JASA ARTICLE

Perceptual effects of reducing algorithmic latency on deep-learning based noise reduction[a]

Eric W. Healy,[1,2,b] Sarah E. Yoho,[1,2] Kian Fallah,[1] Ashutosh Pandey,[3,4] and DeLiang Wang[2,3]

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

[2]Center for Cognitive and Brain Sciences, The Ohio State University, Columbus, Ohio 43210, USA

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

[4]Meta Reality Labs, Redmond, WA 98052, USA

ABSTRACT:

Low latency is an essential requirement for noise reduction in real-world devices such as hearing aids and cochlear implants. Reducing the algorithmic latency of a deep neural network charged with noise reduction allows additional time for other processing. However, a larger analysis window may be advantageous to the performance of the network. This trade-off is currently examined with regard to human speech-intelligibility performance. The algorithmic latency of the attentive recurrent network (ARN) was modified by reducing the size of the analysis time frame. The ARN model was talker, noise, and recording-channel independent, and fully causal. Listeners with hearing loss and with normal hearing heard sentences in babble at various signal-to-noise ratios. Large increases in intelligibility were observed as a result of noise reduction, especially for the listeners with hearing loss and at less favorable signal-to-noise ratios. Slightly larger objective measures of network performance were observed at larger latencies. But more critically, human performance was essentially unchanged as algorithmic latency was reduced from 20 to 10 or 5 ms. These results are discussed in the context of overall design and implementation of deep-learning based noise reduction, and information on latency requirements for human listeners is summarized.

https://doi.org/10.1121/10.0037197

(Received 3 March 2025; revised 19 June 2025; accepted 30 June 2025; published online 14 July 2025)

[Editor: Li Xu] Pages: 380–390

I. INTRODUCITON

The primary auditory concern of individuals with hearing loss involves poor speech understanding when background noise is present. Accordingly, considerable attention has been paid to noise reduction for several decades. An approach based on artificial intelligence [deep neural networks (DNNs)] has proven to be promising. Deep-learning based noise reduction has evolved over the past decade from its introduction as a highly-limited laboratory proof of concept (Healy *et al.*, 2013; Wang and Wang, 2013), to elegant networks that can operate in the real world (for reviews, see Wang, 2017; Healy *et al.*, 2023). They have recently been introduced into commercially available hearing aids, including devices from Oticon (2021 deployment; Santurette *et al.*, 2020; Bramsløw and Beck, 2021) and Phonak (2024 deployment; Hasemann and Kryloa, 2024; Wright *et al.*, 2024).

Any signal-processing scheme intended for people in the real world carries concerns about processing latency. In the case of a DNN, the overall latency is the sum of two distinct elements: intrinsic network latency (algorithmic latency) and processing latency. Algorithmic latency results from the design of the algorithm itself and corresponds to the size of the time frame or analysis window. Processing latency results from the computational effort needed to process a frame.

Whereas processing latency depends entirely on the particular hardware and implementation employed, algorithmic latency is independent of hardware and implementation and is instead intrinsic to algorithm design. Unlike processing latency, algorithmic latency may impact the performance of a noise-reduction/speech-enhancement algorithm. Accordingly, the impact of algorithmic latency on the intelligibility of noise-reduced/enhanced speech for human listeners was examined currently.

A wide variety of factors affect the detectible or tolerable auditory delay for human listeners, and these factors govern the latency requirements for communication devices. These factors include the task (speech perception vs production), the device (hearing aid, cochlear implant, or phone), and the fitting, if the device is a hearing aid (open vs closed). The science is fairly clear with regard to acceptable delays across these situations. For hearing aids, detection/disruption thresholds generally range from 15–30 ms for individuals with mild-moderate hearing loss and 20–40 ms for individuals with moderate-severe hearing loss (Stone and Moore, 1999, 2002, 2005; Goehring *et al.*, 2018). Smaller delay tolerances (e.g., 5–6 ms) have been observed but

[a]Portions of this work were presented in "Impact of reduced latency on AI-based noise reduction," Proceedings of the 52nd Annual Scientific and Technology Conference of the American Auditory Society, Scottsdale, AZ, February 2025.

[b]Email: healy.66@osu.edu

attributed to artifacts associated with simulating dynamic processing of hearing aids in normal-hearing (NH) listeners (Stone *et al.*, 2008). For voice or video calls, larger acceptable values are found (150 ms; ITU, 2003). For cochlear implant users, delay detection thresholds are driven by auditory-visual synchrony and are quite large (approximately 225 ms; Conrey and Pisoni, 2006; Hay-McCutcheon *et al.*, 2009; Baskent and Bazo, 2011). With regard to speech production, delayed auditory feedback can produce disfluencies above 50 ms and slowed speech above 25 ms (Stuart *et al.*, 2002), but suppression of the own-voice signal is commonly implemented in modern hearing aids, which mitigates this effect. Although latency around 10 ms is sometimes considered to be a requirement for hearing devices, such low values do not appear to be well grounded in the literature.

In the current study, the attentive recurrent network (ARN, Pandey and Wang, 2022) was employed. The ARN is a time-domain model for speech enhancement that directly processes sequences of raw noisy frames to generate enhanced (noise-reduced) frames. This method bypasses time-frequency representations such as the short-time Fourier transform, instead adopting an end-to-end approach that leverages supervised training to improve speech enhancement. The latency of such a network is defined by the length of the time frame used for processing, because the network outputs the estimate of clean (noise-free) speech at the end of each frame.

In Healy *et al.* (2023), the frame length of a causal ARN was set to 20 ms, and intelligibility for hearing-impaired (HI) and NH listeners was assessed. Hearing in noise test (HINT; Nilsson *et al.*, 1994) sentences were mixed with speech-shaped noise and multi-talker babble at various signal-to-noise ratios (SNRs), and intelligibility prior to and following noise reduction was assessed. Intelligibility increases resulting from ARN noise reduction averaged 51% points for HI listeners and 14 points for NH listeners.

The question addressed currently involves whether a DNN charged with noise reduction can be modified to produce a lower algorithmic latency without hindering human intelligibility. Shorter algorithmic latencies can be desirable because they allow additional time to be allocated to processing, hence alleviating hardware requirements, or to other stages of the speech-processing path. In contrast, a larger analysis window can be advantageous to the performance of the network – a larger time frame contains additional contextual information within each frame and allows for a greater statistical pooling effect with the same frame shift (see Sec. III). Although low latency as a requirement for real-world deep-learning based noise reduction has been considered by others, and the performance of human listeners has been evaluated when using low-latency systems (e.g., Goehring *et al.*, 2017; Bramsløw *et al.*, 2018), the impact on human intelligibility of manipulating the latency of an otherwise identical neural network has received less attention.

