

# A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation

Eric W. Healy<sup>a)</sup>

*Department of Speech and Hearing Science, and Center for Cognitive and Brain Sciences,  
The Ohio State University, Columbus, Ohio 43210, USA*

Masood Delfarah

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

Eric M. Johnson

*Department of Speech and Hearing Science, and Center for Cognitive and Brain Sciences,  
The Ohio State University, Columbus, Ohio 43210, USA*

DeLiang Wang

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

(Received 6 December 2018; revised 6 February 2019; accepted 19 February 2019; published online 15 March 2019)

For deep learning based speech segregation to have translational significance as a noise-reduction tool, it must perform in a wide variety of acoustic environments. In the current study, performance was examined when target speech was subjected to interference from a single talker and room reverberation. Conditions were compared in which an algorithm was trained to remove both reverberation and interfering speech, or only interfering speech. A recurrent neural network incorporating bidirectional long short-term memory was trained to estimate the ideal ratio mask corresponding to target speech. Substantial intelligibility improvements were found for hearing-impaired (HI) and normal-hearing (NH) listeners across a range of target-to-interferer ratios (TIRs). HI listeners performed better with reverberation removed, whereas NH listeners demonstrated no difference. Algorithm benefit averaged 56 percentage points for the HI listeners at the least-favorable TIR, allowing these listeners to perform numerically better than young NH listeners without processing. The current study highlights the difficulty associated with perceiving speech in reverberant-noisy environments, and it extends the range of environments in which deep learning based speech segregation can be effectively applied. This increasingly wide array of environments includes not only a variety of background noises and interfering speech, but also room reverberation.

© 2019 Acoustical Society of America. <https://doi.org/10.1121/1.5093547>

[JL]

Pages: 1378–1388

## I. INTRODUCTION

The acoustic environments encountered by listeners vary widely, and so speech must be understood in the presence of a wide variety of different background interferences. This places enormous demands on speech-segregation or noise-reduction technology. The task is perhaps simplest when the background consists of only non-speech noise, especially when that noise is steady. Under these conditions, the target speech and the background noise form distinct classes, which can be exploited by noise-reduction algorithms. The task is more challenging when the background consists of the babble of multiple talkers, because the distinctiveness between target and background is reduced. Also challenging are more ecologically valid backgrounds that contain a variety of different noise types, including concurrent speech and non-speech sounds (e.g., recordings from a busy cafeteria). Although the magnitude of the challenge

associated with these situations varies, the task is the same—target speech must be extracted from a background, allowing intelligibility to be increased.

Fortunately, deep learning algorithms based on time-frequency (T-F) masking have been successful at improving intelligibility under these varied conditions, and particularly for the population of greatest need—those with sensorineural hearing loss who wear hearing aids. Intelligibility improvements have been observed for backgrounds consisting of steady-state noise (Healy *et al.*, 2013; Healy *et al.*, 2014; Monaghan *et al.*, 2017; Zhao *et al.*, 2018), speech babble (Healy *et al.*, 2013; Healy *et al.*, 2014; Healy *et al.*, 2015; Chen *et al.*, 2016; Monaghan *et al.*, 2017; Bentsen *et al.*, 2018; Zhao *et al.*, 2018), and cafeteria-noise recordings (Healy *et al.*, 2015; Chen *et al.*, 2016; Zhao *et al.*, 2018).

The task is somewhat different when the background consists of speech from an interfering talker (speaker separation).<sup>1</sup> This is particularly true when the competing speech itself is fully intelligible. In the human auditory system, this type of segregation likely relies on different mechanisms.

<sup>a)</sup>Electronic mail: healy.66@osu.edu

But deep learning has again provided substantial intelligibility improvements for hearing-impaired (HI) listeners under these conditions (Healy *et al.*, 2017; Bramsløw *et al.*, 2018).

Another fundamental aspect that characterizes the everyday acoustic environment involves room reverberation. Reverberation often occurs along with background noise/interference,<sup>2</sup> making it important that hearing technology be able to address these concurrent corruptions. But the two types of distortion disrupt the acoustic speech signal in different ways, leading to different effects on speech perception (producing for example different patterns of vowel-perception errors in human listeners, Nábělek and Dagenais, 1986). Likely due to the considerable challenge associated with addressing these concurrent distortions, only recently reported is the first demonstration of intelligibility improvement resulting from a single-microphone (monaural) algorithm in reverberant-noisy conditions (Zhao *et al.*, 2018). In this report, deep learning was used to estimate T-F masks for sentences corrupted by reverberation plus speech-shaped noise or reverberation plus multi-talker babble. Substantial intelligibility improvements were observed following processing for HI listeners, which allowed their performance to approach or match that of young normal-hearing (NH) listeners without processing.

When speech enhancement or speaker separation is performed on reverberant-noisy speech, a decision must be made with regard to what signal should be segregated from the corrupted speech. Said differently, what signal should the algorithm be trained to extract? In Zhao *et al.* (2018), the decision was made to remove both reverberation and noise, and aim for reverberation-free, noise-free speech. This method, which can be referred to as the direct-sound (DS) target, is one of two processing strategies employed in the current study. But alternative approaches exist. These alternatives involve removing just background interference, while preserving some or all of the reverberation corresponding to the speech of interest (Roman and Woodruff, 2013; Li *et al.*, 2015).

The alternative approach assessed currently involved the removal of interfering speech but no attempt to remove reverberation from the speech of interest. This may be referred to as the reverberant (R) target. The motivation for this approach is twofold. First, when background interference is absent, human listeners can tolerate a substantial amount of reverberation. This is particularly true for NH listeners, but also to some extent for those with hearing loss or cochlear implants (see the brief discussion by Zhao *et al.*, 2018). Second, the computational challenge associated with accurately estimating the R target is likely reduced relative to that associated with accurately estimating the DS target, because de-reverberation of the speech of interest is not performed in the former and the output sound more closely resembles the original input sound. Said differently, it might be easier to accurately estimate the R target than the DS target (see Zhao *et al.*, 2014).

Another alternative was also considered, which involved keeping the early reverberation reflections, typically defined as those occurring within 50 ms, of the direct sound (Roman and Woodruff, 2013; Li *et al.*, 2015). A computational investigation, however, showed that the segregation performance of this alternative lies between those of the DS and R targets.

So the effects of different targets, if present, should be revealed by comparing the DS and R targets. Therefore, the current study does not evaluate this early-reverberant alternative.

The implementation of any of these training targets (DS, early reverberation, or R) is straightforward in a deep learning context. During a training phase, the algorithm receives acoustic features extracted from the corrupted speech. In the current study, this corruption involved reverberation plus an interfering talker. Because the learning is supervised, the algorithm also receives the desired outcome, which in our work has involved the ideal T-F mask for each particular target sentence. These masks are ideal because they employ knowledge of the separate speech and interference signals as well as their reverberations, i.e., they are oracle masks. The different training targets are implemented simply by providing the deep learning network with the desired mask. In the case of the DS target, the mask provided during training involves the reverberation-free, interference-free speech of interest. In the case of the R target, the mask provided during training involves the fully reverberant but interference-free speech of interest. Had the choice been made to implement an early-reverberant target, then the mask provided during training would have involved the speech of interest containing only early reflections but no interference. Through many training trials, the network learns to estimate the mask that it is trained on. Once trained, it can estimate that mask when provided with only features extracted from novel (unseen during training) corrupted speech, with no knowledge of the separate target or interfering signals or their reverberation components. It is important to recognize that network training is completed prior to operation, and so training duration is unrelated to operational efficiency.

