JASA ARTICLE

Progress made in the efficacy and viability of deep-learning-based noise reduction

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

ABSTRACT:
Recent years have brought considerable advances to our ability to increase intelligibility through deep-learning-based noise reduction, especially for hearing-impaired (HI) listeners. In this study, intelligibility improvements resulting from a current algorithm are assessed. These benefits are compared to those resulting from the initial demonstration of deep-learning-based noise reduction for HI listeners ten years ago in Healy, Yoho, Wang, and Wang [(2013). J. Acoust. Soc. Am. **134**, 3029–3038]. The stimuli and procedures were broadly similar across studies. However, whereas the initial study involved highly matched training and test conditions, as well as non-causal operation, preventing its ability to operate in the real world, the current attentive recurrent network employed different noise types, talkers, and speech corpora for training versus test, as required for generalization, and it was fully causal, as required for real-time operation. Significant intelligibility benefit was observed in every condition, which averaged 51% points across conditions for HI listeners. Further, benefit was comparable to that obtained in the initial demonstration, despite the considerable additional demands placed on the current algorithm. The retention of large benefit despite the systematic removal of various constraints as required for real-world operation reflects the substantial advances made to deep-learning-based noise reduction. © *2023 Acoustical Society of America*.
https://doi.org/10.1121/10.0019341
(Received 6 August 2022; revised 17 April 2023; accepted 17 April 2023; published online 3 May 2023)
[Editor: Pavel Zahorik] Pages: 2751–2768

I. INTRODUCTION

Difficulty understanding speech in background noise remains the primary auditory complaint of hearing-impaired (HI) listeners (Kramer *et al.*, 1998; Dillon, 2012). This problem persists despite considerable technological advances made to hearing aids and other devices over several decades. A noise-reduction approach that has shown considerable promise involves deep learning. In this approach, a deep neural network (DNN) is trained to isolate target speech from various interferences including background noise, interfering speech, and/or room reverberation, allowing substantial increases in target-speech intelligibility. The current study was conducted to establish advances made toward implementing single-microphone deep learning noise reduction[1] into hearing devices to solve the speech-in-noise problem. There were two specific purposes: First, intelligibly increases for HI and normal-hearing (NH) listeners resulting from a state-of-the-art deep learning algorithm were examined. This model was trained and tested using different noises (noise independent), different talkers and speech recordings (speaker and channel independent), and used only past and the current time frame, allowing real-time operation (causal, see Pandey and Wang, 2022). Accordingly, it was free of all fundamental constraints preventing operation in the real world. The second purpose of the current study was to compare these results to those obtained from the initial demonstration of deep-learning-based noise reduction (Healy *et al.*, 2013). In sharp contrast to the current model, that initial model was highly constrained. It used the same noise segments from which training and test noise samples were randomly chosen, and training and test utterances were selected from the same talker and recording. This high level of similarity in training and test would likely cause it to fail if presented with different noises and talkers, and it used future time frames, preventing real time operation. The use of similar subject populations and test conditions facilitates comparison across the decade.

Prior studies of deep-learning-based noise reduction that we are aware of and in which human intelligibility was assessed are listed in Table I. The current focus is on human-subjects intelligibility (particularly HI or cochlear implant, CI, listeners) rather than objective measures, the latter of which involve acoustic comparison of the clean versus processed noisy speech. Although objective measures are far more efficient than human-subjects testing, and therefore invaluable for guiding the design of various model architectures, these values are often only loosely related to actual human intelligibility. This is especially true when HI

a)Electronic mail: healy.66@osu.edu
b)Also at: the Division of Communication Sciences and Disorders, West Virginia University, Morgantown, WV 26506, USA.
c)Also at: Meta Reality Labs, Redmond, WA 98052, USA.

TABLE I. Characteristics of various deep-learning models employed to examine human intelligibility. SE, Speech Enhancement; SS, Speaker Separation; SSN, Speech-Shaped Noise; SMN, Speech-Modulated Noise.

	Fundamental Requirements							Characteristics					
	Noise Segment Independent	Noise Independent	Nonstationary Environmental Noise Recordings	Talker Independent	Corpus/Recording Channel Independent	Causal	Small Network	Interference Plus Room Reverberation	Cross Language	Complex Domain	Task	Listener Type	Noises Employed
Healy *et al.* (2013)	NO	NO	YES	NO	NO	NO	NO	NO	NO	NO	SE	NH/HI	SSN, Babble
Healy *et al.* (2015)	YES	NO	YES	NO	NO	NO	NO	NO	NO	NO	SE	NH/HI	Babble, Cafeteria
Chen *et al.* (2016b)	—	YES	YES	NO	NO	NO	NO	NO	NO	NO	SE	NH/HI	Babble, Cafeteria
Goehring *et al.* (2016)	YES	NO	YES	NO	NO	YES	YES	NO	NO	NO	SE	HI	SSN, Babble
Goehring *et al.* (2017)	YES	NO	YES	YES	YES	YES	YES	NO	NO	NO	SE	CI	Babble, Traffic
Healy *et al.* (2017)	—	—	—	NO	NO	NO	NO	NO	NO	NO	SS	NH/HI	Interfering Speech
Monaghan *et al.* (2017)	YES	NO	YES	NO	NO	YES	YES	NO	NO	NO	SE	NH/HI	SSN, Babble
Bensten *et al.* (2018)	YES	NO	YES	NO	NO	YES	YES	NO	NO	NO	SE	NH	Babble
Bramsløw *et al.* (2018)	—	—	—	NO	NO	YES	NO [a]	NO	NO	NO	SS	NH/HI	Interfering Speech
Lai *et al.* (2018)	—	YES	YES	NO	NO	NO	NO	NO	NO	NO	SE	CI	Two-talker, Construction
Zhao *et al.* (2018)	YES	NO	YES	NO	NO	NO	NO	YES	NO	NO	SE	NH/HI	SSN, Babble
Goehring *et al.* (2019)	YES [b]	NO [b]	YES	YES	YES	YES	YES	NO	NO	NO	SE	NH/CI	SSN, SMN, Babble
Healy *et al.*, 2019	—	—	—	NO	NO	NO	NO	YES	NO	NO	SS	NH/HI	Interfering Speech
Keshavarzi *et al.* (2019)[c]	YES	NO	YES	YES	YES	YES	YES	NO	NO	NO	SE	NH/HI	Babble
Healy *et al.* (2020)	—	—	—	YES	YES	NO	NO	YES	NO	YES	SS	NH/HI	Interfering Speech
Healy *et al.* (2021a)	—	—	—	YES	YES	NO	NO	YES	YES	YES	SS	NH	Interfering Speech
Healy *et al.* (2021c)	—	YES	YES	NO	NO	YES[d]	NO	NO	NO	YES	SE	NH/HI	Babble, Cafeteria
Healy *et al.* (2021b)	—	—	—	YES	YES	YES	NO	YES	NO	YES	SS	NH/HI	Interfering Speech
Li *et al.* (2021)	—	YES	YES	NO	NO	NO	YES	NO	NO	NO	SE	CI	Two-talker, Construction
Current Study	—	YES	YES	YES	YES	YES	NO	NO	NO	YES	SE	NH/HI	SSN, Babble

[a]Low latency model.
[b]Different noise recordings of the same type (babble or traffic) used for training and test.
[c]Subject preference assessed.
[d]Effectively causal.

https://doi.org/10.1121/10.0019341

(or CI) listeners are involved or when complex acoustic backgrounds are employed (see, e.g., Gustafsson *et al.*, 1998; Zhao *et al.*, 2018). Other more technical reviews of deep-learning-based speech enhancement and related systems that are less focused on human intelligibility assessments may be found elsewhere (e.g., Wang and Chen, 2018; Nossier *et al.*, 2020; Ochieng, 2022).

The implementation of deep-learning-based noise reduction into hearing technology can be divided into two broad considerations: efficacy and viability. The former refers to the ability of an algorithm to improve intelligibility for a wide variety of listeners (particularly HI), across a wide variety of acoustic environments. The latter involves the ability of an algorithm to operate in real time on an actual device. So, whereas the former consideration involves the question, "can it work in principle?" the latter involves the question, "can it work in practice?"

A. Efficacy considerations

The question, "can it work in principle?" necessarily means, "does the algorithm increase target-speech intelligibility across a large variety of environmental conditions for a wide variety of HI listeners?" The key to efficacy is the ability of an algorithm to generalize to conditions not encountered during network training—to tolerate a training-test mismatch. This is critical because it is obviously impossible to train a network on all conditions that will be encountered by a listener in the real world. Models that are not trained using sufficiently varied noises and speech will tend to overfit the training conditions and fail to generalize.

Table I shows the progression that has occurred with regard to generalization. If a study included any conditions that satisfy the listed requirement, then the requirement is marked "YES." On a positive note, these works largely employed actual recordings of environmental sounds and a range of signal-to-noise ratios (SNRs). They also generally employed sentence materials and different sentences for training versus tests were always employed. However, other aspects of generalization were more challenging. As a result, these aspects formed the focus of systematic study.

The first aspect to be addressed involved noise independence—the use of different noise types and recordings for network training versus test. As mentioned previously, the initial introduction of deep-learning-based noise reduction (Healy *et al.*, 2013) was highly constrained in this aspect, using the same brief noise recordings for both training and test. That network was even trained on speech mixed with noise at the same SNR as that used for the test mixtures. This highly matched approach was taken for a reason: The earliest implementations of deep learning used far more rudimentary techniques than those employed today (ideal binary mask, perceptrons and restricted Boltzmann machines, sub-band classifiers, noisy phase, etc.), and large intelligibility benefit for human listeners would likely not have been achieved had more mismatched training versus test conditions been employed.

The first step taken toward noise independence involved the use of different segments of the same noise recording for training versus test (e.g., an 8-min segment for training and a different 2-min segment for test). Table I shows that noise-segment independence was adopted early on and, once achieved, was retained in subsequent studies (Healy *et al.*, 2015; Goehring *et al.*, 2016; Goehring *et al.*, 2017; Monaghan *et al.*, 2017; Bensten *et al.*, 2018; Zhao *et al.*, 2018; Goehring *et al.*, 2019; Keshavarzi *et al.*, 2019).

The next step involved full noise independence—the use of entirely different noise recordings and types for training versus test. Successful increases in HI intelligibility were achieved using a noise-independent model by Chen *et al.* (2016), by training on a database of 10 000 different noises, then testing using noises not in the training set. The noise types in this study were also expanded to be more similar to those in the real world—in addition to multitalker babble, an environmental recording containing multiple sound sources was used (cafeteria noise, which contains speech babble, impact sounds from dishes, etc.). The data-driven approach of large-scale training allows the training set to have sufficient variability so that the network learns the concept of "noise" more generally and is not tied to any specific examples of it. However, despite this early success and the large intelligibility benefit observed, this aspect of generalization is challenging and, as Table I shows, was not adopted routinely in subsequent papers until several years later. Lai *et al.* (2018), Healy *et al.* (2021c), and Li *et al.* (2021) employed fully noise-independent noise-reduction models.

Another primary generalization aspect involves talker independence (or speaker independence). Whereas the early studies listed previously tended to focus on noise independence rather than talker independence, Goehring *et al.* (2017), Goehring *et al.* (2019), and Keshavarzi *et al.* (2019) adopted the opposite approach. Using models that were talker independent, but not noise independent, Goehring *et al.* observed improved intelligibility in noise for CI users, and Keshavarzi *et al.* showed that HI listeners displayed a preference for processed speech.

