JASA ARTICLE

A causal and talker-independent speaker separation/dereverberation deep learning algorithm: Cost associated with conversion to real-time capable operation

Eric W. Healy,[1,a)] Hassan Taherian,[2] Eric M. Johnson,[1,b)] and DeLiang Wang[2,c)]

[1]*Department of Speech and Hearing Science, The Ohio State University, Columbus, Ohio 43210, USA*

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

ABSTRACT:

The fundamental requirement for real-time operation of a speech-processing algorithm is causality—that it operate without utilizing future time frames. In the present study, the performance of a fully causal deep computational auditory scene analysis algorithm was assessed. Target sentences were isolated from complex interference consisting of an interfering talker and concurrent room reverberation. The talker- and corpus/channel-independent model used Dense-UNet and temporal convolutional networks and estimated both magnitude and phase of the target speech. It was found that mean algorithm benefit was significant in every condition. Mean benefit for hearing-impaired (HI) listeners across all conditions was 46.4 percentage points. The cost of converting the algorithm to causal processing was also assessed by comparing to a prior non-causal version. Intelligibility decrements for HI and normal-hearing listeners from non-causal to causal processing were present in most but not all conditions, and these decrements were statistically significant in half of the conditions tested—those representing the greater levels of complex interference. Although a cost associated with causal processing was present in most conditions, it may be considered modest relative to the overall level of benefit. © *2021 Acoustical Society of America.*

https://doi.org/10.1121/10.0007134

(Received 17 May 2021; revised 19 October 2021; accepted 22 October 2021; published online 23 November 2021)

[Editor: James F. Lynch] Pages: 3976–3986

I. INTRODUCTION

Real-time operation represents a critical requirement for deep learning based solutions to improve speech intelligibility in hearing technology. The critical requirement for real-time operation is causality—that an algorithm operate without using future-frame information, which introduces processing delays. The other requirement for real-time operation involves computational complexity and the burden that a neural network places on hardware. But unlike causality, this aspect is not fundamental. It is instead directly related to the ever-advancing computational power of the hardware on which it operates.

It is important to understand (i) if real-time capable deep learning can improve intelligibility and (ii) what is the performance cost associated with real-time capability—by how much does benefit decline when an algorithm is made causal. Evidence exists to indicate that causal deep learning can indeed improve intelligibility for hearing-impaired (HI) or cochlear implant listeners in the presence of background noise or other interference (Goehring *et al.*, 2017; Monaghan *et al.*, 2017; Bramsløw *et al.*, 2018; Goehring *et al.*, 2019; Keshavarzi *et al.*, 2019; Healy *et al.*, 2021).

What is more poorly understood is the cost, particularly in terms of human intelligibility and especially in HI individuals, associated with transitioning an algorithm from non-causal to causal operation. This is because studies using human performance to examine causal operation have often also used smaller networks (fewer layers in the deep neural network and fewer units in each layer) to produce a network that is overall more easily implementable. Whereas this is an admirable goal and increases the challenge substantially, removing future time-frame information and reducing network size can both serve to reduce algorithm performance, and so this combined approach confounds the effects of the two manipulations.

Also poorly understood is the extent to which real-time capable deep learning can improve intelligibility in the presence of complex interference characteristic of real-world listening. One such example of complex interference involves competing speech and concurrent room reverberation. These two interferences corrupt the signal of interest in very different ways, but they often occur concurrently in real-world environments. And it is well known that these concurrent interferences can disrupt speech intelligibility substantially in normal-hearing (NH) listeners and especially in HI listeners (Plomp, 1976; Culling *et al.*, 2003; Moore, 2007; Healy *et al.*, 2019; Healy *et al.*, 2020). The current study aims to increase our understanding of these issues.

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

[b)] Also at: Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210, USA.

[c)] Also at: Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, OH 43210, USA. ORCID: 0000-0001-8195-6319.

https://doi.org/10.1121/10.0007134

In the current study, a deep computational auditory scene analysis (deep CASA) algorithm (Liu and Wang, 2019, 2020) was used to increase intelligibility for HI and NH listeners by isolating a target talker from a competing talker and removing room reverberation. Importantly, the network was fully causal. But otherwise, the network, stimuli, and experimental procedures were highly similar to a previous work involving a non-causal version of deep CASA (Healy *et al.*, 2020). This allows the performance cost associated with causality to be directly established. The algorithm employed Dense-UNet (Liu and Wang, 2019) and temporal convolutional network (TCN) (Lea *et al.*, 2016) architectures. It was talker independent (trained and tested on different talkers) as well as corpus/recording-channel independent. This latter aspect reflects the fact the algorithm was trained and tested on speech from different speech corpora, recorded using different apparatus in different environments (see Healy *et al.*, 2020; Pandey and Wang, 2020). Finally, the algorithm operated in the complex domain (estimated both the real and imaginary parts) allowing both the magnitude and phase of the signal of interest to be estimated (Williamson *et al.*, 2016). More comprehensive discussions of the overall problem and the current solution may be found in Healy *et al.* (2020), and additional technical details on causal deep CASA may be found in Liu and Wang (2020).

II. METHOD

A. Subjects

Both HI and NH listeners participated. None had any prior exposure to the target or interferer sentences employed. Ten HI listeners were recruited from The Ohio State University Speech-Language-Hearing Clinic and surrounding community to represent a range of typical bilateral hearing aid users with sensorineural hearing loss. Hearing losses ranged from mild to profound and were moderate on average. Configurations ranged from flat to sloping. Pure-tone average audiometric thresholds (PTAs) based on 500, 1000, and 2000 Hz and averaged across ears ranged from 33 to 75 dB HL, with a mean of 53. Their ages ranged from 21 to 79 years (mean = 62 years), and six were male; four were female. The HI listeners each received a monetary incentive for participating. Figure 1 displays pure-tone audiograms (ANSI, 2004, 2010a) performed on day of test. These listeners were numbered in order of increasing PTA.

Ten NH listeners also participated. They were recruited from undergraduate courses in The Ohio State University Department of Speech and Hearing Science, and each received course credit for participating. They were native speakers of American English, with ages ranging from 18 to 27 years (mean = 21). Two were male, and eight were female. All produced pure-tone audiometric thresholds of 20 dB HL or lower at octave frequencies from 250 to 8000 Hz on the day of the test (ANSI, 2004, 2010a).