Here, ARNs having algorithmic latencies of 20, 10, and 5 ms were employed. The network architecture and training were identical, with this one exception. Frame shift was also held constant to produce a relatively fixed computational demand. The ARN is a modern time-domain DNN, representative of time-domain deep-learning models for speech enhancement. Objective measures based on acoustic characteristics of the signals themselves were performed to determine the extent to which the algorithm is able to output a more acoustically veridical speech signal when frames are longer or alternatively, if the frames can be reduced in size without acoustic hindrance. More critically, intelligibility testing revealed the extent to which any difference in the acoustic accuracy of the output signal has perceptual ramifications for HI and NH listeners.

II. METHODS

A. Subjects

A group of ten listeners with hearing loss was recruited from The Ohio State University Speech-Language-Hearing Clinic to represent typical patients. All were bilateral hearing-aid users and had sensorineural hearing loss of likely cochlear origin. Ages ranged from 38 to 84 years, six were female, and four were male. Pure-tone air-conduction audiometric thresholds are shown in Fig. 1. The hearing losses varied, but can generally be characterized as symmetric and sloping. Average degree of loss (3- or 4-frequency pure-tone average) varied from mild to moderately-severe.

A second group of ten listeners with NH also participated. These individuals were recruited from courses at The Ohio State University. All passed a pure-tone air-conduction audiometric screening from 250 to 8000 Hz at 20 dB hearing level (HL) on the day of the test. Their ages ranged from 18 to 22 years (average = 19.7) and all were female. All participants (HI and NH) were native speakers of American English, having no previous exposure to the test sentences used, and all received either course credit or a monetary incentive for participating. Note that the lack of age matching allows for the comparison across typical hearing-aid users and young “ideal” ears.

B. Stimuli

The stimuli used to test human listeners were sentences from the HINT. The standard recordings produced by a male talker speaking standard American English were used. The background noise was a 20-talker babble recording from Auditec (http://www.auditec.com, approximately 10 min in duration). This noise was selected to represent a common real-world background, and our prior work indicates that the noise-independent model is capable of generalizing to different untrained nonstationary noises with similar effectiveness. To create each test stimulus, a sentence was mixed with the babble, using a random start point in the babble file and a duration that matched that of the sentence. The SNRs used for testing the HI listeners were –2, 0, and 3 dB, and those used for the NH listeners were –5, –2, and 0 dB.

The speech stimuli used for neural-network training came from the Librispeech corpus (Panayotov *et al.*, 2015). This corpus contains approximately 1000 h of speech recordings from over 2000 speakers. It is primarily used for research in large-vocabulary continuous speech recognition systems and sources its data from the LibriVox project (Kearns, 2014). LibriVox features a diverse collection of audiobook recordings contributed by volunteers worldwide. The varied recording conditions within Librispeech, including different microphones and room acoustics, make it an ideal dataset for training corpus-independent speech enhancement algorithms, as noted by Pandey and Wang (2020a,b). This diversity is crucial for preventing overfitting to specific acoustic characteristics of the recording hardware or environment.

The background noise stimuli used for neural-network training included 10 000 non-speech sounds from a sound-effects library (www.sound-ideas.com). During training, clean and noisy speech pairs were generated by randomly selecting an utterance, a noise segment, and an SNR level from a range of –5 to 0 dB. Additionally, a set of 150 validation mixtures was prepared using utterances from six speakers from the Wall Street Journal Corpus (WSJ0, Paul and Baker, 1992) combined with factory noise from the NOISEX dataset (Varga and Steeneken, 1993).

C. ARN

A noisy speech signal $\boldsymbol{y}$, recorded by a microphone, can be described as the sum of the target speech signal $\boldsymbol{s}$ and the background noise signal $\mathbf{n}$ as

$$\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n}, \qquad (1)$$

where $\{\boldsymbol{y}, \boldsymbol{s}, \boldsymbol{n}\} \in \mathbb{R}^{1\times N}$. A speech enhancement algorithm aims to process the observed noisy signal $\boldsymbol{y}$ to produce an accurate estimate, $\hat{\boldsymbol{s}}$, of $\boldsymbol{s}$ as

$$\hat{\boldsymbol{s}} = f(\boldsymbol{y}); \qquad (2)$$

here, $f : \mathbb{R}^{1\times N} \to \mathbb{R}^{1\times N}$ is a function that takes a noisy signal as input and outputs an enhanced signal of the same length. In the realm of DNN-based speech enhancement, f denotes a DNN model.

Typically, speech enhancement algorithms operate as short-time processing systems. In this setup, the input $\boldsymbol{y}$ is segmented into T overlapping frames $[\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_{T-2}, \boldsymbol{y}_{T-1}]$, with tth frame defined as

$$\boldsymbol{y}_t = \boldsymbol{y}\big[(H\cdot t) : (H\cdot t + L)\big], \qquad (3)$$

where $\boldsymbol{y}[a : b] \in \mathbb{R}^{1\times(b-a)}$, is a vector formed by extracting elements from $\boldsymbol{y}$ starting at index a and ending at b, H is hop size (frame shift), and L is frame size. Next, a speech enhancement function is applied to enhance all the noisy frames as

$$\hat{\boldsymbol{s}}_t = f_1(\boldsymbol{y}_{t-M}, \ldots, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t+N}), \qquad (4)$$

where f_1 is a function taking M past frames, N future frames, and the current frame $\boldsymbol{y}_t$ to estimate $\hat{\boldsymbol{s}}_t$. Finally, overlap-add is applied to combine enhanced frames to obtain an enhanced waveform.

The function defined in Eq. (4) leverages future frames to estimate a current frame, thereby improving robustness and performance. However, this method necessitates waiting for future frames, resulting in latency between the desired signal and the output signal. To eliminate the need to wait for future frames, a common strategy is to design the algorithm in a causal manner. This approach ensures that the

