JASA ARTICLE

A talker-independent deep learning algorithm to increase intelligibility for hearing-impaired listeners in reverberant competing talker conditions

Eric W. Healy,[1,a)] Eric M. Johnson,[1] Masood Delfarah,[2] and DeLiang Wang[2,b)]
[1]*Department of Speech and Hearing Science, and Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA*
[2]*Department of Computer Science and Engineering, The Ohio State University, Columbus, Ohio 43210, USA*

ABSTRACT:
Deep learning based speech separation or noise reduction needs to generalize to voices not encountered during training and to operate under multiple corruptions. The current study provides such a demonstration for hearing-impaired (HI) listeners. Sentence intelligibility was assessed under conditions of a single interfering talker and substantial amounts of room reverberation. A talker-independent deep computational auditory scene analysis (CASA) algorithm was employed, in which talkers were separated and dereverberated in each time frame (simultaneous grouping stage), then the separated frames were organized to form two streams (sequential grouping stage). The deep neural networks consisted of specialized convolutional neural networks, one based on U-Net and the other a temporal convolutional network. It was found that every HI (and normal-hearing, NH) listener received algorithm benefit in every condition. Benefit averaged across all conditions ranged from 52 to 76 percentage points for individual HI listeners and averaged 65 points. Further, processed HI intelligibility significantly exceeded unprocessed NH intelligibility. Although the current utterance-based model was not implemented as a real-time system, a perspective on this important issue is provided. It is concluded that deep CASA represents a powerful framework capable of producing large increases in HI intelligibility for potentially any two voices.
 https://doi.org/10.1121/10.0001441
(Received 16 December 2019; revised 28 May 2020; accepted 29 May 2020; published online 23 June 2020)
[Editor: Karen S. Helfer] Pages: 4106–4118

I. INTRODUCTION

One of the most important characteristics of a successful speech separation or noise-reduction algorithm involves its ability to operate in a wide variety of acoustic environments. This ability is important for machine learning algorithms, including deep learning, because they typically undergo training involving exposure to one or more acoustic environments. If not properly trained, the algorithm can "overfit" its training environments (its training data) and have difficulty performing well in acoustic environments that differ from those environments.

One of the greatest algorithmic challenges involves generalization to untrained noisy backgrounds. This is because different noises can vary so widely in their acoustic characteristics, and altogether, "noise" can be considered to occupy a vast space. Accordingly, a recent line of studies involving deep learning noise reduction for hearing-impaired (HI) listeners focused initially on this aspect of generalization. These studies demonstrated intelligibility improvements for HI listeners across a progression that first involved overlapping noise segments for both training and testing (Healy *et al.*, 2013), to novel (untrained or "unseen") segments of the same noise type (Healy *et al.*, 2015; Monaghan *et al.*, 2017; Zhao *et al.*, 2018; Keshavarzi *et al.*, 2019), to entirely novel noise types for training and testing (Chen *et al.*, 2016).

Closely related to this ability to generalize to untrained noises is the ability to deal with a variety of acoustic interferences. Accordingly, studies have progressed from improving HI speech intelligibility in steady-state noise (Healy *et al.*, 2013; Healy *et al.*, 2014; Monaghan *et al.*, 2017; Zhao *et al.*, 2018), to nonstationary noises including 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), to complex noises containing a variety of different sound sources (Healy *et al.*, 2015; Chen *et al.*, 2016; Zhao *et al.*, 2018). More recently, improvements in compound interferences have been shown, including concurrent background noise and reverberation (Zhao *et al.*, 2018).

Other studies have focused on speaker separation, which involves the isolation of one talker from one or more interfering talkers. Despite this representing a somewhat different computational task, deep learning approaches have also proven effective at improving HI-listener intelligibility (Healy *et al.*, 2017; Bramsløw *et al.*, 2018). These results

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

have also held when reverberation was added to single-talker interference (Healy *et al.*, 2019), which requires the algorithm to perform both speaker separation and dereverberation.[1]

Another important aspect of generalization involves whether the particular target talker is employed during algorithm training. Two views exist in this regard. In the first view, the tendency for an algorithm to perform optimally with a particular frequent communication partner is seen as a virtue. This is related to the general concept that effective communication between frequent communication partners holds particular value (e.g., Tye-Murray *et al.*, 2016). To achieve this, an algorithm would simply be trained to favor the voice of one or more frequent partners. But perhaps more ideally, a highly effective algorithm would perform equally well when targeting a voice that it had not previously encountered during training. This is referred to as talker independence (or speaker independence). Perhaps the simplest approach that yields talker independence involves "large scale" training of the network (Chen *et al.*, 2016) using many different talkers so that the network learns to identify speech more generally and is not overfit to any particular voice.

Some work has shown that deep learning can be effective in a talker-independent context. Chen and Wang (2017) trained a recurrent neural network (RNN) on speech from 83 talkers in order to generalize to untrained talkers. Six untrained talkers of both primary genders were used for testing. Although human intelligibility was not assessed, considerable increases were observed using a standard objective intelligibility metric based on acoustic analysis. Goehring *et al.* (2017) trained a neural network using speech from 36 talkers and tested using a single talker not employed for training. They observed improved intelligibility in noise for cochlear-implant (CI) users in stationary noises but not in multi-talker babble. Goehring *et al.* (2019) then increased the training set to 80 different talkers and again tested using a talker not employed for training. They observed improved intelligibility in multi-talker babble for CI users and for normal hearing (NH) listeners in CI simulation, but not for CI users tested in traffic noise. Keshavarzi *et al.* (2019) trained an RNN using 80 talkers and used 6 untrained talkers for testing. Although human intelligibility was not assessed, improvements in objective intelligibility measures were obtained, and HI listeners displayed a slight subjective intelligibility preference for the processed relative to the unprocessed speech in multi-talker babble.

These prior studies suggest that deep learning can be effectively used to improve the intelligibility of noisy speech in talker-independent settings, at least for CI users and in some noise conditions. The current study was designed to provide a clear demonstration of improved intelligibility for HI listeners in two-talker mixtures and room reverberation. HI listeners show a large intelligibility gap from NH listeners when interfering speech is present (Festen and Plomp, 1990; Moore, 2007; Healy *et al.*, 2017). Further, room reverberation characterizes daily listening environments and, when combined with interfering speech, is a major hurdle for target speech perception, particularly for HI listeners (Plomp, 1976; Culling *et al.*, 2003; Healy *et al.*, 2019). The current study thus involved speaker separation in reverberant conditions in order to further challenge the algorithm and to increase the ecological validity of the acoustic environment. The eventual goal is to provide an algorithm that can separate any two voices, even in suboptimal conditions.

One challenge particular to separating concurrent talkers not employed during algorithm training is known as the permutation problem (for more detail, the interested reader is directed to Hershey *et al.*, 2016; Yu *et al.*, 2017; Liu and Wang, 2019). Briefly, a deep neural network (DNN) involves an input layer that receives features of the input signal, a number of hidden layers, and an output layer. During training, given outputs in the output layer are assigned to given target sounds. Learning fails when the assignment of outputs to targets is arbitrary and indeterminate, as it is during training for talker independence. More specifically, the permutation problem states that output layers of a DNN cannot be straightforwardly assigned to individual untrained talkers, which stands in contrast to talker-dependent separation (Du *et al.*, 2014; Huang *et al.*, 2015; Healy *et al.*, 2017; Healy *et al.*, 2019). This problem can be addressed through the technique of deep clustering (Hershey *et al.*, 2016), which combines DNN based feature transformations (known as embeddings) and clustering. In deep clustering, a DNN learns to transform (or expand) each time-frequency (T-F) unit to an embedding vector so that the DNN produces similar embedding vectors for the T-F units that are dominated by the same speaker. The DNN then estimates the ideal binary mask (Wang, 2005) by clustering embedding vectors (i.e., T-F units) into different talkers. Another solution involves permutation-invariant training (Kolbaek *et al.*, 2017), in which all possible permutations between network outputs and target sounds are considered during training. The best output-to-target assignment is determined, then error is minimized given that assignment. In the current study, the permutation problem was addressed by extending deep computational auditory scene analysis (deep CASA), which integrates permutation-invariant training at the frame level and sequential clustering (Liu and Wang, 2019). Deep CASA has been shown to produce state-of-the-art talker-independent speaker separation results in anechoic conditions.