B. Stimuli

To facilitate direct comparison, the stimuli used for algorithm training and testing were identical to those employed for the non-causal version of deep CASA by Healy *et al.* (2020). The training dataset was based on the Wall Street Journal Continuous Speech Recognition Corpus (WSJ0) (Paul and Baker, 1992). Two-talker mixtures were generated by selecting sentences from various pairs of talkers in the si_tr_s folder of the WSJ0 corpus. Either talker could be either male or female. For sentence pairs of unequal duration, the longer was trimmed to match the duration of the shorter. Prior to mixing, the signals were equalized to the same root mean square (RMS) level. The sampling rate for all signals was 16 kHz.

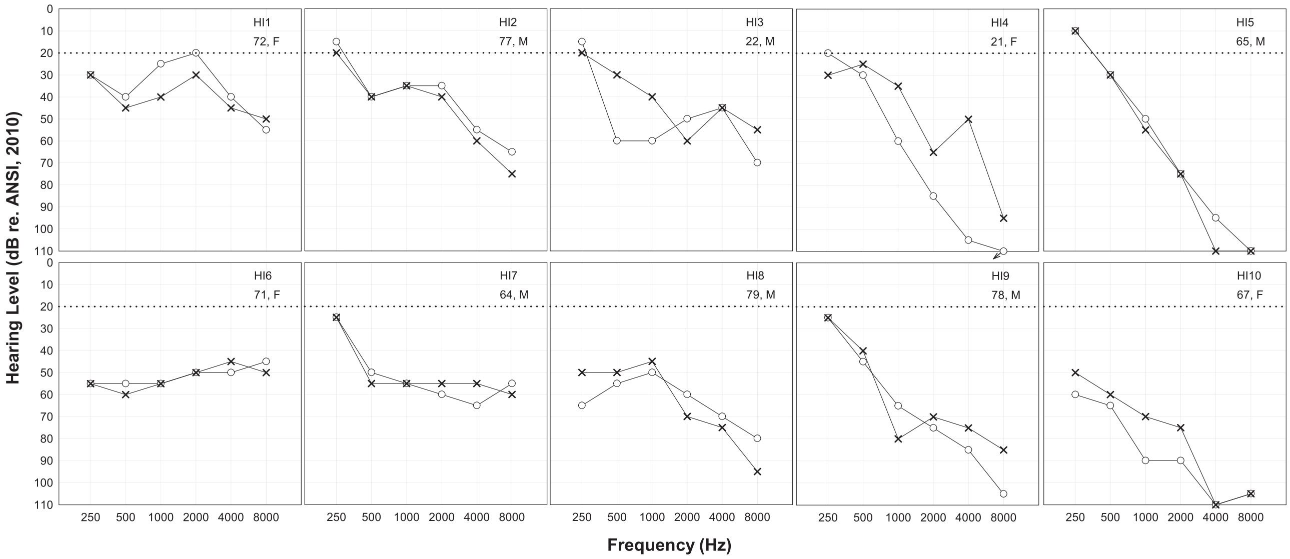

FIG. 1. Audiometric thresholds for the HI listeners. Pure-tone air-conduction thresholds for right ears are represented by circles and those for left ears are represented by X's. The arrow indicates threshold exceeding audiometer limits. The 20 dB HL limit for NH is represented by a horizontal dotted line. Listeners are numbered in order of increasing degree of hearing loss. Also provided are identifying numbers, ages in years, and sexes.

To generate reverberant utterances, the signals were convolved with room impulse responses (RIRs). The RIRs were generated based on the image method (Allen and Berkley, 1979; Habets, 2020) in a simulated room having dimensions of $6 \times 7 \times 3$ m. The microphone was placed at the position of (3, 4, 1.5) m, and the reverberation time (T_{60}) was selected randomly in the range of 0.3 to 1.0 s. Talker locations were selected randomly from 36 evenly distributed angles around the microphone, with the target talker 1 m from the microphone and the interfering talker 2 m away. The elevation of talkers matched that of the microphone. The target-to-interferer (TIR) ratio was set to 0 dB for the training data. In total, 200 500 mixtures were created for training, with 500 reserved for cross-validation.

The test stimuli were drawn from the Institute of Electrical and Electronics Engineers (IEEE) Revised List of Phonetically Balanced Sentences (Harvard Sentences) (IEEE, 1969). The exact unprocessed reverberant mixtures were identical to those employed by Healy *et al.* (2020). Test stimuli consisted of 160 target sentences mixed with 160 interfering sentences, plus reverberation. Sentences were formed into pairs based on similarity in duration. The average duration difference between members of a pair was 5 ms, and the difference did not exceed 10 ms for any pair. The target sentences were distinct from the interfering sentences, and both were distinct from the training sentences. All target sentences were spoken by the same male talker (average fundamental frequency = 132 Hz, standard deviation = 41 Hz), and all interfering sentences were spoken by the same female talker (average fundamental frequency = 209 Hz, standard deviation = 42 Hz). Neither of these talkers is present in the WSJ0 corpus recordings. The TIRs employed for HI listeners were 0 and 5 dB, and those for the NH listeners were –5 and 0 dB. These TIRs matched those of Healy *et al.* (2020) and were selected to produce a variety of unprocessed intelligibility scores reasonably free of floor and ceiling effects.

The same reverberation procedures were applied as for the training set, except that the talker positions were shifted by 5° to ensure that the test RIRs were different from those used for training. The T_{60} values for the test sentences were 0.3 and 0.6 s, which are representative of realistic 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 0.3 s falls within that limit. In total, 160 reverberant two-talker mixtures were generated for testing in each of the six test conditions (3 TIRs and 2 T_{60} values). Additional stimuli detail can be found in Healy *et al.* (2020).

It is potentially noteworthy that different sexes were used for the concurrent talkers to reduce confusion for the human listeners with regard to which talker was the target. Algorithm performance has been shown to be similar when the two concurrent talkers are the same versus different sexes (Liu and Wang, 2019).

C. Algorithm description

Given a two-talker mixture signal recorded in reverberant conditions, the talker-independent speaker separation problem in the short-time Fourier transform (STFT) domain is formulated as

$$Y(t,f) = H_1(t,f)S_1(t,f) + H_2(t,f)S_2(t,f), \tag{1}$$

where $Y(t,f)$, $S_1(t,f)$, and $S_2(t,f)$ denote the complex STFT values of the mixture signal, direct-sound target and interferer signals at time frame t and frequency f, respectively. $H_i(t,f)$, $i = 1, 2$, represents the RIR in the STFT domain corresponding to talker i. The goal is to estimate $S_1(t,f)$, designated as the target talker, based on the reverberant mixture signal $Y(t,f)$.

To address this problem, deep CASA was employed (Liu and Wang, 2019), which achieves high performance in both anechoic and reverberant conditions (Healy *et al.*, 2020). Deep CASA breaks down the speaker separation task into two stages: simultaneous and sequential grouping, motivated by auditory scene analysis principles (Bregman, 1990; Wang and Brown, 2006). In the simultaneous grouping stage, separation and dereverberation of the source signals are performed at the frame level. In the sequential grouping stage, the separated frames are organized to one of two talker streams.