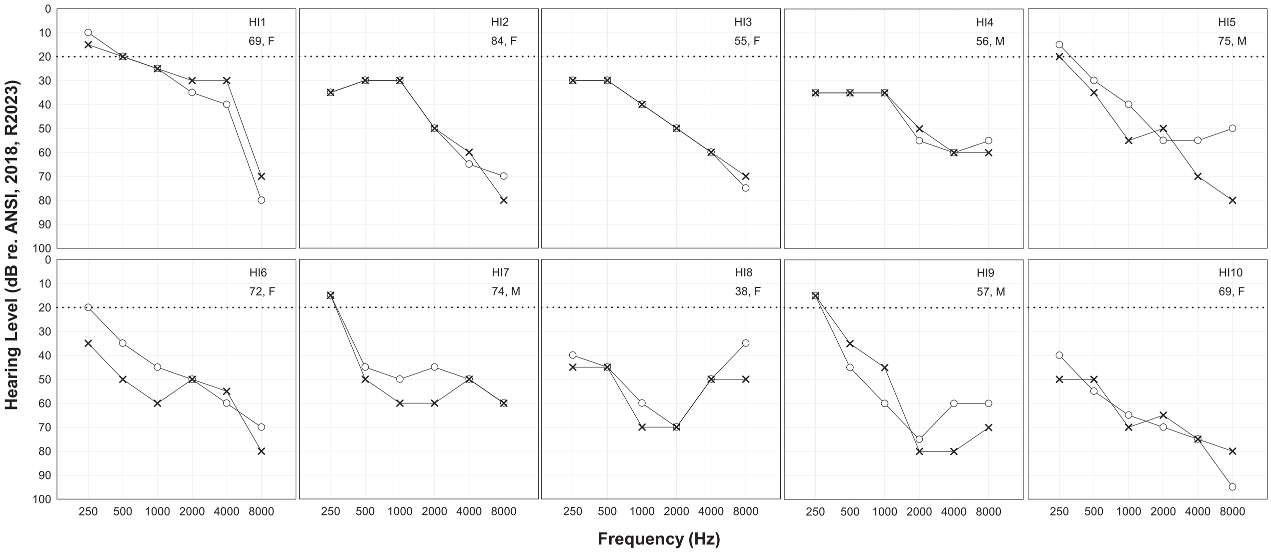

FIG. 1. Pure-tone air-conduction audiometric thresholds for the listeners with hearing loss. Listeners are numbered in order of increasing degree of hearing loss. Right ears, circles; left ears, ×'s. Listener numbers, ages in years, and sexes are also provided.

output for any given frame relies solely on the current and previous frames, as given by

$$\hat{\boldsymbol{s}}_t = f_2(\boldsymbol{y}_{t-M}, \ldots, \boldsymbol{y}_t). \quad (5)$$

Modern deep learning architectures, such as recurrent neural networks (RNNs) and transformer models (Vaswani etal., 2017), are adept at processing sequential inputs. These models can ingest a sequence of frames and leverage the entire sequence of past frames rather than a fixed number of past frames. This approach allows them to estimate a sequence of enhanced frames, as described in the following:

$$[\hat{\boldsymbol{s}}_0, \ldots, \hat{\boldsymbol{s}}_t] = f_3(\boldsymbol{y}_0, \ldots, \boldsymbol{y}_t). \quad (6)$$

Algorithmic latency associated with a causal speech enhancement algorithm, as specified in Eqs. (5) and (6), stems from short-time processing and corresponds to the frame size (L). In the current study, ARNs having frame sizes of 20, 10, and 5 ms were employed. Following the insights from Pandey and Wang (2020b), the hop size was adjusted from 2 to 1 ms, aiming to optimize performance for reduced latencies. Pandey and Wang (2020b) highlighted that a smaller hop size can effectively enhance the generalizability of speech enhancement models to new and untrained acoustic scenarios.

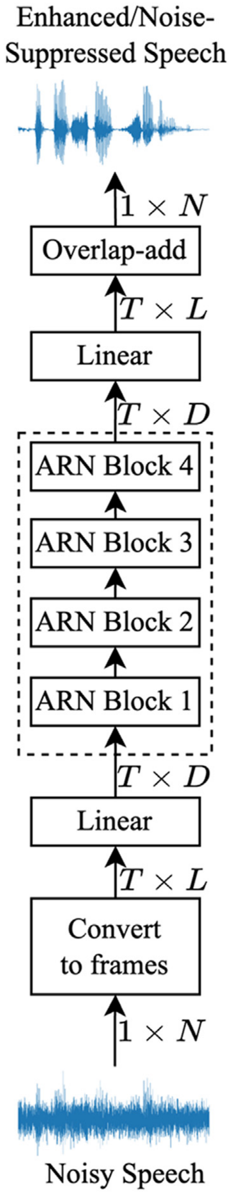

FIG. 2. The ARN model employed for time-domain speech enhancement/noise reduction.

The employed ARN model for speech enhancement is shown in Fig. 2. First, noisy speech $\boldsymbol{y}$ is segmented into frames using a frame size of L and a hop size of H, resulting in T frames. Next, a linear layer is used to transform (expand) them into a representation of size D. This representation is then processed using a series of four ARN blocks. After this, another linear layer is applied to reduce the size of the output from D to L. Finally, overlap-add is applied to reconstruct the enhanced waveform.

The ARN model is fundamentally built around an ARN block, designed to handle inputs $\in \mathbb{R}^{T\times D}$ and produce outputs $\in \mathbb{R}^{T\times D}$, thus preserving the data's original dimensionality. This characteristic is crucial as it allows for the seamless integration of multiple ARN blocks in series to enhance performance without dimensional discrepancies. Illustrated in Fig. 3, the ARN block's architecture includes three integral components: an RNN block, which processes temporal dependencies within the data; an attention block, which strategically emphasizes important features of the input; and a feedforward block, which performs non-linear transformations.

A schematic diagram of the three blocks within an ARN block is shown in Fig. 4. The RNN block comprises layer normalization (Ba *et al.*, 2016) followed by a long short-term memory (LSTM; Hochreiter and Schmidhuber, 1997) RNN. Layer normalization is used to stabilize the training process, enhance convergence speed, and improve the generalization of models by normalizing the inputs across features within each layer independently. LSTM is employed to effectively capture the temporal dependencies within a sequence of frames in a causal manner.

The attention block refines the RNN output using self-attention, effectively addressing the limitations of LSTM in handling long sequences. Whereas LSTM excels in modeling temporal sequences, it compresses sequence information into a single vector (hidden state), potentially losing details in longer sequences. Self-attention complements this by identifying correlations across the sequence to enhance contextual understanding. However, unlike LSTM, self-attention does not naturally maintain the sequential order. By combining LSTM's sequential processing with self-attention's refined processing of past information, the model

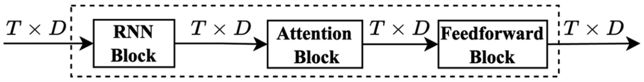

FIG. 3. The building blocks of the ARN. It is composed of an RNN block, an attention block, and a feedforward block.

captures both the order and the rich contextual details of speech.

The input to the attention block undergoes normalization through two distinct layer normalizations, each with its own set of trainable parameters for scale and bias. The output from the first normalization serves as the query ($\boldsymbol{Q}$), while the output from the second normalization acts as the key ($\boldsymbol{K}$) and value ($\boldsymbol{V}$) for the subsequent self-attention module. This module takes as input three matrices, $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V} \in \mathbb{R}^{T\times D}$, and produces a single output matrix $\boldsymbol{A} \in \mathbb{R}^{T\times D}$. The self-attention mechanism includes three trainable vectors, $\boldsymbol{q}$, $\boldsymbol{k}$, and $\boldsymbol{v} \in \mathbb{R}^{1\times D}$. The vector $\boldsymbol{q}$ refines $\boldsymbol{Q}$ through the operation $\boldsymbol{Q}' = \mathrm{Lin}(\boldsymbol{Q}) \odot \sigma(\boldsymbol{q})$, where σ represents the sigmoid function and $\odot$ denotes element-wise multiplication. Similarly, $\boldsymbol{K}$ is refined as $\boldsymbol{K}' = \boldsymbol{K} \odot \sigma(\boldsymbol{k})$, and $\boldsymbol{V}$ is refined through $\boldsymbol{V}' = \boldsymbol{V} \odot [\sigma(\mathrm{Lin}(\boldsymbol{v}) \odot \mathrm{Tanh}(\mathrm{Lin}(\boldsymbol{v}))]$, where Tanh is the hyperbolic tangent function.