In addition to background noise, individuals with hearing impairment also have substantial difficulty understanding speech in the presence of an interfering talker. In Healy *et al.* (2017), a DNN was trained to separate a target talker from a single interfering talker of opposite sex,[2] and the target sentence was passed along to the listener (speaker separation). As Table I shows, Healy *et al.* (2020), Healy *et al.* (2021a), and Healy *et al.* (2021b) introduced talker independence to speaker separation. The approach taken in these studies to accomplish talker independence is similar to that employed for noise—by training using many talkers, the network learns what "speech" is more generally and not tied to any particular talker. Because different speech corpora are typically used for training and testing talker-independent networks, these models are also speech-corpus or recording-channel independent (see Sec. IV).

Another corruption common to everyday acoustic environments involves room reverberation. Room reverberation

can be highly detrimental to speech perception, especially for HI listeners, and especially when combined with background noise or interfering speech. Zhao *et al.* (2018) demonstrated the ability of a deep learning algorithm to improve intelligibility for HI listeners in the presence of background noise and concurrent room reverberation. This initial study was noise-segment independent, but not talker independent. This work was extended to speaker separation involving interfering speech and concurrent room reverberation using talker dependent (Healy *et al.*, 2019) and talker-independent models (Healy *et al.*, 2020).

Recently, the concept of generalization was extended to perhaps a limit, by examining cross-language generalization (Healy *et al.*, 2021a). A talker-independent DNN was used to isolate a target talker from an interfering talker of the opposite sex and simultaneously remove large amounts of room reverberation. Training was performed using English speech materials, but testing was performed using Mandarin speech materials. These two languages were selected based on their high prevalence of speakers and lack of known common ancestry, the latter of which produces large linguistic differences. The observation that the benefit in cross-language conditions was comparable to that observed in within-language conditions reflects the vast capability of modern networks to generalize to acoustic conditions not encountered during training.

B. Viability considerations

The question, "can it work in practice?" necessarily means, "can the trained network be implemented into hearing technology and operate in real-time (while retaining the ability to produce large intelligibility benefit)?" There is only one fundamental requirement for viability—causality—that a network operate on only past and present time frames and not delay output in order to take advantage of future time frames.[3]

There are two approaches that can be taken with regard to network design. One first emphasizes efficacy, whereas the other first emphasizes viability. The advantages and disadvantages of these approaches are discussed in Sec. IV. Table I shows the different approaches. The earliest studies emphasized efficacy (e.g., Healy *et al.*, 2013; Healy *et al.*, 2017; Chen *et al.*, 2016). Accordingly, benefit for human HI listeners remained high in all conditions as progressive steps were taken toward generalization (in those studies, noise independence). Other early studies emphasized viability. The networks of Goehring *et al.* (2016), Monaghan *et al.* (2017), and Bensten *et al.* (2018) were causal to be real-time capable and smaller to produce less burden on hardware. Improved intelligibility in noise for HI or NH listeners was observed in most but not all conditions. Keshavarzi *et al.* (2019) also employed a small and causal model. Goehring *et al.* (2017) and Goehring *et al.* (2019) extended this work to CI users and showed that a small and causal model could improve speech-reception thresholds in noise in most but not all conditions. The network of Bramsløw *et al.* (2018) was causal and low latency, but not small in size. These authors observed large intelligibility increases in a speaker-separation task involving HI listeners. The network of Healy *et al.* (2021b) was also causal.

"Effectively causal" (Healy *et al.*, 2021c) involves the use of future time frames that produce delays below the human detection or disruption threshold, which may be used without hindrance to the listener (Stone and Moore, 1999; 2005; Goehring *et al.*, 2018) but with potential benefit to the network. These delays may be introduced so long as they do not increase the overall latency of the system beyond the human tolerance thresholds. Using 20 ms of future information (and a large complex-domain network), Healy *et al.* (2021c) observed significant intelligibility increases in all effectively causal conditions.

C. The current study

In the current study, HI and NH intelligibility benefit resulting from a current deep-learning noise-reduction algorithm was assessed. Unlike prior models, the current network was free of the constraints of the past and performed extensive generalization (see Table I). The network was trained using 10 000 naturally occurring noises and tested on entirely different noises (noise-independent). It was also trained using over 2000 different talkers and tested using a talker not included in this set and from an entirely different corpus (talker and corpus independent). Further, the current model employed only past and the current time frame, allowing it to be fully causal as required for real-time operation.

Once this assessment was completed, the results were compared to those associated with the initial study (Healy *et al.*, 2013). The comparison model is both the first intelligibility demonstration of deep-learning-based noise reduction and highly constrained. To allow direct comparison, the test-speech recordings, test-noise types, some SNRs, test procedures, and listener populations were the same across studies. In sharp contrast to the current network, the initial model employed the same talker and sentence corpus for training and test (talker and corpus dependent) and the same brief noise segments for training and test (noise dependent). Also, the initial algorithm employed an analysis window involving both past and future time frames (non-causal).

The overall goals of the current study were to establish the intelligibility benefit for HI listeners resulting from a state-of-the-art network and to establish the benefit that was lost/gained across the decade-long span of steps taken toward the creation of this constraint-free, real-time capable deep-learning-based noise-reduction algorithm for hearing technology.

II. METHOD

A. Subjects

Two groups of listeners participated. The HI group consisted of 12 listeners, representing typical hearing aid users with bilateral sensorineural hearing loss. All were binaural hearing aid users, and seven were female. These listeners

https://doi.org/10.1121/10.0019341

were recruited from The Ohio State University Speech-Language-Hearing Clinic and ranged in age from 20 to 85 years (mean = 57). Pure-tone audiometry (ANSI, 2004, 2010) was used to verify hearing losses on the day of the test. In accord with our desire to recruit representative patients, the degree of hearing loss varied across frequencies and listeners, who generally had sloping hearing losses that ranged in degree from mild to profound. Pure-tone average audiometric thresholds (PTAs) (means across thresholds at 500, 1000, and 2000 Hz across ears) ranged from 27 to 76 dB hearing level (HL) with a mean of 56 dB HL. Three of the listeners had audiometric thresholds within normal limits (20 dB HL or lower) for at least one frequency in at least one ear (see dotted horizontal line in Fig. 1), but otherwise, all thresholds were elevated in both ears for all audiometric frequencies. Sensorineural loss was defined as air-bone gaps of 10 dB or less at 500, 1000, 2000, and 4000 Hz, bilaterally. Although not selected based on the symmetry of hearing losses, only one subject (HI5) had an asymmetric loss (15 dB or greater between ears at three or more contiguous frequencies). No subjects were excluded based on results of cognitive testing. Figure 1 displays audiograms for each of these listeners, who are numbered from HI1 to HI12 in order of ascending PTA.

The NH group consisted of 12 listeners (all female) with 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, 2010). The exception was NH10, whose threshold at 8000 Hz in the right ear was 25 dB HL. Recruited from undergraduate courses at The Ohio State University, they ranged in age from 19 to 26 years (mean = 21) and represented young listeners with "ideal" hearing abilities. All participants (HI and NH) were native speakers of American English having no previous exposure to the test sentences used in the current study, and all received either course credit or a monetary incentive for participating.

B. Stimuli

Target sentences for evaluation purposes were drawn from the standard recordings of the Hearing in Noise Test (HINT, Nilsson *et al.*, 1994) and were produced by a male talker in general American English. Speech-shaped noise and multi-talker babble were used as background noises for

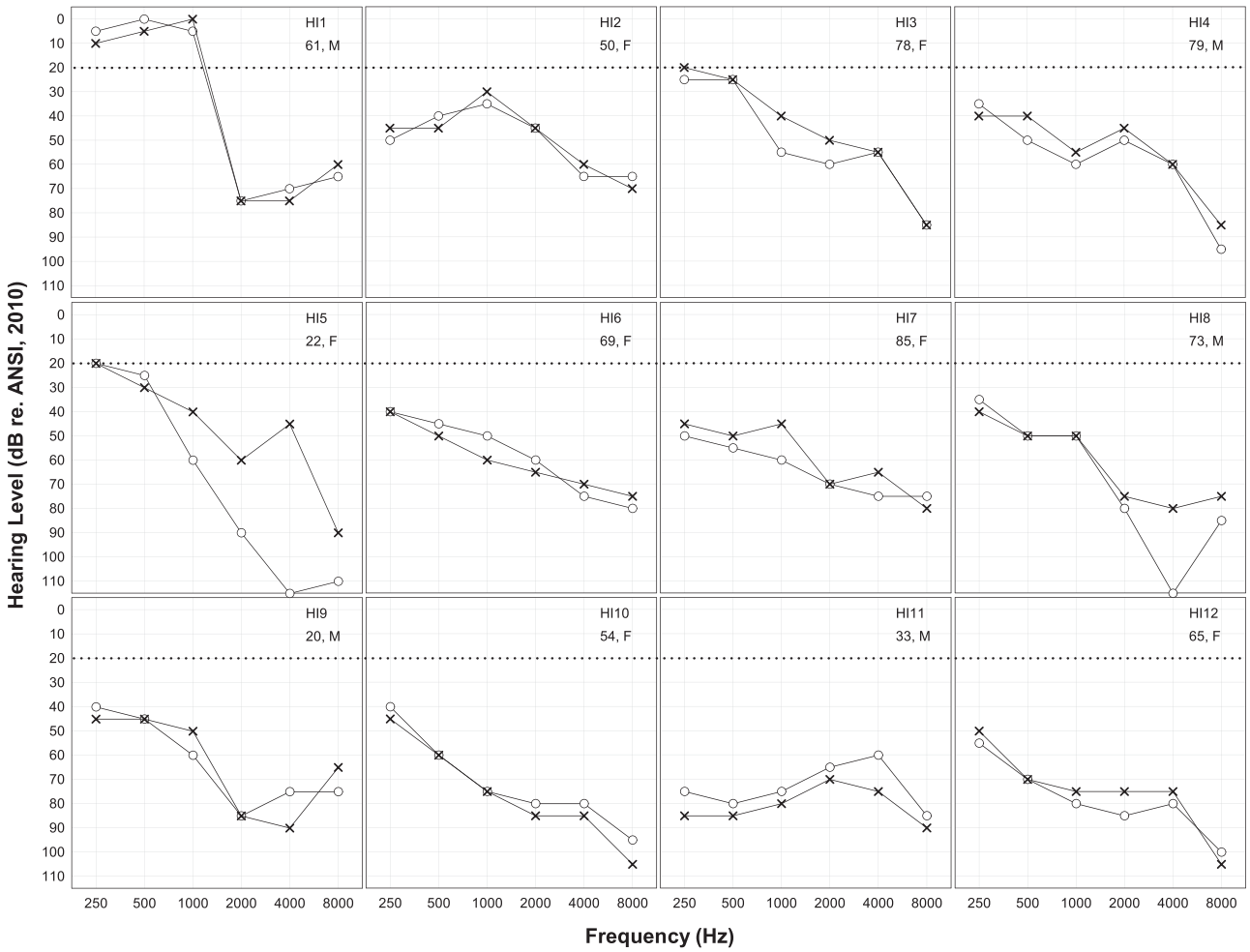

FIG. 1. Pure-tone air-conduction audiograms for the HI listeners. Circles represent right ears and ×'s represent left ears. The horizontal dotted line in each panel represents the NH limit of 20 dB HL. Listener ages and sexes are also listed.

the evaluation. The babble was a standard 9 min 52 s recording from auditec (http://www.auditec.com). The HI listeners were tested at –2 and –5 dB SNR in SSN and at 0 and –2 dB SNR in babble. The NH listeners were tested at –2 and –5 dB SNR in both SSN and babble. Stimuli were presented as unprocessed noisy sentences and as algorithm-processed (enhanced) versions of noisy sentences.

