JASA ARTICLE

An effectively causal deep learning algorithm to increase intelligibility in untrained noises for hearing-impaired listeners

Eric W. Healy,[1,a)] Ke Tan,[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:
Real-time operation is critical for noise reduction in hearing technology. The essential requirement of real-time operation is causality—that an algorithm does not use future time-frame information and, instead, completes its operation by the end of the current time frame. This requirement is extended currently through the concept of "effectively causal," in which future time-frame information within the brief delay tolerance of the human speech-perception mechanism is used. Effectively causal deep learning was used to separate speech from background noise and improve intelligibility for hearing-impaired listeners. A single-microphone, gated convolutional recurrent network was used to perform complex spectral mapping. By estimating both the real and imaginary parts of the noise-free speech, both the magnitude and phase of the estimated noise-free speech were obtained. The deep neural network was trained using a large set of noises and tested using complex noises not employed during training. Significant algorithm benefit was observed in every condition, which was largest for those with the greatest hearing loss. Allowable delays across different communication settings are reviewed and assessed. The current work demonstrates that effectively causal deep learning can significantly improve intelligibility for one of the largest populations of need in challenging conditions involving untrained background noises. © *2021 Acoustical Society of America*.
https://doi.org/10.1121/10.0005089

(Received 20 November 2020; revised 9 May 2021; accepted 10 May 2021; published online 7 June 2021)
[Editor: G. Christopher Stecker] Pages: 3943–3953

I. INTRODUCTION

Two primary challenges associated with implementing deep learning-based noise reduction into hearing devices involve (i) generalization to conditions not encountered during training and (ii) real-time operation. The current study provides a demonstration of deep learning-based intelligibility improvement for hearing-impaired (HI) listeners in the context of these challenges.

Generalization refers to the ability of a deep learning algorithm to operate effectively on conditions not encountered during training. This aspect has formed the focus of a series of studies (see Healy *et al.*, 2020). One of the greatest challenges involves the generalization to noises not employed during training (unseen noises). The challenge originates from the fact that "noise" varies widely in its acoustic characteristics, more widely than other sources of acoustic variability, such as different talkers. But despite the challenge, advances have been made with regard to improving intelligibility for HI listeners in these environments. The progression involved the same noise segments used for both training and testing (Healy *et al.*, 2013; Healy *et al.*, 2014), to novel 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).

With regard to real-time operation, the fundamental requirement is causality—that the algorithm operates on only past and current time frames and not future time frames. The other primary aspect of real-time operation involves the computational complexity of the algorithm (i.e., the size of the neural network) and the demand that it places on hardware. But this other aspect is directly related to the computational power of the system on which it runs, which is constantly advancing, and so this would represent no fundamental barrier.

Some work exists to suggest that deep learning noise reduction can improve intelligibility when implemented as a causal system. This is particularly true for listeners who use cochlear implants (CIs). Goehring *et al.* (2017) employed CI listeners and a DNN (deep neural network) using no future time frames and reduced computational complexity to separate sentences from noise. Different segments of the same noise were used for training and testing, and both talker-dependent (same talker used for training and test) and talker-independent (different talkers used for training and test) models were tested. It was found that DNN processing produced speech-reception threshold improvements for CI listeners in all but one condition with mean improvements in those conditions of 1.4–6.4 dB. Goehring *et al.* (2019)

[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.

also employed DNN-based noise reduction. The challenge associated with untrained noise was increased by requiring generalization to a different noise of the same type used for training (either babble or traffic). The model was also talker independent. Speech-reception thresholds were significantly improved in babble (group mean improvement of 3.4 dB) but not in traffic noise.

The challenge is perhaps somewhat greater when the largest population of need is considered—those with sensorineural hearing loss and who wear hearing aids (referred to here as HI listeners). This is because HI listeners usually display better performance in background noise than do CI listeners. Accordingly, baseline (unprocessed) scores tend to be higher, and so algorithm benefit can be more elusive.[1] Despite this overall challenge, some success has been observed. Monaghan *et al.* (2017) employed DNNs using no future time frames and reduced computational complexity to separate sentences from speech-shaped noise (SSN) and babble. Different segments of the same noise were used for training and testing, and the model was talker dependent. Significant intelligibility increases for HI listeners were observed in five of eight conditions. Keshavarzi *et al.* (2019) also employed a DNN using no future frames and reduced computational complexity to separate sentences from multitalker babble. Different segments of the same noise were used for training and testing, and a large set of training talkers was used to produce generalization to untrained talkers. Although human intelligibility was not reported, HI listeners expressed slight but statistically significant preferences for DNN-processed sentences over unprocessed sentences in a paired-comparison task. Bramsløw *et al.* (2018) demonstrated increased intelligibility for HI listeners in a somewhat different task by using a causal and talker-dependent DNN to separate concurrent talkers (speaker separation, rather than speech enhancement).

Here, we extend the concept of causality to "effectively causal." If the target user is human, then future time-frame information that is below the threshold for delay detection or disturbance can be employed with no detriment to the user but with a possible computational advantage. The human delay tolerances that define the limits of effectively causal vary quite widely depending on the communication setting. Non face-to-face communication allows for relatively large delays. For example, the maximum recommended delay across individuals engaged in voice or video calls is 150 ms (ITU, 2003).

The delay tolerances for face-to-face communication by individuals who use hearing technology also vary quite widely. For those who use a CI, auditory-visual synchrony is typically assumed to dictate tolerable delay. There exists a window over which humans perceive auditory and visual information to be synchronous, despite physical delays in one modality relative to the other. These audiovisual synchrony detection thresholds for speech by normal-hearing (NH) individuals are typically in the range of 200 ms (for auditory delayed relative to visual). Thresholds are approximately the same for CI users, causing the tolerable delay for typical CI users to be on the order of 200–250 ms (Hay-McCutcheon *et al.*, 2009).

The delay tolerance is lower for individuals who use hearing aids. This is because in addition to auditory-visual synchrony, both speech perception and speech production need to be considered, and the auditory transmission route during these activities can include both air conduction and bone conduction. When these various factors are all considered, delay tolerances of only 20–30 ms are obtained (Stone and Moore, 1999; 2005; Goehring *et al.*, 2018). This value of 20 ms represents one of the smallest tolerable human communication delays and is used currently to define effectively causal. The various issues surrounding human delay tolerances are further discussed in Sec. IV.