Before multiplication, the vectors $\boldsymbol{q}$, $\boldsymbol{k}$, and $\boldsymbol{v}$ are broadcast to $\mathbb{R}^{T\times D}$ to align with $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$, respectively. All linear layers involved produce outputs of size D. It is important to note that the term $[\sigma(\mathrm{Lin}(\boldsymbol{v}) \odot \mathrm{Tanh}(\mathrm{Lin}(\boldsymbol{v}))]$ is a constant vector since $\boldsymbol{v}$ is constant. This aspect of the model is relevant only during training, as it facilitates the optimization of vector $\boldsymbol{v}$ (Merity, 2019). During evaluation, the constant value derived from trained $\boldsymbol{v}$ is utilized.

The output $\boldsymbol{A}$ from the attention module is computed using the following set of equations:

$$W = \frac{\boldsymbol{Q}'\boldsymbol{K}'^{\mathbb{T}}}{\sqrt{D}}, \qquad (7)$$

$$\boldsymbol{W}' = \mathrm{Mask}(\boldsymbol{W}), \qquad (8)$$

$$W'(i,j) = \begin{cases} W(i,j), & \text{if } i \le j, \\ -\infty & \text{otherwise}, \end{cases} \qquad (9)$$

$$\boldsymbol{P} = \mathrm{Softmax}(\boldsymbol{W}'), \qquad (10)$$

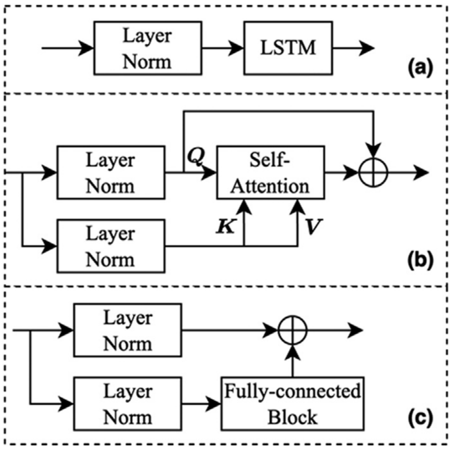

FIG. 4. The three building blocks within an ARN block. (a) RNN block, (b) attention block, and (c) feedforward block. Layer norm denotes layer normalization and $\oplus$ is the elementwise addition operator.

$$\mathrm{Softmax}(W)(i,j) = \frac{e^{W(i,j)}}{\sum_{j=1}^{T} e^{W(i,j)}}, \qquad (11)$$

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{V}', \qquad (12)$$

where $\mathbb{T}$ is the transpose operator. First, pairwise correlation scores are computed between the rows of $\boldsymbol{Q}'$ and $\boldsymbol{K}'$, represented as $\{\boldsymbol{Q}'_i, \boldsymbol{K}'_j\}$, for $i, j \in \{1, \ldots, T\}$, through matrix multiplication as detailed in Eq. (7). Subsequently, a masking function described in Eq. (9) is applied to set the correlation of any given row i with a subsequent row j (where $j > i$) to $-\infty$. This critical step ensures that the influence of future rows is nullified, preparing the scores for transformation into probabilities via the Softmax function. The Softmax operation, which utilizes an exponential function followed by a normalization step in the denominator, effectively converts any $-\infty$ values to zero. This transformation is essential for maintaining the causality of the algorithm, as it prevents future frames from influencing the current frame's output. The attention output $\boldsymbol{A}$ is then calculated by multiplying the attention probabilities $\boldsymbol{P}$ with the values matrix $\boldsymbol{V}'$, as specified in Eq. (12). The process ends with the addition of $\boldsymbol{A}$ to $\boldsymbol{Q}$, establishing a residual connection to facilitate gradient flow during training (He *et al.*, 2016).

The feedforward block takes the output from the attention block and further refines it through nonlinear transformations. This block begins by applying two separate layer normalizations to its input. The first normalized input is then directed through a fully connected block. Within this block, a linear layer initially projects the input to a higher-dimensional space. This projection is immediately followed by the application of a Gaussian error linear unit (GELU) nonlinearity (Hendrycks and Gimpel, 2016) and a dropout layer to enhance model robustness. Subsequently, the expanded output is reduced back to its original size by dividing it into four separate vectors, which are then summed. This condensed output is combined with the second normalized input to produce the final output of the feedforward block, effectively integrating and refining the output from the attention block.

All utterances were resampled to a standard 16 kHz (16-bit). The employed frame sizes of 20, 10, and 5 ms, correspond to L values of 320, 160, and 80, respectively. The frame shift of 1 ms corresponded to $H = 16$. Within the RNN block, an LSTM with a hidden size of 1024 was employed, and a dropout rate of 5% was applied in the fully connected block to prevent overfitting.

The ARN model underwent training over 100 epochs with batches comprising 32 utterances each. To standardize input length, all utterances within a batch were either truncated or zero-padded to achieve a uniform duration of 4 s. The input was root-mean-square (RMS) normalized, and the target clean speech was appropriately scaled to maintain a consistent SNR.

For optimization, the Adam optimizer (Kingma and Ba, 2014) was used with an initial learning rate of 0.0002. After the first 33 epochs, the learning rate underwent exponential decay, reaching 0.00002 by the final epoch. The models were developed using PyTorch. To enhance training efficiency, mixed precision training (Micikevicius *et al.*, 2017) was employed, and each batch was distributed across GPU using PyTorch's DataParallel module.

In terms of model size, the ARN models with the frame lengths of 20, 10, and 5 ms have 55.3, 55.0, and 54.8 × 10^6 parameters, respectively. The corresponding computational complexities are 59.4, 59.1, and 58.9 GFLOPs. As the current focus was on the perceptual effects of inherent or algorithmic latency, no attempt was made to compress or otherwise reduce the computational complexity of the models as would be required for deployment on mobile devices. We point out that techniques exist that substantially reduce DNN models with minimal degradation of speech enhancement performance (see, e.g., Tan and Wang, 2021).

D. Procedure

Each listener heard 15 HINT sentences in each of 12 conditions (3 SNRs × 4 processing conditions). The conditions were blocked by SNR, and the processing conditions within SNR, as well as the SNRs, were presented in random order for each listener. The sentences were presented in the same order for each listener, providing a random sentence-to-condition correspondence for each listener.