A major limitation of deep CASA for real-world deployment is that prior implementations were not causal: the algorithm used future information as long as 9 s. Recently, Liu and Wang (2020) proposed a causal version of deep CASA to address this limitation in anechoic environments. Several aspects of simultaneous and sequential grouping were modified so that no future information was used throughout the algorithm. These modifications reduced the operational latency to one frame of STFT and enabled real-time processing.

In the current study, causal deep CASA was extended to perform speaker separation in reverberant conditions. In what follows, the two stages of deep CASA are described, as are modifications needed for causal processing. We note that deep CASA is a dedicated speaker separation algorithm. The network also performed de-reverberation, but noise removal was not addressed.

1. Simultaneous grouping

This stage uses a deep neural network (DNN) to estimate two complex ratio masks $cRM_i(t,f)$, $i = 1, 2$, based on the real and imaginary parts of $Y(t,f)$. The masks are multiplied with the reverberant mixture to generate the reconstructed sources in the complex domain (Williamson *et al.*, 2016),

$$\hat{S}_{u_i}(t,f) = cRM_i(t,f) \otimes Y(t,f), \tag{2}$$

where $\hat{S}_{u_i}$ denotes the unorganized frames that are separated and dereverberated. Symbol $\otimes$ denotes point-wise complex multiplication. The DNN is trained with a frame-level permutation invariant training (tPIT) (Yu *et al.*, 2017) criterion, which chooses optimal output-talker assignment based on the pairing that minimizes the l_1 norm over all possible talker permutations. After organizing the frames using tPIT, the optimally organized talker frames $\hat{S}_{o_i}(t,f)$ are converted to a time-domain signal $\hat{s}_{o_i}(t)$ via inverse STFT. Finally, a signal-to-noise ratio (SNR) loss J^{SNR} is used to optimize the network,

$$J^{SNR} = -10 \sum_{i=1,2} \log \frac{\sum_t s_i(t)^2}{\sum_t \left[s_i(t) - \hat{s}_{o_i}(t) \right]^2}. \quad (3)$$

The simultaneous grouping DNN is based on a Dense-UNet architecture. It includes downsampling layers and upsampling layers interleaved with dense convolutional blocks. The dilation factor in the dense convolutional blocks was increased from 1 to 8 to account for the reverberation effect.

Three changes were made to make Dense-UNet causal. First, downsampling and upsampling layers were applied only in the frequency dimension and not across time frames so as to avoid using future frames. Second, the dense convolutional blocks were modified to include causal temporal convolution operations that only rely on past information. Third, the normalization method was changed from standard layer normalization to batch normalization so that recalculation of statistics is not needed during inference.

2. Sequential grouping

With unorganized frames from the previous stage, the sequential grouping stage performs a temporal organization by assigning each separated frame to one talker. Specifically, a DNN was used to estimate an embedding vector $\boldsymbol{V}(t) \in \mathbb{R}^d$ for each time frame. The training target was a two-dimensional indicator vector $\boldsymbol{A}(t)$ that represents the optimal assignment. The label was set to $\boldsymbol{A}(t) = [1, 0]$, if the order of $\hat{S}_{u_1}$ and $\hat{S}_{u_2}$ was matched correctly with the order of talkers; otherwise the label was $\boldsymbol{A} = [0, 1]$, to mean that the order should be switched. The DNN model was optimized over a sequence of T frames with a weighted objective function using $\boldsymbol{A}\ (T \times 2)$ and $\boldsymbol{V}(T \times d)$ matrices,

$$J^{DC} = ||\boldsymbol{W}(\boldsymbol{V}\boldsymbol{V}^T - \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{W}||_F^2, \quad (4)$$

where $\boldsymbol{W}$ denotes $T \times T$ diagonal weight matrix, $||.||_F$ represents the Frobenius norm, and DC stands for deep clustering (Hershey *et al.*, 2016). The diagonal entries of $\boldsymbol{W}$ correspond to a frame-level weight vector $w(t) = |LD(t)|/\sum_t |LD(t)|$, where $LD(t)$ represents the simultaneous grouping tPIT loss difference (LD) between the two possible talker assignments. The idea of using the weight matrix $\boldsymbol{W}$ was to emphasize those frames where the two outputs are substantially different so that the wrong order of outputs increases the tPIT loss.

Instead of using noncausal k-means clustering during inference, the estimated embedding vectors $\boldsymbol{V}(t)$ were assigned into two groups using a causal clustering algorithm. Two first-in-first-out (FIFO) queues were initialized, and the embedding vector of the first frame was pushed to the first queue. Starting from the second frame, the similarity of the embedding vectors between the current frame and the previous frame was computed. If the similarity was higher than a predefined threshold, then the embedding vector was pushed to the first queue. This process continued until one frame did not meet the threshold, and it was then assigned to the second queue. Once the second queue loaded the first embedding vector, the mean values of the two queues were tracked. Next, each embedding vector with significant energy was assigned to the queue whose mean value was closer to the embedding vector. After each assignment, the mean values were updated based on the 20 most recent items in the queues; that is, the length of the two FIFO queues was 20. At the end, the queues were used for organizing the features $\hat{S}_{u_2}(c,f)$ and $\hat{S}_{u_2}(c,f)$.

A TCN was used for sequential grouping. The real, imaginary, and magnitude STFT of the reverberant mixture, as well as the outputs of the simultaneous grouping stage, were stacked to form the input to the TCN, which consisted of eight dilated convolutional blocks each comprising three convolutional layers. Layer normalization and convolution operations were modified to their causal version (Liu and Wang, 2020). The simultaneous grouping and sequential grouping modules were trained in turn separately with the Adam optimizer (Kingma and Ba, 2014).

Although computational complexity and the burden that a neural network places on hardware is not a fundamental constraint, it nevertheless represents an important consideration. Accordingly, the computational complexity of the current causal deep CASA model was calculated in terms of floating-point operations (FLOPs), a common metric for evaluating DNN model complexity. The current model has 12.8 M parameters and requires 147.54 G FLOPs to process a 1-s input signal, using the current 32-ms frames with 8-ms shift. Causal deep CASA has also been assessed in terms of the real time factor (RTF) on a single NVIDIA V100 GPU (Liu and Wang, 2020). RTF is defined as the ratio of processing time to input signal duration, and so values up to 1.0 represent real-time capable operation. The RTF of causal deep CASA was 0.011.