In the current study, deep learning was used to segregate sentences from noise. The primary contributions involve the observation of improved intelligibility for HI listeners in all conditions in the context of an effectively causal system, which operates on entirely untrained noises. A new neural network—a gated convolutional recurrent network (GCRN) was employed. The GCRN was trained using sentences from the Institute of Electrical and Electronics Engineers (IEEE) corpus mixed with 10 000 different noises. The training signal-to-noise ratios (SNRs) did not fully encompass the test SNRs. The model was talker dependent to restrict the generalization challenge to untrained noises. The test noises were both spectro-temporally complex and consisted of a variety of sound sources (multitalker babble and cafeteria noise). Complex mapping was also employed in which the GCRN was used to estimate both the real and imaginary components of the noise-free speech from those of the noisy speech. Using this complex-domain representation, both the magnitude and phase of the noise-free speech signal were estimated. This approach contrasts with that of most other deep learning speech enhancement studies in which only the magnitude of the noise-free speech is estimated and then combined with the phase from the original noisy signal ("noisy phase") to construct the enhanced/noise-reduced speech signal. Typical HI listeners were tested, and significant intelligibility improvements were observed in all conditions. In addition, effectively causal limits for various communication settings were reviewed, and the impact of various amounts of future frame information on performance of the current network was examined.

II. METHOD

A. Subjects

Twelve HI listeners received a monetary incentive for participation. They were recruited from The Ohio State University Speech-Language-Hearing Clinic to represent a diverse sample of typical hearing aid patients. Ages ranged from 62 to 88 years (mean = 72 years), half were female, and all were native speakers of American English. None had previous exposure to the sentence materials used currently.

https://doi.org/10.1121/10.0005089

Otoscopy, tympanometry (ANSI, 1987), and pure-tone audiometry (ANSI, 2004, 2010) were performed for all listeners. Otoscopy was unremarkable for all listeners. Middle-ear peak pressures and compliances were within normal limits for all listeners except HI10, who presented with flat tympanograms. However, the lack of significant air-bone gaps indicated a likely cochlear site of lesion. Listeners generally had bilateral sloping hearing losses of likely cochlear origin that ranged in degree from mild to profound. Pure-tone average audiometric thresholds (PTAs; means across 500, 1000, and 2000 Hz and ears) ranged from 27 to 58 dB hearing level (HL) with a mean of 45 dB HL.

All listeners were binaural hearing aid users with the exceptions of HI1 and HI6. Rather than binaural hearing aids, HI1 uses an amplification system with bilateral microphones and contralateral routing of one of the signals (BiCROS; Harford, 1966). HI6 uses a single, monaural hearing aid in the left ear. Only the better ears of HI1 (right ear) and HI6 (left ear) were tested currently in accord with the hearing aid input they receive. For these two listeners, PTA was also based only on the ear receiving input. Figure 1 displays audiometric thresholds for all listeners, who are numbered in ascending order of degree of hearing loss as reflected by the PTAs.

The decision was made to focus the current study on HI listeners and not test listeners with NH because the latter are remarkably robust to background noise and have perhaps less need for speech-processing technology like that examined here. Further, their robustness to noise causes baseline performance in unprocessed conditions to be quite high, limiting the opportunity to observe benefit.

B. Stimuli

The speech stimuli consisted of IEEE sentences (IEEE, 1969) produced by a male talker having a general Midwestern-American dialect. The corpus of 720 IEEE sentences was split into 3 sets, which included 500, 60, and 160 sentences for training, validation, and testing, respectively.

The noises used for algorithm training consisted of 10 000 nonspeech sounds from a sound-effect library (Richmond Hill, ON, Canada[2]). A large training set of 320 000 mixtures was created by mixing each of the 500 training sentences with a random segment from the set of

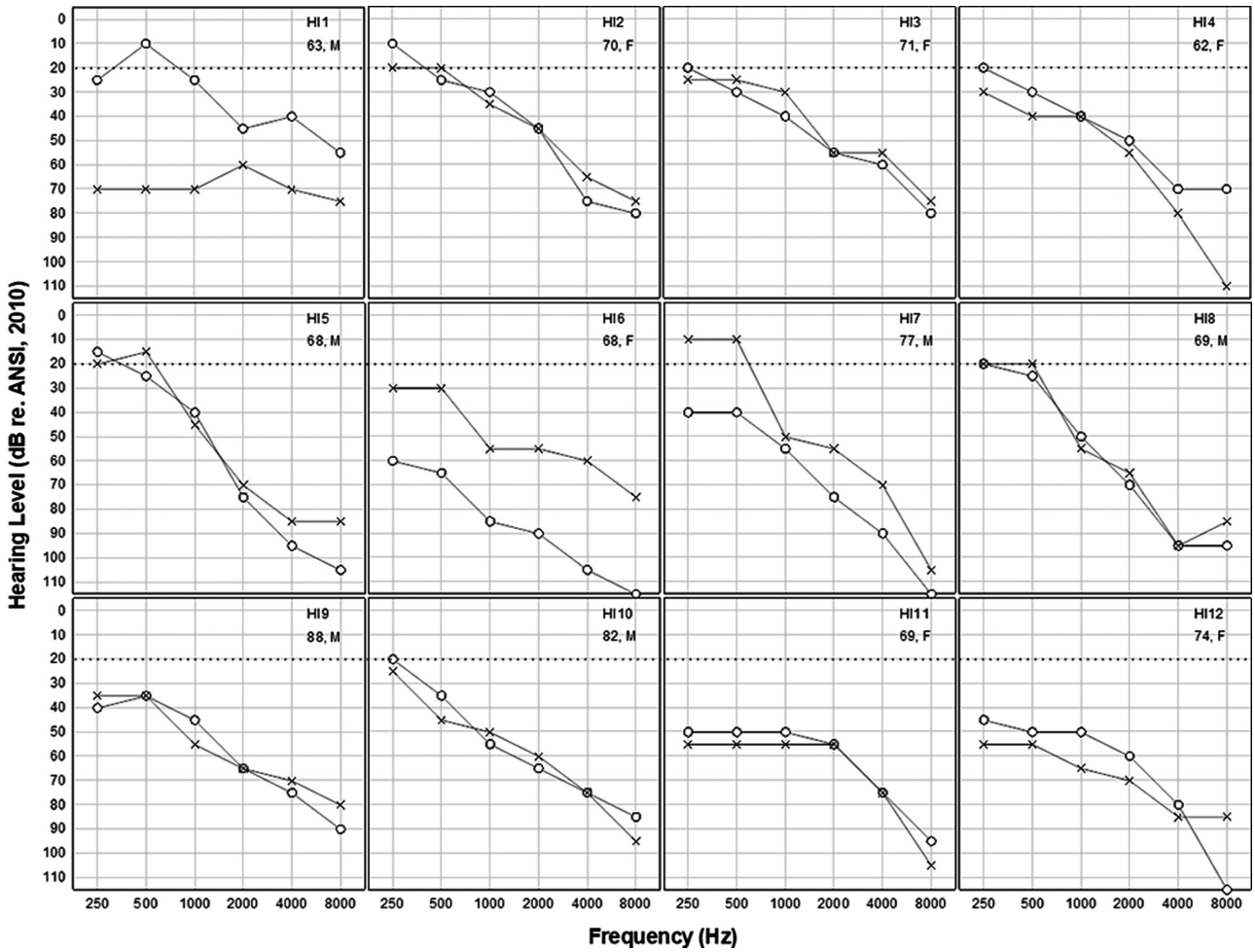

FIG. 1. Pure-tone air conduction audiometric thresholds for each of the 12 listeners, who are numbered in order of increasing pure-tone average audiometric threshold (PTA, means across 500, 1000, and 2000 Hz and ears). Ages and genders are also displayed. Right ears are represented by circles and left ears are represented by X's. The horizontal dotted line in each panel represents the NH limit of 20 dB HL.