The current study aimed to improve the intelligibility of a target talker in the presence of a single interfering talker and substantial amounts of room reverberation. Thus, both speaker separation and speech dereverberation had to be performed. In addition to the primary generalization involving untrained talkers, the current implementation required generalization to untrained intensity relationships between target and interfering speech (in half of the conditions for each listener group), and untrained reverberation characteristics (room impulse responses, RIRs) in all conditions. Perhaps importantly, the use of different speech corpora for training and testing also required a "cross-corpus"

generalization. HI and NH listeners were employed, and large increases in intelligibility were observed in all conditions. Whereas speech understanding in most background interferences is challenging for HI listeners, understanding one voice in the presence of another is also challenging for NH listeners when reverberation is present. This allowed large benefits to also be observed for these listeners.

The algorithm was a variation of deep CASA (Liu and Wang, 2019). In this approach, a simultaneous grouping stage precedes a sequential grouping stage. The rationale comes from classic theories of auditory scene analysis, in which the human analysis of complex sound scenes is hypothesized to involve such stages (Bregman, 1990), and from computational auditory scene analysis (Wang and Brown, 2006), which often employs such stages. In the current study, simultaneous grouping was performed by separating the acoustic elements corresponding to the two sound sources in each time frame using deep learning. Specifically, this stage employed a U-Net convolutional neural network with densely-connected layers (Dense-UNet; Liu and Wang, 2019). In the sequential grouping stage, these separated frames were organized to form two streams, one for each voice. This process was performed using a temporal convolutional network (TCN; Bai *et al.*, 2018; Lea *et al.*, 2016). The rationale for deconstructing speech separation into these two stages involves the fact that they are different operations, requiring different considerations, and potentially different techniques. By separating the grouping processes, each could be optimized separately, which led to high performance.

Further, previous studies have largely concentrated on estimating only the amplitude representation of the signal of interest. This isolated amplitude information is then combined with the phase of the original unprocessed sound mixture ("noisy phase") to reconstruct the isolated signal of interest. In the current study, both the amplitude and phase of the signal of interest were obtained by working in the complex domain. In this approach, the real and imaginary parts are both estimated by the DNN, which allows both the amplitude and phase of the signal of interest to be calculated.

II. METHOD

A. Subjects

Two groups of listeners participated. The HI group consisted of ten listeners, representing typical hearing aid users with bilateral sensorineural hearing loss. All were binaural hearing aid users, with one exception described below. They ranged in age from 62 to 88 years (mean = 73), and four were female. These listeners were recruited from The Ohio State University Speech-Language-Hearing Clinic and surrounding community. Otoscopy, tympanometry (ANSI, 1987), and pure-tone audiometry (ANSI, 2004, 2010a) were used to verify the listeners' hearing losses on day of test. Otoscopy was unremarkable for all listeners. Middle-ear peak pressures and compliances were within normal limits for all listeners except for HI10, who presented with flat tympanograms. However, bone-conduction thresholds showed no significant air-bone gap for this listener (or any other), indicating a likely cochlear site of lesion. Figure 1 presents audiograms for each of these listeners, who are numbered in order of increasing pure-tone average audiometric thresholds (PTAs; means across thresholds at 500, 1000, 2000 Hz, and ears). Each panel also indicates ages and genders. PTAs ranged from 27 to 52 dB hearing level (HL) with a mean of 43 dB HL. The HI listeners generally had sloping hearing losses that ranged in degree from mild to profound.

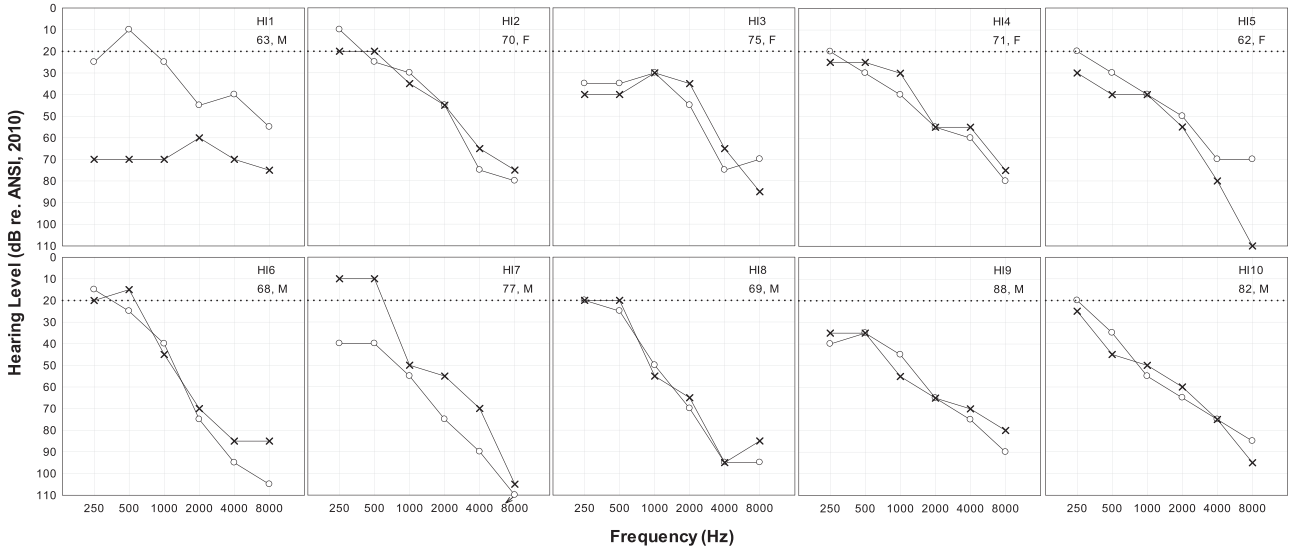

FIG. 1. Pure-tone air-conduction audiometric thresholds for the listeners with hearing impairment. Listeners are numbered in order of increasing degree of hearing loss. Right ears are represented by circles, and left ears are represented by X's. An arrow indicates a threshold exceeding audiometer limits.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

https://doi.org/10.1121/10.0001441

Rather than binaural hearing aids, HI1 uses an amplification system with bilateral microphones and contralateral routing of one of the signals (BiCROS; Harford, 1966). BiCROS systems are sometimes recommended for patients with bilateral hearing loss, but with one ear deemed "unaidable." With this type of fitting, the better ear receives an amplified mixture of the sound arriving at both ears, whereas the poorer ear receives no input from the instrument. Only the better (right) ear of HI1 was tested currently, in accord with the input he receives in everyday listening. Thresholds for both of his ears are presented in Fig. 1, but PTA was calculated based on the ear receiving input.

The NH group consisted of ten listeners (all female) with pure-tone audiometric thresholds of 20 dB HL or lower at octave frequencies from 250 to 8000 Hz on day of test (ANSI, 2004, 2010a). Recruited from undergraduate courses at The Ohio State University, they ranged in age from 18 to 21 years (mean = 19.7) and represented young listeners with "ideal" hearing abilities. All participants (HI and NH) were native speakers of American English having no previous exposure to the test sentences used in the current study. They received either extra course credit or a monetary incentive for participating.

B. Stimuli

The algorithm was trained using materials from the Wall Street Journal Continuous Speech Recognition Corpus (WSJ0; Paul and Baker, 1992). This corpus was developed to support research on large-vocabulary continuous speech recognition systems and contains recordings of many talkers reading Wall Street Journal newspaper articles dating from 1987 to 1989 (approximately 39 000 recorded sentences totaling approximately 80 h of audio). The WSJ0 corpus is commonly used for training and testing talker-independent speaker separation (Hershey *et al.*, 2016). Training sentences for the current study were drawn from the si_tr_s folders of the WSJ0 corpus, which contain recordings from 49 male and 52 female talkers, each of whom produced an average of 124 sentences. To generate a reverberant two-talker mixture, two sentences produced by different talkers were selected and equalized to the same root mean square level. Each of the talkers comprising a sentence pair could be male or female, but they were always two different talkers. If the sentences were different in duration, the longer ones were trimmed to match the shorter to avoid long periods containing a single talker. All training and test signals used for processing were sampled at 16 kHz with 16 bit resolution.