Figure 2 displays spectrogram images of various signals. Panel (a) displays the two-talker reverberant mixture, mixed at a TIR of –5 dB in a room with $T_{60} = 0.6$ s. This represents the input to the algorithm as well as the unprocessed signal used for human-subjects testing. Panels (b) and (c) display the individual clean anechoic utterances, panels (d) and (e) display these utterances extracted from the reverberant two-talker mixture using the non-causal version of deep CASA employed by Healy *et al.* (2020), and panels (f) and (g) display these utterances extracted from the reverberant two-talker mixture using the current fully causal version of deep CASA.

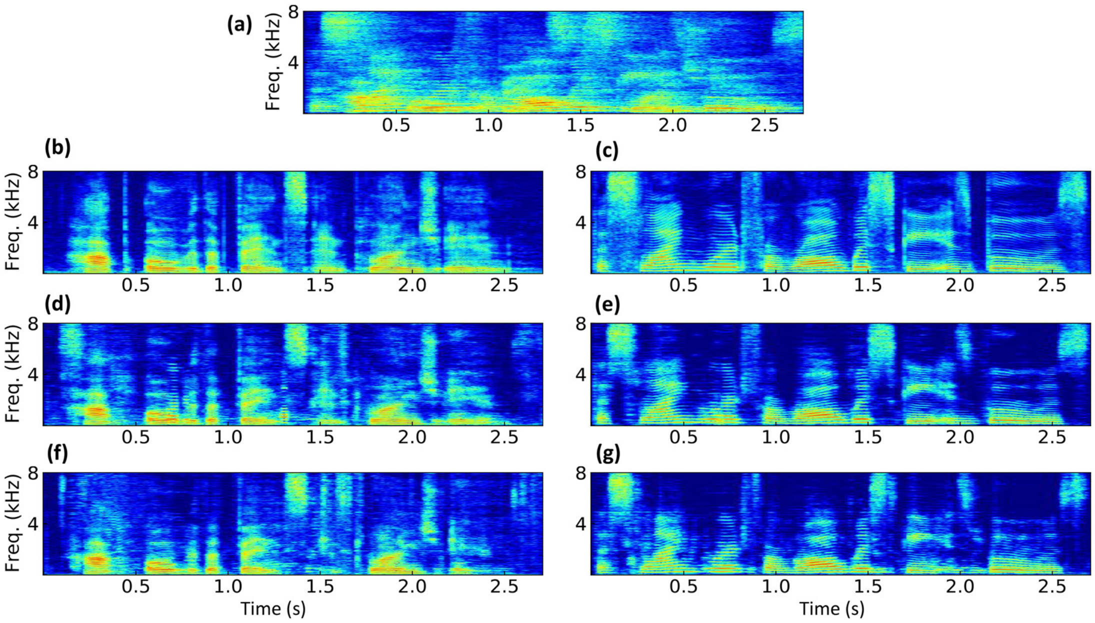

FIG. 2. (Color online) Spectrogram images representing the separation of a target sentence from a reverberant mixture of two talkers (TIR = –5 dB, $T_{60} = 0.6$ s) using deep CASA. Panel (a) reverberant two-talker mixture, (b) clean anechoic target utterance, "Hop over the fence and plunge in," (c) clean anechoic interfering utterance, "The slang word for raw whisky is booze," (d) non-causal deep CASA output 1 (target), (e) non-causal deep CASA output 2 (interferer), (f) causal deep CASA output 1 (target), and (g) causal deep CASA output 2 (interferer).

D. Procedure

The procedure was largely identical to that employed by Healy *et al.* (2020) involving the non-causal version of deep CASA, in order to facilitate direct comparison. Each subject heard a total of eight conditions (2 TIRs × 2 T_{60}s × 2 unprocessed/processed), with 20 target sentences per condition. Unprocessed conditions refer to the concurrent reverberant sentence mixtures and processed conditions refer to these mixtures following processing by the fully causal deep CASA speech separation/dereverberation algorithm. The TIR-T_{60} conditions were blocked and presented in random order for each listener. The comparison of greatest interest is that between the unprocessed and processed condition at each TIR-T_{60}, and so these conditions were presented in juxtaposed and random order for each listener. The use of a single fixed presentation order for the sentence materials allowed a random correspondence between sentence pairs and conditions. No sentence was presented more than once for any listener.

The stimuli were played back from a Windows PC using an RME Fireface UCX digital-to-analog converter (Haimhausen, Germany), through a Mackie 1202-VLZ mixer (Woodinville, WA), and presented diotically using Sennheiser HD 280 Pro headphones (Wedemark, Germany). The overall RMS level of each stimulus was set to 65 dBA in each ear using a sound-level meter and flat-plate coupler (Larson Davis models 824 and AEC 101, Depew, NY). For the HI listeners, additional frequency-specific gains were applied to compensate for the hearing loss of each individual listener using the NAL-RP hearing-aid fitting formula (Byrne *et al.*, 1990). These gains were implemented using a RANE DEQ 60 L digital equalizer (Mukilteo, WA), as described in Healy *et al.* (2015). Accordingly, these listeners were tested with hearing aids removed. The final presentation level for the HI listeners ranged from 79.1 to 93.4 dBA (mean = 87.2 dBA).

Twenty-five practice stimuli were presented prior to formal testing, consisting of five stimuli in each condition: (1) clean sentences with no interference; (2) processed sentences at the higher of the two TIRs for each listener group and a T_{60} value of 0.3 s; (3) processed sentences at the lower TIR for each listener group and a T_{60} value of 0.6 s; (4) unprocessed mixtures at the higher TIR and a T_{60} of 0.3 s; and (5) unprocessed mixtures at the lower TIR and a T_{60} of 0.6 s. During this familiarization, the HI listeners were asked about the loudness of the signals, and no listener reported the level to be uncomfortable.

Listeners then heard the eight blocks of conditions while seated in a double-walled sound booth. They were instructed to attend to the male voice, to repeat back each sentence as best they could, and to guess if unsure of the content of the sentence. The listeners were blind to the condition under test, but the experimenter was not. The experimenter controlled the presentation of each stimulus and scored keywords correctly reported. The 20 target sentences presented in each condition each contained five keywords, for a total of 100 keywords in each condition. Sentence recognition was expressed as the percentage of these keywords correctly reported.