nonspeech sounds at an SNR that was randomly sampled from {−5,−4,−3,−2,−1,0} dB. The noises used for algorithm validation included a factory noise and destroyer engine-room noise (from the NOISEX-92 dataset, Varga and Steeneken, 1993) and a SSN generated using VOICEBOX (Brookes, 2005). A set of 60 validation mixtures was created by mixing the 60 validation sentences with a randomly selected portion of 1 of the 3 noises at −5 dB SNR. The noises used for testing included a 20-talker babble and cafeteria noise, each approximately 10 min in duration and both from an Auditec compact disc (St. Louis, MO[3]). The cafeteria noise consisted of three overdubbed recordings from a busy hospital cafeteria and, consequently, involved a variety of sound sources, including multiple voices, impact noises from dishes, etc. The SNRs used for testing were −1 and +3 dB. The testing set, therefore, consisted of 160 sentences × 2 noises × 2 SNRs = 640 mixtures.

C. Algorithm description

As mentioned in Sec. I, conventional speech enhancement involves estimation of only the magnitude of noise-free speech and the use of noisy phase to construct the enhanced/noise-reduced signal. This approach stems, in part, from the fact that the phase spectrogram of speech exhibits no clear structure, rendering its direct estimation unfeasible. However, both the real and imaginary spectrograms of speech do exhibit clear spectro-temporal structure (Williamson *et al.*, 2016). Accordingly, they can both be estimated using deep learning from which both the magnitude and phase of the target noise-free speech can be estimated. Such a complex mapping approach (Fu *et al.*, 2017; Tan and Wang, 2020) is employed currently.

In the current study, a GCRN was employed, which operated in the complex domain (was used to estimate both the real and imaginary parts). The network accepted speech-plus-noise input from a single microphone and so was a monaural system. The model was based on the convolutional recurrent network (CRN) proposed by Tan and Wang (2018), and then extended to the complex domain by Tan and Wang (2020). The algorithm is extended currently to be effectively causal by allowing the use of a small amount of future time-frame information within the delay tolerance of the human auditory system.

Figure 2 depicts a simple system diagram for the proposed algorithm. All speech and noise were sampled at 16 kHz for processing. A 20-ms Hamming window was employed to segment the signals into a series of time frames with a 10-ms window shift. To derive the time-frequency (T-F) representations of the signals, a 320-point fast Fourier transform was applied to each time frame, yielding 161-dimensional complex spectra.

The original CRN (Tan and Wang, 2018) was essentially an encoder-decoder architecture with long short-term memory (LSTM) between the encoder and decoder. Specifically, the encoder was composed of five convolutional layers, and the decoder was composed of five deconvolutional layers. Between the encoder and decoder, two LSTM layers were employed to model temporal dependencies. The encoder-decoder architecture has a symmetric structure, where the number of kernels progressively increases in the encoder and decreases in the decoder. A stride of two was adopted along the frequency dimension in all convolutional and deconvolutional layers to aggregate context along the frequency dimension. Hence, the frequency dimensionality of feature maps was halved layer by layer in the encoder and doubled layer by layer in the decoder. Such a design ensures that the output spectrogram has the same resolution as the input spectrogram. Moreover, skip connections were used to concatenate the output of each encoder layer to the input of the corresponding decoder layer. Note that all convolutions and deconvolutions were causal so that this CRN did not use any future time-frame information for estimation.

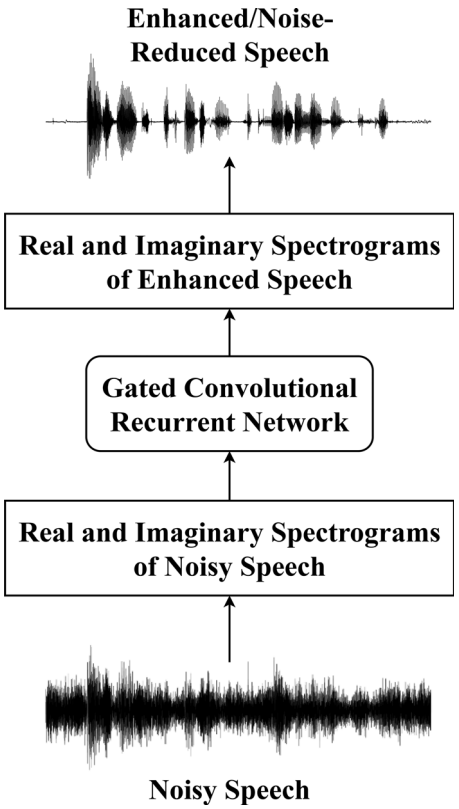

FIG. 2. A system diagram of the proposed GCRN-based speech-segregation algorithm.

To extend this model to perform complex spectral mapping, Tan and Wang (2020) added gated linear units (GLUs; Dauphin *et al.*, 2017). The resulting network is referred to as a GCRN. Gating mechanisms control the flow of information through the network, potentially allowing it to model more complex interactions. A convolutional gated linear unit block (ConvGLU) is illustrated on the left side of Fig. 3. A deconvolutional gated linear unit block (DeconvGLU) is analogous except that the convolutional layers are replaced by deconvolutional layers as shown on the right side of Fig. 3.

https://doi.org/10.1121/10.0005089

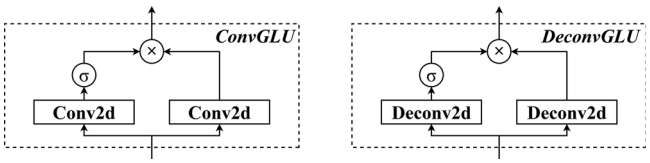

FIG. 3. Diagrams of a ConvGLU and DeconvGLU, where σ denotes a sigmoid function.

Figure 4 depicts the GCRN architecture. Note that the real and imaginary spectrograms of noisy speech are treated as two different input channels. As shown in Fig. 4, the encoder and LSTMs were shared across the estimates of real and imaginary components, whereas two decoders were used to estimate the real and imaginary spectrograms of clean speech, respectively. Each ConvGLU or DeconvGLU block was coupled with a batch normalization operation (Ioffe and Szegedy, 2015) and an exponential linear unit (ELU; Clevert *et al.*, 2016) activation function. The GLU blocks in the encoder had 16, 32, 64, 128, and 256 output channels successively, and those in the decoders had 128, 64, 32, 16, and 1 output channels successively. For all convolutional and deconvolutional layers, the kernel size was set to 1×3. Moreover, a linear layer was stacked on top of each decoder to project the learned features to the real or imaginary spectrograms.