Some literature exists regarding intelligibility produced by DS versus R masks when implemented as binary and ideal, and presented to NH listeners (Roman and Woodruff, 2011; Roman and Woodruff, 2013). Some literature also exists regarding these masks, including algorithm-estimated versions, when assessed using objective metrics on a speaker-identification task (Zhao *et al.*, 2014). But the results of these studies suggest no simple relationship between DS and R masks. Instead, factors including noise-rejection threshold and algorithm-estimation accuracy play roles. Zhao *et al.* (2018) demonstrated that a deep learning algorithm based on the ideal ratio mask (IRM, Srinivasan *et al.*, 2006; Narayanan and Wang, 2013; Wang *et al.*, 2014) and trained to estimate the DS target can effectively remove concurrent reverberation and noise to improve intelligibility for HI listeners. But the relative performance of human listeners, particularly HI, on algorithms trained to estimate DS versus R targets is not known, and direct assessment is important.

In the current study, a single-microphone deep learning algorithm based on T-F masking was employed to improve the intelligibility of a target talker in the presence of reverberation and a single interfering talker. Likely because this interference situation poses a substantial computational challenge, it has not yet been addressed. To compare the relative intelligibilities resulting from two training targets, a deep neural network was trained to estimate either the DS or R target. Both HI and NH listeners were employed in this

study. Unlike previous studies that used feedforward neural networks (Healy *et al.*, 2017; Zhao *et al.*, 2018), the current study employed a recurrent neural network (RNN) with bidirectional long short-term memory (BLSTM) to estimate the IRM. Such an RNN encodes temporal contexts via memory cells, capturing information from a much wider context over a sentence. This property makes RNNs capable of modeling the temporal correlations intrinsic to reverberant speech.

## II. METHOD

### A. Subjects

A first group of listeners was composed of ten individuals with bilateral sensorineural hearing loss of likely cochlear origin. All were bilateral hearing-aid wearers recruited from The Ohio State University Speech-Language-Hearing Clinic. These listeners were selected to represent typical HI patients. The age range was 51 to 73 yr (mean = 68), and five were female. Prior diagnoses were 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 listeners. Bone-conduction audiometry helped establish the cochlear site of lesion. Figure 1 displays audiograms for each of these listeners, who are numbered in ascending order of pure-tone average audiometric thresholds (PTAs; means across thresholds at 500, 1000, and 2000 Hz and ears). In accord with our desire to recruit listeners representative of typical patients, the degree of hearing loss varied across listeners and frequencies, but less so across ears. Seven of the listeners had audiometric thresholds within normal limits [20 dB hearing level (HL) or lower] for at least one frequency. But all listeners had elevated thresholds (in at least one ear) for at least half of the audiometric frequencies. These degrees of hearing losses ranged from mild to profound. PTAs ranged from 19 to 72 dB HL with an average of

34 dB HL. The primary configuration of hearing loss was sloping (gently to precipitously).

A second group of listeners was composed of ten individuals with NH, defined by audiometric thresholds of 20 dB HL or below at octave frequencies from 250 to 8000 Hz (ANSI, 2004, 2010). The exceptions were two listeners (NH1 and NH5) who had thresholds of 25 dB HL at 250 Hz in the right ear. These individuals were recruited from undergraduate courses at The Ohio State University. Their ages ranged from 18 to 21 yr (mean = 19.8), and all were female. All listeners received a monetary incentive or course credit for participating. As in our previous studies, age matching between HI and NH listeners was not performed because the goal was to assess the performance of typical (older) HI listeners relative to the ideal performance of young NH listeners. No listeners in either group had any prior exposure to the sentence materials employed.

### B. Stimuli

The stimuli were drawn from the Institute of Electrical and Electronics Engineers (IEEE) revised list of phonetically balanced sentences (IEEE, 1969). This set is composed of 720 grammatically and semantically correct sentences, each having five scoring keywords. The stimuli used for current human-subjects testing consisted of 153 target sentences, each mixed with a single interfering sentence. 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, and all interfering sentences were spoken by the same female talker. The 44.1 kHz, 16-bit sentence recordings were down-sampled to 16 kHz for processing and presentation.

Each target and interfering sentence was convolved with a different room impulse response (RIR). The image method (Allen and Berkley, 1979) was used to generate the RIRs.<sup>3</sup> A simulated room having a reverberation time ( $T_{60}$ ) of 600 ms

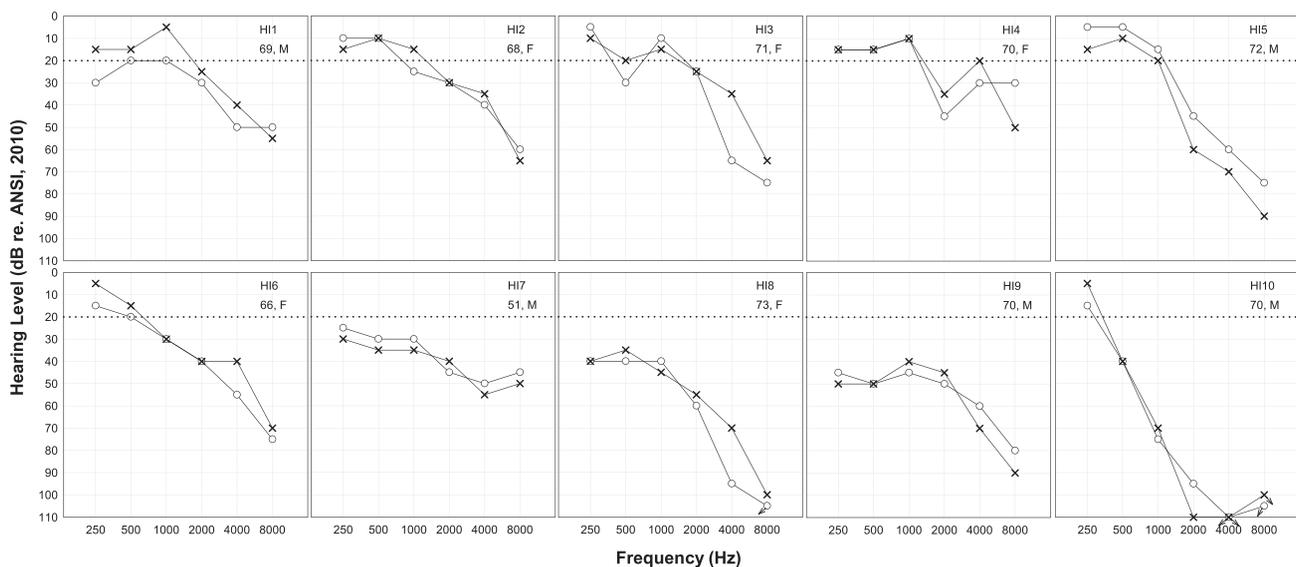


FIG. 1. Pure-tone air-conduction audiometric thresholds for the listeners with hearing impairment. Listeners are numbered in order of increasing degree of hearing loss. Right ears are represented by circles and left ears are represented by X's. Arrows indicate thresholds exceeding audiometer limits. The normal-hearing limit of 20 dB HL is represented by a horizontal dotted line in each panel. Subject numbers, listener ages in years, and genders are also provided.

and dimensions of  $6\text{ m} \times 7\text{ m} \times 3\text{ m}$  ( $L \times W \times H$ ) was employed. The microphone was placed at a fixed position within the room at (3.5, 4, 1.7) m. The target-talker position was chosen randomly from a set of 36 positions uniformly spaced on a 1-m radius circle centered on the microphone, at the same elevation. The interfering talker was placed on one of 36 uniformly spaced positions on a circle having a 2-m distance from the microphone, also at the same elevation. These room characteristics and dimensions were chosen to represent a large living room having a relatively high reverberation time, and the positions of the talkers and listener were chosen to represent a conversation taking place in the room. It is noteworthy that, unlike Zhao *et al.* (2018), reverberation was applied to both the target speech and the interference.

After convolution, test signals were generated by mixing the reverberant target and interfering signals into pairs at one of the following target-to-interferer ratios (TIRs):  $-6$ ,  $-3$ ,  $0$ ,  $3$ , and  $6$  dB; the TIR definition used the reverberant target as the reference signal. Target and interfering sentences were paired such that each was approximately equal in duration. No duration difference between members of a pair exceeded 0.01 s. Target and interfering sentences were mixed so that their onsets aligned, and target signals tended to be slightly longer in duration.