III. RESULTS AND DISCUSSION

A. Human performance

1. HI listeners

Sentence recognition for individual HI listeners is shown in Fig. 3. These listeners are numbered and plotted in

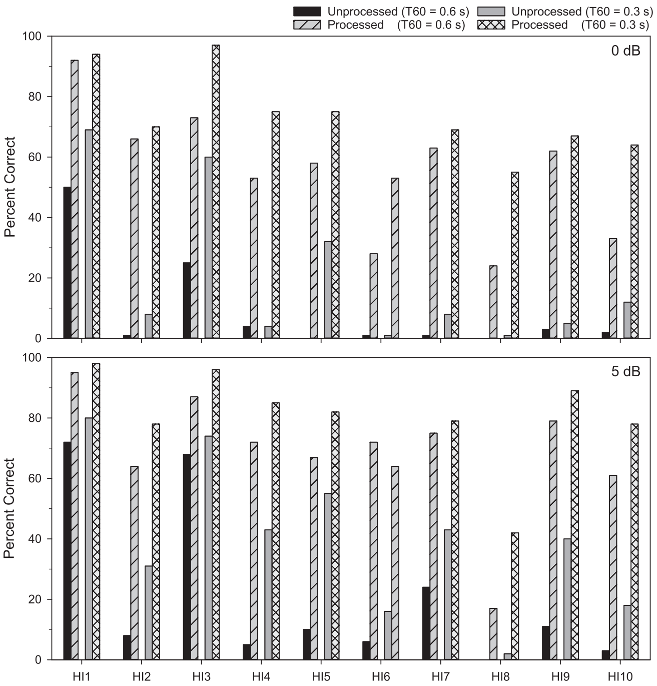

FIG. 3. Sentence-intelligibility scores for individual HI listeners. The solid black and solid gray columns represent scores for unprocessed, reverberant concurrent sentences. The hatched columns represent scores following algorithm processing to isolate the target talker and remove room reverberation. Fully causal algorithm benefit is then represented by the difference between a solid column and the immediately adjacent hatched column. The target-to-interferer ratios of 0 and 5 dB are displayed in separate panels, and the T_{60} times of 0.6 and 0.3 s are represented in the legend. Listeners are numbered in order of increasing degree of hearing loss, as in Fig. 1.

order of increasing PTA, as in Fig. 1. Sentence recognition scores at different TIRs are plotted in separate panels. Within each panel, unprocessed and processed scores at each T_{60} are shown as solid and hatched columns, respectively. For any given listener, algorithm benefit is then represented by the difference between a solid column and the hatched column directly to its right. The missing columns for HI5 and HI8 reflect an inability to correctly report any keywords in those conditions.

Algorithm processing enhanced speech recognition for every HI listener in every condition. Considering the 40 unprocessed-processed pairs across HI listeners and conditions, algorithm benefit ranged from 17 to 71 percentage points. It equaled or exceeded 40 percentage points in 70% of cases, 50 percentage points in 48% of cases, and 60 percentage points in 25% of cases. The grand-mean algorithm benefit for HI listeners across all conditions was 46.4 percentage points.

There appears to be some tendency for overall scores to decrease from left to right in each panel of Fig. 3, suggesting that scores overall could be associated with listener PTA. However, Spearman rank-order correlations failed to reveal any significant relationships between PTA and unprocessed scores averaged across the four TIR-T_{60} combinations $[|r_s(8)| = 0.48,\ p = 0.15]$. Similarly, no significant rank-order correlations were found between PTA and mean processed scores $[|r_s(8)| = 0.40,\ p = 0.23]$ or between PTA and mean algorithm benefit $[|r_s(8)| = 0.32,\ p = 0.35]$.

The right half of Fig. 5 shows group-mean scores and standard errors of the mean (SEMs) for the HI listeners in each condition. As in Fig. 3, benefit in each condition is reflected by the difference between a solid column and the hatched column directly to its right. Group-mean unprocessed scores increased monotonically as TIR and T_{60} became more favorable (from left to right in the right half of Fig. 5). In the least favorable condition (TIR = 0 dB, $T_{60} = 0.6$ s), the lowest unprocessed score of 8.7% rose to 55.2%, for an algorithm benefit of 46.5 percentage points. The next condition, also in the center panel, was at the same TIR but with an improved T_{60} of 0.3 s. This mean unprocessed score of 20.0% rose to 71.9% after processing, producing a benefit of 51.9 percentage points. The right-most panel of Fig. 5 displays the more favorable TIR of 5 dB. At the T_{60} of 0.6 s, the score increased from 20.7% to 68.9% for a benefit of 48.2 percentage points. Finally, the most favorable TIR-T_{60} condition (5 dB-0.3 s) produced the highest unprocessed score for the HI group (40.2%), which rose to 79.1%, resulting in the smallest HI group-mean benefit of 38.9 percentage points.

Planned comparisons consisting of uncorrected two-tailed paired *t*-tests on rationalized arcsine units (RAUs) (Studebaker, 1985) were performed to examine algorithm benefit for HI listeners in each condition. Processed scores were found to be significantly higher than unprocessed scores for each of the four combinations of TIR and T_{60} [each $t(9) \geq 7.7$, $p \leq 0.0001$]. These results remain significant after Bonferroni correction.

2. NH listeners

Figure 4 displays sentence-recognition scores for the individual NH listeners in each condition. As the figure shows, the algorithm also provided benefit for each NH listener in each condition, with one exception where the unprocessed score equaled the processed score (NH7, TIR = 0 dB, $T_{60} = 0.6$ s). Predictably, the NH-listener unprocessed scores were higher than the corresponding HI-listener scores. These higher unprocessed scores resulted in less opportunity for benefit from algorithm processing. Accordingly, benefit tended to be smaller for these listeners. Across all NH listeners and conditions (40 cases), benefit was 20 percentage points or greater in 53% of cases and 30 percentage points or greater in 30% of cases. These proportions are somewhat higher when only the lower TIR of –5 dB is considered, where unprocessed scores were lower. Across all conditions, the grand-mean algorithm benefit for the NH listener group was 21.5 percentage points.