The grouping strategy of Gao *et al.* (2018) was employed to reduce model complexity and increase efficiency. In this strategy, the input and hidden layers are split into groups, and features are learned separately within each group, substantially reducing the number of connections between layers and, thus, model complexity as illustrated in Fig. 5. Then, to recover the important dependencies across groups, a representation-rearrangement layer is employed between the recurrent layers. This grouping strategy was employed for the LSTM layers, given that most trainable parameters in the GCRN reside in the LSTM layers.

The model was trained using the AMSGrad optimizer (Reddi *et al.*, 2018) with a learning rate of 0.001. The mean squared error (MSE) was used as the objective function. The minibatch size was set to 16 at the utterance level. Within a minibatch, all training examples were padded with zeros to have the same number of time steps as the longest example. The best model was selected by cross-validation. During inference, the real and imaginary spectrograms of the estimated noise-free speech were used to resynthesize the waveform via an inverse short-time Fourier transform (iSTFT).

The fully causal GCRN was slightly modified for the current study to use two future time frames. Specifically, the kernel sizes of the convGLU and deconvGLU layers preceding and following the LSTM module were set to 3×3 rather than 1×3. The incorporation of two future frames, each 20 ms in duration with 10-ms frameshift, added 20 ms of processing delay. This delay adds to the inherent system delay of 10 ms, corresponding to the size of the current time frame.

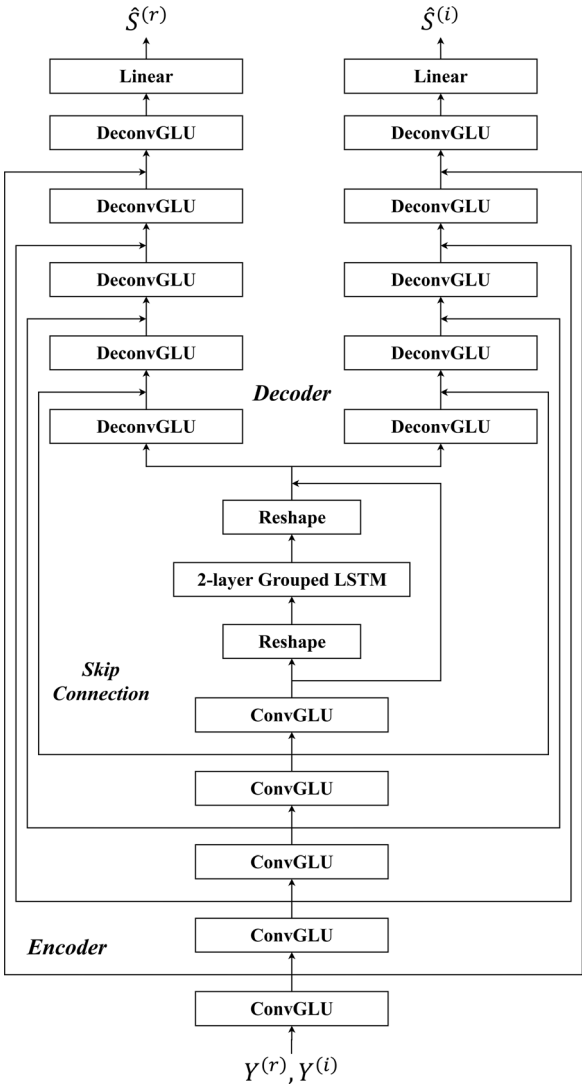

FIG. 4. The network architecture of the GCRN for complex spectral mapping, where $Y^{(r)}$ and $Y^{(i)}$ denote the real and imaginary spectrograms, respectively, of noisy speech, and $\hat{S}^{(r)}$ and $\hat{S}^{(i)}$ denote the real and imaginary spectrograms, respectively, of enhanced (noise-reduced) speech.

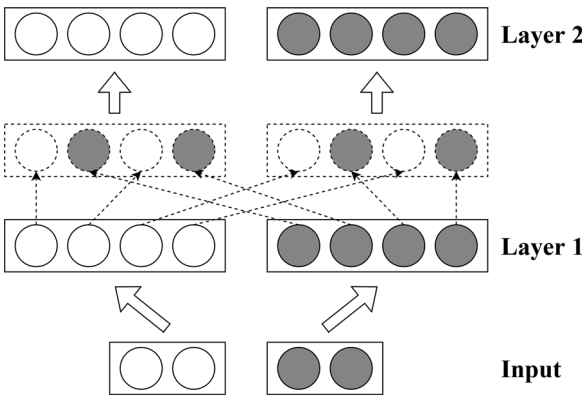

FIG. 5. Illustration of the grouping strategy for LSTMs in which the input features and units in each LSTM layer were split into two groups. Displayed between layers 1 and 2 is a representation-rearrangement layer employed between the recurrent layers.

D. Procedure

The two noise types, two SNRs, and two processing conditions (unprocessed and algorithm processed) yielded eight conditions. Each listener heard 20 sentences in each condition for a total of 160 sentences.[4] The comparison of greatest interest was that between the unprocessed and corresponding processed conditions, and so those conditions were presented in juxtaposed and random order for each noise type and SNR. The correspondence between the sentence list and condition was random, and no sentence was used more than once for any listener.

The signals were presented using a Windows personal computer (Microsoft, Redmond, WA), Echo Digital Audio Gina 3G digital-to-analog converter (Santa Barbara, CA), Mackie 1202-VLZ mixer (Woodinville, WA), and Sennheiser HD 280 Pro headphones (Wedemark, Germany). The stimuli were each scaled to the same total root mean square level and presented at 65 dBA plus individualized frequency-specific gains as prescribed by the NAL-RP hearing aid fitting formula (Byrne *et al.*, 1990). Because the NAL-RP formula does not prescribe gains for 125 or 8000 Hz, gains for 250 and 6000 Hz, respectively, were applied to these two frequencies. This listener-specific amplification was applied using a RANE DEQ 60L digital equalizer (Mukilteo, WA). Listeners were tested diotically using gains based on audiometric thresholds averaged across the two ears. The exceptions were HI1 and HI6, who were tested monaurally in accord with their hearing aid input. NAL-RP gains for these two listeners were calculated using the audiometric thresholds for the test ear. Because hearing aid gains were applied using the experimental apparatus, listeners were tested with their hearing aids removed. Presentation levels were calibrated at the beginning of each test session using a sound-level meter and flat-plate headphone coupler (Larson Davis 824 and AEC 101, Depew, NY).

Testing began with a brief familiarization. Listeners heard 25 practice sentences equally divided into 5 blocks. The order of the practice conditions were (1) clean speech, (2) algorithm-processed 3 dB SNR babble, (3) algorithm-processed -1 dB SNR cafeteria noise, (4) unprocessed 3 dB SNR babble, and (5) unprocessed -1 dB SNR cafeteria noise. These practice sentences were drawn from the IEEE set used for algorithm training, and so they were distinct from those used for formal testing.