The stimuli were scaled to the same total RMS amplitude, then 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 average level at each earphone was set to 65 dBA using a sound-level meter and flat-plate coupler (Larson Davis models 824 and AEC 101; Larson Davis, Depew, NY). Individualized frequency-specific gains were added to this presentation level for the HI listeners. The NAL-RP hearing-aid fitting formula (Byrne *et al.*, 1990) and a RANE DEQ 60 L digital equalizer (RANE, Mukilteo, WA) were used to implement these gains. The use of the linear prescription formula is in accord with the use of linear amplification. The gains prescribed for 250 and 6000 Hz were applied to 125 and 8000 Hz, respectively, because the NAL-RP does not prescribe gains at these lowest and highest audiometric frequencies. Because amplification was provided, HI listeners were tested with hearing aids removed.

Listeners were tested individually while seated in a double-walled audiometric booth. The experimenter was seated outside of the booth, using a large viewing window and two-way intercom to maintain contact. Testing began with practice conditions consisting of five HINT sentences (not used for testing) for each of the following conditions: in quiet, processed using the middle latency and SNR, then in noise at the middle SNR. The HI listeners were asked during this practice if the signals sounded comfortably loud. One listener indicated that the signals sounded loud, but comfortable once reduced by 5 dB. Following practice, the listeners heard the 180 test sentences. They were instructed to repeat each back as best as they could, guessing if unsure. The experimenter recorded the responses and controlled the presentation of sentences. In accord with standard HINT scoring procedures, component words were scored as correct if repeated exactly, apart from verb tense (e.g., is/was) and article (a/the) variations. No feedback was provided during testing, and no sentence was repeated for any listener.

III. RESULTS

A. Objective measures

Table I displays commonly employed objective measures of model performance for three algorithmic delays, where the measures are based on acoustic analyses of the signals themselves. Included are short-time objective intelligibility (STOI) (Taal *et al.*, 2011), which reflects the correlation between the amplitude envelopes of (a) the original clean speech prior to mixing with noise and (b) the enhanced speech extracted from the noisy speech by the network. The scale ranges from 0 to 100, where higher scores represent higher predicted intelligibility. Perceptual evaluation of speech quality (PESQ) (Rix *et al.*, 2001) is an objective measure of sound quality and also reflects a comparison between clean and processed speech. The scale ranges from −0.5 to 4.5, and higher values represent better predicted sound quality. These measures were developed and validated for NH (and not HI) listeners. Finally, SNR is the estimated SNR of the signal prior to and following processing.

Apparent from Table I is that substantial increases in objective values as a result of noise-reduction processing were obtained at each SNR tested (compare Mixture versus X ms values). These data also show that the largest objective values (in bold) were obtained in the 20-ms latency condition—this is true for each of the measures. This slight advantage in the 20-ms condition over the 10- or 5-ms conditions reflects the slight algorithmic advantage associated with longer time frames.

TABLE I. Objective measures.

Test SNR		−5 dB	−2 dB	0 dB	3 dB
STOI	Mixture	54.0	62.6	68.4	76.6
	20 ms	**77.9**	**85.4**	**88.6**	**92.1**
	10 ms	76.2	83.9	87.5	91.4
	5 ms	74.0	82.8	86.6	90.8
PESQ	Mixture	1.23	1.43	1.59	1.77
	20 ms	**2.05**	**2.40**	**2.59**	**2.82**
	10 ms	1.99	2.33	2.51	2.76
	5 ms	1.87	2.25	2.45	2.71
SNR	Mixture	−5.0	−2.0	0.0	3.0
	20 ms	**6.7**	**9.0**	**10.3**	**12.2**
	10 ms	6.1	8.3	9.7	11.6
	5 ms	5.3	7.6	8.9	10.9

TABLE II. Objective measures.

Test SNR		−2 dB	0 dB	3 dB
HASPI	Mixture	0.139	0.256	0.534
	20 ms	**0.749**	0.776	0.796
	10 ms	0.742	**0.781**	**0.802**
	5 ms	0.715	0.773	**0.802**
HASQI	Mixture	0.091	0.121	0.184
	20 ms	**0.416**	**0.470**	**0.537**
	10 ms	0.394	0.449	0.519
	5 ms	0.377	0.434	0.507

Table II contains values for the objective measures Hearing-Aid Speech Perception Index (HASPI v2; Kates and Arehart, 2021) and the Hearing-Aid Speech Quality Index (HASQI v2, Kates and Arehart, 2014), for the SNRs heard by the HI listeners. As with STOI and PESQ, these indices involve a comparison between a test condition and a clean-speech reference. However, these indices include a model of the auditory periphery that can be adjusted to represent hearing loss. Both range from 0.0 to 1.0, with higher values representing higher predicted intelligibility/sound quality. Values were calculated using audiometric thresholds averaged across the HI listeners (20 ears).

Apparent from Table II are large increases in predicted intelligibility and sound quality as a result of algorithm processing. For sound quality (bottom half of Table II), the largest objective values are again obtained at a latency of 20 ms. However, for predicted intelligibility (top half of Table II), this is not uniformly the case. Instead, values are similar overall, and the largest predicted intelligibility values are distributed among the different latencies.

B. Human intelligibility

Sentence recognition (intelligibility) was expressed as the percentage of words correctly reported, and these scores were transformed to rationalized arcsine units (RAUs; Studebaker, 1985) to normalize variance prior to statistical analysis. Figures 5 and 6 display mean intelligibility scores for individual HI and NH listeners, respectively. It is apparent from Fig. 5 that appreciable increases in scores were observed from speech-in-noise to the processed conditions, so long as baseline scores in quiet were low enough to allow improvement. Also, apparent is no regular pattern to the different algorithmic-latency scores. Apparent from Fig. 6 is the better overall performance of the NH listeners.