### C. Algorithm description

The general operation of the algorithm is depicted in Fig. 2. A complementary set of acoustic features was extracted from the concurrent reverberant sentences and delivered to a neural network (an RNN with BLSTM). During training, the network received these features from a given sentence pair, along with the IRM for that pair. The IRM consists of a time-frequency matrix of values, ranging from 0 to 1. These values can be considered attenuation values, where values closer to 1 represent T-F units corresponding primarily to the desired signal and values closer to 0 represent units corresponding primarily to the undesired signal. The IRM provided during training was that representing either the DS or R signal. So the desired signal was either the DS speech of interest or that speech plus its reverberation. Following training with many such examples, the algorithm entered a test (operation) phase. During this phase, the network again received features derived from a sentence pair, but it output the estimated IRM directly from these features. The network was trained separately to estimate the IRM for the DS signal and the IRM for the R signal.

Of the remaining 567 sentences from each talker (720 total minus 153 used for algorithm testing), 537 were used for training and 30 were used for cross validation. Thus, the test stimuli were not used during algorithm training or validation. The same method used to generate the test stimuli was used to generate each training mixture, except that the RIRs corresponded to different sets of 36 points on the same circles, and these points were interleaved with those used to create the test signals (i.e., with 5-deg separation). Training TIRs were selected from the range  $-12.5$  to  $12.5$  dB with a step size of 1 dB. Note that no RIR or TIR was the same

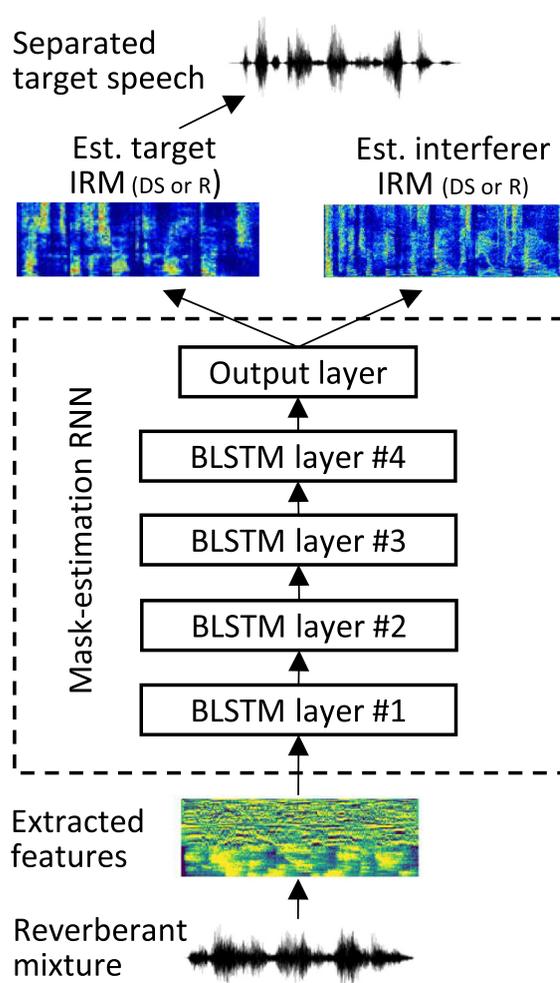


FIG. 2. (Color online) Schematic diagram of the speech-separation framework. Est. = network-estimated, IRM = ideal ratio mask, RNN = recurrent neural network, BLSTM = bidirectional long short-term memory.

between the test and training utterances. To obtain utterance pairs for training that fully overlapped in time, the interference sentence was repeated if needed for each mixture until it matched the duration of the target sentence. The training set consisted of 100 000 such mixtures (approximately 74 h of audio material) and the cross validation set consisted of 500 mixtures (approximately 21 min of audio material).

The acoustic features delivered to the RNN were generated by dividing each signal in the training set into 20-ms frames with 10-ms frame shift. In each time frame, a combination of features was extracted, consisting of a 40-dimensional (40-D) log-mel filterbank, a 31-D gammatone frequency cepstral coefficient (GFCC, Shao and Wang, 2008), and a 31-D power-normalized cepstral coefficient (PNCC, Kim and Stern, 2016). These together formed a 102-D feature vector. This combination of features has been shown to form a complementary set and outperform other feature combinations (Delfarah and Wang, 2017). The mean ( $\mu$ ) and standard deviation ( $\sigma$ ) of the feature vectors were computed across the entire training set. Then, every feature vector  $F$  was normalized as follows:

$$\tilde{F} = \frac{F - \mu}{\sigma}, \quad (1)$$

where  $\tilde{F}$  is the normalized 102-D feature vector. This normalization has been shown to facilitate deep neural network training (Delfarah and Wang, 2017).

Let the IRM corresponding to the reverberation-free, interference-free target speech be  $IRM_{DS}$ , and the IRM corresponding to the fully reverberant, interference-free target speech be  $IRM_R$ . These IRMs served as the training targets in the current study (see Wang *et al.*, 2014). Let  $x_1$  and  $x_2$  represent the direct target and interferer signals, respectively. The reverberant mixture signal  $y(t)$  can be described as

$$y(t) = \tilde{x}_1(t) + \tilde{x}_2(t) = x_1(t) * h_1(t) + x_2(t) * h_2(t), \quad (2)$$

where  $h_1(t)$  and  $h_2(t)$  represent the RIR for each speaker, the symbol  $*$  denotes convolution, and  $\tilde{x}_1(t)$  and  $\tilde{x}_2(t)$  denote the reverberant signals. These two versions of the IRM (see Wang *et al.*, 2014; Huang *et al.*, 2015) are

$$IRM_{DS} = \left[ \frac{S(x_1)}{S(x_1) + S(y - x_1)}, \frac{S(x_2)}{S(x_2) + S(y - x_2)} \right], \quad (3)$$

$$IRM_R = \left[ \frac{S(\tilde{x}_1)}{S(\tilde{x}_1) + S(\tilde{x}_2)}, \frac{S(\tilde{x}_2)}{S(\tilde{x}_1) + S(\tilde{x}_2)} \right], \quad (4)$$

where  $S(\cdot)$  represents the magnitude short-time Fourier transform (STFT) of a signal. A fast Fourier transform on frames of length 20 ms with 10-ms overlap, aligned with acoustic-feature extraction, was applied to generate the 161-D STFT representation of the signals.

The goal of masking-based speech segregation involves a function  $G$  that maps features  $\tilde{F}$  to the IRM (see Wang *et al.*, 2014). In the current algorithm,  $G$  was the RNN consisting of four hidden layers with BLSTM (Hochreiter and Schmidhuber, 1997). A BLSTM layer consisted of one long short-term memory (LSTM) that processed feature frames from the beginning to end of the sentence, and another LSTM operating in the reverse direction. The use of memory cells in the RNN with BLSTM eliminated the need for features in the neighboring frames to predict a frame of the IRM, as commonly done in feedforward networks (e.g., Zhao *et al.*, 2018). There were 600 hidden neurons in each layer (300 neurons per direction). The last hidden layer was connected to a feedforward output layer having 322 units (corresponding to the dimensions of one target IRM frame and one interferer IRM frame) and the sigmoidal activation function. The network therefore contained approximately  $7.6 \times 10^6$  trainable parameters. The Adam optimizer (Kingma and Ba, 2014) was used to minimize the following mean square error (MSE) loss function:

$$\mathcal{L}(IRM; \Theta) = (IRM - G_{\Theta}(\tilde{F}))^2, \quad (5)$$

where  $\Theta$  represents network parameters.