During familiarization, listeners were asked if the signals were clearly audible and comfortable in level. Five listeners (HI4, HI7, HI8, HI10, and HI12) reported that the stimuli sounded loud. After reducing presentation level by 5 dB, three of these listeners reported that the signals were clear and comfortable. HI4 reported that the signals were clear and comfortable following an additional 1-dB reduction. HI8 reported that the signals were not loud enough following the 5-dB reduction, but that they were clear and comfortable following a subsequent 3 dB gain increase. The overall presentation level after the application of NAL-RP gains and comfort adjustment ranged across listeners from 73.6 to 87.3 dBA (mean = 81.3 dBA).

Following familiarization, the listeners heard the eight blocks of experimental conditions. They were instructed to repeat back each sentence as accurately as possible and guess if unsure. The listeners were blind to the condition under test, but the experimenter was not. Listeners were seated in a double-walled audiometric booth with the experimenter who controlled the presentation of each stimulus and scored keywords correctly reported. Each sentence was played only once during testing. The total duration of testing was approximately 45 min for each listener.

III. RESULTS AND DISCUSSION

Each sentence contained 5 keywords for a total of 100 keywords in each condition. Intelligibility for each condition was based on the proportion of these keywords that were correctly reported. Figures 6 and 7 display the intelligibility for each individual HI listener in each condition: Fig. 6 for the multitalker babble conditions and Fig. 7 for the cafeteria-noise conditions. The two SNRs employed for

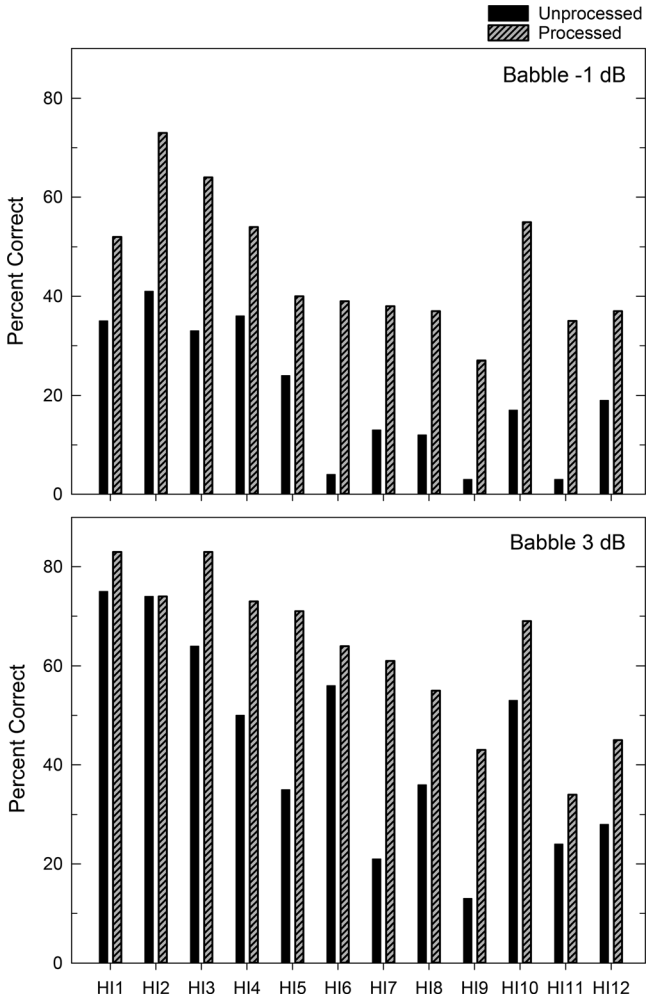

FIG. 6. The sentence intelligibility for each HI listener in multitalker babble. The two SNRs employed are displayed in separate panels. The unprocessed speech-in-babble scores are represented by solid columns and algorithm-processed scores are represented by hatched columns. Algorithm benefit for each listener is thus represented by the difference between these corresponding columns.

https://doi.org/10.1121/10.0005089

each noise type are plotted in separate panels in Figs. 6 and 7. Note that HI3 was unavailable to participate in cafeteria noise, and so her data are not plotted in Fig. 7. In each panel, the unprocessed and processed conditions are represented by different columns. The algorithm benefit for each listener thus corresponds to the difference between each solid column (unprocessed) and corresponding hatched column (processed).

As Fig. 6 shows, every HI listener received algorithm benefit at the less favorable babble SNR and all but one received benefit at the more favorable babble SNR. As Fig. 7 shows, all but two listeners received algorithm benefit at the less favorable cafeteria-noise SNR, and all but one received benefit at the more favorable cafeteria-noise SNR. Considering all listeners and conditions (46 cases), benefit was 10 percentage points or greater in 67% of cases, 20 percentage points or greater in 37% of cases, and 30 percentage points or greater in 22% of cases.

As expected, unprocessed scores in Figs. 6 and 7 show a general trend toward decreasing from left to right as the degree of hearing loss (PTA) increased. Because algorithm benefit is a function of unprocessed scores, these values are necessarily related with lower unprocessed scores tending to be associated with greater benefit. A correlation between listener PTA and mean benefit score, averaged across conditions and based on rationalized arcsine units (RAUs; Studebaker, 1985), revealed that PTA was significantly related to benefit [$r(10) = 0.65$, $p = 0.02$] with those having greater degrees of hearing loss displaying a greater benefit. However, similar correlations performed for each condition separately suggested that the overall effect was largely driven by the pattern of scores in cafeteria noise at an SNR of 3 dB as this was the only significant correlation of these four.

Figure 8 displays the group-mean scores and standard errors for each condition. The top panel of Fig. 8 represents the two babble SNRs. The less favorable babble SNR (top panel of Fig. 8, left) produced the lowest group-mean

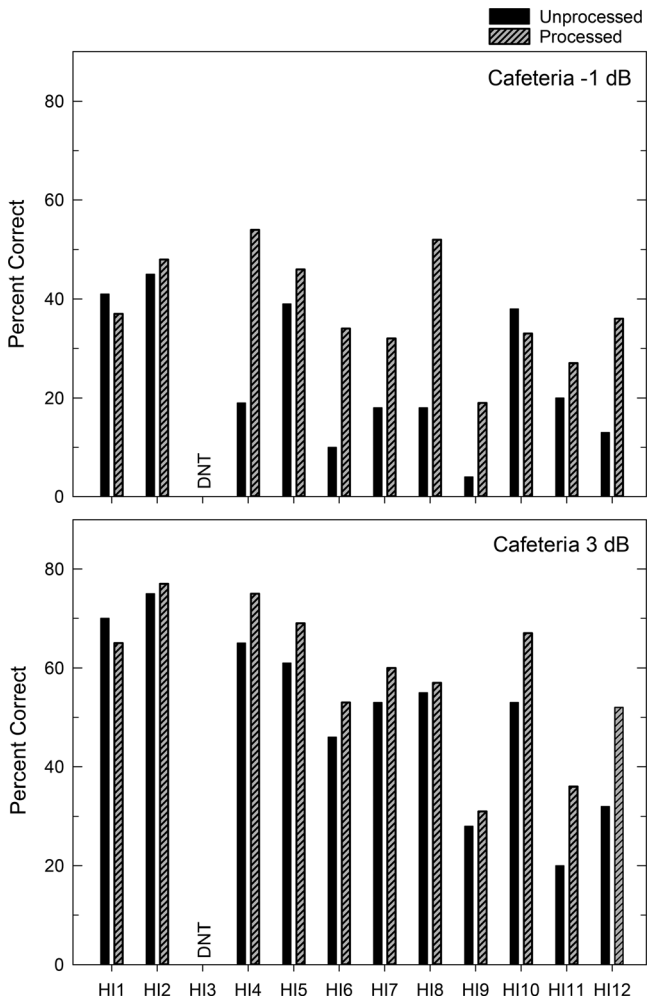

FIG. 7. As Fig. 6 but for the cafeteria-noise conditions. DNT, did not test.