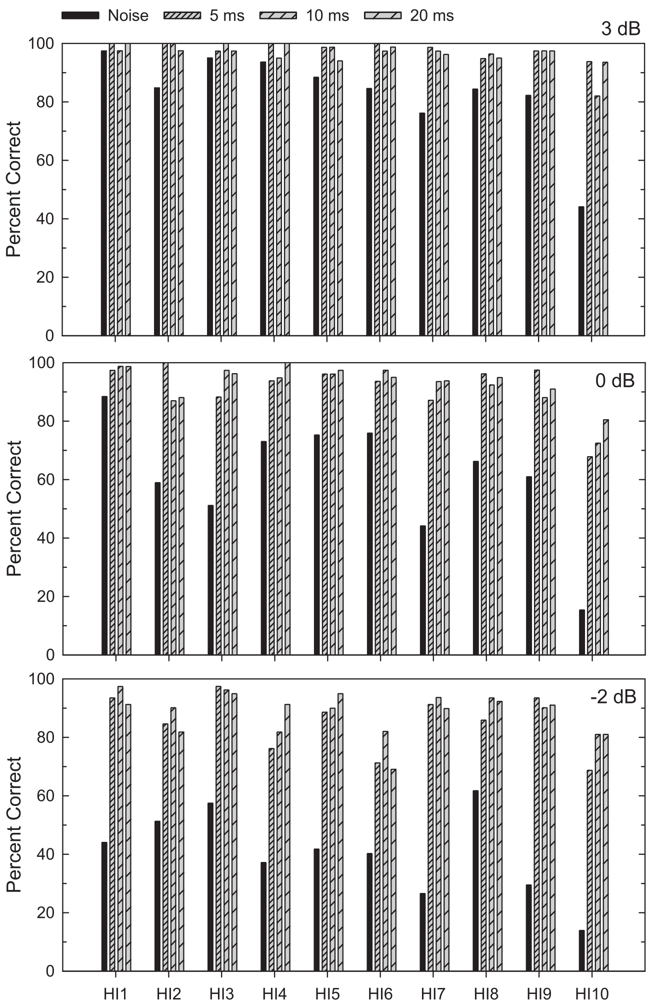

FIG. 5. Displayed are scores for individual listeners with hearing loss. Percent-correct sentence intelligibility is shown for speech in noise and when processed by the noise-reduction algorithm having an algorithmic latency of 20, 10, or 5 ms. Scores for each of the SNR conditions are displayed in different panels.

Figure 7 displays group mean scores for the HI and NH listeners in each condition. The average intelligibility improvements for the HI listeners were 47, 31, and 14% points at the three SNRs (−2, 0, 3 dB). These results generally replicate those of Healy *et al.* (2023), where an ARN having a frame length of 20 ms (but a 2-ms hop) also increased HINT scores in babble. The intelligibility increases in that prior study were larger (58 and 53% points at −2 and 0 dB SNR), likely attributable in part to the participation of a different group of HI listeners having a greater average degree of hearing loss and consequent lower baseline scores in noise.

Of particular relevance for the current study, scores in the algorithmic-latency conditions are similar. The three latency scores at each SNR are within 3% points of one another, on average across SNRs.

Group mean benefit for NH listeners was far smaller, as a direct result of higher baseline scores (where benefit = intelligibility following noise reduction minus that in noise). One notable facet of the data from the NH listeners involves the lower baseline score at the least favorable SNR, as expected, but also lower algorithm-processed scores. Low SNRs required to produce low baseline scores for NH listeners (e.g., −5 dB) produce challenges for noise-reduction systems, as observed currently. Fortunately, such low SNRs are not common in real-world environments (see Sec. III).

The data from the HI listeners were subjected to a 2-way repeated-measures analysis of variance (ANOVA) (3 SNRs × 4 processing conditions). The test revealed significant main effects of SNR [$F(2, 18) = 38.2$, $p < 0.001$] and

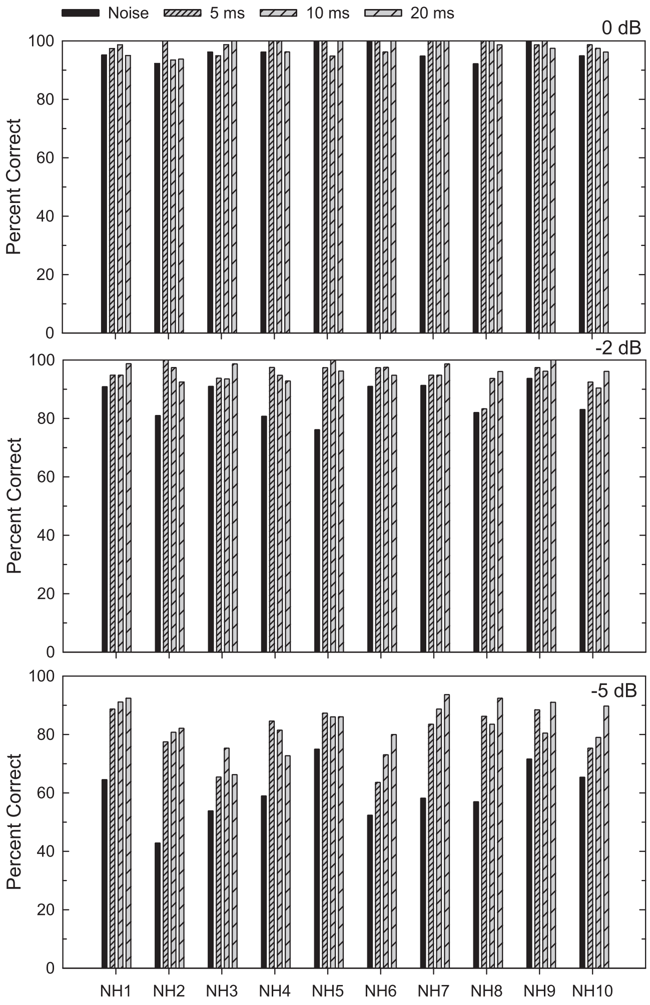

FIG. 6. As Fig. 5, but for individual listeners with normal hearing.