To generate reverberation, a simulated room with dimensions of 6 m × 7 m × 3 m was used. The virtual microphone (representing the listener) was placed at a fixed position in the room located at (3, 4, 1.5) m. The room T_{60} for each sentence pair was a randomly selected value between 0.3 and 1.0 s. Each talker was randomly positioned at one of 36 angles evenly distributed around the microphone, with the target talker 1 m from the microphone and the interfering talker 2 m away. Both talkers had the same elevation as the microphone. Then, each individual recorded sentence of the pair was convolved with an RIR to generate reverberant speech, using an RIR generator[2] that implements the image method (Allen and Berkley, 1979). The two reverberant recordings were mixed in the time domain to generate the reverberant two-talker mixture. A single target-to-interferer ratio (TIR) of 0 dB was used to create the training data. In total, 200 500 such training mixtures were generated, from which 500 mixtures were set aside for cross validation.

The test stimuli were drawn from a different speech corpus. Specifically, the test material consisted of one male talker and one female talker reading the Institute of Electrical and Electronics Engineers (IEEE) revised list of phonetically balanced sentences (IEEE, 1969), with the male designated as the target talker. Neither of these talkers is in the WSJ0 corpus, which was used during algorithm training. Different-gender talkers were employed for testing to reduce confusion for the human subjects with regard to which talker was the target. Algorithm performance has been shown to be similar when target and interfering talkers are different genders versus when they are both male or both female (see Table V from Liu and Wang, 2019).

To generate the test stimuli, the same procedure was followed as for the training set, except for additional test TIRs and that the set of talker positions was shifted by 5 degrees to ensure that the test RIRs were different from those used for training. T_{60} values for the test sentences were 0.6 and 0.9 s. These values are representative of considerable amounts of room reverberation. The value of 0.6 s corresponds to the upper limit for acceptable room reverberation in classrooms (ANSI, 2010b), whereas the value of 0.9 exceeds that limit. Each test stimulus consisted of two sentences, one spoken by each talker. The sentences comprising each pair were selected so that they matched in duration without trimming. The average duration difference between members of a pair was 5 ms, and the difference did not exceed 10 ms for any pair. In total, 160 such two-talker mixtures were generated, and each was prepared in the various TIR and room T_{60} conditions. Note that, in addition to the different RIRs used for training and test, the fixed training TIR of 0 dB matched one test TIR used for each listener group but was different from the other TIR used for each group.

C. Algorithm description

As stated in Sec. I, the current solution was built upon the principles of auditory scene analysis and CASA, the latter of which is a traditional approach to speech separation. This traditional approach typically addresses the speech-separation problem in two organizational stages: (1) simultaneous grouping and (2) sequential grouping. The current algorithm was a variation of deep CASA, which was recently proposed for talker-independent speaker separation in anechoic conditions (Liu and Wang, 2019). Deep CASA first involves a simultaneous grouping stage, in which the source signals are separated in each time frame, and then a

sequential grouping stage, in which the separated frames are organized to form two streams, each corresponding to one talker. In the current study, deep CASA was extended to speaker separation in reverberant conditions by performing both frame-level segregation and dereverberation during the simultaneous grouping stage. The estimated direct-sound signal for each talker was then obtained during the sequential grouping stage.

As also stated in Sec. I, working in the complex domain allows both the amplitude and phase of the segregated signals to be estimated. This use of the complex domain is reflected in both the training target and in the features extracted from the original sound mixtures and fed to the DNN. The training target refers to the goal of the deep learning algorithm—the output it is trained to produce. This output typically represents the interference-free speech. Previous studies have employed the ideal ratio mask (IRM) as the training target. When the IRM is employed, only the amplitude information of the interference-free speech is obtained. As indicated earlier, this information is then combined with the noisy phase to construct the interference-free speech output. In the current study, the training target was the complex ideal ratio mask (cIRM, Williamson *et al.*, 2016). In this scheme, both the real and imaginary parts are estimated by the network, allowing both the amplitude and phase of the interference-free speech to be calculated. Further, when the original sound mixture contains reverberation, the training target can either be the anechoic (reverberation-free) target speech, fully reverberant target speech, or target speech containing only some of the reverberation components. The choice was made currently to dereverberate the segregated signals because HI listeners achieve higher intelligibility in two-talker reverberant conditions when dereverberation accompanies signal separation (Healy *et al.*, 2019).

The features employed currently involved the real and imaginary components of the complex short-time Fourier transform (STFT). They were selected because they are demonstrated effective complex-domain features (Liu and Wang, 2019). This selection contrasts with the complementary feature set used in other studies (e.g., Healy *et al.*, 2019), which are in the real domain and used to estimate the IRM defined in the magnitude domain.

The talker-independent speaker separation problem in reverberant conditions was defined as extracting two anechoic talkers $s_1(t)$ and $s_2(t)$ from the mixture signal $y(t)$,

$$y(t) = h_1(t) * s_1(t) + h_2(t) * s_2(t), \tag{1}$$

where $h_1(t)$ and $h_2(t)$ are RIRs for specific locations in the reverberant room, $*$ denotes convolution, and the two test talkers were not employed during algorithm training. Thus, this separation goal involving simultaneous signal separation and dereverberation contrasts with the goal of separating the two signals, but not performing dereverberation, which would be accomplished by estimating $h_1(t) * s_1(t)$ and $h_2(t) * s_2(t)$. The two stages of the current deep CASA algorithm for reverberant conditions are presented in the following two sections. Additional technical details can be found in Liu and Wang (2019).

1. Simultaneous grouping

A feedforward DNN was used for this stage with an architecture referred to as Dense-Unet (Liu and Wang, 2019). Dense-Unet is motivated by recent successes with DenseNet (Huang *et al.*, 2017) and U-Net (Ronneberger *et al.*, 2015), and consists of a U-Net having densely-connected layers. U-Net is a specialized convolutional neural network composed of an encoder and a decoder that together form a U-shaped architecture. The encoder in the first half of the model maps lower-level input features into a higher level of abstraction. It does so by decreasing the resolution of the input-feature representation at each stage of the encoder path. In the current implementation, this portion consisted of four downsampling layers with dense convolutional layers interleaved between every two layers. The decoder in the second half of the model projects the encoded features, now having lower resolution, back to their original resolution. Currently, this portion consisted of four upsampling layers interleaved with dense convolutional blocks.

Given a reverberant mixture signal, an STFT was applied having frames of length 32 ms with the frame shift of 8 ms (i.e., 75% overlap). The real and imaginary STFT features $M(m,f)$ were extracted and delivered to the DNN just described, where m represents the frame index and f is the frequency channel. This network was used to estimate two cIRMs, one for each talker. These masks were pointwise multiplied by $M(m,f)$ in the complex domain to form two STFT signals $\hat{S}_{u_1}(m,f)$ and $\hat{S}_{u_2}(m,f)$. These signals represent the separated and dereverberated talker frames, yet to be organized over time.

Frame-level permutation invariant training (tPIT; Kolbaek *et al.*, 2017) criteria were used for network optimization. Specifically, two loss functions were calculated per time frame as follows:

$$l_1(m) = \Sigma_f|\hat{S}_{u_1}(m,f) - S_1(m,f)| + \Sigma_f|\hat{S}_{u_2}(m,f) - S_2(m,f)|, \tag{2}$$

$$l_2(m) = \Sigma_f|\hat{S}_{u_1}(m,f) - S_2(m,f)| + \Sigma_f|\hat{S}_{u_2}(m,f) - S_1(m,f)|, \tag{3}$$

where $S_1(m,f)$ and $S_2(m,f)$ are the concatenated real and imaginary STFT features of the two source talkers and $|\cdot|$ denotes magnitude spectrogram. The loss values $l_1(m)$ and $l_2(m)$ were used to obtain the optimally organized talker signals $\hat{S}_{o_1}(m,f)$ and $\hat{S}_{o_2}(m,f)$,

$$\hat{S}_{o_1}(m,f),\ \hat{S}_{o_2}(m,f) = \begin{cases} \hat{S}_{u_1}(m,f),\ \hat{S}_{u_2}(m,f) & \text{if } l_1(m) \le l_2(m) \\ \hat{S}_{u_2}(m,f),\ \hat{S}_{u_1}(m,f) & \text{otherwise.} \end{cases} \tag{4}$$

https://doi.org/10.1121/10.0001441

The above equation selects the output-to-talker assignment with the smaller loss. Then, via inverse STFT, $\hat{S}_{o_1}(m,f)$ and $\hat{S}_{o_2}(m,f)$ were converted into time-domain signals $\hat{s}_{o_1}(t)$ and $\hat{s}_{o_2}(t)$. Finally, the network was optimized to minimize the signal-to-noise ratio loss function J^{SNR} as

$$J^{SNR} = -10 \sum_{i=1,2} \log \frac{\sum_t s_i(t)^2}{\sum_t \left[s_i(t) - \hat{s}_{o_i}(t)\right]^2} \tag{5}$$

using the Adam optimizer (Kingma and Ba, 2014). Figure 2(a) depicts this simultaneous grouping stage.