During training for the  $IRM_{DS}$  and  $IRM_R$  targets, the RNN was unrolled over 100 time frames, and the data were processed in batches of size 32 samples. The learning rate was initially set to 0.0003 for 30 epochs. At the completion of training, the  $G_{\Theta}$  having the least error on the validation set was chosen and used in the test phase. For the  $IRM_{DS}$

target,  $G_{\Theta}$  was obtained after 16 h of training on a single graphics processing unit (GPU), and for the  $IRM_R$  target,  $G_{\Theta}$  was obtained after 19 h of training on the same GPU.

During the test phase, features were extracted directly from each sentence-pair mixture, normalized by  $\mu$  and  $\sigma$  as given in Eq. (1), and passed through the RNN to generate the estimated IRM (either DS or R) for that mixture. The estimated IRM for the mixture was then point-wise multiplied by  $S(y)$  for that mixture to generate the estimated magnitude STFT for that target speech. Finally, the time-domain signal was resynthesized using this estimated target magnitude, the mixture-signal phase, and the overlap-add method. The algorithm trained to estimate the  $IRM_{DS}$  will be referred to as the DS algorithm, and the algorithm trained to estimate the  $IRM_R$  will be referred to as the R algorithm.

Figure 3 displays spectrograms of a sentence-pair mixture at each of several processing stages. Panel (a) shows the target sentence, and panel (b) displays the interfering sentence, both without reverberation. Panels (c) and (d) show these same utterances following the application of reverberation, and panel (e) shows these reverberant sentences mixed at  $-6$  dB TIR. Panel (f) shows the  $IRM_{DS}$ , and panel (g) shows the algorithm-estimated  $IRM_{DS}$ . Panel (h) shows the result of applying the estimated  $IRM_{DS}$  in panel (g) to the mixture in panel (e). Panel (i) shows the  $IRM_R$ , and panel (j) shows the algorithm-estimated  $IRM_R$ . Finally, panel (k) shows the result of applying the estimated  $IRM_R$  in panel (j) to the mixture in panel (e). The accuracy of the algorithm-estimated IRM relative to its ideal counterpart can be assessed by comparing panels (f) versus (g), and (i) versus (j). How veridically the output of the algorithm represents the original target speech can be assessed by comparing panels (a) versus (h), and (c) versus (k).

## D. Procedure

Each listener heard nine conditions (3 processing conditions  $\times$  3 TIRs). The processing conditions included (1) unprocessed reverberant concurrent sentences, (2) reverberant concurrent sentences processed by the DS algorithm, and (3) reverberant concurrent sentences processed by the R algorithm. The TIRs for HI listeners were 0, 3, and 6 dB, and those for the NH listeners were  $-6$ ,  $-3$ , and 0 dB. Each listener heard 17 pairs of concurrent sentences in each condition. Condition order was blocked by TIR to allow unprocessed, DS, and R conditions to appear juxtaposed for each TIR. The order of TIRs, as well as for processing condition within each TIR, was pseudorandomized for each listener. The sentence pairs were presented to each listener in the same order to allow a pseudorandom correspondence between sentence pair and condition. No sentence was repeated for any listener (either as a target sentence or as an interfering sentence).

The signals were converted to analog form using Echo Digital Audio Gina 3G digital-to-analog converters (Santa Barbara, CA), routed through a Mackie 1202-VLZ mixer (Woodinville, WA), and presented diotically over Sennheiser HD 280 Pro headphones (Wedemark, Germany). The overall root-mean-square (RMS) level of each stimulus was set to 65 dBA in each ear for the NH listeners using a sound-level meter

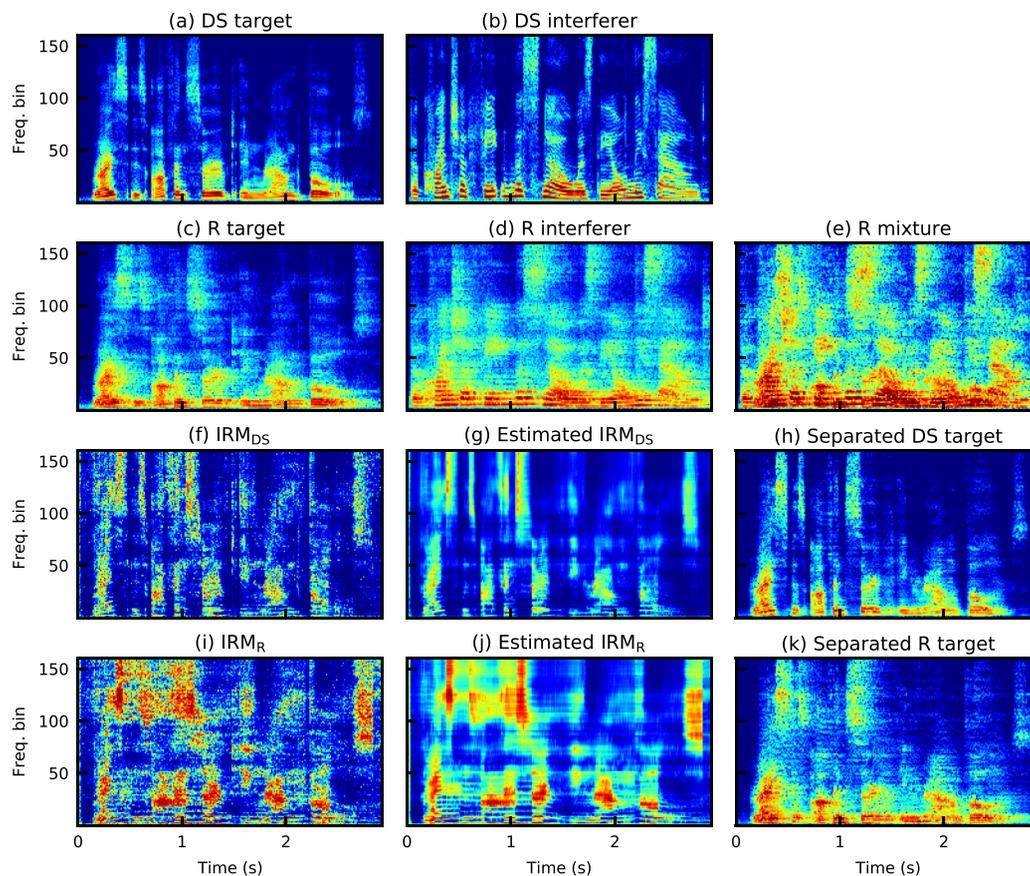


FIG. 3. (Color online) Spectrogram images of a single target and a single interfering sentence at each of several processing stages. Panel (a) direct-sound target sentence, panel (b) direct-sound interfering sentence, panel (c) reverberant target sentence, panel (d) reverberant interfering sentence, panel (e) mixture of reverberant target and interfering sentences at  $-6$  dB, panel (f) direct-sound IRM, panel (g) algorithm-estimated direct-sound IRM, panel (h) direct-sound target sentence segregated from mixture (e) using algorithm-estimated IRM (g), panel (i) reverberant IRM, panel (j) algorithm-estimated reverberant IRM, panel (k) reverberant target sentence segregated from mixture (e) using algorithm-estimated reverberant IRM (j).

and flat-plate coupler (Larson Davis models 824 and AEC 101, Depew, NY). The same 65 dBA level was used for the HI listeners, plus frequency-specific gains as prescribed by the NAL-RP hearing-aid fitting formula (Byrne *et al.*, 1990). Gains were determined separately for each HI listener and implemented using a RANE DEQ 60L digital equalizer (Mukilteo, WA), as described in Healy *et al.* (2015). The NAL-RP formula does not provide prescription gains at 125 or 8000 Hz, and so the gains applied to 250 and 6000 Hz (respectively) were also applied to these two most extreme frequencies. In accord with the use of hearing-aid amplification for individual HI listeners, these listeners were tested with their actual hearing aids removed.