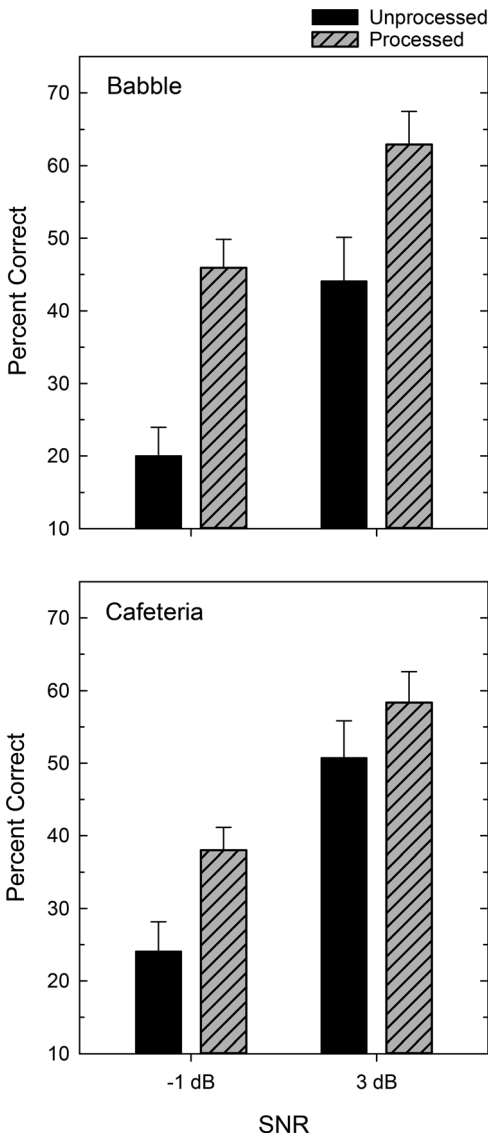

FIG. 8. The group-mean intelligibility scores and standard errors for the HI listeners. The two noise types are displayed in separate panels, and the two SNRs employed are displayed in each panel. Unprocessed and processed scores are again represented by separate columns, and the algorithm benefit is reflected by the difference between corresponding columns as in Fig. 6.

unprocessed score (20%), which rose (to 46%) to produce the largest mean benefit of 26 percentage points. The more favorable babble SNR produced an unprocessed score of 44% and a benefit of 19 percentage points. The bottom panel of Fig. 8 represents the two cafeteria-noise SNRs. The less favorable cafeteria-noise SNR produced an unprocessed score of 24% and a benefit of 14 percentage points. The more favorable SNR produced the highest group-mean unprocessed score (51%) and the correspondingly smallest algorithm benefit of 8 percentage points.

Planned comparisons consisting of uncorrected two-tailed paired *t*-tests on RAUs were performed to examine algorithm benefit in each condition. Scores for algorithm-processed conditions were significantly higher than the corresponding unprocessed scores in every condition (two babble SNRs [$t(11) \geq 5.4$, $p < 0.001$], two cafeteria-noise SNRs [$t(10) \geq 3.3$, $p < 0.01$]). These results all survive Bonferroni correction for multiple comparisons.

A supplementary statistical analysis was performed using a linear mixed-effects model. The outcome variable was the RAU-transformed percent-correct scores for each listener in each condition. The fixed effects for this model were processing condition, SNR, and noise type, plus each of the two-way interactions between the first-order effects. The model also included random intercepts for listener. Deviation coding was used to represent the variables of processing condition, SNR, and noise type. Deviation coding specifies contrasts that are analogous to factors in traditional analyses of variance and thus facilitates model interpretation. The unprocessed condition was coded as −0.5 (baseline), and the algorithm-processed condition was coded as 0.5 (comparison). The lower SNR of −1 dB was coded as −0.5 (baseline) and the higher SNR of 3 dB was coded as 0.5 (comparison). Cafeteria noise was arbitrarily assigned as the baseline condition and coded as −0.5, whereas babble was designated as the comparison and coded as 0.5. Visual inspection of residual plots did not reveal any clear violations of homoscedasticity or normality. The degrees of freedom were based on Satterthwaite's approximation. The analysis was performed using *R* 3.5.1 (*R* Core Team, 2019) with the *lme4* (Bates *et al.*, 2015) and lmerTest (Kuznetsova *et al.*, 2020) packages.

Most importantly, the fixed effect of algorithm processing was highly significant, reflecting that the algorithm was successful in increasing the listeners' overall intelligibility [$\beta = 16.6$ RAU, $t(73.9) = 8.9$, $p < 0.0001$]. The fixed effect of SNR was also large and significant, simply reflecting the overall higher intelligibility in the more favorable SNR conditions [$\beta = 21.7$ RAU, $t(73.9) = 11.7$, $p < 0.0001$]. Finally, the fixed effect of noise type was nonsignificant, indicating no overall difference in intelligibility between the babble and cafeteria-noise conditions [$\beta = -0.09$ RAU, $t(74.5) = -0.05$, $p = 0.96$]. The interaction between SNR and processing condition was significant, reflecting a greater algorithm benefit for the lower SNR conditions [$\beta = -11.2$ RAU, $t(73.9) = -3.0$, $p < 0.005$]. The interaction between noise type and processing condition was also significant, reflecting a greater algorithm benefit in the babble conditions [$\beta = 8.24$ RAU, $t(73.9) = 2.2$, $p < 0.05$]. Finally, the interaction between the noise type and SNR was nonsignificant, indicating no significant differential effect of SNR between noise types [$\beta = 4.46$ RAU, $t(73.9) = -1.2$, $p = 0.23$].