The left half of Fig. 5 shows group-mean scores and SEMs for the NH listeners in each condition. For these listeners, mean benefit was largely a function of mean unprocessed scores, where higher baseline scores limited benefit. Group-mean algorithm benefit was largest for the NH listeners in the condition with the greatest interference (TIR = –5 dB, $T_{60} = 0.6$ s), at 32.1 percentage

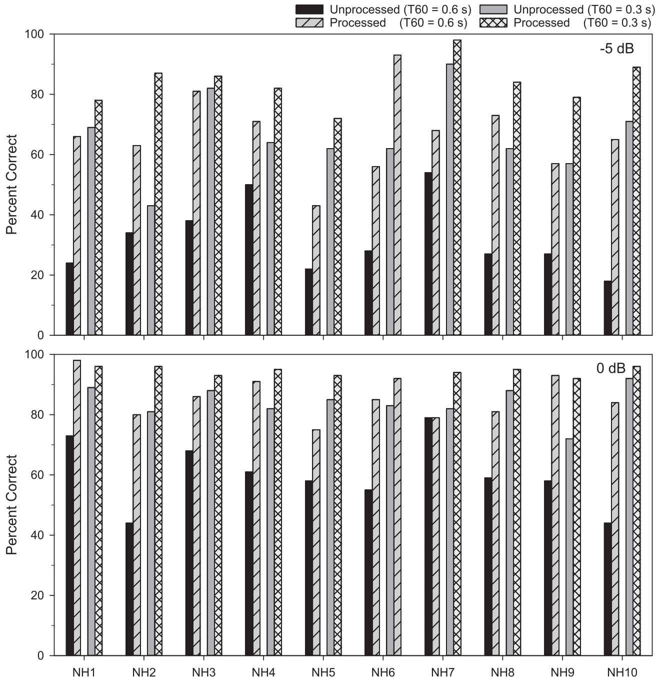

FIG. 4. As Fig. 3, but for the NH listeners. Note the different target-to-interferer ratios employed for these listeners.

points. The second highest group-mean benefit for the NH listeners (25.3 percentage points) was also obtained at the T_{60} of 0.6 s, but it came at the more favorable TIR of 0 dB. This was followed by a benefit of 18.6 percentage points at TIR = –5 dB, T_{60} = 0.3 s. The smallest benefit was obtained in the most favorable listening condition for the NH listeners (TIR = 0 dB, T_{60} = 0.3 s), where the mean unprocessed score of 84.2% correct rose 10.0 percentage points to 94.2% correct with the aid of algorithm processing.

As for the HI listeners, the algorithm provided significant benefit at all four TIR-T_{60} conditions in which the NH listeners were tested, as revealed by four planned two-tailed paired *t*-tests on RAUs [each $t(9) \geq 5.3$, $p < 0.001$]. These results also remain significant after Bonferroni correction.

Another question of interest involves whether HI listeners, when aided by the algorithm, can approximate the unaided speech-recognition performance of ideal young NH listeners in otherwise identical conditions. In other words, can this fully causal algorithm restore NH speech recognition abilities to HI listeners in complex interference? This question was assessed by comparing the unprocessed scores (solid columns) of the NH group to the corresponding processed scores (hatched columns) of the HI group at the common TIR of 0 dB, displayed in the center panel of Fig. 5. This was done separately for each T_{60} value. The mean NH unprocessed score was above the HI processed score at both T_{60} times. Values were within 5 percentage points at $T_{60} = 0.6$ s but differed by 12 percentage points at $T_{60} = 0.3$ s. Planned two-tailed Welch's independent samples *t*-tests were conducted on RAU-transformed scores. These *t*-tests revealed no significant differences between the unprocessed scores of the NH listeners and the processed scores of the HI listeners with either 0.6 s [$t(13.3) = 0.53$, $p = 0.61$] or 0.3 s [$t(11.8) = 2.1$, $p = 0.06$] of reverberation time. Therefore, the algorithm-aided speech recognition performance of the HI listeners appeared to approximate that of the young NH listeners in the less favorable interference and be below but not significantly in the more favorable interference.

3. Comparison to the non-causal algorithm

To examine the human-subjects performance costs associated with making the non-causal deep CASA algorithm fully causal, the current data were compared to those of Healy *et al.* (2020) in which a highly similar but non-causal algorithm was used to separate and dereverberate the same concurrent sentence pairs used currently. Figure 6 displays group-mean speech recognition scores and SEMs for the HI and NH listeners. The pairs of columns labeled "non-causal" are from Healy *et al.* (2020), and those labeled "causal" are replotted from Fig. 5. The TIR values displayed were common across the two studies, as was the T_{60} value

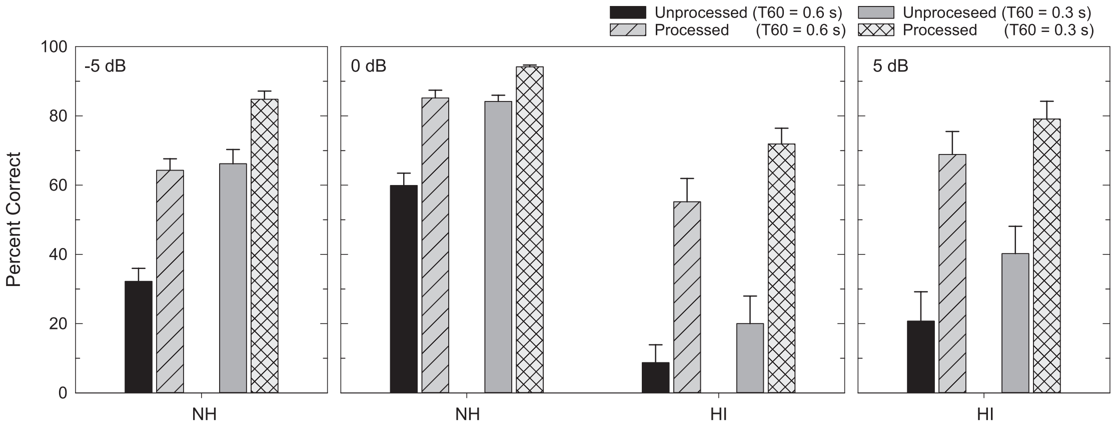

FIG. 5. Group-mean sentence-intelligibility scores (and standard errors) for each condition. As in Figs. 3 and 4, the different target-to-interferer ratios (–5, 0, and 5 dB) are displayed in separate panels, and the different reverberation T_{60} values are represented by different column pairs in each panel. Means for the HI and NH listeners are presented separately. Mean algorithm benefit resulting from the fully causal deep CASA algorithm is then represented as the difference between a solid column and the immediately adjacent hatched column.

