, this is the first monaural (single-microphone) algorithm that provides substantial speech intelligibility improvements for HI listeners in the presence of interfering speech. Finally, the somewhat different implementation considerations that exist for telecommunications versus hearing aids and CIs are discussed.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication Disorders (Grant No. R01 DC015521 to E.W.H. and Grant No. R01 DC012048 to D.L.W.) and the Air Force Office of Scientific Research (Grant No. FA9550-12-1-0130 to D.L.W.). We gratefully acknowledge computing resources from the Ohio Supercomputer Center.

[1]One subject (HI 9) indicated that she was uncomfortable listening to unprocessed sentences in the presence of the competing talker and therefore did not complete these conditions in their entirety. This subject had the second-poorest PTA and reported being unable to hear any target words even when informed of the content of the target sentence during practice. This subject's scores in the unprocessed conditions were therefore based on the maximum number of sentence presentations that could be tolerated. This included 20 presentations at −6 dB SNR, 8 presentations at −3 dB, and 3 presentations at the least-favorable −9 dB, which was presented last. No responses to these conditions were provided at any point during the experiment. To assess the impact of including this subject, algorithm benefit was recalculated after excluding her and was found to have changed by less than 3%-points on average across the three SNRs, and by less than 5%-points at all SNRs. Further, no statistical conclusion was changed.

ANSI (**1987**). S3.39 (R2012), *American National Standard Specifications for Instruments to Measure Aural Acoustic Impedance and Admittance* (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).

Apoux, F., and Healy, E. W. (**2011**). "Relative contribution of target and masker temporal fine structure to the unmasking of consonants in noise," J. Acoust. Soc. Am. **130**, 4044–4052.

Apoux, F., and Healy, E. W. (**2013**). "A glimpsing account of the role of temporal fine structure information in speech recognition," in *Basic Aspects of Hearing: Physiology and Perception*, edited by B. C. J. Moore, R. Patterson, I. M. Winter, R. P. Carlyon, and H. E. Gockel (Springer, New York), pp. 119–126.

Apoux, F., Yoho, S. E., Youngdahl, C. L., and Healy, E. W. (**2013**). "Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners," J. Acoust. Soc. Am. **134**, 2205–2212.

Ardoint, M., Sheft, S., Fleuriot, P., Garnier, S., and Lorenzi, C. (**2010**). "Perception of temporal fine-structure cues in speech with minimal envelope cues for listeners with mild-to-moderate hearing loss," Int. J. Audiol. **49**, 823–831.

Arehart, K. H., King, C. A., and McLean-Mudgett, K. S. (**1997**). "Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss," J. Speech Lang. Hear. Res. **40**, 1434–1444.

Arehart, K. H., Rossi-Katz, J., and Swensson-Prutsman, J. (**2005**). "Double-vowel perception in listeners with cochlear hearing loss: Differences in fundamental frequency, ear of presentation, and relative amplitude," J. Speech Lang. Hear. Res. **48**, 236–252.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

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.

Buss, E., Hall, J. W. III, and Grose, J. H. (**2004**). "Temporal fine-structure cues to speech and pure tone modulation in observers with sensorineural hearing loss," Ear Hear. **25**, 242–250.

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.

Carhart, R., and Tillman, T. W. (**1970**). "Interaction of competing speech signals with hearing losses," Arch. Otolaryng. **91**, 273–279.

Chen, J., and Wang, D. L. (**2016**). "Long short-term memory for speaker generalization in supervised speech separation," in *Proceedings of INTERSPEECH*, pp. 3314–3318.

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. Proc. **22**, 1993–2002.

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.

Culling, J. F., and Darwin, C. J. (**1993**). "Perceptual separation of simultaneous vowels: Within and across-formant grouping by F_0," J. Acoust. Soc. Am. **93**, 3454–3467.

Delfarah, M., and Wang, D. L. (**2016**). "A feature study for masking-based reverberant speech separation," in *Proceedings of INTERSPEECH*, pp. 555–559.

Du, J., Tu, Y., Xu, Y., Dai, L.-R., and Lee, C.-H. (**2014**). "Speech separation of a target speaker based on deep neural networks," in *Proceedings of ICSP*, pp. 473–477.

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.

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.