The algorithm was trained using speech materials from the LibriSpeech corpus (Panayotov *et al.*, 2015). LibriSpeech is a corpus of approximately 1000 h of speech from more than 2000 speakers. It is primarily used for research on large-vocabulary continuous speech recognition systems. The data in LibriSpeech are derived from the LibriVox project (Kearns, 2014), which contains audiobook recordings created using volunteers from across the globe. Pandey and Wang (2020a,b) found LibriSpeech to be highly effective for training corpus-independent speech enhancement algorithms. Since LibriSpeech is recorded by thousands of volunteers in diverse environments, recording conditions vary considerably within the dataset, which is a key to avoiding overfitting corpus-specific characteristics, such as recording microphones and room acoustics.

The training noises consisted of 10 000 nonspeech sounds from a sound-effect library (Richmond Hill, ON, Canada[4]). Pairs of clean and noisy speech were created during training by randomly selecting an utterance, a noise segment, and an SNR from { –5, –4, –3, –2, –1, 0} dB. A set of 150 validation mixtures was generated using utterances from 6 speakers in the Wall Street Journal Corpus (WSJ0; Paul and Baker, 1992) and a factory noise from the NOISEX dataset (Varga and Steeneken, 1993).

C. Algorithm description

Given a speech signal $\boldsymbol{s}$ and a noise signal $\boldsymbol{n}$ with N samples, a noisy speech signal $\boldsymbol{y}$ is modeled as

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n}, \tag{1}$$

where $\{y, s, n\} \in \mathbb{R}^{1\times N}$. A speech enhancement algorithm is concerned with improving the intelligibility and quality of noisy speech $\boldsymbol{y}$ by obtaining a good estimate of $\boldsymbol{s}$ (i.e., $\hat{\boldsymbol{s}}$) from $\boldsymbol{y}$. The current model was a time-domain algorithm. For such an algorithm, the input feature is the time-domain signal $\boldsymbol{y}$ instead of a time-frequency representation, such as short-time Fourier transform (STFT), and the estimated output is the time-domain signal $\hat{\boldsymbol{s}}$. This approach alleviates the need to perform fast Fourier transformations and then the inverse transforms to return to the time domain.

In the current algorithm, an attentive recurrent network (ARN) was employed for mapping a noisy waveform to an enhanced waveform (Pandey and Wang, 2022). The ARN accepts speech-plus-noise input and is a monaural (single-microphone) system. The algorithm is made causal by restricting the use of time-frame information to the current and past frames. A block diagram of the ARN for time-domain speech enhancement is shown in Fig. 2. Details on the ARN, including comparisons across various implementations and comparisons to other approaches, may be found in Pandey and Wang (2022).

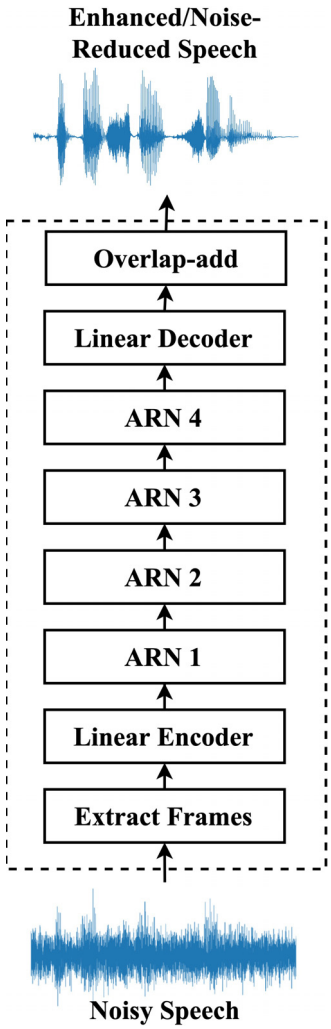

FIG. 2. (Color online) The employed ARN for time-domain speech enhancement.

We note that the LibriSpeech training utterances are very different from the HINT sentences used for testing, and the 10 000 noises used for training also differed from those used for testing. The development of a corpus-independent DNN in low SNR conditions has been found to be particularly challenging due to the recently revealed cross-corpus generalization issue in DNNs (Pandey and Wang, 2020b).

The current algorithm represents a large improvement in terms of speech enhancement problem formulation. In the initial study, speech enhancement was formulated as magnitude-only enhancement and trained subband DNNs were used to estimate the ideal binary mask (IBM). The goal of the algorithm was to obtain an accurate estimate of the IBM and not the clean speech. This early formulation requires the use of noisy phase for reconstruction but also leads to limited magnitude enhancement. The current study formulates speech enhancement in the time domain, where the goal is to directly estimate the enhanced waveform from the noisy waveform, and as a result, the magnitude and the

https://doi.org/10.1121/10.0019341

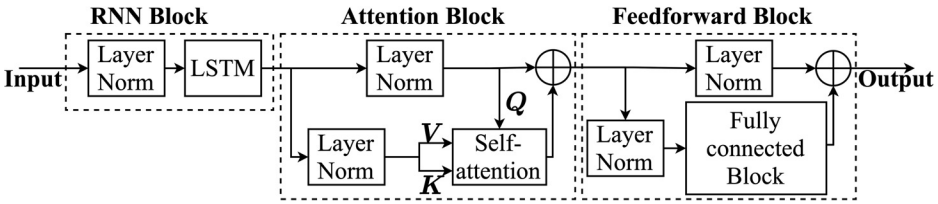

FIG. 3. The components of an ARN block. It is composed of an RNN block, an attention block, and a feedforward block. Layer Norm denotes layer normalization and $\oplus$ is the elementwise addition operator.

phase are jointly enhanced. In the case of ideal estimation, a time-domain algorithm will output the clean speech, whereas the IBM will output intelligible speech but with variable quality. Moreover, a time-domain algorithm does not require feature extraction at the input or waveform reconstruction at the output. Healy *et al.* (2013) used a set of complementary features (from Wang *et al.*, 2013) at the input and performed waveform reconstruction at the output using gammatone filterbanks.

The current algorithm also represents an advance in terms of model architecture compared to the initial study, which employed multilayer perceptrons with pretraining based on restricted Boltzmann machines (see Wang and Wang, 2013). The current algorithm employs modern DNN building blocks suitable for effective sequential processing with contextual information, such as long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997) and self-attention (Vaswani *et al.*, 2017). More specifically, the ARN algorithm combines the power of recurrent neural networks (RNNs) and self-attention for processing a sequence of noisy speech frames. The RNN is naturally suited for sequential modeling, where a sequence of input frames is processed one by one in temporal order. The output for a given input frame in the sequence is a function of the input frame as well as RNN hidden states that encode past input frames. However, RNNs suffer from the problem of exploding and vanishing gradients and have difficulty encoding longer-term correlations. The LSTM used in the ARN is a form of RNN that alleviates this problem to some extent. The mechanism of self-attention, on the other hand, uses all the input frames in a sequence to estimate the output of a given input frame by correlating the given frame with all the input frames and utilizing (attending to) the input frames that are most helpful for the estimation. It was recently found to be superior to RNN for a plethora of sequential modeling tasks. However, self-attention does not exploit sequential order, i.e., it does not care about the temporal order of input frames. The ARN is designed to leverage the capabilities of RNN sequential processing and self-attention's longer contextual modeling.

In the ARN currently employed for speech enhancement, a noisy input $\boldsymbol{y}$ is first segmented into overlapping frames using a frame size of L samples and a frame shift of H samples to get $\boldsymbol{Y} \in \mathbb{R}^{T\times L}$, where T is the number of frames. Next, all the frames are projected to a higher dimension of size D using a linear encoder, which serves the purpose of dimension expansion. The output from the encoder is processed using a stack of four ARN blocks. The output from the final ARN block is projected back to the original frame size of L using a linear decoder at the output. Finally, overlap-and-add is applied to the sequence of enhanced frames to obtain the enhanced waveform.

The building blocks of the ARN are shown in Fig. 3. The ARN is composed of an RNN block, an attention block, and a feedforward block. The input to the RNN block is first normalized using a layer normalization (Ba *et al.*, 2016) and then processed using an LSTM RNN. Layer normalization is used for faster training convergence and improved generalization. LSTM is used to model the temporal dependency between the sequence of frames in a causal fashion. The input and the output of the RNN block are of size $T \times D$.

The RNN block is followed by the attention block. The input to the attention block is normalized using two separate layer normalizations having different scale and bias parameters. The first output is used as query ($\boldsymbol{Q}$), and the second output is used as key ($\boldsymbol{K}$) and value ($\boldsymbol{V}$) for a following self-attention mechanism, where $\{\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\} \in \mathbb{R}^{T\times D}$.

A schematic of the self-attention mechanism is shown in Fig. 4. It involves three trainable vectors $\{\boldsymbol{q}, \boldsymbol{k}, \boldsymbol{v}\} \in \mathbb{R}^{1\times D}$. The rows in $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are refined using the following gating mechanism:

$$\boldsymbol{K}' = \boldsymbol{K} \odot \sigma(\boldsymbol{k}), \tag{2}$$

$$\boldsymbol{Q}' = \mathrm{Lin}(\boldsymbol{Q}) \odot \sigma(\boldsymbol{q}), \tag{3}$$

$$\boldsymbol{V}' = \boldsymbol{V} \odot [\sigma(\mathrm{Lin}(\boldsymbol{v})) \odot \mathrm{Tanh}(\mathrm{Lin}(\boldsymbol{v}))], \tag{4}$$

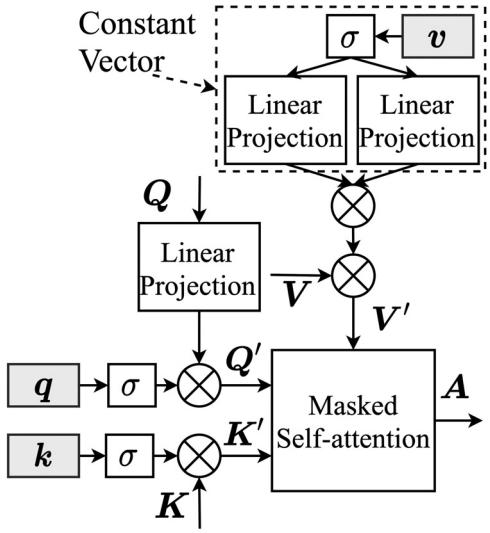

FIG. 4. The self-attention mechanism inside the attention block. The inputs to self-attention are **Q**, **K** and **V**, and the final output is **A**. Vectors **q**, **k**, **v** and parameters of linear projections are trainable.

where σ is the sigmoidal nonlinearity, Lin is a linear layer and $\odot$ is elementwise multiplication. Before the elementwise multiplication, vectors $\boldsymbol{q}$, $\boldsymbol{k}$, and $\boldsymbol{v}$ are broadcast to match with the shape of matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, respectively. Given that $\boldsymbol{v}$ is a fixed vector, $\sigma(\text{Lin}(\boldsymbol{v})) \odot \text{Tanh}(\text{Lin}(\boldsymbol{v}))$ is also a fixed vector. This operation is only required at training time for the optimization of $\boldsymbol{v}$ (Merity, 2019). A precomputed constant vector from the best model after training is used during evaluation.