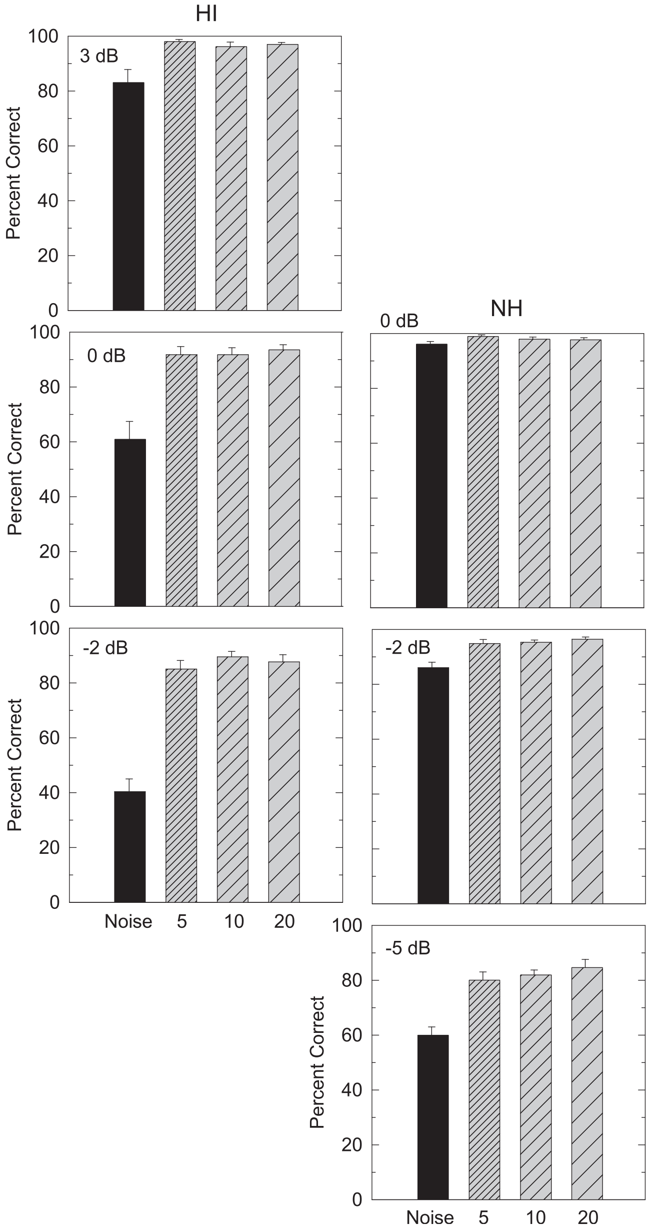

FIG. 7. Group-mean sentence intelligibility (and std. errors) for speech in noise and when processed by the noise-reduction algorithm having an algorithmic latency of 20, 10, or 5 ms. Means for the listeners with hearing loss are displayed in the left panels and those for listeners with normal hearing are displayed in the right panels. The different SNR conditions are displayed in different panels.

processing [$F(3, 27) = 88.1$, $p < 0.001$], as well as a significant interaction [$F(6, 54) = 10.4$, $p < 0.001$]. Incremental *post hoc* tests (Holm-Sidak) indicated that scores in the noisy condition were different from those in each of the corresponding processed conditions ($p < 0.001$), and that scores in the three processing-latency conditions did not differ significantly from one another ($p > 0.05$).

A similar ANOVA for the NH-listener scores yielded the same pattern (significant main effects of SNR [$F(2, 18) = 100.0$, $p < 0.001$] and processing [$F(3, 27) = 28.5$, $p < 0.001$], and a significant interaction [$F(6, 54) = 5.3$, $p < 0.001$]). Also, as for the HI listeners, *post hoc* tests revealed that scores in the noisy conditions were different from those in the corresponding processed conditions ($p < 0.001$), and that scores in each of the three algorithmic-latency conditions did not differ significantly from one another ($p > 0.05$). For the NH listeners, this pattern held for SNRs of −5 and −2 dB, but not at 0 dB where scores did not differ significantly in any condition ($p > 0.05$).

A priori planned comparisons (uncorrected paired *t*-tests) were also performed, as a way to probe possible differences between algorithmic-latency scores with greater sensitivity. For the HI listeners, intelligibility in every processed condition was higher than in the corresponding noise condition (i.e., benefit, $p < 0.001$). Scores in the three latency conditions did not differ from one another at any SNR ($p > 0.05$). The exception was at an SNR of −2 dB between 5 and 10 ms [$t(9) = 2.8$, $p = 0.02$].

Planned comparisons performed on the data for the NH listeners revealed a similar pattern. Intelligibility in every processed condition was higher than in the corresponding noise condition ($p < 0.001$), and scores in the three latency conditions did not differ from one another at any SNR ($p > 0.05$). The exception occurred at the most favorable SNR of 0 dB, where baseline scores were high and benefit was significant ($p < 0.05$) for only the 5-ms latency condition.

An additional comparison was made between unprocessed speech in noise for NH listeners versus algorithm-processed scores for HI listeners. This comparison reflects the extent to which we are able to eliminate the speech-in-noise deficit for HI listeners by producing conditions that allow their performance to match that of young, healthy ears. It also mimics a communication situation in which a typical hearing-aid user and a NH individual are communication partners in a particular setting—can the older HI individual perform as well as their young NH counterpart, if given access to the current DNN noise reduction? At the common SNR of 0 dB, intelligibility for the HI listeners, averaged across processing latencies, was within four percentage points (below) of the intelligibility produced by the NH listeners (HI, 92.4%; NH, 96.2%). At the common SNR of −2 dB, where scores are farther from the performance ceiling, averaged intelligibility for the HI listeners was within two percentage points (above) of that produced by NH listeners (HI: 87.5%, NH: 86.1%). Planned comparisons (uncorrected *t*-tests) on RAU-transformed scores indicated that the difference between NH intelligibility in noise versus HI intelligibility after algorithm processing was nonsignificant at each latency for both SNRs ($p > 0.05$).

IV. DISCUSSION

The current results replicate those of Healy *et al.* (2023), where a frame length of 20 ms was used. They also show that the latency of the DNN can be reduced by a factor of 2 or 4 without affecting intelligibility for human listeners. The objective measures in Tables I and II indicate that the waveforms output by the longer-frame ARN are slightly more veridically accurate (more similar to the original speech prior to mixing with noise) relative to the waveforms output by the shorter-frame ARNs. This was true for most, but not all, metrics. However, the critical question is whether these slight increases in acoustic accuracy are perceptually meaningful for human listeners—do they result in better intelligibility? That answer appears to be no.

Longer time frames can have algorithmic advantages. The larger frame offers additional contextual cues for the neural network to exploit within each frame. This context advantage exists despite the use of RNNs that integrate prior time-frame information (which also provides context). In frequency-domain processing, such as short-time Fourier transform, larger time frames produce greater frequency resolution, at the expense of resolution in the time domain. This higher frequency resolution is often desirable despite the trade-off, as the extraction of speech cues from noise often requires good frequency resolution (e.g., Apoux and Healy, 2009). A second potential advantage of longer frames exists when the frame shift is held constant. This is what was done currently in order to compare algorithms having similar computational complexity. A given time point of a signal will fall within a larger number of frames when those frames are larger, following the overlap-add step in reconstructing the output noise-reduced waveform signal (see Fig. 2). Thus, the output signal at that time point will result from the average of a larger number of estimates—a statistical pooling effect. However, despite these potential advantages associated with longer time frames, the current study suggests that the currently employed smaller frames appear to be equivalent to longer time frames for human intelligibility.