Immediately prior to formal testing, listeners participated in a brief familiarization involving seven reverberation-free sentences spoken by the male target talker. This was followed by 21 pairs of concurrent sentences from the male target talker and the female interfering talker, seven as DS-algorithm processed stimuli, seven as R-algorithm processed stimuli, then seven as unprocessed reverberant mixtures. These familiarization sentences were all drawn from the pool of sentences used for algorithm training. The TIR employed in familiarization was 3 dB higher than the most favorable TIR employed for each group in testing. During familiarization, the HI listeners were asked if the signal was uncomfortably loud. Two (HI8 and HI10) reported that the signal was loud, and so the overall

level was reduced by 5 dB. After adjustment, both of these listeners indicated that the signal was comfortable. Overall total RMS presentation levels for the HI listeners with NAL-RP amplification and adjustment ranged from 72.5 to 94.6 dBA (mean = 78.2 dBA).

After familiarization, listeners heard the nine blocks of experimental conditions. They were seated with the experimenter in a double-walled sound booth, instructed to repeat back each sentence spoken by the male talker, and to guess if unsure. The experimenter controlled stimulus presentation and scored keywords correctly reported.

### III. RESULTS AND DISCUSSION

#### A. Human performance

Sentence intelligibility was based on percentage of sentence keywords reported. Figure 4 displays these intelligibility scores produced by each individual HI listener in each condition, and Fig. 5 displays the corresponding data for the NH listeners. Figure 6 displays group-mean scores and standard errors of the mean in each condition. In each of these figures, the HI and NH listeners are plotted separately. Also in each figure, the processing conditions are represented by different columns, and the TIRs are displayed in separate panels.

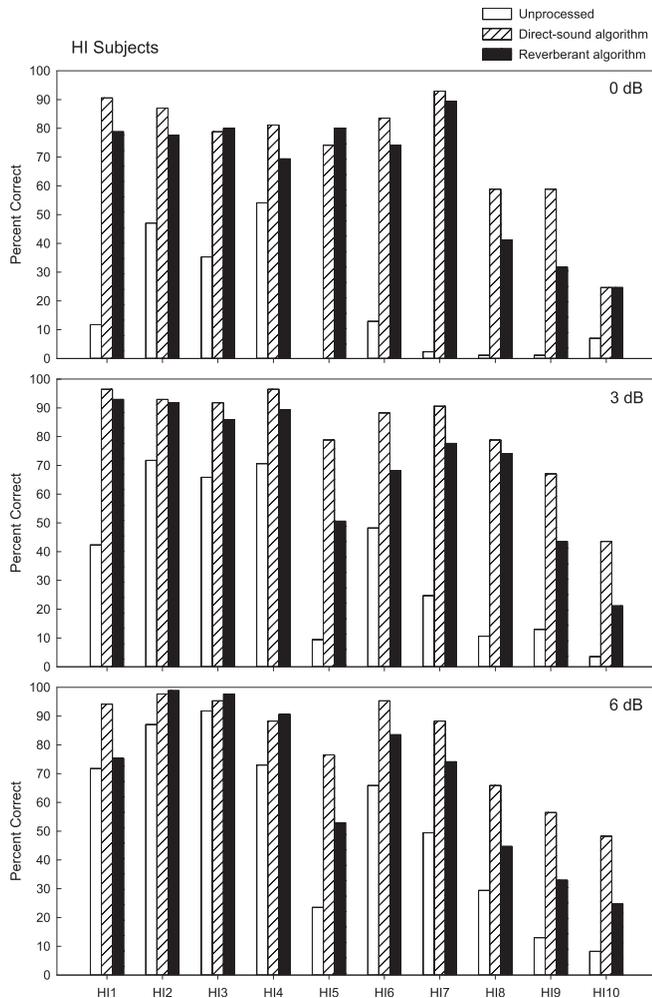


FIG. 4. Sentence-intelligibility scores for each individual hearing-impaired listener in the presence of reverberation and a single interfering talker. The one unprocessed and two algorithm-processed conditions are represented by different columns. Algorithm processing targeted the reverberation-free, interference-free target speech (direct-sound algorithm) or the reverberant interference-free target speech (reverberant algorithm). The three target-to-interferer ratios are displayed in separate panels.

### 1. HI listeners

Apparent in each panel of Fig. 4 is the result of plotting the HI listeners in order of increasing degree of hearing loss (increasing PTA). Scores tend to decline from left to right. The amount of algorithm benefit is largely a function of the scores in the unprocessed condition. Accordingly, benefit was largest at the least favorable TIR, where unprocessed scores were lowest. At the TIR of 0 dB, mean unprocessed scores averaged 17.3% (see Fig. 6). This value increased to 73.1% and 64.7% for the DS and R algorithms, respectively, producing mean algorithm benefits of 55.8 and 47.4 percentage points. As the TIR improved to 3 dB, mean unprocessed scores reached 36.0%. But scores were considerably higher in the DS- and R-algorithm conditions, and mean benefit was 46.5 and 33.5 percentage points. Finally, at a TIR of 6 dB, the mean unprocessed score reached 51.3%, and DS- and R-algorithm benefit averaged 29.3 and 16.2 percentage points.

The more-favorable TIRs highlight another important aspect of algorithm performance, namely that the processing

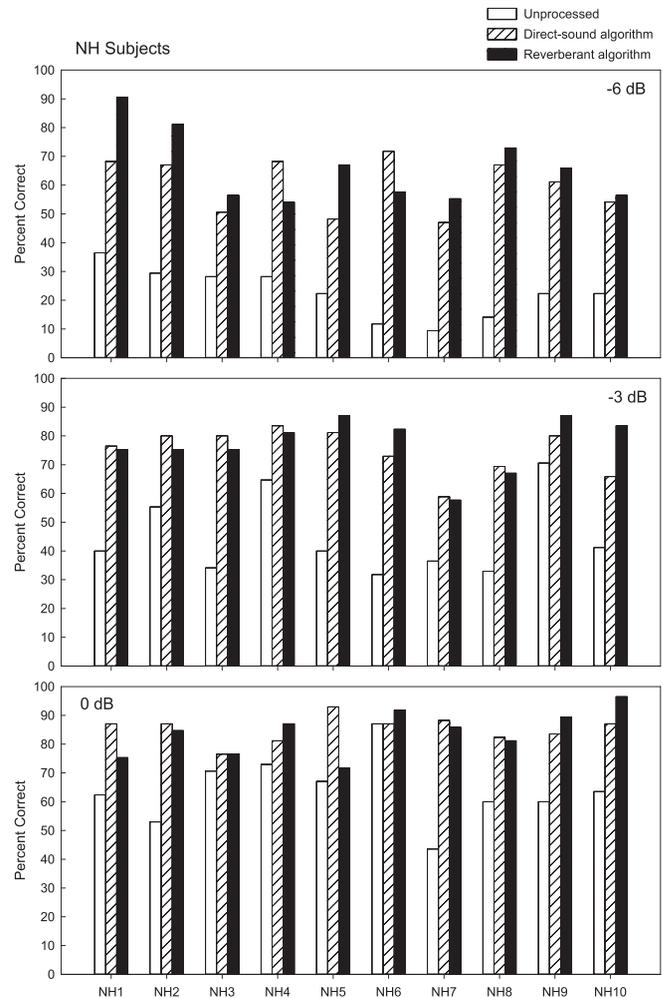


FIG. 5. As Fig. 4, but for the normal-hearing listeners.

does not degrade performance in conditions where benefit is not needed. This is apparent in the middle and bottom panels of Fig. 4—no reductions were observed from unprocessed to processed scores in any condition for any listener, despite that unprocessed scores exceeded 70% in six cases and were as high as 92%.

The primary statistical approach involved a linear mixed effects model (on RAU-transformed scores, [Studebaker, 1985](#)), using planned pairwise comparisons to assess differences across the conditions.<sup>4</sup> This approach was employed to address the effect of the algorithms at various TIR levels while controlling for within-subject variability. The standardized mean differences (differences between z-scores) from the model are shown in Table I, which presents the differences between unprocessed and algorithm scores across TIR levels and at each TIR level. These comparisons indicated that algorithm benefit was large (in terms of effect size) and significant for the HI listeners in all conditions for both algorithms. DS-algorithm scores were significantly greater than unprocessed scores across TIR levels ( $p < 0.001$ ) and at each TIR ( $p < 0.001$ ). R-algorithm scores were also significantly greater than unprocessed scores across TIR levels ( $p < 0.001$ ) and at each TIR ( $p < 0.001$ ).