The final output of the attention block, $\boldsymbol{A} \in \mathbb{R}^{T\times D}$, is computed using the following set of equations:

$$W = \frac{\boldsymbol{Q}'\boldsymbol{K}'^{\mathbb{T}}}{\sqrt{D}}, \tag{5}$$

$$\boldsymbol{W}' = \text{Mask}(\boldsymbol{W}), \tag{6}$$

$$W'(i,j) = \begin{cases} W(i,j) & \text{if} \quad i \le j \\ -\infty & \text{otherwise}, \end{cases} \tag{7}$$

$$\boldsymbol{P} = \text{Softmax}(\boldsymbol{W}'), \tag{8}$$

$$\text{Softmax}(W)(i,j) = \frac{e^{W(i,j)}}{\sum_{j=1}^{T} e^{W(i,j)}}, \tag{9}$$

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{V}', \tag{10}$$

where $\mathbb{T}$ is the transpose operator.

First, correlation scores between pairs of rows in $\boldsymbol{Q}'$ and $\boldsymbol{K}'$, $\{\boldsymbol{Q}'_i, \boldsymbol{K}'_j\}$, where $i, j \in \{1, \ldots, T\}$, are computed using Eq. (5). Next, correlation scores of future frames are masked or ignored by using a mask operator defined in Eq. (7), which sets the correlation scores of future frames to $-\infty$. Then, the softmax operator in Eq. (9) is applied to convert correlation scores to probability values $\boldsymbol{P} \in \mathbb{R}^{T\times D}$. The softmax operator uses exponential followed by summation in the denominator. The exponential converts a $-\infty$ (from mask) to 0, and hence, the contribution of future frames in the total sum of denominator becomes zero, which makes the algorithm fully causal. Finally, the attention output, $\boldsymbol{A} \in \mathbb{R}^{T\times D}$, is computed using Eq. (10). The final output from the attention block is obtained by adding $\boldsymbol{A}$ to $\boldsymbol{Q}$, which provides a residual connection for improved gradient flow during training (He *et al.*, 2016).

The output from the attention block is processed using a feedforward block. The feedforward block provides additional representation power to the preceding attention block (Vaswani *et al.*, 2017). The input to the feedforward block is normalized using two separate layer normalizations. The first normalized input is processed using a fully connected block shown in Fig. 5. In the fully connected block, a linear layer is first used to project its input of size D to a higher dimension of size $4D$, which is followed by Gaussian error linear unit (GELU) nonlinearity (Hendrycks and Gimpel, 2016) and dropout. The output of size $4D$ is then collapsed to size D by splitting it into four different vectors and adding them together. Finally, the collapsed output is added to the second normalized input to get the final output of the feedforward block.

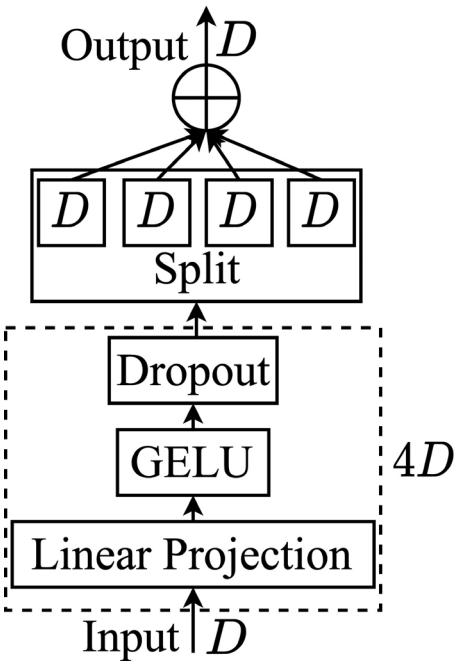

FIG. 5. The fully connected block inside the feedforward ARN block.

All stimuli were resampled to 16 kHz for processing. Prior to mixing, the target (clean) speech was scaled to achieve the desired SNR. Next, the input to the network was root-mean-square (RMS) normalized. A frame size of 20 ms ($L = 320$) and a frame shift of 2 ms ($H = 32$) were used. The use of a smaller frame shift was inspired by earlier studies on corpus-independent speech enhancement (Pandey and Wang, 2020a,b), where a smaller frame shift led to improved generalization on untrained corpora. The RNN block used LSTM with a hidden size of 1024. The parameter D was set to 1024. A dropout of 5% was used in the fully connected block.

The ARN was trained for 100 epochs with a batch size of 32 utterances. The pairs of clean and noisy utterances were dynamically generated during training by adding random segments of speech to random segments of noise. All the utterances within a batch were either truncated or padded with zeros to have length of 4 s.

The Adam optimizer was used for training (Kingma and Ba, 2014). The learning rate was set to 0.0002 for the first 33 epochs after which it is exponentially decayed every epoch using a scale that resulted in a learning rate of 0.00002 in the final epoch. All the models were developed in PyTorch. Mixed precision training was used to increase efficiency (Micikevicius *et al.*, 2017). The ARN was trained using two Nvidia V100 32GB GPUs, with each batch distributed over two GPUs using PyTorch's DataParallel module.

Figure 6 displays waveforms and spectrograms for clean, noisy, and algorithm-processed versions of a HINT sentence used in the present study.

https://doi.org/10.1121/10.0019341

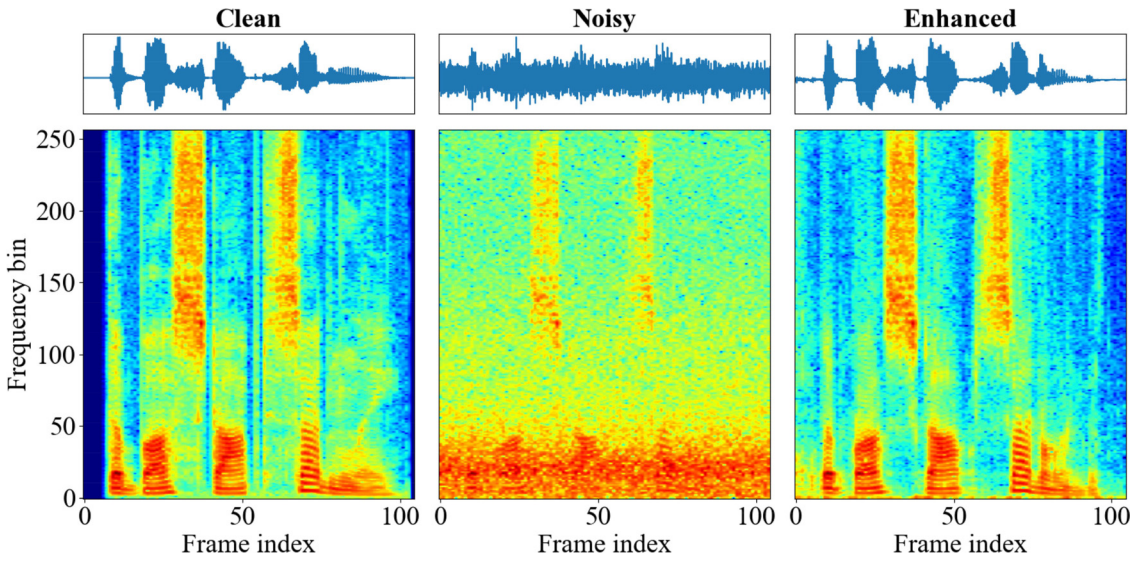

FIG. 6. (Color online) Enhancement of a HINT utterance corrupted by speech-shaped noise at –5 dB SNR using the employed ARN.

D. Procedure

There were eight conditions for each listener (two processing conditions × two noise types × two SNRs). Again, processing conditions were unprocessed speech-in-noise and algorithm processed speech-in-noise, and the noise types were SSN and babble. SNRs for HI were –2 and –5 dB in SSN and 0 and –2 dB in babble. SNRs for NH listeners were –2 and –5 dB in both noise types. Each listener heard 160 sentences, blocked by condition, with 20 sentences in each block. For each combination of noise type and SNR, unprocessed and processed conditions were presented juxtaposed. The presentation order of the four noise type-SNR combinations was randomized for each listener, as was the order of the two processing conditions within each noise type-SNR block. By presenting the sentences in the same fixed order to all listeners, but randomizing conditions, the correspondence between the sentence list and condition was random. No sentence was used more than once for any listener.

The stimuli were played back from a Windows PC using an RME Fireface UCX digital-to-analog converter (RME, Haimhausen, Germany), through a Mackie 1202-VLZ mixer (Mackie, Woodinville, WA), and presented diotically using Sennheiser HD 280 Pro headphones (Sennheiser, 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 using the NAL-RP hearing-aid fitting formula (Byrne *et al.*, 1990). This formula does not prescribe gains at 125 or 8000 Hz, and so the gains applied to 250 and 6000 Hz, respectively, were also applied to these two most extreme standard audiometric frequencies. These gains were implemented using a RANE DEQ 60 L digital equalizer (RANE, Mukilteo, WA), as described in Healy *et al.* (2015). Accordingly, these listeners were tested without their hearing aids.

Twenty-five practice stimuli were presented to each listener prior to formal testing, consisting of five stimuli in each of the following five conditions: (1) sentences in quiet, (2) processed sentences in babble at the higher of the two SNRs for each listener group, (3) processed sentences in SSN at the lower SNR for each listener group, (4) unprocessed sentences in babble at the higher SNR, and (5) unprocessed sentences in SSN at the lower SNR. During this familiarization, listeners were instructed to repeat back each sentence as best they could and guess if unsure of the content of the sentence. Hearing-impaired listeners were also asked to report about the loudness of the signals. All but two reported that the stimuli sounded audible and comfortable. HI10 reported that they sounded comfortable after reducing the presentation level by 5 dB. HI12 reported that the stimuli sounded comfortable and audible after a 10-dB reduction in presentation level. The final presentation level for the HI listeners ranged from 81.0 to 98.6 dBA (mean = 91.2).

Listeners then heard the 160 test stimuli while seated in a double-walled sound booth. They were again instructed to repeat back each sentence to their best ability, guessing if unsure. The experimenter controlled the presentation of each stimulus and scored words correctly reported. For a word to be scored as correct, it had to be repeated exactly apart from verb tense (is/was, are/were, and has/had) and article (a/the) variations. The 20 target sentences presented in each condition each contained three to seven words, for a total of 103 to 110 words in each condition. Sentence recognition was expressed as the percentage of words correctly reported, and these percent-correct scores were transformed to rationalized arcsine units (RAUs; Studebaker, 1985) prior to statistical analysis.