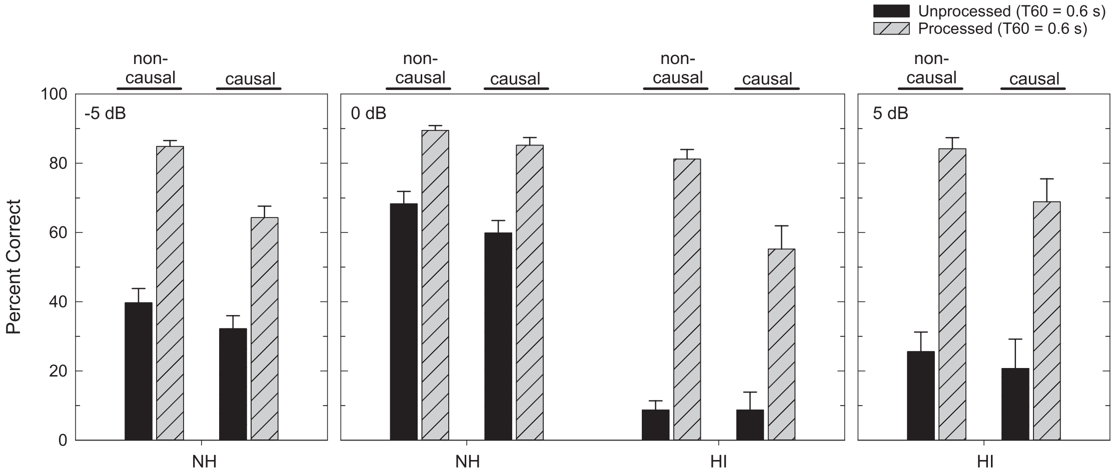

FIG. 6. Comparison between non-causal and causal processing: Plotted are group mean (and standard error) intelligibility resulting from the non-causal version of deep CASA (data from Healy *et al.*, 2020), along with values from Fig. 5 for the fully causal version of the algorithm, at the common conditions of $T_{60} = 0.6$ s. The speech stimuli and experimental procedures were identical across studies. The algorithm was also largely identical, with the exception of conversion from utterance-based processing to a fully causal time-frame-based model. Different groups of listeners were employed across the two studies, likely accounting for the difference in unprocessed scores. Again, benefit is represented by each difference between unprocessed (solid column) and processed (adjacent hatched column) scores.

of 0.6 s. As with the previous figures, benefit is reflected as the difference between each unprocessed (solid column) and corresponding processed score (hatched column). This is particularly important to note in Fig. 6, where baseline unprocessed scores were not identical across studies.

For NH listeners at –5 dB TIR (left panel), the group-mean non-causal algorithm benefit equaled 45.2 percentage points, which was reduced to 32.1 percentage points for the causal algorithm, amounting to a "causality cost" of 13.1 percentage points. At 0 dB TIR (center panel, left half), the causal algorithm produced more benefit than the non-causal algorithm for NH listeners (a negative cost), with benefit rising 4.2 percentage points from 21.2 to 25.3.

The greatest causality cost (26 percentage points) was observed for HI listeners at 0 dB TIR (center panel, right half), where the large non-causal benefit of 72.5 percentage points was reduced to 46.5 percentage points for causal algorithm benefit. The cost of causality for HI listeners was smaller at the TIR of 5 dB (right panel), where a benefit of 58.6 percentage points for the non-causal algorithm was reduced to 48.2 percentage points for the causal algorithm, resulting in a causality cost of 10.4 percentage points.

Planned comparisons consisting of four uncorrected two-tailed Welch's *t*-tests on RAUs were used to assess the causality cost for each listener group at each TIR. For NH listeners, causality significantly reduced algorithm benefit at –5 dB TIR [$t(17.9) = 3.07$, $p = 0.007$] but the difference (the negative cost) was not significant at 0 dB TIR [$t(18.0) = –0.51$, $p = 0.61$]. For the HI listeners, causality significantly reduced algorithm benefit at 0 dB TIR [$t(17.3) = 3.76$, $p = 0.0015$] but not at 5 dB TIR [$t(17.9) = 1.22$, $p = 0.24$]. The two significant results survive Bonferroni correction with a family size of four comparisons. Performance costs associated with making the algorithm causal were therefore greater and only significant at the less-favorable TIR tested for each listener type.

B. Objective measures

The current results were evaluated in terms of extended short-time objective intelligibility (ESTOI) (Jensen and Taal, 2016), perceptual evaluation of speech quality (PESQ) (Rix *et al.*, 2001), and signal-to-distortion ratio improvement (ΔSDR) (Vincent *et al.*, 2006). ESTOI (typical scale = 0 to 100%) is a predictor of speech intelligibility resulting from essentially a correlation between acoustic amplitude envelopes of clean target speech and that same speech following corruption and processing to remove corruption. PESQ (scale = –0.5 to 4.5) is an objective measure of speech sound quality and also represents the acoustic relationship between clean target speech and speech following corruption and processing. ΔSDR (dB) reflects the signal-to-noise ratio improvement following processing. The average scores in each TIR condition are displayed in Table I. Values are provided for the current fully causal version of deep CASA as well as for the non-causal version from Healy *et al.* (2020), in the common conditions of $T_{60} = 0.6$ s.

As expected, causal deep CASA performed more poorly than the non-causal version, due to the lack of future

TABLE I. Average ESTOI, PESQ, and ΔSDR at different target-to-interferer ratio conditions for the target speaker in reverberant two-talker mixtures with $T_{60} = 0.6$ s. The signals were processed by non-causal and fully causal deep CASA.

TIR (dB)		–5	0	5	Average
ESTOI (%)	Noisy	20.48	27.84	36.26	28.19
	Non-causal	66.55	74.79	81.95	74.43
	Causal	51.81	66.78	75.33	64.64
PESQ	Noisy	1.23	1.47	1.72	1.47
	Non-causal	2.45	2.69	2.97	2.70
	Causal	1.89	2.41	2.72	2.34
ΔSDR (dB)	Non-causal	13.95	11.78	10.50	12.07
	Causal	10.46	9.76	8.08	9.43

information. Compared to the anechoic condition (Liu and Wang, 2020), the degradation resulting from causal processing was larger in the reverberant environment. This can be attributed to the decline of dereverberation performance by removing the future context in causal processing (Zhao *et al.*, 2020). Nevertheless, substantial improvements were observed in all conditions after the reverberant mixtures were separated with fully causal deep CASA. On average, ESTOI, PESQ, and ΔSDR scores were improved by 36.45%, 0.87 and 9.43 dB, respectively.

IV. GENERAL DISCUSSION

The current results demonstrate that substantial increases in intelligibility can be obtained for both HI and NH listeners, in conditions of complex interference involving a competing talker and concurrent room reverberation, using a fully causal deep learning algorithm. The efficacy of the current algorithm, as well as that of the non-causal version, can be visualized in Fig. 2. The top panel (a) shows that the acoustic energy of the two-talker reverberant mixture is considerably spread over time and that acoustic landmarks characteristic of clear speech are considerably distorted, particularly when viewed in relation to the clean individual signals in panels (b) and (c). But the panels below (b) and (c) show that both algorithms output signals that appear highly similar in acoustic structure to the original clean speech. Also, apparent is the similarity in structure between the outputs of the non-causal and causal versions of the algorithm.

It is particularly important to observe that substantial benefit remains during fully causal processing in realistic amounts of room reverberation. The reverberant spread or duplication of acoustic energy occurs forward in time (the echo follows the event). Accordingly, access to this future time-frame information could be especially beneficial in reverberant environments as the network seeks to perform dereverberation. Effective operation without this information is notable.