With regard to the comparison between the DS and R algorithms, scores were higher on average in the DS conditions for the HI listeners, with differences averaging 8.4,

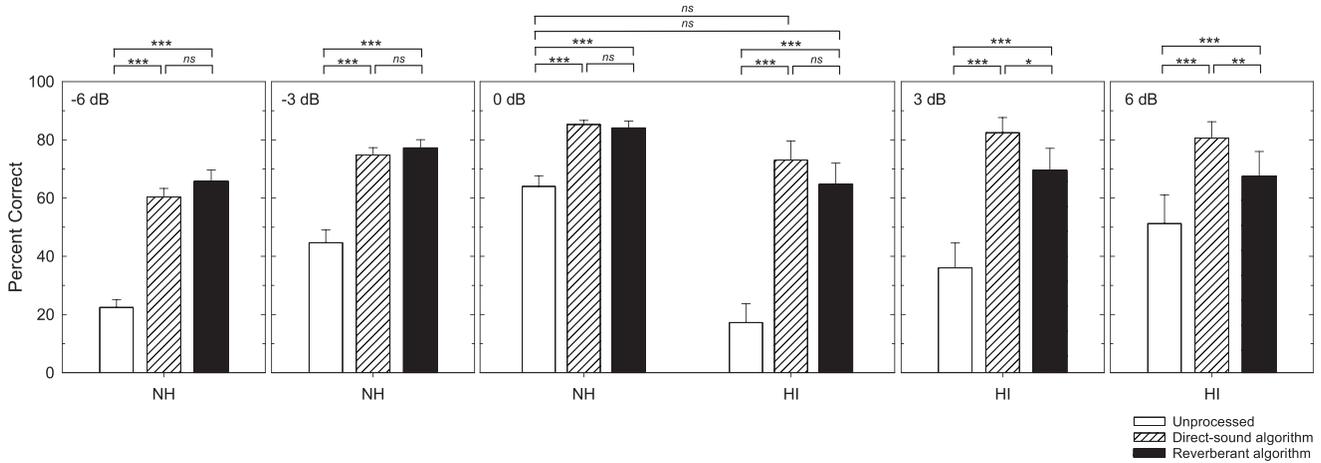


FIG. 6. Group-mean sentence-intelligibility scores and standard errors of the mean for each condition. The unprocessed, direct-sound-algorithm, and reverberant-algorithm conditions are represented by separate columns, as in Figs. 4 and 5. Means for the normal-hearing (NH) and hearing-impaired (HI) listeners are presented separately. Note the different target-to-interferer ratios required across the two listener groups. *ns* = nonsignificant, (\*)  $p < 0.05$ , (\*\*)  $p < 0.01$ , (\*\*\*)  $p < 0.001$ .

12.9, and 13.1 percentage points at 0, 3, and 6 dB TIR, respectively. The DS score exceeded the R score in 24 of 30 cases for these HI listeners (10 listeners  $\times$  3 TIRs). The linear mixed effects model indicated that these differences were significant across the levels of TIR ( $p = 0.002$ ). At 3 and 6 dB (but not at 0 dB), the DS-algorithm scores were moderately (effect sizes  $\geq 0.4$ ) and significantly ( $p < 0.02$ ) higher than the R algorithm scores. When the *a priori* DS-R comparison was examined using planned comparison as a supplement to the linear mixed effects model (two-tailed, uncorrected, paired t test), the DS-R difference at 0 dB was also significant [ $t(9) = 3.0$ ,  $p < 0.02$ ].

## 2. NH listeners

As with the HI listeners, algorithm benefit was observed for the NH listeners at each TIR, and this benefit was largest at the least-favorable TIR (see Figs. 5 and 6). At the TIR of  $-6$  dB, the unprocessed score averaged 22.5%, which increased to 60.4% and 65.8% for the DS and R algorithms, respectively, resulting in algorithm benefits of 37.9 and 43.3 percentage points. From an unprocessed mean of 44.7% (TIR =  $-3$  dB), DS and R benefits were 30.1 and 32.5 percentage points. And from an unprocessed mean of 64.0% (TIR = 0 dB), DS and R benefits were 21.3 and 20.0 percentage points.

Also as found for the HI listeners, NH algorithm benefit was significant in every condition (see Table I). DS-algorithm scores were significantly higher than unprocessed scores across TIR levels ( $p < 0.001$ ) and at each of the three TIRs ( $p < 0.001$ ). The same was true for R-algorithm scores relative to unprocessed scores (across TIRs,  $p < 0.001$ ; at each TIR,  $p < 0.001$ ). Again as for the HI listeners, algorithm processing did not produce a reduction in scores for any listener in any condition.

Unlike the HI listeners, the NH listeners did not display a substantial difference between DS- and R-algorithm scores. The DS score exceeded the R score in only 13 of the 30 cases, which is approximately expected by chance. The differences in mean scores (DS rel. to R) were  $-5.4$ ,  $-2.4$ , and 1.3 percentage points at  $-6$ ,  $-3$ , and 0 dB, respectively.

These differences had small standardized mean differences (effect sizes  $\leq 0.26$ ) and none were significant.

Another comparison of interest involves the performance of HI listeners having access to the algorithm versus NH listeners without the algorithm, in conditions of identical background interference. As shown in Fig. 6, the TIR of 0 dB was common to both listener groups. In this interference condition, the NH unprocessed mean was 64.0%, whereas the HI listeners averaged 73.1% for the DS algorithm and 64.7% for the R algorithm. Pairwise comparisons indicated that, whereas the HI listeners achieved numerically higher scores than the NH listeners when using either algorithm, these NH versus HI comparisons did not differ significantly ( $p > 0.43$ ).

## B. Objective measures of intelligibility and sound quality

This subsection evaluates the RNN with BLSTM using objective metrics, which should help the interested reader who intends to replicate the reported segregation results.

### 1. Intelligibility

Short-time objective intelligibility (STOI, Taal *et al.*, 2011) was calculated based on all 153 sentence-pair stimuli

TABLE I. Linear mixed effects model standardized mean differences for various comparisons.  $p$ -values are based on Satterthwaite approximation to degrees of freedom. The  $p$ -values have been adjusted for family-wise comparisons using Tukey HSD. (\*)  $p < 0.05$ , (\*\*)  $p < 0.01$ , (\*\*\*)  $p < 0.001$ .

Pairwise comparison	Across TIRs	At specific TIR				
		$-6$ dB	$-3$ dB	0 dB	3 dB	6 dB
Hearing impaired						
DS—unprocessed	1.45***	—	—	1.87***	1.52***	0.97***
R—unprocessed	1.08***	—	—	1.60***	1.08***	0.57***
DS—R	0.37**	—	—	0.27	0.44*	0.40**
Normal hearing						
DS—unprocessed	1.37***	1.71***	1.34***	1.05***	—	—
R—unprocessed	1.48***	1.98***	1.47***	1.00***	—	—
DS—R	$-0.12$	$-0.26$	$-0.13$	0.05	—	—

TABLE II. Objective scores based on acoustic signals for direct-sound (DS) algorithm performance. Direct-sound target speech was used as the reference.