III. RESULTS AND DISCUSSION

A. HI listeners

Figures 7 and 8 display intelligibility scores for each individual HI listener in each condition. Results for the SSN

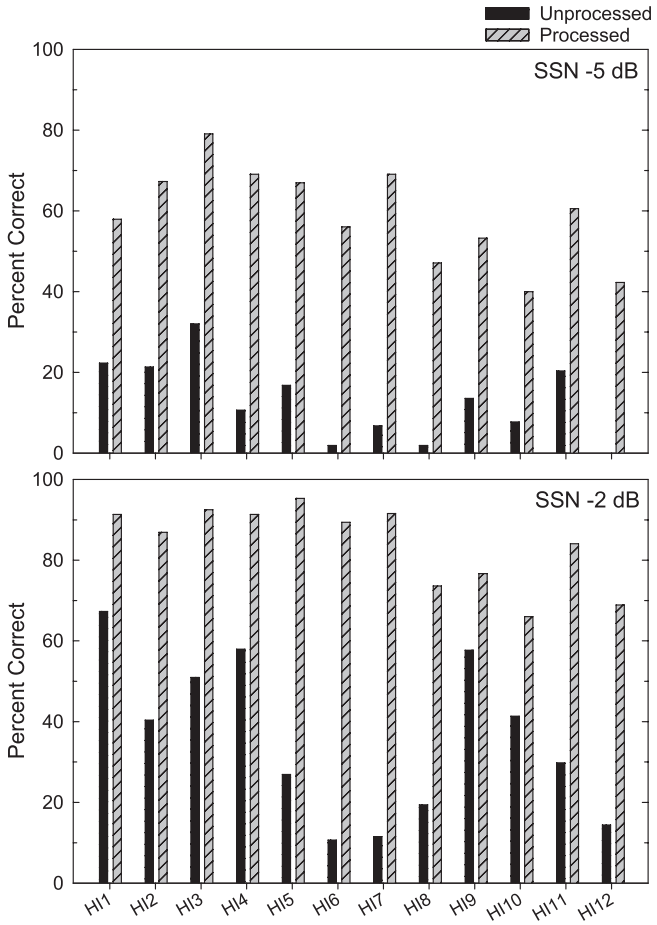

FIG. 7. Sentence intelligibility scores for individual HI listeners. Black columns represent unprocessed speech-in-noise conditions and shaded/hatched columns represent these same conditions following algorithm processing. Algorithm benefit is then represented as the difference between a shaded/hatched column and the solid column directly to its left. The background was speech-shaped noise (SSN) at the SNRs indicated.

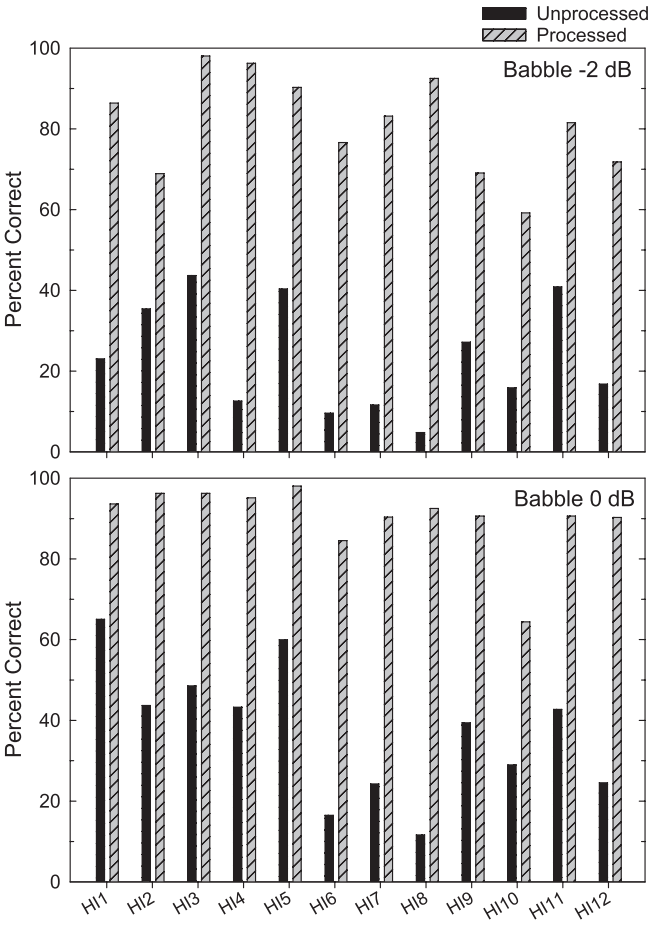

FIG. 8. As Fig. 7, but for individual HI listeners and speech in multitalker babble.

conditions are displayed in Fig. 7, and those for the babble conditions are displayed in Fig. 8. Each panel corresponds to a different SNR, as indicated. The black and shaded/hatched columns represent scores before and after algorithm processing, respectively. The absence of a black column for HI12 in SSN at –5 dB SNR reflects that this subject was unable to correctly report any words in that unprocessed condition. The algorithm benefit for each listener in each condition corresponds to the difference in height between a shaded/hatched column and the black column immediately to the left of it.

Apparent from Fig. 7 is that the algorithm benefitted all HI listeners at both SNRs in SSN. At least half of the HI listeners received benefit exceeding 45 and 50% points for the SNRs of –5 and –2 dB, respectively. Algorithm benefit in SSN exceeded 30% points in 21 out of the 24 cases (12 HI listeners $\times$ 2 SNRs). Apparent from Fig. 8 is that all HI listeners also received benefit at both SNRs in babble, with at least half of them receiving benefit exceeding 50% points at both SNRs. The benefit in babble exceeded 30% points in 23 of the 24 cases.

Planned comparisons consisting of four two-tailed paired *t*-tests on RAUs between unprocessed and processed scores revealed significant algorithm benefit for HI listeners at each noise types and SNR [each $t(11) \geq 7.9$, each p value < 0.0001, each Cohen's $d > 3.02$]. These significant results survive Bonferroni correction.

B. NH listeners

Figures 9 and 10 display intelligibility for the individual NH listeners. Results for the SSN conditions are displayed in Fig. 9, and those for the babble conditions are displayed in Fig. 10. Note that the NH listeners were tested at the same SNRs as the HI listeners in SSN, but overlapped at only one SNR in babble. As anticipated, the performance of the NH listeners exceeded that of the HI listeners in unprocessed conditions. The mean NH scores for unprocessed stimuli were 64% and 86% correct for the two SSN SNRs (–5 and –2 dB) and 57% and 82% correct for the two babble SNRs (also –5 and –2 dB). Accordingly, the algorithm benefit was considerably smaller for the NH than for the HI listeners. However, some benefit was observed in 20 of the 24 cases for SSN and in all 24 of the cases for babble. Planned comparisons consisting of two-tailed paired *t*-tests on RAUs for each of the four conditions of Figs. 9 and 10 revealed

https://doi.org/10.1121/10.0019341

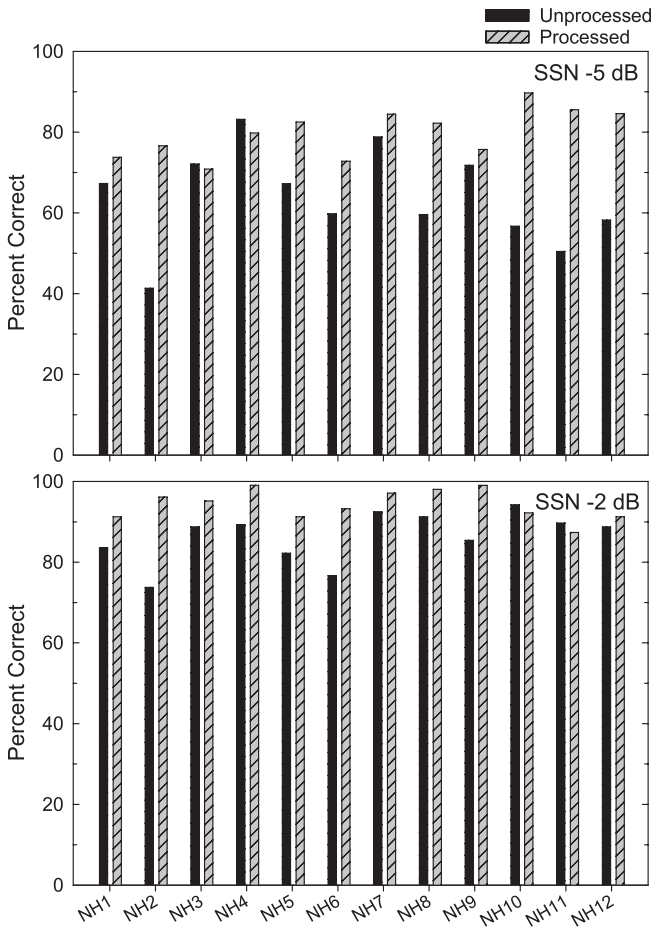

FIG. 9. As Fig. 7, but for individual NH listeners and speech in SSN.

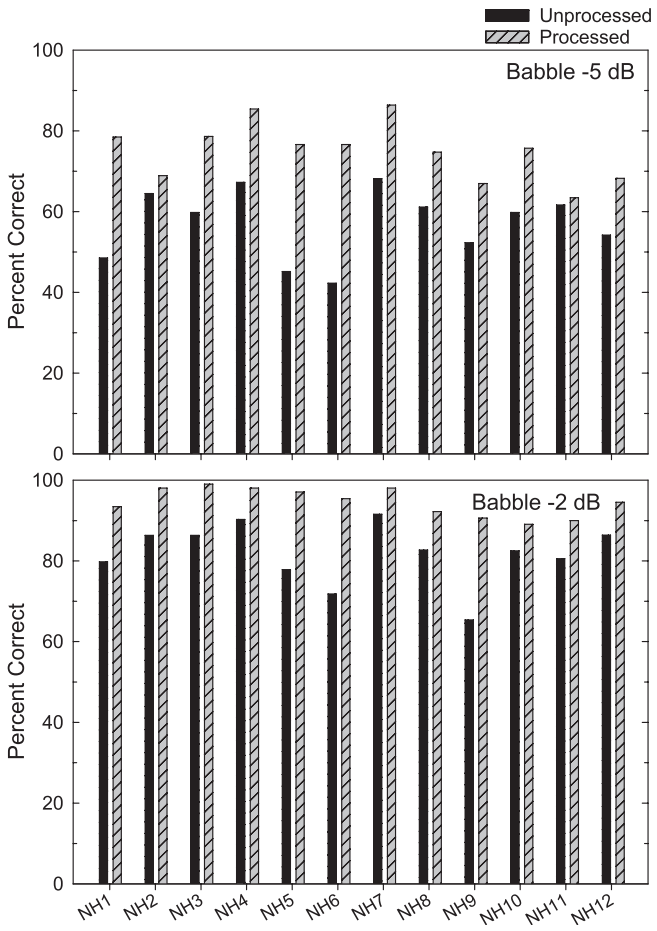

FIG. 10. As Fig. 7, but for individual NH listeners and speech in multitalker babble.

significant benefit [each $t(11) \geq 3.9$, each p value < 0.01, each Cohen's $d > 1.61$]. This set of significant results also survives Bonferroni correction.

C. Comparison between HI and NH listeners

Figure 11 displays group-mean sentence intelligibility scores and standard errors of the mean (SEMs) for both HI and NH listeners, plotted separately, in each condition. Again, noise types and SNRs are plotted in separate panels, black columns represent unprocessed scores, and shaded/hatched represent processed scores. The group-mean algorithm benefit for the HI listeners was 46 and 48% points for SSN at –5 and –2 dB SNR, respectively, and 58 and 53% points for babble at –2 and 0 dB SNR, respectively. When benefit was expressed in RAUs to control for ceiling and floor effects, these values increased slightly to 50 units for both SSN SNRs, and 60 and 57 units in babble. The figure also shows that the manipulation of SNR yielded the desired baseline (unprocessed) scores for the HI listeners (scores free of strong floor and ceiling effects). The mean baseline intelligibilities were 13% and 36% correct for SSN and 24% to 37% correct for babble. For the NH listeners, group-mean benefit values were 16 and 8% points at –5 and –2 dB SNR, respectively, in SSN, and 18 and 13% points at these SNRs in babble. Conversion to RAUs produced no change in benefit at the lower SNR for either noise type and an increase to 13 and 19 units at the higher SNR in SSN and babble respectively.