IV. GENERAL DISCUSSION

The current study demonstrates that effectively causal deep learning noise reduction can lead to consistent intelligibility improvements for HI listeners. Benefit was significant in every condition and largest in the babble background at the less favorable SNR and for those with the greatest degree of hearing loss. As described in Sec. I, the current model was talker dependent to restrict the generalization challenge to untrained noises. The use of a talker-dependent model is in accord with the work of Goehring *et al.* (2017), Monaghan *et al.* (2017), Bentsen *et al.* (2018), and Bramsløw *et al.* (2018). However, it is unlike the talker-independent models of Goehring *et al.* (2017), Goehring *et al.* (2019), and Keshavarzi *et al.* (2019). The current use of entirely novel noise types for testing makes the current study unlike previous works, which employed different segments of the same noise (Goehring *et al.*, 2017; Monaghan *et al.*, 2017; Bensten *et al.*, 2018; Keshavarzi *et al.*, 2019) or different noises of the same type (Goehring *et al.*, 2019) for training and testing. The algorithmic advances over prior studies include the use of a GCRN and complex mapping to estimate both the magnitude and phase of the target speech and the introduction of an effectively causal model.

As described in Sec. I, it is important to note that the tolerable delay for human listeners is not fixed across communication settings and, instead, is multifaceted. At one end of the continuum is non-face-to-face communication, such as voice calls or radio communication. These delay tolerances can be quite large, at least the 150 ms specified by ITU (2003), and perhaps considerably longer.

The tolerable delay is also large for CI users because little if any auditory information is obtained through air- or bone-conducted sound, and essentially all of the auditory input is through the implant. Accordingly, speech production is typically not considered when assessing the impact of CI processing delay because delayed auditory feedback from non-device air conduction or bone conduction of the talkers own voice is absent. Similarly, speech perception does not occur mixed with non-device air-conducted sound (which can occur, for example, in an open hearing aid fitting as described below). Thus, audiovisual synchrony is typically assumed to define CI delay tolerance. The tolerable delay is considerably larger when the auditory signal is delayed relative to the visual signal, relative to the reverse, which is convenient for audio-processing technology. The delay detection thresholds for CI users obtained by Hay-McCutcheon *et al.* (2009; for delayed audio) averaged 261 ms for middle-aged CI users and 294 ms for elderly CI users. Almost exactly the same values were observed for

age-matched NH comparison subjects. See Conrey and Pisoni (2006) for a review of human auditory-visual synchrony detection.

Within a hearing aid, the situation is more multifaceted and the delay value judged to be disruptive can vary depending on the fitting, communication direction (speech production versus perception), and other factors. Stone and Moore conducted a series of studies to examine tolerable group delays associated with speaking while wearing a hearing aid. The issue arises because one's own voice is heard without delay through bone conduction (and air conduction in an open fitting) but also through the hearing aid following any processing delay, potentially producing delayed auditory feedback. It was found that delays needed to be 20–30 ms to be judged "disturbing" in simulated mild or moderate hearing losses and 40 ms for moderately severe losses (Stone and Moore, 1999). Using HI listeners, delays of 20 ms produced judgments below disturbing for all subject groups (Stone and Moore, 2005).

The perception of others' voices is also an important concern for hearing aid wearers, especially when an open fitting is employed. This fitting provides no seal in the ear canal, and so the communication partner's voice will be received through the air without delay if the hearing loss is mild enough but also amplified through the hearing aid following any processing delay. Stone *et al.* (2008) examined the impact of processing delay in this context using NH listeners and simulated mild to moderate hearing losses. Far smaller delay tolerances of only several ms were observed, but the authors suggested that the use of wide dynamic range compression plus simulated loudness recruitment could have interacted with the effects of delay, complicating interpretation. Goehring *et al.* (2018) examined the effect of hearing aid delays without any nonlinear processing such as compression. Group-mean ratings by HI listeners below the midpoint on a seven-point "annoyance" scale (toward a "not annoying at all" end point) were obtained at 30–40 ms for both self-produced and external voices. These authors also found that the delay tolerance was greater for HI than for NH listeners and greater for HI listeners having larger degrees of hearing loss.

Together, this body of work suggests that a multitude of factors must be considered when determining an allowable processing delay. But this complexity also provides opportunity. Algorithms designed for telecommunications or CIs might take advantage of the relatively high tolerance and consider substantial future time-frame information as a way to improve performance. Turning to hearing aids, advances in technology can potentially cause the acceptable delay values to shift. For example, modern devices can detect when the wearer is speaking and decrease amplification, thus reducing the importance of own-voice considerations. As a caveat, we note that future-frame delays add to other delays inherent in system processing to produce overall delays.

It is important to recognize that the impact of incorporating future time frames likely depends on the particular network architecture and training employed. That said, an analysis was conducted to assess the impact of various amounts of future-frame information in the current model. Algorithms were constructed that were identical to that employed currently, except that future frames totaling 0, 20, 80, 150, 200, and 500 ms were employed. A noncausal, utterance-based model was also constructed. Training was identical for each model as described in Sec. II.

Specifically, the kernel size in the first encoder GLU block was changed to 13×3, 27×3, 37×3, and 97×3, respectively, corresponding to 80, 150, 200, and 500 ms future-frame delays. The GCRN with an utterance-length delay was derived by replacing unidirectional LSTMs in the fully causal GCRN with bidirectional LSTMs. These models were evaluated using four objective metrics, including (1) short-time objective intelligibility (STOI; Taal *et al.*, 2011), (2) extended short-time objective intelligibility (ESTOI; Jensen and Taal, 2016), (3) perceptual evaluation of speech quality (PESQ; Rix *et al.*, 2001), and (4) signal-to-noise ratio improvement (ΔSNR). Scores were averaged over the two test noises for this analysis.

Table I displays the objective measures of intelligibility and sound quality for these models. Apparent in these values is one clear increment in scores from 0 to 20 ms of future-frame information, then another clear increment from 500 ms to the utterance-length model. Although these steps represent the most substantial score increases, the increases resulting from the remaining steps were positive in 26 of 32 instances. These objective results clearly indicate that the use of an effectively causal model can result in a processed

TABLE I. Average STOI, ESTOI, PESQ, and ΔSNR values produced by systems with different future time-frame delays.

SNR	−1 dB				3 dB			
	STOI (%)	ESTOI (%)	PESQ	ΔSNR (dB)	STOI (%)	ESTOI (%)	PESQ	ΔSNR (dB)
Unprocessed	62.89	30.33	1.63	0.00	72.70	42.41	1.89	0.00
0 ms	81.70	60.19	2.35	9.50	87.44	69.90	2.62	7.61
20 ms	83.43	63.26	2.42	10.05	88.55	72.24	2.68	8.04
80 ms	83.51	63.46	2.44	9.99	88.44	72.21	2.70	7.85
150 ms	83.88	63.73	2.45	10.11	88.60	72.30	2.72	8.01
200 ms	84.01	64.00	2.45	10.16	88.82	72.66	2.70	8.15
500 ms	84.04	64.27	2.46	10.22	88.95	73.00	2.71	8.20
Utterance	85.48	67.14	2.54	10.57	90.07	75.34	2.80	8.54

signal that is more acoustically similar to the original clean speech than that produced by a fully causal model.