The current results also allow the cost associated with conversion to causal processing to be known. It is first important to note that the causal costs established currently apply to the current deep CASA algorithm, and different implementations might display different costs. Figure 6 displays the intelligibility comparison across non-causal and fully causal deep CASA, in otherwise essentially identical conditions. It is clear that causal processing does not come without cost in most conditions. The decrement associated with causal processing was found to be significant in half of the conditions examined. But it is also clear that substantial benefit remains even in the case of fully causal processing. The HI listeners received 46 percentage points of benefit on average across the current fully causal conditions. And the NH listeners received 22 percentage points of benefit on average across the fully causal conditions, with their benefit likely reduced due to high baseline scores. Thus, it can be concluded that the cost associated with the current fully causal processing is significant in some conditions, but modest in magnitude relative to the overall level of benefit.

The current network was talker independent and required to generalize across different TIRs and RIRs. It was also corpus independent and recording-channel independent, which are highly related and can challenge generalization, particularly at lower SNRs/TIRs. The different speech corpora used for algorithm training versus testing possess linguistic and speech-production differences that can potentially challenge generalization. The different recording equipment and recording environments used to create stimuli for training versus test impart acoustic differences that can also challenge generalization. Specifically, the frequency response of the microphone and the background noise and reverberation environment of the recording session impart a constant acoustic influence on the speech signal that a deep learning algorithm could be sensitive to. Pandey and Wang (2020) examined this issue in speech enhancement and found that differences in recording-channel characteristics dominate the cross-corpus/channel generalization challenge.

The current network also used complex representations, which allowed both the magnitude and phase of the signal of interest to be estimated. This technique was introduced by Williamson *et al.* (2016) through the concept of the complex ratio mask. It represents an advancement over the more traditional approach (ratio mask) in which only the magnitude of the signal of interest is estimated by the network, then combined with the phase corresponding to the unprocessed speech-plus-interference ("noisy phase") to construct the estimated signal of interest. Initial investigations of complex-domain phase estimation focused on speech quality, which is known to be impacted by phase and is in fact substantially improved through the addition of estimated "clean" phase. But the past several years have brought an increased understanding that estimation of both magnitude and phase can serve to improve speech separation in a far broader sense [e.g., Choi *et al.* (2019) and Gu *et al.* (2021)]. The dramatic advantage associated with complex representations is illustrated by the fact that the deep CASA-estimated complex ideal ratio mask can actually outperform the ideal ratio mask (non-complex), despite that the former is an algorithmic estimate made with no knowledge of the separate target speech and interference signals, whereas the latter is an oracle mask calculated using knowledge of these unmixed signals (Liu and Wang, 2019).

One limitation of the current study involves the use of separate listener groups to assess benefit in non-causal versus causal conditions. Further, the hearing loss of the current HI group was 10 dB HL greater than that employed by Healy *et al.* (2020), which could possibly serve to increase benefit in the current study and reduce the causal cost. But mitigating these effects is the fact that benefit was always measured within subjects as the processed-unprocessed difference. And although benefit has been related to PTA in other works, this relationship was not found in the current data, suggesting that any influence of hearing-loss difference between studies is small if present.

This work follows the philosophy that optimal algorithm performance should be targeted first, to the near exclusion of other considerations. Accordingly, our prior works have targeted performance without much concern over causality or network size. The rationale is that high performance affords room for performance reductions as implementation concerns are addressed. The other rationale is that the establishment of high performance benchmarks allows any performance reductions associated with implementation modifications to be known.

The current study involves the next step, which involves the only fundamental requirement for real-time operation–causality. The high performance benchmark associated with the non-causal network afforded room for reductions to occur and benefit to remain large. Subsequent works will address network size and computational burden. One approach to this issue is to simply reduce the number of layers and units per layer in a traditional neural network. But this typically results in substantial performance decrements. Fortunately, techniques are being developed to allow small, more readily implementable networks to operate with efficacy similar to larger networks [see, e.g., Tan and Wang (2021)].

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., R01 DC012048 to D.L.W., and F32 DC019314 to E.M.J.). We gratefully acknowledge computing resources from the Ohio Supercomputer Center, algorithmic assistance from Yuzhou Liu, and manuscript preparation assistance from Victoria Sevich.

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

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).

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* (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.

Choi, H.-S., Kim, J.-H., Huh, J., Kim, A., Ha, J.-W., and Lee, K. (**2019**). "Phase-aware speech enhancement with deep complex U-Net," in *Proceedings of ICLR*.

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.

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.

Gu, R., Zhang, S.-X., Zou, Y., and Yu, D. (**2021**). "Complex neural spatial filter: Enhancing multi-channel target speech separation in complex domain," IEEE Sign. Proc. Lett. **28**, 1370–1374.

Habets, E. (**2020**). ehabets/RIR-Generator: RIR Generator (v2.2.20201022), Zenodo, https://doi.org/10.5281/zenodo.4117640 (Last viewed September 28, 2021).

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., Johnson, E. M., Delfarah, M., and Wang, D. L. (**2020**). "A talker-independent deep learning algorithm to increase intelligibility for hearing-impaired listeners in reverberant competing talker conditions," J. Acoust. Soc. Am. **147**, 4106–4118.

Healy, E. W., Tan, K., Johnson, E. M., and Wang, D. L. (**2021**). "An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners," J. Acoust. Soc. Am. **149**, 3943–3953.

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.

Hershey, J., Chen, Z., Le Roux, J., and Watanabe, S. (**2016**). "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proceedings of ICASSP*, pp. 31–35.

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.

Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (**2016**). "Temporal convolutional networks: A unified approach to action segmentation," in *Proceedings of ECCV*, pp. 47–54.

Liu, Y., and Wang, D. L. (**2020**). "Causal deep CASA for monaural talker-independent speaker separation," IEEE/ACM Trans. Audio Speech Lang. Proc. **28**, 2109–2118.

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).

Pandey, A., and Wang, D. L. (**2020**). "On cross-corpus generalization of deep learning based speech enhancement," IEEE Trans. Audio, Speech, Lang. Proc. **28**, 2489–2499.

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*, 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*, pp. 749–752.

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

Tan, K., and Wang, D. L. (**2021**). “Compressing deep neural networks for efficient speech enhancement,” in *Proceedings of ICASSP-21*, pp. 8358–8362.

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., and Brown, G. J. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley & IEEE Press, 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., Xu, B., and Zhang, T. (**2020**). “Monaural speech dereverberation using temporal convolutional networks with self attention,” IEEE/ACM Trans. Audio Speech Lang. Process **28**, 1598–1607.