To address the question of whether the algorithm can restore NH speech-in-noise recognition abilities to these HI listeners, the performance of the HI listeners following algorithm processing was compared to the performance of young NH listeners without processing, in the conditions common to both groups. As Fig. 11 shows, the HI listeners approached within 1% point of the NH listeners' performance in one condition (babble at –2 dB SNR) and within 5% points in the remaining two conditions (SSN at –5 and –2 dB SNR). Although we note that nonsignificant differences do not imply equality, three planned comparisons (two-tailed Welch's independent-samples t-tests on RAUs) between the algorithm-processed scores for the HI listeners and the unprocessed scores for the NH listeners in the three common conditions were conducted. These differences were not significant [each $t \leq 1.1$, each $p > 0.3$, each Cohen's $d < 0.41$, dfs adjusted using the Welch-Satterthwaite method ranged 17–22].

D. Comparison to Healy *et al.* (2013)

To examine the human-subjects performance differences associated with upgrading to the current modern

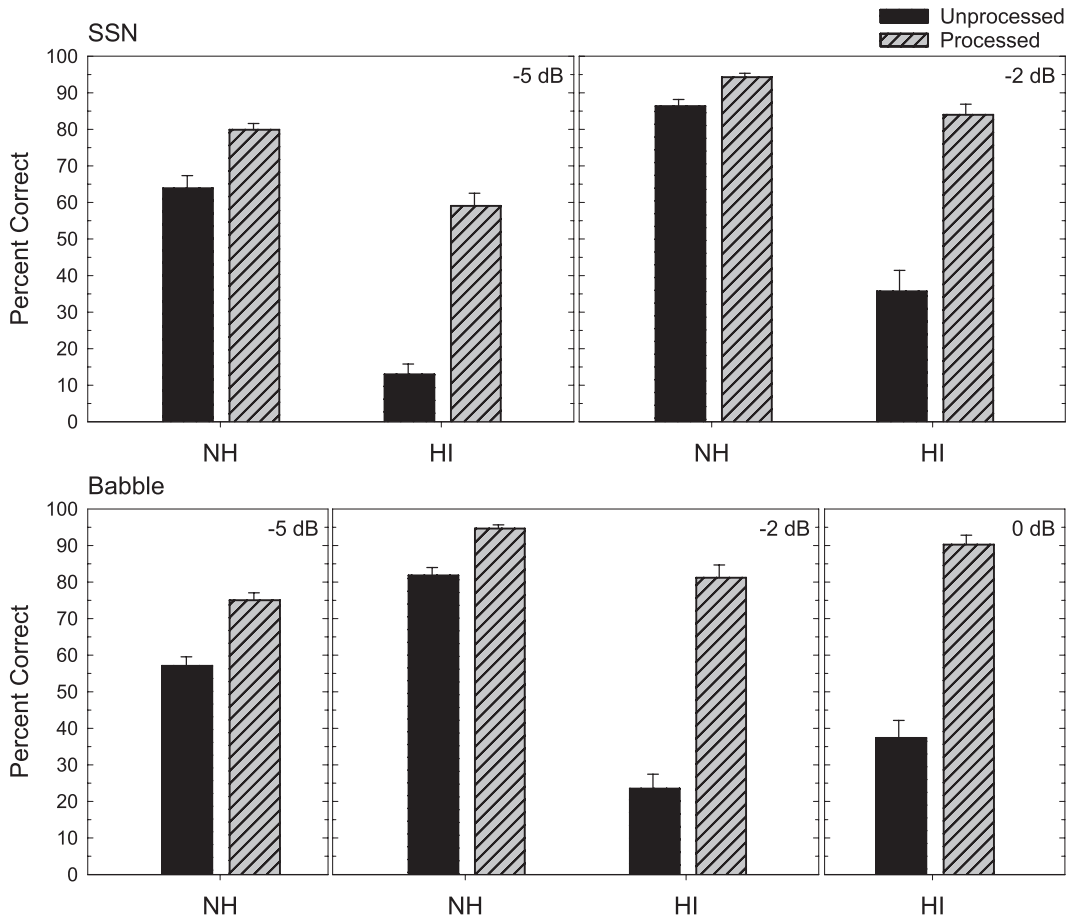

FIG. 11. Group mean (and standard error) sentence intelligibility for the HI and NH listener groups. The background noise type and SNR are indicated, and the HI and NH groups are plotted separately. As in Figs. 7–10, benefit is reflected as the difference between a shaded/hatched column and the black column immediately to its left.

problem formulation and network architecture while substantially increasing the demands placed on the algorithm, the present results were compared with those of Healy *et al.* (2013). The present study was identical to Healy *et al.* (2013) in terms of the speech recordings used, the noise types and (most) SNRs employed, the populations from which subjects were drawn, the numbers of subjects, the testing procedures, and the inclusion criteria for each listener group. The particular noises differed but were of the same type (SSN and babble). The primary difference across studies was the algorithm used for processing and the demands that had to be met.

Again, the 2013 algorithm was both trained and tested using the same talker and 10-s noise segments and it operated on future time frames, meaning it was neither talker, noise, nor corpus independent, nor was it causal. The present algorithm was required to generalize to an untrained talker, from an untrained corpus, and to untrained noise types, as well as being fully causal. Figures 12 and 13 display group-mean sentence intelligibility scores and SEMs for HI and NH listeners in the conditions common to both studies, with SSN in the upper panels and babble in the lower panels of each figure. Pairs of columns labeled "2023" represent the current results and are replotted from Fig. 11, whereas pairs of columns labeled "2013" are from Healy *et al.* (2013). As with the previous figures, benefit is represented as the difference between each unprocessed (solid column) and the corresponding processed score (hatched column). It is noted that baseline (unprocessed) scores differed across studies, likely attributable to the different samples of listeners employed and the use of different noises of the same type (different SSNs and different babbles).

Figure 12 displays scores for the HI listeners. In SSN at –5 dB SNR (top left panel), the group-mean algorithm benefit was the same across studies (46% points). However, group-mean algorithm benefit was considerably higher currently relative to 2013 (48 vs 24% points) at –2 dB SSN (top right panel). For HI listeners in babble, mean benefit was slightly higher in the current study compared to 2013 at both –2 dB SNR (58 vs 55% points, bottom left panel) and 0 dB SNR (53 vs 50% points, bottom right panel).

Figure 13 displays scores for the NH listeners. In SSN, the SNR common across studies was –5 dB. In this condition, group-mean benefit was slightly lower in the current study compared to 2013 (16 vs 17% points, top panel). Normal-hearing benefit was more noticeably lower for the current study than for 2013 in babble at –5 dB SNR (18 vs 35% points, bottom left panel) and at –2 dB SNR (13 vs

https://doi.org/10.1121/10.0019341

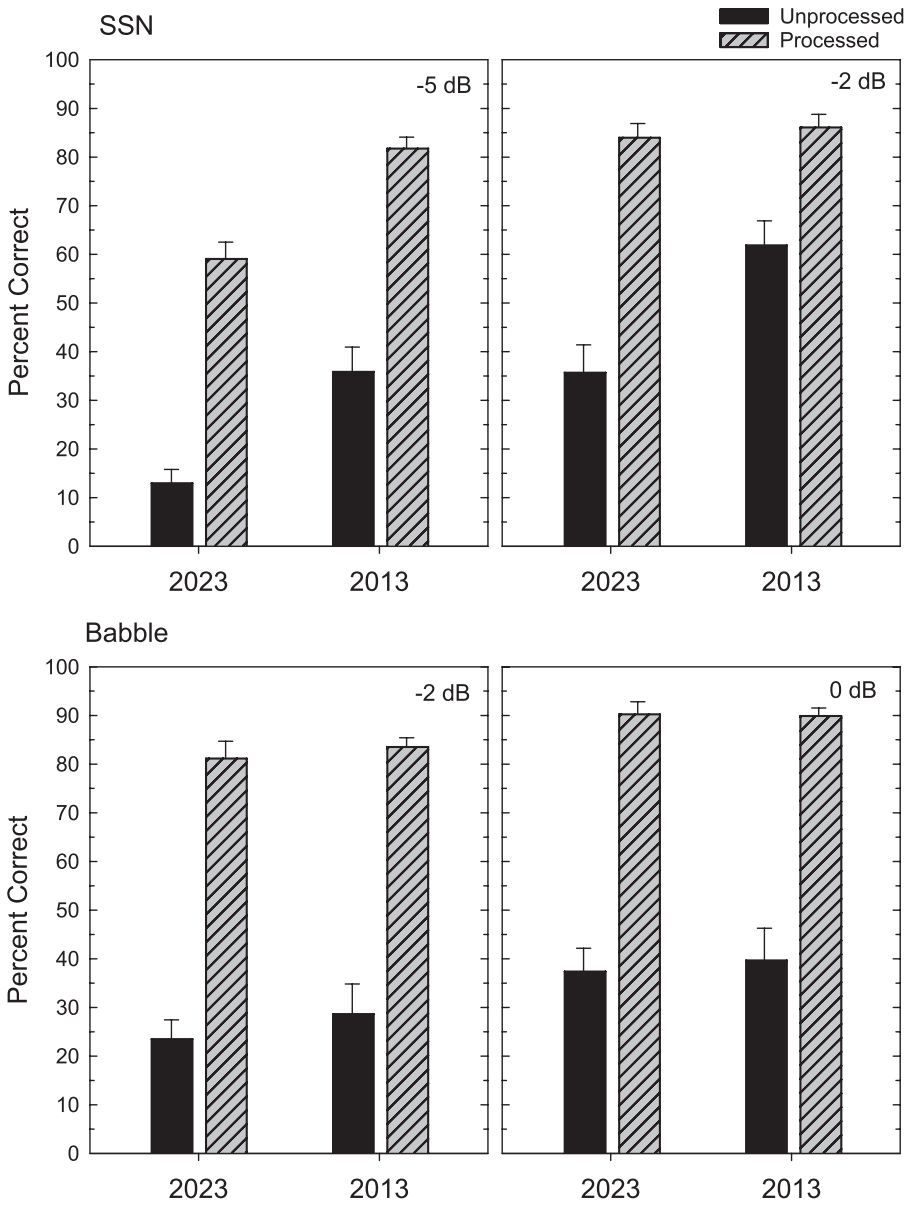

FIG. 12. Comparison between the current study (2023) and the initial study (Healy *et al.*, 2013). Shown are sentence intelligibility scores (and standard errors) for HI listeners, in two noise types, at different SNRs, both before and after algorithm processing. The comparison is between benefit obtained currently versus that obtained in 2013. Benefit is reflected as the height difference between each shaded/hatched column and the adjacent black column. The speech recordings, noise types, testing procedures, and subject populations were identical across studies. The primary difference was the demand placed on the algorithm and the network architecture.

21% points, bottom right panel), perhaps partly reflecting the fact that the unprocessed scores were higher in the current study.

Planned comparisons consisting of two-tailed Welch's *t*-tests on RAUs were used to assess differences in group-mean algorithm benefit between the two studies. For HI subjects, mean benefit was numerically higher for the present (2023) algorithm in all 4 conditions, despite the far greater demands placed on it, but this difference was only significant in SSN at –2 dB [$t(18.6) = 3.1$, $p < 0.01$, $d = 1.28$].

For NH subjects, there was no significant difference in benefit between the 2023 vs 2013 results in two out of the three conditions common to both studies (SSN at –5 dB SNR and babble at –2 dB SNR). In babble at –5 dB SNR, benefit was significantly higher for the 2013 algorithm [$t(20.1) = 3.68$, $p = 0.0015$, $d = 1.50$].