Input TIR	Unprocessed STOI (%)	Processed STOI (%)	Benefit STOI	Unprocessed PESQ	Processed PESQ	Benefit PESQ
−6 dB	45.56	78.69	33.13	1.54	2.30	0.76
−3 dB	50.83	81.43	30.60	1.60	2.41	0.81
0 dB	56.09	83.56	27.47	1.66	2.50	0.84
3 dB	61.00	85.15	24.15	1.76	2.58	0.82
6 dB	65.11	86.27	21.16	1.88	2.67	0.79
<b>Mean</b>	<b>55.72</b>	<b>83.02</b>	<b>27.30</b>	<b>1.69</b>	<b>2.49</b>	<b>0.80</b>

used for human-subjects testing. STOI is a widely used objective measure of speech intelligibility, which is based on the acoustic signals and correlates well to NH human performance. It essentially represents a correlation between the amplitude envelopes of unaltered target speech and corrupted target speech following processing. Accordingly, its scale typically ranges from 0.0 to 1.0 (or 0% to 100%).

Table II displays STOI values for the unprocessed mixtures and for the sentence pairs processed by the DS algorithm, at each TIR employed for both listener groups. The direct-sound target speech was used as the reference in these calculations. Values for unprocessed mixtures averaged 55.7%, which increased to 83.0% following algorithm processing, representing a benefit of 27.3 percentage points. The benefits formed a monotonic function ranging from 33.1 to 21.2 percentage points as TIR increased from −6 to 6 dB.

Table III displays the corresponding values for the R algorithm. Here, the reverberant target speech was used as the reference. These STOI values are similar—unprocessed mixtures averaged 59.4%, which increased to 86.1% following algorithm processing, representing a gain of 26.7 points. These benefits also formed a monotonic function ranging from 36.9 to 15.6 points across TIRs of −6 to 6 dB. An examination of the STOI benefit values across DS and R algorithms reveals that the range across TIRs was larger for the R algorithm, but the mean benefit across TIRs was within 1 percentage point.

## 2. Sound quality

To gain insight into potential sound-quality benefits of the current algorithm, the perceptual evaluation of speech quality (PESQ; Rix *et al.*, 2001) was also calculated across the same 153 sentence pairs used for human subjects. Like STOI, PESQ reflects a comparison between unaltered target speech and corrupted speech following processing. It is a standard measure of speech sound quality based on acoustic measurement, having a scale ranging from −0.5 to 4.5.

Table II displays the PESQ values for the unprocessed mixtures and for the sentence pairs processed by the DS algorithm, at each TIR employed. Values for unprocessed mixtures averaged 1.7, which increased to 2.5 following algorithm processing. This amount of benefit was similar across the TIRs employed. Table III displays the corresponding values for the R algorithm. These values for unprocessed mixtures averaged 2.0, which increased to 3.0 following algorithm processing. These benefits were also similar across TIRs. An examination across DS and R algorithms reveals that mean PESQ benefit across TIRs was somewhat larger for the R than for the DS algorithm.

## IV. GENERAL DISCUSSION

Consistent with previous work (Nábělek and Mason, 1981; George *et al.*, 2010; Hazrati and Loizou, 2012) the current study highlights the challenge that concurrent reverberation and background interference can present for the human perception of speech. When these same IEEE sentences were assessed under conditions in which reverberation was absent but the single-talker interference was the same (Healy *et al.*, 2017), TIR values had to be 9 dB less favorable than employed currently for the HI listeners and 6 dB less favorable for the NH listeners in order to obtain unprocessed scores relatively free of floor and ceiling effects. There were no overlapping conditions across studies for the HI listeners. There was, however, one such condition for NH listeners (−6 dB TIR). Under essentially identical single competing-talker conditions, NH listeners scored close to 90% when reverberation was absent (Healy *et al.*, 2017), but below 25% when concurrent reverberation was present (current study). So the substantial disruption produced by these concurrent corruptions is not restricted to only HI listeners and instead impacts NH listeners quite substantially. The addition of reverberation to single-talker interference appears to be particularly disruptive for both listener types, as the

TABLE III. Objective scores based on acoustic signals for reverberant (R) algorithm performance. Reverberant target speech was used as the reference.

Input TIR	Unprocessed STOI (%)	Processed STOI (%)	Benefit STOI	Unprocessed PESQ	Processed PESQ	Benefit PESQ
−6 dB	41.51	78.43	36.92	1.59	2.57	0.98
−3 dB	50.42	83.00	32.58	1.73	2.77	1.04
0 dB	59.74	86.83	27.09	1.94	2.97	1.03
3 dB	68.70	89.93	21.23	2.16	3.16	1.00
6 dB	76.66	92.28	15.62	2.41	3.36	0.95
<b>Mean</b>	<b>59.41</b>	<b>86.09</b>	<b>26.69</b>	<b>1.97</b>	<b>2.97</b>	<b>1.00</b>

addition of reverberation to multi-talker babble (compare Healy *et al.*, 2015 to Zhao *et al.*, 2018) had a smaller effect.

The current deep learning algorithm was found to be effective at improving intelligibility for both listener groups. These benefits were largest at the least-favorable TIRs, where unprocessed scores were lowest and room to improve was largest. Algorithm benefit was considerably larger for the HI listeners, as has been consistently observed previously. Over half of the HI listeners received a benefit of 58 percentage points or better from the current DS algorithm at the least-favorable TIR. Finally, both algorithms allowed HI performance to numerically exceed that of young NH listeners under identical interference conditions (0 dB TIR plus reverberation). This final comparison is an important one, because it simulates a listening situation in which a typical HI individual is with a young NH listener in a difficult listening environment. Consider perhaps an older individual with hearing loss together with a young-adult relative or interacting with a young-adult staff member in a noisy restaurant. If the HI individual had access to a comparable algorithm implemented in a hearing aid, then their performance could rival or even exceed that of their young NH conversation partner.

It was found that an algorithm trained to extract the reverberation-free, interference-free speech of interest (DS) performed better for the HI listeners than one trained to target the reverberant, interference-free speech of interest (R). But mean NH-listener scores were similar across these two conditions. This highlights the greater negative effects of reverberation on HI listeners compared to NH listeners. The implementation of these DS and R algorithms was highly similar. But the performance of the DS algorithm indicates that the deep learning framework can perform simultaneous de-reverberation and segregation with effectiveness sufficient to produce substantial intelligibility improvements.

Most of our focus has been on the performance of HI listeners, who are challenged most by speech interference and room reverberation. But the performance of NH listeners observed currently suggests that there might be a need for algorithms like the current one to be implemented into technology targeted toward NH individuals.

The decision regarding which algorithm (DS or R) would be preferable to deploy in hearing technology is impacted by both the target audience and implementation constraints. For the more obvious target audience, which involves HI listeners (e.g., hearing aids and cochlear implants), then the superior performance of the DS algorithm may warrant its likely greater computational challenge. However, if the target audience were to be NH listeners (e.g., cellular telephones), then the current results suggest that the R algorithm is sufficient to realize maximum performance, at least for the current interference conditions.

With regard to sound quality, the PESQ scores indicate that the current algorithm also improved this aspect of speech. This is important, particularly for HI listeners—poor sound quality can lead to poor device satisfaction and compliance (e.g., Ng and Loke, 2015). Although the increase in predicted sound quality was found to be larger for the R than for the DS algorithm, the reference signals are different. This is, the reference for the R algorithm was selected to be reverberant target

speech, whereas that for the DS algorithm was anechoic target speech. But regardless of the difference between algorithms, the PESQ increase associated with both is clear.

Given the considerable challenge associated with improving intelligibility in conditions of concurrent degradation, focus has been placed on performance. The current study serves as a demonstration that deep learning can be effective in these conditions. Once demonstrated, attention can begin to turn toward implementation into auditory devices. BLSTM, used currently, involves both the forward and backward directions across each sentence pair and is therefore not causal. Future work will make such algorithms causal, e.g., by employing unidirectional LSTM.

The current study expands the scope of listening situations in which deep learning based algorithms are effective. In addition to their ability to improve intelligibility in the presence of various background interferences, they are able to effectively address combined reverberation and interference. The current conditions expand this scope to include interference from a single talker plus concurrent reverberation applied to both the target and interfering speech, which represents a particular challenge for both humans and machines. To our knowledge, the current study represents the first demonstration of a single-microphone (monaural) algorithm capable of producing speech-intelligibility improvements in both competing speech and room reverberation.