2. Sequential grouping

Once the two talkers have been separated and dereverberated in each time frame, the frames need to be organized across time into two streams corresponding to the two utterances. A temporal convolutional network (TCN) (Bai *et al.*, 2018; Lea *et al.*, 2016) was used as the DNN in this stage. TCNs have recently been proposed to replace RNNs based on their performance on a variety of tasks. A TCN involves a series of convolutional layers stacked to form a deep network. Its architecture allows it to capture long-range contextual information, which is desirable for tracking a talker over a long utterance. The current TCN had eight dilated convolutional blocks, each composed of three convolutional layers. As in the previous simultaneous grouping stage, this sequential stage used the Adam optimizer (Kingma and Ba, 2014) during training.

As seen in Eqs. (2)–(5), the training of the previous simultaneous grouping stage uses $s_1(t)$ and $s_2(t)$ to optimize the network with the optimal frame-speaker assignments. Because these signals are not available at test time, the sequential grouping stage was trained to predict a temporal organization given the unorganized signals $\hat{S}_{u_1}(m,f)$ and $\hat{S}_{u_2}(m,f)$. The input features to this stage were $|M(m,f)|$, $|\hat{S}_{u_1}(m,f)|$, and $|\hat{S}_{u_2}(m,f)|$. The training target was a 2×1 vector $\boldsymbol{A}$. Specifically, $\boldsymbol{A} = [1,0]$ indicates that $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$ correctly represent frames for talker 1 and talker 2, respectively, while $\boldsymbol{A} = [0,1]$ means that the frames need to be switched. Correct prediction of $\boldsymbol{A}$ over the entire utterance optimally organizes the separated frames $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$. To this end, the network predicted an embedding vector $\boldsymbol{V}(c) \in \mathbb{R}^d$ per time frame and optimized the loss function (Hershey *et al.*, 2016; Liu and Wang, 2019),

$$J^{DC} = \|\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{A}\boldsymbol{A}^T\|_F^2, \tag{6}$$

in which $\|\cdot\|_F$ denotes the Frobenius norm. The first term of Eq. (6) represents the predicted frame-talker assignment and the second term represents the optimal frame-talker assignment. The loss minimization in Eq. (6) modifies the network so as to approach the optimal assignment.

Once the embeddings $\boldsymbol{V}(c)$ were predicted, a K-means algorithm clustered these vectors into two groups, labeled as $\hat{A}(m) = \{0,1\}$, which were used to organize $\hat{S}_{u_1}(m,f)$ and $\hat{S}_{u_2}(m,f)$. Finally, the inverse STFT was applied to yield the time-domain signals $\hat{s}_1(t)$ and $\hat{s}_2(t)$, which are the estimated anechoic talker signals. Figure 2(b) depicts this sequential grouping stage.

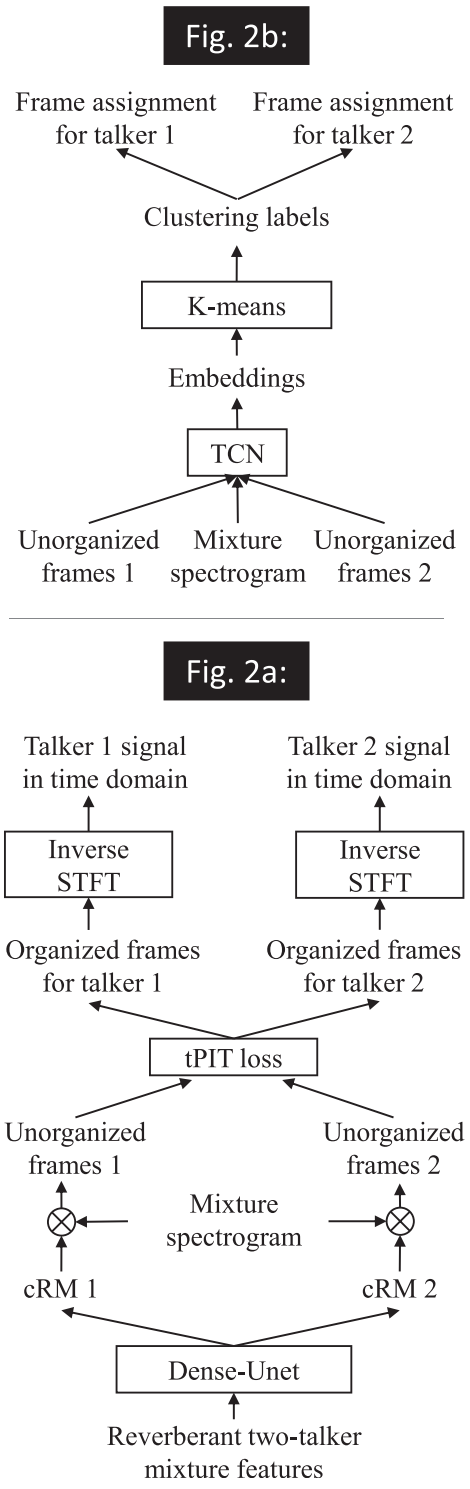

FIG. 2. Schematic of the current deep CASA algorithm for talker-independent speaker separation in reverberant conditions. The processing flow is from bottom to top, and so the lower panel (a) represents the first stage (simultaneous grouping) and the upper panel (b) represents the second stage (sequential grouping). cRM stands for an estimate of the cIRM.

In the current study, the simultaneous grouping stage was trained for 20 epochs, and the parameters with the lowest cross-validation loss in Eq. (5) were chosen and used during training of the sequential-grouping stage. Then, the sequential-grouping TCN was trained for 15 epochs, and the parameters with the smallest number of sequential organization errors on the cross-validation set were selected and used during the inference phase. The lowest cross-validation losses occurred before these epoch numbers.

Figure 3 displays spectrogram images of various stages of processing for an example reverberant two-talker mixture. Figure 3(a) displays the mixture. The target utterance was, "All sat frozen and watched the screen" and the interferer utterance was, "Hold the hammer near the end to drive the nail." They were mixed at a TIR of –5 dB in a room with $T_{60} = 0.9$ s. Figures 3(b) and 3(c) display these individual utterances prior to mixing and the addition of reverberation. Figures 3(d) and 3(e) display these utterances separated from the reverberant two-talker mixture [Fig. 3(a)] using the deep CASA algorithm. A visual indication of algorithm accuracy can be obtained by comparing Fig. 3(b) to Fig. 3(d) and Fig. 3(c) to Fig. 3(e), all relative to Fig. 3(a).

D. Procedure

There were eight conditions for each listener (two processing conditions × two TIRs × two T_{60}s). The processing conditions consisted of the concurrent reverberant sentences prior to and following processing by the algorithm to target the anechoic speech of the male talker (unprocessed/processed). The TIRs were 0 and 5 dB for the HI listeners and –5 and 0 dB for the NH listeners. The T_{60} values of 0.6 and 0.9 s were used for both listener groups. Each listener heard 160 sentence pairs, blocked by condition, with 20 sentences in each condition. Unprocessed/processed conditions were presented juxtaposed within each TIR-T_{60} block. The order of the TIR-T_{60} blocks was randomized for each listener, as was the order of the two processing conditions within each TIR-T_{60} block. The sentence materials were presented to each listener in a fixed order to ensure a random correspondence between sentence pairs and conditions. No sentence was used more than once for any listener, including that no sentence was used for both target and interferer.

Test stimuli were played back using a Windows PC and an Echo Digital Audio Gina 3 G digital-to-analog converter (Santa Barbara, CA), routed through a Mackie 1202-VLZ mixer (Woodinville, WA), and presented diotically over Sennheiser HD 280 Pro headphones (Wedemark, Germany). The stimuli were scaled to the same total root-mean-square level and presented at 65 dBA in each ear for the NH listeners, as measured by a sound-level meter and flat-plate coupler (Larson Davis models 824 and AEC 101, Depew, NY).