To determine the ability of these *t*-tests to detect difference between the results of the present study and those of the 2013 study, a set of *post hoc* power analyses was conducted using the MKmisc package (Kohl, 2022) in R 4.2.2 (R Core Team, 2020). It was determined that each of the seven comparisons had a power greater than or equal to 0.8 to detect an effect size of 1.2. Accordingly, these tests had sufficient power to detect large differences, should they exist.

E. Objective measures of intelligibility and sound quality

Table II displays objective measures obtained from acoustic analyses of the current stimuli. Scores for both noisy mixtures and processed mixtures are provided. Short-time objective intelligibility (STOI; Taal *et al.*, 2011), represents a correlation between the amplitude envelope of the clean speech and that of the enhanced speech from the mixture. Extended short-time objective intelligibility (ESTOI;

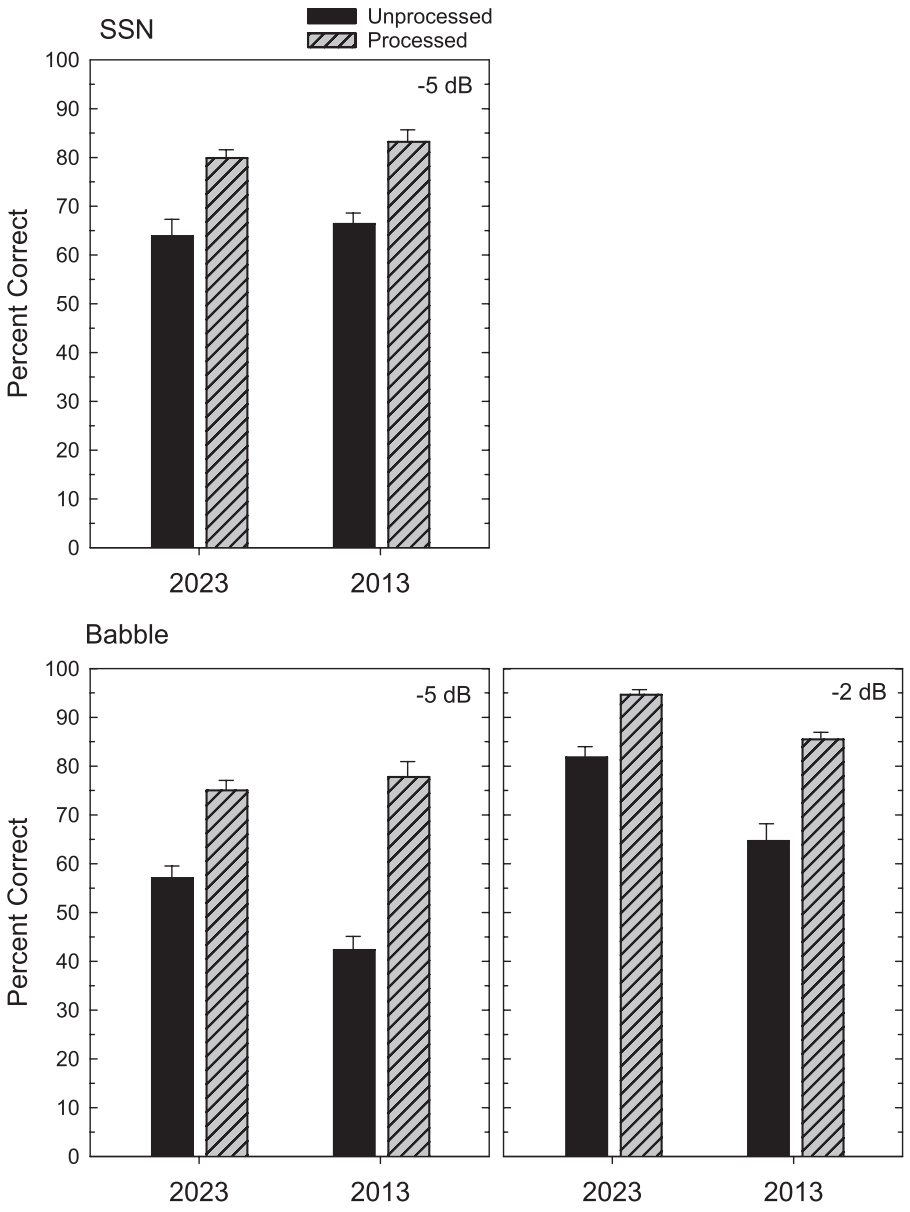

FIG. 13. As Fig. 12, but for the NH listeners.

Jensen and Taal, 2016), is similar and designed to better handle fluctuating noisy signals. Perceptual evaluation of speech quality (PESQ; Rix *et al.*, 2001) is an objective measure of sound quality and ranges from −0.5 to 4.5. These measures have been developed and validated for NH (and not HI) listeners. Finally, the scale invariant signal-to-noise ratio (SI-SNR; Le Roux *et al.*, 2018) is an SNR estimate of the noisy and processed signals. STOI increased from noisy to ARN-processed by 20 points when averaged across conditions. ESTOI increased by an average of 32 points. PESQ increased by an average of 0.8, and SI-SNR increased by an average of 10 dB.

TABLE II. Improvement in objective scores for sentences in two noise types at different SNRs.

		STOI (%)		ESTOI (%)		PESQ		SI-SNR (dB)	
		Noisy	ARN	Noisy	ARN	Noisy	ARN	Noisy	ARN
Babble	−2 dB	62.6	84.9	38.4	72.6	1.47	2.37	−2.0	8.1
	0 dB	68.4	88.4	45.5	78.1	1.59	2.56	0.0	9.6
SSN	−5 dB	59.0	77.9	32.7	62.9	1.37	2.01	−5.0	6.2
	−2 dB	67.2	85.3	43.0	73.4	1.54	2.33	−2.0	8.7

IV. GENERAL DISCUSSION

The current results demonstrate that state-of-the-art deep-learning-based noise reduction can produce large intelligibility improvements for HI and NH listeners, despite the considerable demands associated with extensive generalization and fully causal operation. Intelligibility benefit resulting from processing averaged 46 to 58% points across conditions for the HI listeners. Benefit was lower for NH listeners (8 to 18 points), who typically experience less benefit. These benefits were statistically significant in all conditions.

The other goal of the current study was to compare the ability of the current algorithm to improve intelligibility

https://doi.org/10.1121/10.0019341

relative to the performance of the initial demonstration (Healy *et al*., 2013). This comparison provides a cumulative assessment of algorithm performance, or possible performance loss, resulting from the removal over time of various real-world constraints including talker, corpus, and noise dependence, as well as from causal operation. Whereas the removal of these constraints increases the challenge, it is necessary for operation in the real world. For HI listeners, it was found that the benefit observed currently matched or exceeded that observed in the initial study (but only significantly exceeded in one condition). For NH listeners, benefit was lower currently than in the initial study (but only significantly so in one condition).

It was noted that baseline intelligibility in unprocessed conditions differed somewhat across the current study and the initial demonstration (see Figs. 12 and 13). However, all scores are free of strong floor and ceiling effects, and so they largely fall in the linear portion of the psychometric function relating intelligibility to information content. This allows benefit (differences across two points on this largely linear function) to be compared with reasonable confidence. Nevertheless, the comparison of exact benefit values across studies should not be emphasized.

It is concluded that modern deep-learning-based noise reduction can produce large intelligibility benefit for HI listeners (and also some benefit for NH listeners), despite the removal of multiple constraints and the resultant demanding test conditions. No decrement in benefit was observed for the HI listeners (but some decrement was observed for the NH listeners) resulting from the current algorithm relative to the initial demonstration, thus illustrating the advancements made to neural network design and deep-learning-based noise reduction since 2013.[5]

As mentioned in Sec. I, there are two main approaches to model design. In the efficacy-first approach, the focus is on network performance and obtaining large intelligibility benefit across a wide variety of generalizations. Causality and network size (viability) is addressed once this efficacy is achieved (network size refers to computational complexity and the burden placed on host hardware). This is the approach that we have taken in this area. In the viability-first approach, causal networks are employed that are often also small in size. Examples of this approach include Goehring *et al*. (2016), Goehring *et al*. (2017), Goehring *et al*. (2019), Monaghan *et al*. (2017), Bensten *et al*. (2018), and Keshavarzi *et al*. (2019). Each approach has advantages and disadvantages. An early emphasis on efficacy produces networks that are unable to be implemented, but that produce vast benefit. An early emphasis on viability provides networks that are more easily implemented, but that typically produce smaller benefit. The logic for each is as follows: The emphasis on efficacy retains intelligibility benefit, while advances in network design allow this algorithm performance to remain as steps are taken toward implementability. The emphasis on viability retains implementability, while advances in network design allow intelligibility benefit to increase over time.

One advantage to the former approach is that a large intelligibility-benefit benchmark is established, and this benchmark can be monitored as generalization is systematically advanced and steps are taken toward viability. Accordingly, the performance "costs" associated with each step are known and alternative approaches can be sought if that cost is high. An example of this approach involves the comparison in Healy *et al*. (2021b), who examined the cost associated with conversion to causal operation. The large intelligibility-benefit benchmark established using a non-causal speaker-separation/de-reverberation network (Healy *et al*., 2020) was used to directly establish the "causal cost" by modifying the 2020 model to be fully causal. A decrement in HI intelligibility benefit was observed in most but not all conditions, and these decrements were statistically significant in half of the conditions tested. However, benefit resulting from the large network remained high despite the removal of future time frames. It was concluded that a cost associated with causal processing was present in most conditions, but may be considered modest relative to the overall level of benefit. Thus, that path forward can be followed, resulting in a speaker separation/de-reverberation algorithm that is fully causal and yet produces large intelligibility benefit for HI listeners.

A generalization challenge that was revealed more recently involves the concept of speech-corpus or recording-channel independence (Pandey and Wang, 2020b). The use of the same fixed recording apparatus and environment for the training and test speech can lead to overfit even when different speakers are recorded. This is because the network may be sensitive to characteristics of the microphone and recording room and may learn to rely on them. In practice, the use of a large multi-talker corpus for training and a separate standardized corpus for testing produces not only talker independence but also corpus/recording channel independence (e.g., see, Healy *et al*., 2020). Pandey and Wang (2020b) attributed the challenge of cross-corpus generalization largely to differences in recording channels and proposed techniques to successfully address cross-corpus generalization.

The cross-language study of Healy *et al*. (2021a) perhaps best exemplifies the ability of modern networks to generalize. The speaker-separation network not only generalized across different languages, it also removed large amounts of room reverberation and generalized across speech corpus/recording channels, target-to-interferer energy ratios, room impulse responses (reverberation), and talkers. Despite only NH listeners being tested (they typically produce smaller benefit than do HI listeners), intelligibility increases averaged 44% points and were comparable to those observed in within-language conditions. Modern network training requires the availability of vast amounts of training data, including many examples of speech and noise. The largest speech corpora are typically in English, which could potentially hinder the deployment of deep-learning-based noise reduction in non-English-speaking countries. However, the study of Healy *et al*. (2021a) demonstrates

that this is not an issue, and instead large generalizations are possible including across different languages. This work further suggests that the learning that takes place by modern networks transcends particular languages and is instead perhaps more centered on the commonality of sounds that humans produce.