Hopkins, K., and Moore, B. C. J. (**2009**). "The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise," J. Acoust. Soc. Am. **125**, 442–446.

Hopkins, K., and Moore, B. C. J. (**2011**). "The effects of age and cochlear hearing loss on temporal fine structure sensitivity, frequency selectivity, and speech reception in noise," J. Acoust. Soc. Am. **130**, 334–349.

Hopkins, K., Moore, B. C. J., and Stone, M. A. (**2008**). "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," J. Acoust. Soc. Am. **123**, 1140–1153.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (**2014**). "Deep learning for monaural speech separation," in *Proceedings of ICASSP*, pp. 1562–1566.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (**2015**). "Joint optimization of masks and deep recurrent neural networks for

monaural source separation," IEEE/ACM Trans. Audio Speech Lang. Proc. **23**, 2136–2147.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Jackson, H. M., and Moore, B. C. J. (**2013**). "Contribution of temporal fine structure information and fundamental frequency separation to intelligibility in a competing-speaker paradigm," J. Acoust. Soc. Am. **133**, 2421–2430.
Kidd, G., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., and Best, V. (**2016**). "Determining the energetic and informational components of speech-on-speech masking," J. Acoust. Soc. Am. **140**, 132–144.
Kressner, A. A., May, T., and Rozell, C. J. (**2016**). "Outcome measures based on classification performance fail to predict the intelligibility of binary-masked speech," J. Acoust. Soc. Am. **139**, 3033–3036.
Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (**2006**). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," Proc. Natl. Acad. Sci. U.S.A. **103**, 18866–18869.
Lunner, T., Hietkamp, R. K., Andersen, M. R., Hopkins, K., and Moore, B. C. J. (**2012**). "Effect of speech material on the benefit of temporal fine structure information in speech for young normal-hearing and older hearing-impaired participants," Ear Hear. **33**, 377–388.
May, T., and Dau, T. (**2014**). "Requirements for the evaluation of computational speech segregation systems," J. Acoust. Soc. Am. **136**, EL398–EL404.
Miller, G. A. (**1947**). "The masking of speech," Psych. Bull. **44**, 105–129.
Moore, B. C. J. (**2016**). "Effects of age and hearing loss on the processing of auditory temporal fine structure," Adv. Exp. Med. Biol. **894**, 1–8.
Nair, V., and Hinton, G. E. (**2010**). "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of ICML*, pp. 807–814.
Qin, M. K., and Oxenham, A. J. (**2003**). "Effects of simulated cochlear implant processing on speech reception in fluctuating maskers," J. Acoust. Soc. Am. **114**, 446–454.
Rosen, S., Souza, P., Ekelund, C., and Majeed, A. A. (**2013**). "Listening to speech in a background of other talkers: Effects of talker number and noise vocoding," J. Acoust. Soc. Am. **133**, 2431–2443.
Shinn-Cunningham, B. G. (**2008**). "Object-based auditory and visual attention," Trends Cogn. Sci. **12**, 182–186.
Stone, M. A., and Moore, B. C. J. (**1999**). "Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses," Ear Hear. **20**, 182–192.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech, Lang., Hear. Res. **28**, 455–462.
Summers, V., and Leek, M. R. (**1998**). "F_0 processing and the separation of competing speech signals by listeners with normal hearing and with hearing loss," J. Speech Lang. Hear. Res. **41**, 1294–1306.
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. Proc. **19**, 2125–2136.
ter Keurs, M., Festen, J. M., and Plomp, R. (**1993**). "Effect of spectral envelope smearing on speech reception. II," J. Acoust. Soc. Am. **93**, 1547–1552.
Tu, Y., Du, J., Xu, Y., Dai, L.-R., and Lee, C.-H. (**2014**). "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *Proceedings of ISCSLP*, pp. 250–254.
Wang, Y., Han, K., and Wang, D. L. (**2013**). "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio Speech Lang. Proc. **21**, 270–279.
Wang, Y., Narayanan, A., and Wang, D. L. (**2014**). "On training targets for supervised speech separation," IEEE/ACM Trans. Audio Speech Lang. Proc. **22**, 1849–1858.
Zhang, X.-L., and Wang, D. L. (**2016**). "A deep ensemble learning method for monaural speech separation," IEEE/ACM Trans. Audio Speech Lang. Proc. **24**, 967–977.