For the HI listeners, the test stimuli were presented at 65 dBA plus individualized frequency-specific gains, as prescribed by the NAL-RP hearing-aid fitting formula (Byrne *et al.*, 1990).[3] In the case of HI1, who was tested monaurally in the right ear, the NAL-RP gains were calculated using audiometric thresholds for that ear only. A RANE DEQ 60 L digital equalizer (Mukilteo, WA) was used to apply the desired frequency-gain response for each HI listener, as described in Healy *et al.* (2015). Since the NAL-RP formula does not prescribe gains for 125 or 8000 Hz, the gains for 250 and 6000 Hz, respectively, were applied to these two frequencies.

During a brief familiarization immediately preceding formal testing, listeners heard 25 IEEE practice sentences

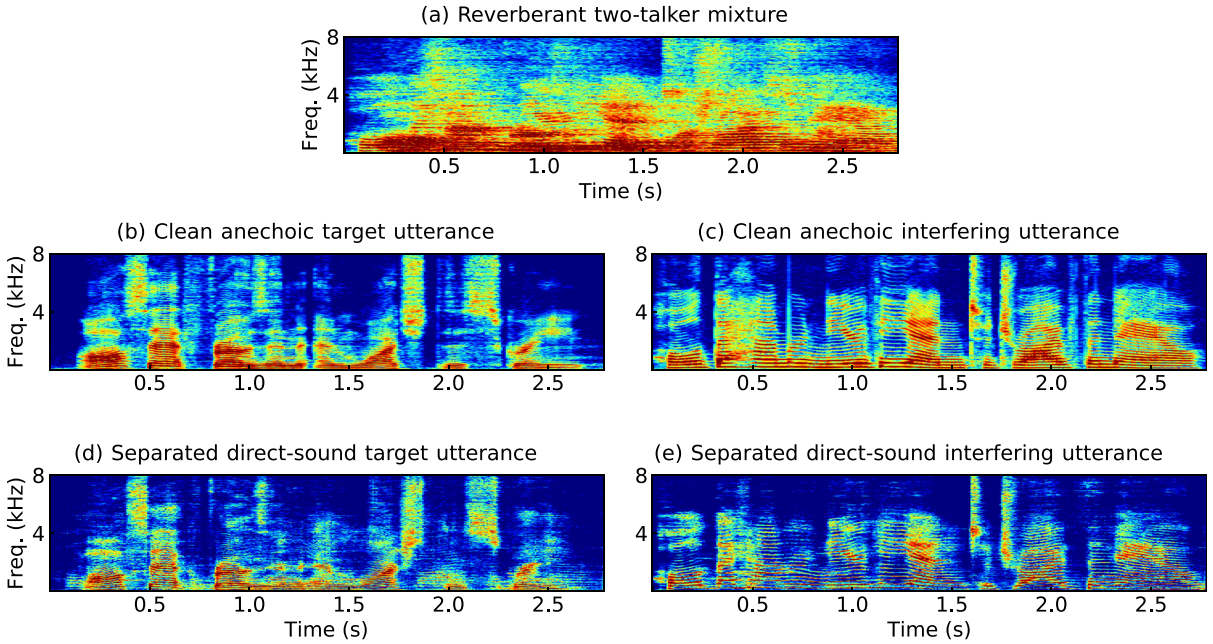

FIG. 3. (Color online) Spectrogram images illustrating the separation of a target sentence from a reverberant mixture of two talkers using the deep CASA algorithm. The utterances were mixed at TIR = –5 dB in a room with $T_{60} = 0.9$ s.

https://doi.org/10.1121/10.0001441

equally divided into five blocks. The practice conditions were: (1) anechoic speech spoken by the target talker only, (2) algorithm processed, in the most favorable TIR-T_{60} condition for each listener type, (3) algorithm processed, in the least favorable condition for each listener type, (4) unprocessed, in the most favorable condition for each listener type, and (5) unprocessed, in the least favorable condition for each listener type. These sentences were distinct from both the targets and interferers used for formal testing.

During this familiarization phase, the HI listeners were asked if the presentation level was comfortable. Two HI listeners (HI5 and HI10) reported that the stimuli sounded loud. After the overall presentation level was reduced by 5 dB, they reported that it was comfortable. The overall presentation level after application of NAL-RP gains and comfort adjustment ranged from 74.7 to 90.5 dBA across HI listeners (mean = 83.3 dBA). Because their individualized hearing-aid gains were provided via the experimental apparatus, HI listeners were tested with their hearing aids removed.

Following familiarization, the listeners heard the eight blocks of experimental conditions while seated in a double-walled audiometric booth with the experimenter. They were instructed to attend to the male voice, to repeat back each sentence as best they could, and to guess if unsure of what was said. No sentences were repeated for any listener, and listeners were blind to the condition being presented. The experimenter controlled the presentation of each stimulus and scored keywords correctly reported. The 20 target sentences presented in each experimental condition contained five keywords each, for 100 keywords in each condition. The total duration of testing was approximately 1 h for each listener.

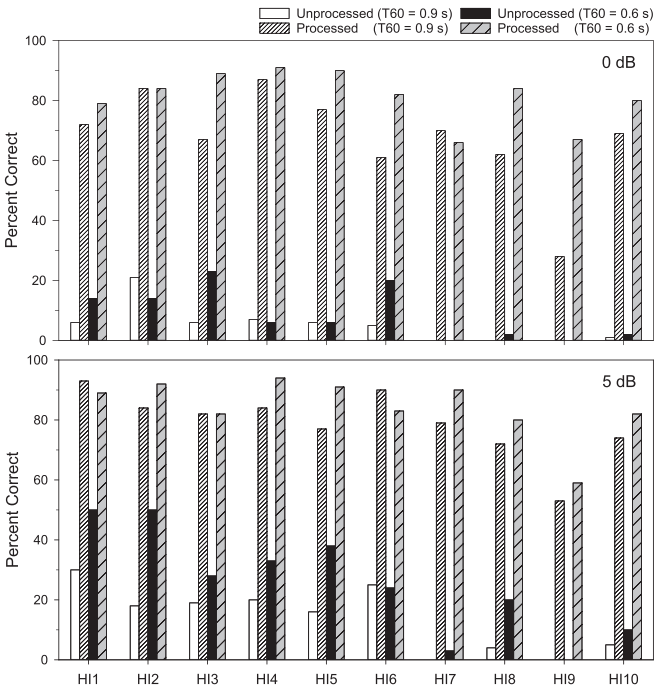

FIG. 4. Sentence-intelligibility scores for each individual HI listener in the presence of a single interfering talker and substantial room reverberation. Listeners are numbered and plotted in order of increasing hearing loss, as in Fig. 1. Each column represents a different condition. The unhatched white and black columns represent scores for unprocessed, reverberant concurrent sentences, and the hatched columns represent scores following algorithm processing, which targeted the interference-free, reverberation-free target speech. The white-filled columns represent scores when the room T_{60} equaled 0.9 s, and the black and shaded columns represent scores when the room T_{60} equaled 0.6 s. The two TIRs of 0 and 5 dB are displayed in separate panels.

III. RESULTS AND DISCUSSION

A. Human performance

Sentence intelligibility was operationalized as the percentage of sentence keywords correctly reported. Figures 4 and 5 display intelligibility for each individual listener in each condition: Fig. 4 for the HI listeners and Fig. 5 for the NH listeners. In Fig. 4, listeners are plotted in order of increasing PTA. Figure 6 displays group-mean scores and standard errors of the mean for each condition, with the HI and NH listeners plotted separately. In each figure, the unprocessed and processed conditions for each T_{60} are represented by different columns, and the TIRs are displayed in separate panels. The algorithm benefit for each listener (or group of listeners) at a given TIR and T_{60} corresponds to the difference between a solid column (unprocessed) and the hatched column directly to the right (processed). The absence of a column indicates that the listener was unable to correctly report any of the 100 keywords in that condition, which occurred for HI7, HI8, and HI9.