One characteristic of some modern networks (see Table I) involves operation in the complex domain or the estimation of both amplitude and phase of the target speech. It had been found that phase could not be estimated directly in frequency-domain/mask-based models, but that estimation of both the real and imaginary parts allowed both the amplitude and phase of the target speech to be estimated (Williamson *et al.*, 2016). In the current time-domain model, the clean target speech is estimated directly, which codes both amplitude and phase. These approaches contrast with the traditional one, in which only the amplitude of the target speech is estimated, then combined with the phase of the speech-plus-noise mixture ("noisy phase"). The use of highly similar testing conditions across Healy *et al.* (2019) and Healy *et al.* (2020) allows direct comparison to be made across two different speaker-separation algorithms. The differences between the models employed were considerable and far from restricted to only complex operation. However, one difference involved the use of noisy phase (2019) versus complex operation (2020). Both studies employed reverberant single-talker interference, the same speech materials, and the common condition [SNR = 0 dB, T_{60} = 0.6 s]. The algorithmic advances over the one-year span allowed HI-listener benefit observed in the latter study (73% points) to exceed that observed in the former study (56% points), despite the additional challenge associated with talker independence in the latter study.

With regard to potential bilateral use, the ARN is a time-domain algorithm that maps a sequence of noisy speech frames to the corresponding sequence of enhanced frames. The algorithm does not alter the scale (level) or timing of clean speech embedded in a noisy signal as it is trained using a scale-preserving MSE (mean squared error) loss measured in terms of signal samples (Pandey and Wang, 2022). In other words, the ARN is expected to maintain time and level cues at the sample-point level. The ramifications of this include that, for example, if applied to left- and right-ear signals independently in bilateral hearing aids, the ARN would preserve interaural time and level differences important for sound localization.

Finally, it is noted that the current network remains large in size. Although computational complexity and the demand that a neural network places on the device on which it runs is not a fundamental aspect of viability, it is nonetheless an important consideration. Fortunately, the emerging field of DNN model compression provides techniques that can be used to reduce the size of a model while retaining high performance (see, e.g., Tan and Wang, 2021).

Real-time deep-learning-based noise reduction has recently been implemented into commercial products, which demonstrates the real-world feasibility of such applications. For example, the software plugin Krisp (https://krisp.ai/) uses artificial neural networks to remove noise in voice and videoconferencing applications such as Zoom. However, its capacity to increase intelligibility, especially for HI listeners, has yet to be determined to our knowledge. More relevant for listeners with hearing loss, a hearing aid containing on-board deep-learning-based noise reduction was made available in 2021—the Oticon More™ line. Using speech in diffuse noise, intelligibility increases of approximately 5% points or 1.2 to 1.5 dB reductions in speech reception thresholds were observed relative to the prior generation Oticon hearing aid (Santurette *et al.*, 2020). Although future advances will undoubtedly produce larger intelligibility benefits, the commercial availability of on-board deep-learning-based noise reduction is a substantial accomplishment and marks a major milestone.

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.

[1]"Noise reduction" is used here as an umbrella term to describe the isolation of target speech from various types of interference including background non-speech noise, speech babble, interfering speech from a single talker, room reverberation, and concurrent interferences involving more than one such interference. More specific terms for these processes include speech enhancement, speaker separation, and de-reverberation. "Single microphone" refers to conditions in which target speech and noise are received by the same single microphone and represents one of the most challenging but broadly applicable noise-reduction approaches.

[2]Talkers of different sex were employed for testing human subjects in order to reduce confusion with regard to which talker was the target. Algorithm performance appears similar when target and interfering talkers are same versus different sex (Liu and Wang, 2019).

[3]Other considerations important for implementation into hearing devices include computational complexity of the network and power consumption. However, whereas causality is fundamental, these other considerations are relative to the ever-advancing processing capabilities of the system on which it runs. But it is clear that smaller/more efficient networks are more readily implementable.

[4]See www.sound-ideas.com (Last viewed 5/21/2021).

[5]Diehl *et al.* (2023) published a study following revision of the current manuscript that involved a talker- and noise-independent DNN that operated in the complex domain and in real time. Speech-reception-threshold benefits for HI listeners were approximately 4 dB. The timely convergence of that work and the present study further highlights the ability of techniques evolved over the past decade to produce substantial intelligibility benefits under unconstrained conditions.

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).

Ba, J. L., Kiros, J. R., and Hinton, G. E. (**2016**). "Layer normalization," arXiv:1607.06450.

Bensten, 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**, e0196924.

Bramsløw, L., Naithani, G., Hafez, A., Barker, T., Pontoppidan, N. H., and Viranen, T. (**2018**). "Improving competing voice segregation for hearing-impaired listeners using a low-latency deep neural network algorithm," J. Acoust. Soc. Am. **144**, 172–185.
Byrne, D., Parkinson, A., and Newall, P. (**1990**). "Hearing aid gain and frequency response requirements for the severely/profoundly hearing impaired," Ear Hear. **11**, 40–49.
Chen, J., Wang, Y., Yoho, S. E., Wang, D. L., and Healy, E. W. (**2016**). "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," J. Acoust. Soc. Am. **139**, 2604–2612.
Dillon, H. (**2012**). *Hearing Aids*, 2nd ed. (Thieme, New York).
Diehl, P. U., Singer, Y., Zilly, H., Schönfeld, U., Meyer-Rachner, P., Berry, M., Sprekeler, H., Sprengel, E., Pudszuhn, A., and Hofmann, V. M. (**2023**). "Restoring speech intelligibility for hearing aid users with deep learning," Sci. Rep. **13**, 2719–2730.
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. 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.
Goehring, T., Yang, X., Monaghan, J. J. M., and Bleek, S. (**2016**). "Speech enhancement for hearing-impaired listeners using deep neural networks with auditory-model based features," in *Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO)*, August 28–September 2, Budapest, Hungary, pp. 2300–2304.
Gustafsson, S., Jax, P., and Vary, P. (**1998**). "A novel psychoacoustically motivated audio enhancement algorithm preserving background noise characteristics," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98*, May 15, Seattle, WA, pp. 397–400.
He, K., Zhang, X., Ren, S., and Sun, J. (**2016**). "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 27–30, Las Vegas, NV, pp. 770–778.
Healy, E. W., Delfarah, M., Johnson, E. M., and Wang, D. L. (**2019**). "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," J. Acoust. Soc. Am. **145**, 1378–1388.
Healy, E. W., Delfarah, M., Vasko, J. L., Carter, B. L., and Wang, D. L. (**2017**). "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," J. Acoust. Soc. Am. **141**, 4230–4239.
Healy, E. W., Johnson, E. M., Delfarah, M., Sevich, V. A., Krishnagiri, D. S., and Wang, D. L. (**2021a**). "Deep learning based speaker separation and dereverberation can generalize across different languages to improve intelligibility," J. Acoust. Soc. Am. **150**, 2526–2538.
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., Taherian, H., Johnson, E. M., and Wang, D. L. (**2021b**). "A causal and talker-independent speaker-separation/dereverberation deep learning algorithm: Cost associated with conversion to real-time capable operation," J. Acoust. Soc. Am. **150**, 3976–3986.
Healy, E. W., Tan, K., Johnson, E. M., and Wang, D. L. (**2021c**). "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.
Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029–3038.
Hendrycks, D., and Gimpel, K. (**2016**). "Gaussian error linear units (GELUS)," arXiv:1606.08415.
Hochreiter, S., and Schmidhuber, J. (**1997**). "Long short-term memory," Neural Comput. **9**(8), 1735–1780.
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. Process. **24**, 2009–2022.
Kearns, J. (**2014**). "LibriVox: Free public domain audiobooks," https://librivox.org/ (Last viewed April 26, 2023).
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.
Kohl, M. (**2022**). "MKmisc: Miscellaneous functions from M. Kohl," R package version 1.9, https://github.com/stamats/MKmisc (Last viewed February 7, 2023).
Kramer, S. E., Kapteyn, T. S., and Festen, J. M. (**1998**). "The self-reported handicapping effect of hearing disabilities," Int. J. Audiol. **37**, 302–312.
Lai, Y.-H., Tsao, Y., Lu, X., Chen, F., Su, Y.-T., Chen, K.-C., Chen, Y.-H., Chen, L.-C., Li, L. P.-H., and Lee, C.-H. (**2018**). "Deep learning-based noise reduction approach to improve speech intelligibility for cochlear implant recipients," Ear. Hear. **39**, 795–809.
Le Roux, J., Wisdom, S., Erdogan, H., and Hershey, J. R. (**2018**). "SDR—Half-baked or well done?," arXiv:1811.02508v1.
Li, L. P.-H., Han, J.-Y., Zheng, W.-Z., Huang, R.-J., and Lai, Y.-H. (**2021**). "Improved environment-aware-based noise reduction system for cochlear implant users based on a knowledge transfer approach: Development and usability study," J. Med. Internet Res. **23**, e25460.
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. Process. **27**, 2092–2102.
Merity, S. (**2019**). "Single headed attention RNN: Stop thinking with your head," arXiv:1911.11423.
Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. (**2017**). "Mixed precision training," arXiv:1710.03740.
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.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.
Nossier, S. A., Wall, J., Moniri, M., Glackin, C., and Cannings, N. (**2020**). "An experimental analysis of deep learning architectures for supervised speech enhancement," Electronics **10**, 17.
Ochieng, P. (**2022**). "Deep neural network techniques for monaural speech enhancement: State of the art analysis," arXiv:2212.00369.
Panayotov, V., Chen, G. D., and Khudanpur, S. (**2015**). "LibriSpeech: An ASR corpus based on public domain audio books," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, April 19–24, Queensland, Australia, pp. 5206–5210.
Pandey, A., and Wang, D. (**2020a**). "Learning complex spectral mapping for speech enhancement with improved cross-corpus generalization," in *Proceedings of Interspeech*, October 25–29, Shanghai, China, pp. 4511–4515.
Pandey, A., and Wang, D. (**2020b**). "On cross-corpus generalization of deep learning based speech enhancement," IEEE/ACM Trans. Audio. Speech. Lang. Process. **28**, 2489–2499.
Pandey, A., and Wang, D. (**2022**). "Self-attending RNN for speech enhancement to improve cross-corpus generalization," IEEE/ACM Trans. Audio. Speech. Lang. Process. **30**, 1374–1385.
Paul, D. B., and Baker, J. (**1992**). "The design for the Wall Street Journal-based CSR corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman*, February 23–26, New York, NY.
R Core Team (**2020**). *R: A Language and Environment for Statistical Computing*, https://www.R-project.org/ (Last viewed February 7, 2023).
Rix, A. W., Beerends, J. G., Hollier, M. P., and Hekstra, A. P. (**2001**). "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 7–11, Salt Lake City, UT, pp. 749–752.

Santurette, S., Ng, E. H. N., Jensen, J. J., and Loong, B. M. K. (**2020**). *Oticon More Clinical Evidence* (Oticon, Somerset, NJ).
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.
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. (**2021**). "Towards model compression for deep learning based speech enhancement," IEEE/ACM Trans. Audio. Speech. Lang. Process. **29**, 1785–1794.
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.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (**2017**). "Attention is all you need," Adv. Neural Inf. Process. Syst. **30**, 5998–6008.
Wang, D., and Chen, J. (**2018**). "Supervised speech separation based on deep learning: An overview," IEEE/ACM Trans. Audio. Speech. Lang. Process. **26**, 1702–1726.
Wang, Y., Han, K., and Wang, D. (**2013**). "Exploring monaural features for classification-based speech segregation," IEEE Trans. Audio. Speech. Lang. Process. **21**, 270–279.
Wang, Y., and Wang, D. (**2013**). "Towards scaling up classification-based speech separation," IEEE Trans. Audio. Speech Lang. Process. **21**, 1381–1390.
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.