## V. CONCLUSIONS

- (1) A recurrent neural network incorporating bidirectional long short-term memory, trained to estimate the ideal ratio mask, was capable of producing substantial sentence-intelligibility improvements for speech corrupted by reverberation and a single interfering talker, for both HI and NH listeners, across a range of target-to-interferer ratios. Group-mean algorithm benefit for the HI listeners was 56 percentage points at the least-favorable target-to-interferer ratio. Intelligibility scores for the HI listeners having access to the algorithm numerically exceeded those of young NH listeners without the algorithm, in identical conditions.
- (2) A comparison was made between deep learning algorithms trained to output either reverberation-free, interference-free target speech (direct-sound algorithm, DS) or interference-free but reverberant target speech (reverberant algorithm, R). HI listeners displayed significantly higher intelligibility for the DS algorithm than for the R algorithm, but the NH listeners displayed similar scores across conditions.
- (3) A deep learning algorithm based on T-F masking can produce substantial intelligibility improvements for speech corrupted by reverberation and interference from another talker. The neural network was able to effectively perform simultaneous de-reverberation and segregation.

## ACKNOWLEDGMENTS

This work was supported in part by grants from the National Institute on Deafness and other Communication

Disorders (R01 DC015521 to E.W.H. and R01 DC012048 to D.L.W.). We gratefully acknowledge computing resources from the Ohio Supercomputer Center.

<sup>1</sup>Although the terms “speaker,” referring to a human producing speech, and “speaker separation,” referring to the process of segregating concurrent voices, have been used extensively in the signal-processing literature, the term “talker” is preferred currently. This usage is intended to differentiate a human producing speech (a talker) from the common term for the device used to transduce an electrical signal into broadcast acoustic energy (a speaker).

<sup>2</sup>The terms “interference” or “background interference” are used currently to encompass the variety of sound sources in the environment that can interfere with the speech signal of interest. These sources can include non-speech noises, speech babble, and a competing talker.

<sup>3</sup><https://github.com/ehabets/RIR-Generator>.

<sup>4</sup>The statistical analysis was also conducted using analysis of variance (ANOVA) with *post hoc* Holm-Sidak pairwise comparisons (also on RAU scores). The use of different TIRs for the two listener groups necessitated three separate ANOVAs—one repeated-measures ANOVA for each listener group and one mixed ANOVA for the across-group comparison. These analyses revealed that the pattern of statistical significance ( $p < 0.05$ ) was identical to that observed for the linear mixed effects model.

Allen, J. B., and Berkley, D. A. (1979). “Image method for efficiently simulating small-room acoustics,” *J. Acoust. Soc. Am.* **65**, 943–950.

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).

Bentsen, T., May, T., Kressner, A. A., and Dau, T. (2018). “The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility,” *PLoS One* **13**(5), e0196924.

Bramsløw, L., Naithani, G., Hafez, A., Barker, T., Pontoppidan, N. H., and Virtanen, T. (2018). “Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm,” *J. Acoust. Soc. Am.* **144**, 172–185.

Byrne, D., Parkinson, A., and Newall, P. (1990). “Hearing aid gain and frequency response requirements for the severely/profoundly hearing impaired,” *Ear Hear.* **11**, 40–49.

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.

Delfarah, M., and Wang, D. L. (2017). “Features for masking-based monaural speech separation in reverberant conditions,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **25**, 1085–1094.

George, E. L. J., Goverts, S. T., Festen, J. M., and Houtgast, T. (2010). “Measuring the effects of reverberation and noise on sentence intelligibility for hearing-impaired listeners,” *J. Speech Lang. Hear. Res.* **53**, 1429–1439.

Hazrati, O., and Loizou, P. C. (2012). “Tackling combined effects of reverberation and masking noise using ideal channel selection,” *J. Speech Lang. Hear. Res.* **55**, 500–510.

Healy, E. W., Delfarah, M., Vasko, J. L., Carter, B. L., and Wang, D. L. (2017). “An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker,” *J. Acoust. Soc. Am.* **141**, 4230–4239.

Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. L. (2015). “An algorithm to increase speech intelligibility for hearing-impaired listeners in novel segments of the same noise type,” *J. Acoust. Soc. Am.* **138**, 1660–1669.

Healy, E. W., Yoho, S. E., Wang, Y., Apoux, F., and Wang, D. L. (2014). “Speech-cue transmission by an algorithm to increase consonant

recognition in noise for hearing-impaired listeners,” *J. Acoust. Soc. Am.* **136**, 3325–3336.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. L. (2013). “An algorithm to improve speech recognition in noise for hearing-impaired listeners,” *J. Acoust. Soc. Am.* **134**, 3029–3038.

Hochreiter, S., and Schmidhuber, J. (1997). “Long short-term memory,” *Neural Comput.* **9**, 1735–1780.

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. Process.* **23**, 2136–2147.

IEEE (1969). “IEEE recommended practice for speech quality measurements,” *IEEE Trans. Audio Electroacoust.* **17**, 225–246.

Kim, C., and Stern, R. M. (2016). “Power-normalized cepstral coefficients (PNCC) for robust speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **24**, 1315–1329.

Kingma, D. P., and Ba, J. (2014). “Adam: A method for stochastic optimization,” arXiv:1412.6980.

Li, J., Xia, R., Fang, Q., Li, A., Pan, J., and Yan, Y. (2015). “Effect of the division between early and late reflections on intelligibility of ideal binary-masked speech,” *J. Acoust. Soc. Am.* **137**, 2801–2810.

Monaghan, J. J. M., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C. M., and Bleeck, S. (2017). “Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners,” *J. Acoust. Soc. Am.* **141**, 1985–1998.

Nábělek, A. K., and Dagenais, P. A. (1986). “Vowel errors in noise and in reverberation by hearing-impaired listeners,” *J. Acoust. Soc. Am.* **80**, 741–748.

Nábělek, A. K., and Mason, D. (1981). “Effect of noise and reverberation on binaural and monaural word identification by subjects with various audiograms,” *J. Speech Hear. Res.* **24**, 375–383.

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.

Ng, J. H., and Loke, A. Y. (2015). “Determinants of hearing-aid adoption and use among the elderly: A systematic review,” *Int. J. Audiol.* **54**, 291–300.

Rix, A., Beerends, J., Hollier, M., and Hekstra, A. (2001). “Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs,” in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 749–752.

Roman, N., and Woodruff, J. (2011). “Intelligibility of reverberant noisy speech with ideal binary masking,” *J. Acoust. Soc. Am.* **130**, 2153–2161.

Roman, N., and Woodruff, J. (2013). “Speech intelligibility in reverberation with ideal binary masking: Effects of early reflections and signal-to-noise ratio threshold,” *J. Acoust. Soc. Am.* **133**, 1707–1717.

Shao, Y., and Wang, D. L. (2008). “Robust speaker identification using auditory features and computational auditory scene analysis,” in *Proceedings of ICASSP*, pp. 1589–1592.

Srinivasan, S., Roman, N., and Wang, D. L. (2006). “Binary and ratio time-frequency masks for robust speech recognition,” *Speech Commun.* **48**, 1486–1501.

Studebaker, G. A. (1985). “A ‘rationalized’ arcsine transform,” *J. Speech, Lang., Hear. Res.* **28**, 455–462.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). “An algorithm for intelligibility prediction of time–frequency weighted noisy speech,” *IEEE Trans. Audio, Speech, Lang. Process.* **19**, 2125–2136.

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.

Zhao, X., Wang, Y., and Wang, D. L. (2014). “Robust speaker identification in noisy and reverberant conditions,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **22**, 836–845.

Zhao, Y., Wang, D. L., Johnson, E. M., and Healy, E. W. (2018). “A deep learning based segregation algorithm to increase speech intelligibility for hearing-impaired listeners in reverberant-noisy conditions,” *J. Acoust. Soc. Am.* **144**, 1627–1637.