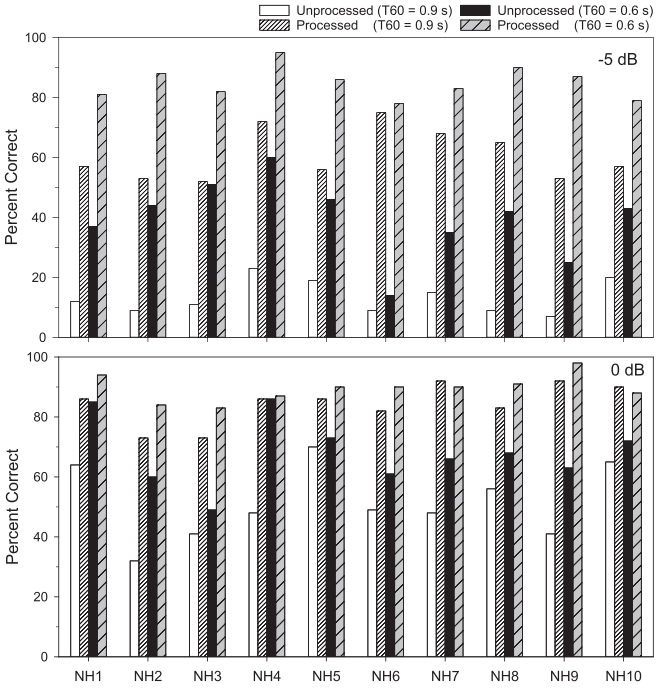

FIG. 5. As Fig. 4, but for the NH listeners, who are numbered and plotted arbitrarily, in the order in which they participated in the experiment.

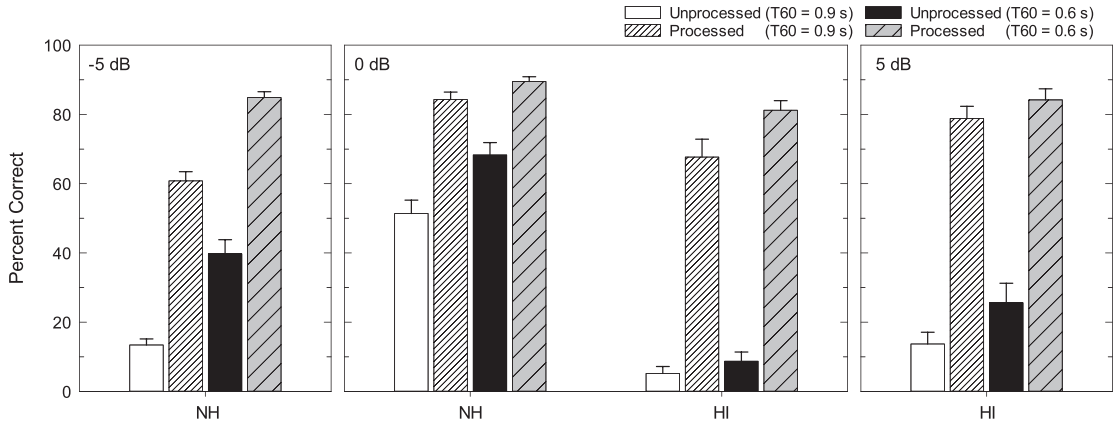

FIG. 6. Group-mean sentence-intelligibility scores and standard errors of the mean for each condition. The unprocessed and processed conditions for each T_{60} are represented by different columns, and the TIRs (−5, 0, and 5 dB) are displayed in separate panels, as in Figs. 4 and 5. Means for the NH and HI listeners are presented separately. Note the different TIRs employed for the two listener groups.

1. HI listeners

As Fig. 4 shows, every HI listener received algorithm benefit in every condition. Considering all listeners and conditions (40 processed–unprocessed cases), benefit was 50 percentage points or greater in 93% of cases, 60 percentage points or greater in 78% of cases, and 70 percentage points or greater in 28% of cases. Algorithm benefit across all conditions for individual HI listeners ranged from 52 percentage points (HI9) to 76 percentage points (HI7), with a grand mean overall benefit of 64.7 percentage points.

As expected, unprocessed scores tended to decrease from left to right in Fig. 4 as degree of hearing loss increased. This is likely due to the difficulty associated with providing adequate amplification in cases of severe to profound hearing loss and the increased suprathreshold deficits associated with greater degrees of sensorineural hearing loss. This association between PTA and unprocessed scores was confirmed using Spearman's rank correlations between the rankings of listener PTAs and scores, for each of the four unprocessed conditions (each $|r_s| \geq 0.74$, $p \leq 0.01$). However, the same consistent decrease in performance with increasing PTA was not as clearly observed in the algorithm-processed conditions. In these conditions, scores tended to be more consistently high in each panel. Spearman's correlations showed no significant relationships between PTA and score for three of the four processed conditions (each $|r_s| \leq 0.57$, $p \geq 0.08$), with (TIR = 5 dB, T_{60} = 0.9 s) being the exception ($|r_s| = 0.78$, $p = 0.005$). Therefore, processed scores were generally high even for those listeners having large amounts of hearing loss and very low unprocessed scores. Consequently, and expectedly, algorithm benefit was largely related to unprocessed scores ($|r| = 0.48$, $p = 0.002$, across all conditions), where lower unprocessed scores produced more room to improve and tended to produce larger benefit.

With regard to HI group-mean scores, conditions progress from least favorable toward the left of Fig. 6 to most favorable at the right. The first pair of conditions had the lowest TIR and highest reverberation time (TIR = 0 dB, T_{60} = 0.9 s). It produced the lowest group-mean HI unprocessed score of 5%, which rose to 68% following processing (benefit = 62.5 percentage points). The next pair, in the same 0 dB TIR panel, produced the largest benefit. This mean HI unprocessed score of 9% rose to 81% following processing (benefit = 72.5 percentage points). In the next panel, the TIR was 5 dB, and both unprocessed and processed scores were higher. The left pair of conditions in this panel produced an HI benefit of 65.1 percentage points. The rightmost pair of conditions had the highest unprocessed score at 26%, and correspondingly, the lowest benefit of 58.6 percentage points.

Planned comparisons consisting of uncorrected paired *t*-tests on rationalized arcsine units (RAUs; Studebaker, 1985) were performed to examine algorithm benefit for the HI listeners in each condition. Scores for unprocessed versus the corresponding processed condition were significantly different for each of the four combinations of TIR and T_{60} [$t(9) \geq 11.7$, $p \leq 0.001$]. These significant results all survive Bonferroni correction for multiple comparisons.

2. NH listeners

As Fig. 5 shows, algorithm benefit was also observed for every NH listener in every condition (albeit very slight in one case). As expected, these listeners produced considerably higher unprocessed scores. Accordingly, algorithm benefit was smaller than for the HI listeners, especially at the more favorable TIR (lower panel) where unprocessed scores were highest. Across all conditions, benefit was 30 percentage points or greater in 70% of cases and 40 percentage points or greater in 45% of cases. When only the less favorable TIR of –5 dB was considered, these proportions rose: 30 percentage points or greater in 100% of cases and 40 percentage points or greater in 75% of cases.

As observed for the HI listeners, greater algorithm benefit was correlated with lower unprocessed scores ($|r| = 0.85$, $p < 0.0001$, across all conditions). At the less favorable TIR of –5 dB, the group-mean benefits from algorithm processing (Fig. 6) were 47.4 and 45.2 percentage points for

https://doi.org/10.1121/10.0001441

the two reverberation conditions (T_{60} = 0.9 and 0.6 s, respectively). At the more favorable TIR of 0 dB, group-mean unprocessed scores were above 50% in both T_{60} conditions and benefits were 32.9 and 21.2 percentage points. Also as found for the HI listeners, NH algorithm benefit was significant in every condition tested, as indicated by planned comparison paired *t*-tests on RAUs [$t(9) \geq 6.3$, $p \leq 0.001$]. These significant results also survive Bonferroni correction.

Another question involves the role of the algorithm under conditions in which unprocessed intelligibility is already high and so assistance is not needed. Under these conditions, it is important to ensure that decrements in performance do not occur as a result of the extensive processing. Although the current conditions did not produce high baseline scores for HI listeners, several NH scores were above 70% to 80% at the more favorable TIR (25% of scores ≥70% and 10% ≥ 80%). In no case was intelligibility reduced by processing.