The decision was made currently to employ an effectively causal system (20-ms future frame delay) rather than a fully causal system because this delay does not hinder human listeners and so its use likely came with no cost to human performance. This also maximized the opportunity to observe benefit as objective scores suggest that a performance advantage may exist. To assess the actual impact of this decision, a fully causal network, identical to that of the main experiment except that no future frames were employed, was used to test listeners. Those with NH were employed because they should be best able to reveal slight differences between the processed speech signals containing little background noise. Ten young-adult NH listeners (pure-tone audiometric thresholds of 20 dB HL or below from 250 to 8000 Hz) heard 15 IEEE sentences at 65 dBA over headphones in each of the following 8 conditions: [2 SNRs (-1 and 3 dB) $\times$ 2 noise types (babble and cafeteria) $\times$ 2 processed conditions (2 versus 0 future frames)]. Thus, all conditions were algorithm processed. No significant difference in scores was observed between two future frames and zero future frames in any of the four conditions (two-tailed paired-comparison t-tests on RAUs, all $p > 0.05$), and the grand-mean average difference between scores in these two processing conditions was less than 1 percentage point. All condition means were free of floor and ceiling effects that could obscure differences. Thus, despite the improvements in objective measures observed with the addition of 20 ms of future-frame information (Table I), the use of 2 versus 0 future time frames had no effect on actual human intelligibility using the current network and conditions.

These results highlight the lack of direct correspondence between acoustic signal similarity as reflected by objective scores and actual human intelligibility. This lack of direct correspondence likely originates, in part, from the fact that Table I documents raw STOI (or ESTOI) scores, which need to be mapped to predicted intelligibility scores with a sigmoidal function whose parameters depend on the speech material used (Taal *et al.*, 2011) and that sigmoidal functions are subject to plateau effects. But clearly, improvements in acoustic similarity were observed for the various effectively causal models corresponding to the different delay tolerances. Further, different network architectures may benefit differently from future-frame information. These together suggest that the concept of effectively causal, tailored to the particular communication setting, remains a potentially useful design option.

It is concluded that effectively causal deep learning can be used to improve intelligibility for HI listeners in the context of entirely untrained complex noises. The current GCRN used complex mapping to obtain estimates of both the magnitude and phase of the noise-free target speech, and intelligibility improvements were observed in all conditions. Future work will be required to add talker independence to the current effectively real-time capable operation. Also, algorithm performance was targeted currently with little concern for model complexity in accord with our overall design and implementation philosophy (described in Healy *et al.*, 2020). Future work will explore techniques to reduce model complexity while maintaining performance. Finally, it is concluded that the concept of effectively causal represents an addition to the array of tools available for the development of deep learning algorithms capable of removing background noise from speech and increasing intelligibility.

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., manuscript preparation also supported by Grant No. F32 DC019314 to E.M.J.) We gratefully acknowledge computing resources from the Ohio Supercomputer Center and thank Victoria Sevich for assistance with the preparation of the manuscript.

[1]These listeners can be considered to fall on a continuum of performance in noise, with CI users at one end (poor performance) and NH listeners at the other (good performance). Although individual variability can be vast, typical HI listeners fall somewhere between these endpoints.
[2]See www.sound-ideas.com (Last viewed 5/21/2021).
[3]See www.auditec.com (Last viewed 5/21/2021).
[4]The exception was HI3 who was unavailable to complete testing and heard all of the babble conditions but none of the cafeteria-noise conditions.

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 (**2010**). S3.6, *American National Standard Specification for Audiometers* (American National Standards Institute, New York).
Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, 1–48.
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.
Brookes, M. (**2005**). "VOICEBOX: Speech processing toolbox for MATLAB," available at http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (Last viewed 01/13/2020).
Byrne, D., Parkinson, A., and Newall, P. (**1990**). "Hearing aid gain and frequency response requirements for the severely/profoundly hearing impaired," Ear Hear. **11**, 40–49.
Chen, J., Wang, Y., Yoho, S. E., Wang, D. L., and Healy, E. W. (**2016**). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," J. Acoust. Soc. Am. **139**, 2604–2612.
Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (**2016**). "Fast and accurate deep network learning by exponential linear units (elus)," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6865–6869.
Conrey, B., and Pisoni, D. P. (**2006**). "Auditory-visual speech perception and synchrony detection for speech and nonspeech signals," J. Acoust. Soc. Am. **119**, 4065–4073.
Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. (**2017**). "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning* Vol. 70, pp. 933–941.

Fu, S.-W., Hu, T.-Y., Tsao, Y., and Lu, X. (**2017**). "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *Proceedings of IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6.

Gao, F., Wu, L., Zhao, L., Qin, T., Cheng, X., and Liu, T.-Y. (**2018**). "Efficient sequence learning with group recurrent networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1 (Long Papers), pp. 799–808.

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., Chapman, J. L., Bleek, S., and Monaghan, J. M. (**2018**). "Tolerable delay for speech production and perception: Effects of hearing ability and experience with hearing aids," Int. J. Audiol. **57**, 61–68.

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.

Hay-McCutcheon, M. J., Pisoni, D. P., and Hunt, K. K. (**2009**). "Audiovisual asynchrony detection and speech perception in hearing-impaired listeners with cochlear implants: A preliminary analysis," Int. J. Audiol. **48**, 321–333.

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., 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.

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

Ioffe, S., and Szegedy, C. (**2015**). "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, pp. 448–456.

ITU-T Recommendation, G.114. (**2003**). *One-Way Transmission Time* (International Telecommunication Union, Geneva, Switzerland).

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.

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. B. (**2020**). "lmerTest: Tests in linear mixed effects models," *R* package version 3.1-3, available at https://CRAN.R-project.org/package=lmerTest (Last viewed 5/21/2021).

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.

R Core Team (**2019**). "R: A language and environment for statistical computing," R Foundation for Statistical Computing, https://www.R-project.org/ (Last viewed 20 May 2021).

Reddi, S. J., Kale, S., and Kumar, S. (**2018**). "On the convergence of adam and beyond," in *International Conference on Learning Representations (ICLR)*.

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 of IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 7–11, Salt Lake City, UT, pp. 749–752.

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. (**2005**). "Tolerable hearing-aid delays: IV. Effects on subjective disturbance during speech production by hearing-impaired subjects," Ear Hear. **26**, 225–235.

Stone, M. A., Moore, B. C. J., Meisenbacher, K., and Derleth, R. P. (**2008**). "Tolerable hearing aid delays. V. Estimation of limits for open canal fittings," Ear Hear. **29**, 601–617.

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.

Tan, K., and Wang, D. L. (**2018**). "A convolutional recurrent neural network for real-time speech enhancement," in *Nineteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 3229–3233.

Tan, K., and Wang, D. L. (**2020**). "Learning complex spectral mapping with a gated convolutional recurrent network for monaural speech enhancement," IEEE/ACM Trans. Audio, Speech, Lang. Process. **28**, 380–390.

Varga, A., and Steeneken, H. J. (**1993**). "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Commun. **12**, 247–251.

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.

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.