Final comparisons of interest involve the performance of HI listeners aided by the algorithm (processed scores) versus NH listeners without the algorithm (unprocessed scores), in conditions of identical background interference. This comparison represents the reception of speech by a typical HI listener having access to the current algorithm versus that of a young-NH conversation partner. These comparisons can be seen in the center panel of Fig. 6, where the TIR was 0 dB and T_{60} was 0.9 and 0.6. At the larger T_{60} value, HI processed intelligibility (68%) exceeded NH unprocessed intelligibility (51.4%), and did so significantly [$t(18) = 2.54$, $p = 0.02$]. This was also true at the smaller T_{60} value, where HI processed intelligibility (81.2%) exceeded NH unprocessed intelligibility (68.3%), significantly [$t(18) = 2.85$, $p = 0.01$]. These significant results also survive Bonferroni correction for the two tests.

B. Objective measures of intelligibility and sound quality

Objective scores are based on the measurement of acoustic signals themselves. They are often employed in the signal-processing literature and can provide a prediction of human performance (although the relationships between these values and actual human performance can be tenuous, especially for some metrics and when concurrent interferences including reverberation are employed; see, e.g., Zhao *et al.*, 2018). These scores are presented currently to facilitate replication and comparison with other methods. They were based on an analysis of all 160 test sentences processed for each condition.

The metrics employed currently are all widely used and included (1) extended short-time objective intelligibility (ESTOI; Jensen and Taal, 2016), (2) short-time objective intelligibility (STOI, Taal *et al.*, 2011), (3) perceptual evaluation of speech quality (PESQ; Rix *et al.*, 2001), and (4) source-to-distortion improvement (ΔSDR; Vincent *et al.*, 2006). ESTOI, like its predecessor STOI, is an intelligibility-prediction metric that reflects how accurately the amplitude envelope of the clean target speech is retained in the corrupted-then-processed speech. Because it is essentially a correlation between envelopes, the scale typically ranges from 0 to 1 (or from 0 to 100%). PESQ is a sound-quality prediction. It also reflects a comparison between clean and processed speech and has a scale ranging from –0.5 to 4.5. Finally, ΔSDR reflects the signal-to-noise ratio improvement in dB resulting from processing. For all four metrics, larger values are preferred, and the direct-sound (anechoic) target signal was used as the reference signal.

Table I displays these objective values in each condition. Substantial improvements were observed for all four metrics as a consequence of algorithm processing. ESTOI improved from an average of 25.2% for the unprocessed reverberant mixtures to 71.7% for the processed stimuli, corresponding to a 46.5 percentage-point improvement. STOI improved from an average of 54.3% to 87.2%, representing an increase of 32.9 percentage points, which is higher than the value obtained in Healy *et al.* (2019). Mean PESQ scores improved from an average of 1.45 to 2.63 for an increase of 1.18, and mean ΔSDR equaled 12.0 dB. These large improvements are in accord with the large human-intelligibility improvements observed in the main experiment.

IV. GENERAL DISCUSSION

The current study provides a demonstration of improved intelligibility for HI (and NH) listeners when deep learning based speaker separation is required to generalize to talkers not employed for training. In addition, the current

TABLE I. Average ESTOI, STOI, PESQ, and ΔSDR values at different room reverberation and TIR conditions for the target speaker in reverberant two-talker mixtures prior to and following processing by deep CASA.

	T_{60} (s)	0.6			0.9			
	TIR (dB)	–5	0	5	–5	0	5	Average
ESTOI (%)	Unprocessed	20.48	27.84	36.26	14.37	23.28	28.99	25.20
	Processed	66.55	74.79	81.95	58.17	70.94	77.75	71.69
STOI (%)	Unprocessed	48.46	56.32	63.71	44.43	53.76	59.29	54.32
	Processed	84.56	88.52	92.32	79.72	87.20	90.91	87.20
PESQ	Unprocessed	1.23	1.47	1.72	1.25	1.42	1.59	1.45
	Processed	2.45	2.69	2.97	2.25	2.60	2.83	2.63
ΔSDR (dB)	Proc-Unp	13.95	11.78	10.50	13.42	11.62	10.87	12.02

algorithm generalized to untrained utterances, TIRs, and reverberant RIRs. However, another generalization employed currently is perhaps underappreciated. It involves the use of different speech corpora for training versus test. This concept of "corpus independence" is not trivial and extends beyond the existence of different utterances. Whereas IEEE is a conventional speech-testing corpus containing phonetically-balanced sentences read in citation style, the WSJ0 corpus consists of read-aloud newspaper articles by a large number of talkers. They were also recorded in different environments using different apparatus. Accordingly, these corpora differ with regard to both linguistic aspects (e.g., speaking style, syntactic structure, semantic predictability, lexical content, phonetic balance, sentence length, etc.) and acoustic aspects (e.g., recording environment, microphone, etc.).

The current study also employed a range of HI-listener characteristics that was wider than employed by us previously. The upper age limit was extended to 88 years (HI9), and one listener had an asymmetric hearing loss and BiCROS (HI1). The benefit experienced by these two listeners was below the mean for all HI subjects, but it was still quite large, despite a floor effect that limited benefit for HI9[4] (58 and 52 percentage points, respectively, when averaged across all conditions). These results suggest that the current approach can be extended to a considerable variety of listeners and hearing losses, including asymmetric.

The current results can be compared directly to those of Healy *et al*. (2019), which was similar in design and procedures, and employed the same IEEE speech materials. Both the previous study and the current study employed a single interfering talker and concurrent reverberation and employed deep learning to isolate the anechoic target talker. But whereas the previous study was performed in talker-dependent fashion, the current model was talker and corpus independent. A different deep learning algorithm was also involved. Both studies employed a TIR = 0 dB and T_{60} = 0.6 s condition for HI listeners. The benefit observed in the previous study (for the direct-sound target) was 55.8 percentage points, whereas the benefit observed currently was 72.5 percentage points. Thus, the current algorithm produced better performance despite the addition of talker independence and cross-corpus generalization.

The primary algorithmic differences between the studies involved different neural networks, different amounts of training data, different input features, and different training targets. With regard to training target, the previous study employed the IRM, whereas the current study employed the cIRM. The superiority of the cIRM is sufficient to allow the deep CASA-estimated cIRM, which is an imperfect estimate of the ideal version, to outperform the actual IRM, which is an oracle mask calculated using knowledge of the unmixed target speech and interference (Liu and Wang, 2019).

Whereas the current study focused on generalization, real-time operation represents another important aspect of implementation for any algorithm. We take this opportunity to provide our perspectives. Perhaps the most fundamental characteristic of a real-time feasible network is causality—that it operates on only past and current time frames of the input signal, and not on future time frames (within the delay tolerance of the human auditory system, e.g., Stone and Moore, 1999, 2002). The other primary characteristic of a real-time feasible network involves its computational cost and the ease with which it can be implanted into portable hardware. Whereas the first characteristic is fundamental, the second characteristic is a direct function of hardware capabilities, which are improving constantly.

Researchers have demonstrated human intelligibility improvements using real-time feasible DNNs. The approach has generally involved (i) operation on non-future time frames and (ii) smaller networks to reduce the computational cost (e.g., Goehring *et al*., 2017; Monaghan *et al*., 2017; Bentsen *et al*., 2018; Bramsløw *et al*., 2018; Goehring *et al*., 2019; Keshavarzi *et al*., 2019). These works provide important proof of concept that small, causal networks can still improve human intelligibility. However, the use of less speech information and smaller networks typically leads to less benefit.

Our approach has been to target maximum benefit. Accordingly, we have not yet addressed real-time operation, and we have been relatively unconcerned with network size. The current network is an example of this approach—it is not causal because its input involves utterance-level features, and little effort has been directed toward minimizing complexity. The rationale is that the performance of large non-causal networks provides an upper bound for benefit and a valuable reference. This way, any performance decrement associated with each step taken toward real-world implementation can be known. Any modification that results in a substantial performance decrement from the reference then has a known cost, and so it can be assessed accordingly and alternatives sought.

We also argue that simply reducing the size of a traditional neural network may not be the optimal approach to real-time operation. Instead, advances in network architecture and insight from the growing field of model compression have the potential to provide new opportunities. The rationale is that these new architectures can be more efficient and capable of performance similar to larger and less efficient traditional networks. The key is to view large performance sacrifices as a nonnecessity.

A closely related point involves the hardware implementation of a neural network into hearing devices. We've advocated previously for implementation on an external smartphone-type device with wireless bidirectional communication to a small earpiece (Healy *et al*., 2017; Wang, 2017). In addition to offering considerable processing power, these devices mitigate the battery limitations of earworn devices. Noteworthy is that smartphone manufacturers have already implemented hardware dedicated to supporting machine learning and neural networks. A challenge moving forward will be to limit the latencies associated with peripheral aspects of processing such as wireless communication, which have the potential to add to algorithm latencies.

https://doi.org/10.1121/10.0001441

Finally, we note that each improvement in hardware capability relaxes the constraints on computational cost.

V. CONCLUSIONS

The current study provided a demonstration that deep learning can produce large intelligibility gains for listeners having typical sensorineural hearing loss (and for NH listeners) in a talker-independent context. A complex interference consisting of a single competing talker and substantial amounts of room reverberation was employed, and cross-corpus generalization was required. The single-microphone deep CASA algorithm produced intelligibility benefits for all listeners in all conditions. Benefit averaged across all conditions and HI listeners was 65 percentage points. This benefit was largest for those who need it most—those with the lowest unprocessed scores, who tended to have the largest degree of hearing loss. It also allowed HI listeners having access to the algorithm to outperform young NH listeners (without the algorithm) in identical interference conditions. Finally, speech quality was not assessed in human listeners, but the PESQ measure suggests that considerable improvements in this aspect were also produced. Because the gender difference between talkers employed currently was introduced to assist the human listeners in their task and is not required for algorithm performance (Liu and Wang, 2019), the current model could theoretically separate any two voices, even in challenging reverberant conditions. The non-causal nature of the current model is in accord with our goal of establishing maximum performance benchmarks, against which models appropriate for real-time implementation can be compared.

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.). We gratefully acknowledge computing resources from the Ohio Supercomputer Center, algorithmic assistance from Yuzhou Liu, and manuscript preparation assistance from Victoria Sevich.

[1]Whereas the aspects of generalization described in Sec. I can be considered primary, other aspects of generalization by deep learning noise reduction or speaker separation algorithms have been demonstrated. Essentially all studies employ different speech utterances for training and testing. Improved intelligibility for HI (or CI) listeners has also been demonstrated in untrained signal-to-noise or TIRs (e.g., Chen *et al.*, 2016; Zhao *et al.*, 2018; Goehring *et al.*, 2019; Healy *et al.*, 2019), and untrained room-impulse-responses when reverberation is involved (e.g., Zhao *et al.*, 2018; Healy *et al.*, 2019).

[2]https://github.com/ehabets/RIR-Generator.

[3]In accordance with our prior work on this topic, all conditions were equated to the same presentation sound pressure level. The equating of level following processing results in increases in the level of target speech in processed conditions, relative to that level in the corresponding unprocessed conditions. However, this technique is preferred because it best mimics an implementation strategy in which gains (hearing aid amplification or CI current mappings) are applied following noise reduction, when noise reduction is turned on.

[4]This lower bound for algorithm benefit observed currently is likely influenced by a floor effect in which subject HI9 was unable to correctly report a single keyword in any of the four unprocessed conditions. His unprocessed scores would likely be lower were it not for the floor at 0% correct, which would result in larger benefit.

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), *Specification for Instruments to Measure Aural Acoustic Impedance and Admittance (Aural Acoustic Immittance)* (American National Standards Institute, New York).

ANSI (**2004**). S3.21 (R2009), *American National Standard Methods for Manual Pure-Tone Threshold Audiometry* (American National Standards Institute, New York).

ANSI (**2010a**). S3.6, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).

ANSI (**2010b**). S12.60 (R2015), *Acoustical Performance Criteria, Design Requirements, and Guidelines for Schools, Part 1: Permanent Schools* (American National Standards Institute, New York).

Bai, S., Kolter, J. Z., and Koltun, V. (**2018**). "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv:1803.01271.

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.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).

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., and Wang, D. L. (**2017**). "Long short-term memory for speaker generalization in supervised speech separation," J. Acoust. Soc. Am. **141**, 4705–4714.

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., Hodder, K. I., and Toh, C. Y. (**2003**). "Effects of reverberation on perceptual segregation of competing voices," J. Acoust. Soc. Am. **114**, 2871–2876.

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*, October 19–23, Hangzhou, China, 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.

Goehring, T., Bolner, F., Monaghan, J. J. M., van Dijk, B., Zarowski, A., and Bleeck, S. (**2017**). "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," Hear. Res. **344**, 183–194.

Goehring, T., Keshavarzi, M., Carlyon, R. P., and Moore, B. C. J. (**2019**). "Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants," J. Acoust. Soc. Am. **146**, 705–718.

Harford, E. (**1966**). "Bilateral CROS: Two-sided listening with one hearing aid," Arch. Otolaryngol. **84**, 426–432.

Healy, E. W., Delfarah, M., Johnson, E. M., and Wang, D. L. (**2019**). "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," J. Acoust. Soc. Am. **145**, 1378–1388.

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.
Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (**2016**). "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proceedings of ICASSP*, March 20–25, Shanghai, China, pp. 31–35.
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.
Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (**2017**). "Densely connected convolutional networks," in *Proceedings of CVPR*, July 21–26, Honolulu, HI, pp. 2261–2269.
IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Jensen, J., and Taal, C. H. (**2016**). "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE/ACM Trans. Audio Speech Lang. Proc. **24**, 2009–2022.
Keshavarzi, M., Goehring, T., Turner, R. E., and Moore, B. C. J. (**2019**). "Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction," J. Acoust. Soc. Am. **145**, 1493–1503.
Kingma, D. P., and Ba, J. (**2014**). "Adam: A method for stochastic optimization," arXiv:1412.6980.
Kolbaek, M., Yu, D., Tan, Z. H., and Jensen, J. (**2017**). "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," IEEE/ACM Trans. Audio Speech Lang. Proc. **25**, 1901–1913.
Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (**2016**). "Temporal convolutional networks: A unified approach to action segmentation," in *Proceedings of ECCV*, October 11–14, Amsterdam, the Netherlands, pp. 47–54.
Liu, Y., and Wang, D. L. (**2019**). "Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation," IEEE/ACM Trans. Audio Speech Lang. Proc. **27**, 2092–2102.
Monaghan, J. J. M., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C. M., and Bleeck, S. (**2017**). "Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners," J. Acoust. Soc. Am. **141**, 1985–1998.
Moore, B. C. J. (**2007**). *Cochlear Hearing Loss*, 2nd ed. (Wiley, Chichester, UK).
Paul, D. B., and Baker, J. M. (**1992**). "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the Workshop on Speech and Natural Language*, February 23–26, Harriman, New York, pp. 357–362.
Plomp, R. (**1976**). "Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of azimuth of a single competing sound source (speech or noise)," Acustica **34**, 200–211.
Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**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*, May 7–11, Salt Lake City, UT, pp. 749–752.
Ronneberger, O., Fischer, P., and Brox, T. (**2015**). "U-Net: Convolutional networks for biomedical image segmentation," arXiv:1505.04597.
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.
Stone, M. A., and Moore, B. C. J. (**2002**). "Tolerable hearing aid delays. II. Estimation of limits imposed during speech production," Ear Hear. **23**, 325–338.
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.
Tye-Murray, N., Spehar, B., Sommers, M., and Barcroft, J. (**2016**). "Auditory training with frequent communication partners," J. Speech Lang. Hear. Res. **59**, 871–875.
Vincent, E., Gribonval, R., and Févotte, C. (**2006**). "Performance measurement in blind audio source separation," IEEE Trans. Audio Speech Lang. Process. **14**, 1462–1469.
Wang, D. L. (**2005**). "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Norwell, MA), pp. 181–197.
Wang, D. L. (**2017**). "Deep learning reinvents the hearing aid," in *IEEE Spectrum* (IEEE, New York), pp. 32–37.
Wang, D. L., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms and Applications* (Wiley-IEEE, Hoboken, NJ).
Williamson, D. S., Wang, Y., and Wang, D. L. (**2016**). "Complex ratio masking for monaural speech separation," IEEE/ACM Trans. Audio Speech Lang. Process. **24**, 483–492.
Yu, D., Kolbaek, M., Tan, Z. H., and Jensen, J. (**2017**). "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proceedings of ICASSP*, March 5–9, New Orleans, LA, pp. 241–245.
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.