Low SNRs were used for the NH listeners in order to reduce baseline scores in quiet and allow benefit to be observed. The lowest SNR value of −5 dB was successful in reducing baseline scores and yielded large benefits of noise reduction, but it also reveals the noise-reduction challenge posed by low SNRs. The relative lack of speech information and dominance of the signal by noise pose difficulty for any noise-reduction system. Accordingly, the noise-reduced scores in this condition were lower than at other SNRs. Fortunately, noise-reduction does not have to operate at very low SNRs in order to provide real-world benefit. This occurs for two reasons. First, typical hearing-aid users display poor speech understanding at higher SNRs (e.g., 0 dB) where the network is able to operate more effectively and improve scores to values over 90% correct (see Fig. 7). Listeners using cochlear implants struggle to understand speech at even higher SNRs (e.g., 10 dB; Gifford and Revit, 2010; Dorman and Gifford, 2017; Abdel-Latif and Meister, 2021), where the noise-reduction task is far easier. Second, low SNRs such as −5 dB do not often occur in the real-world (Pearsons *et al.*, 1977; Smeds *et al.*, 2015; Wu *et al.*, 2018; Benítez-Barrera *et al.*, 2020; Mansour *et al.*, 2021). Although this can be difficult to understand given the high noise levels of some communication environments, it can be understood in terms of the tendency for communication partners to compensate for background noise by modifying speaking intensity and distance. The most extreme environments (e.g., an amplified music concert) provide an example of how individuals correct for background noise—in this environment, the talker will speak loudly into the listener's ear in order to produce an acceptable SNR. Variations of this (albeit to lesser extents) occur constantly (Lombard, 1911; Pearsons *et al.*, 1977). Accordingly, real-world benefits can be obtained despite the algorithmic challenges associated with very low SNRs, which typically do not occur.

It is emphasized that the current study involves monaural or single-microphone noise reduction, in which the speech and noise are picked up by the same single microphone or mixed into a single channel. This arrangement represents the most flexible and universally applicable, but also the most challenging situation for a noise-reduction algorithm. Any additional cues, including those associated with directionality of signals and noises that are not co-located, will add information and make the task easier (Kalkhorani and Wang, 2024). The current results, therefore, apply to situations that are applicable to any environmental situation and that can be considered "worst case" or most challenging for a noise-reduction network.

https://doi.org/10.1121/10.0037197

Deep-learning based solutions have become the future of noise reduction. Modern devices are capable of performing vast computations, which were impossible only a few years ago. A monaural approach has been employed by major hearing aid manufacturers. For example, Oticon uses a monaural DNN for noise reduction that receives input following scene classification and single or dual directional microphones (Oticon, 2022). The pre-processing improves the SNR provided to the monaural DNN. Phonak has adopted a similar approach involving scene classification followed by microphone input adjustment. This signal is then fed to a monaural DNN. Their current frequency-domain DNN operates in the complex domain (involving both real and imaginary parts, Williamson *et al.*, 2016) and has 4.5×10^6 parameters (Hasemann and Kryloa, 2024). The current study, and those like it, may be of value in the design of future devices that best navigate the various constraints that exist when deploying these systems in everyday devices (e.g., Park *et al.*, 2025) by allocating the available processing time most effectively.

ACKNOWLEDGMENTS

This work was supported in part by the National Institute on Deafness and Other Communication Disorders (R01DC015521 and R01DC012048).

AUTHOR DECLARATIONS

Conflict of Interest

The authors declare no conflict of interest.

Ethics Approval

Approval was obtained from The Ohio State University Institutional Review Board. Informed consent was obtained from all participants.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Abdel-Latif, K. H. A., and Meister, H. (**2021**). "Speech recognition and listening effort in cochlear implant recipients and normal-hearing listeners," Front. Neurosci. **15**, 1–13.

Apoux, F., and Healy, E. W. (**2009**). "On the number of auditory filter outputs needed to understand speech: Further evidence for auditory channel independence," Hear. Res. **255**, 99–108.

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

Baskent, D., and Bazo, D. (**2011**). "Audiovisual asynchrony detection and speech intelligibility in noise with moderate to severe sensorineural hearing impairment," Ear Hear. **32**, 582–592.

Benítez-Barrera, C. R., Grantham, D. W., and Hornsby, B. W. Y. (**2020**). "The challenge of listening at home: Speech and noise levels in homes of young children with hearing loss," Ear Hear. **41**, 1575–1585.

Bramsløw, L., and Beck, D. L. (**2021**). "Deep neural networks in hearing devices," Hear Rev. **28**(7), 28–32.

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.

Conrey, B., and Pisoni, D. B. (**2006**). "Auditory-visual speech perception and synchrony detection for speech and nonspeech signals," J. Acoust. Soc. Am. **119**, 4065–4073.

Dorman, M. F., and Gifford, R. H. (**2017**). "Speech understanding in complex listening environments by listeners fit with cochlear implants," J. Speech. Lang. Hear. Res. **60**, 3019–3026.

Gifford, R. H., and Revit, L. J. (**2010**). "Speech perception for adult cochlear implant recipients in a realistic background noise: Effectiveness of preprocessing strategies and external options for improving speech recognition in noise," J. Am. Acad. Audiol. **21**, 441–488.

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.

Hasemann, H., and Kryloa, A. (**2024**). "Revolutionary speech understanding with spheric speech clarity," Phoank Insight.

Hay-McCutcheon, M. J., Pisoni, D. B., 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.

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*, pp. 770–778.

Healy, E. W., Johnson, E. M., Pandey, A., and Wang, D. (**2023**). "Progress made in the efficacy and viability of deep-learning-based noise reduction," J. Acoust. Soc. Am. **153**, 2751–2768.

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**, 1735–1780.

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

Kalkhorani, V. A., and Wang, D. L. (**2024**). "TF-CrossNet: Leveraging global, cross-band, narrow-band, and positional encoding for single- and multi-channel speaker separation," IEEE/ACM Trans. Audio. Speech. Lang. Process. **32**, 4999–5009.

Kates, J. M., and Arehart, K. H. (**2014**). "The hearing-aid speech quality index (HASQI) version 2," J. Audio Eng. Soc. **62**, 99–117.

Kates, J. M., and Arehart, K. H. (**2021**). "The hearing-aid speech perception index (HASPI) version 2," Speech Commun. **131**, 35–46.

Kearns, J. (**2014**). "LibriVox: Free public domain audiobooks" (Last viewed July 8, 2025).

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

Lombard, E. (**1911**). "Le signe de l'elevation de la voix" ("The sign of the elevation of the voice"), Ann. Dis. Ear Larynx. **37**, 101–119.

Mansour, N., Marschall, M., May, T., Westermann, A., and Dau, T. (**2021**). "A method for realistic, conversational signal-to-noise ratio estimation," J. Acoust. Soc. Am. **149**, 1559–1566.

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.

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.

Oticon (**2022**). "MoreSound Intelligence^{TM} for single-microphone custom hearing aids," Oticon Tech Paper.

Panayotov, V., Chen, G., Povey, 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, 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 2020*, 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.

Park, S., Lee, S., Park, J., Choi, H.-S., Lee, K., and Jeon, D. (**2025**). "A real-time speech enhancement processor for hearing aids in 28-nm CMOS," IEEE J. Solid-State Circuits **60**, 1830–1843.

Paul, D. B., and Baker, J. (**1992**). "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of Speech and Natural Language: Proceedings of a Workshop Held at Harriman*, February 23–26, New York.

Pearsons, K. S., Bennett, R. L., and Fidell, S. A. (**1977**). "Speech levels in various noise environments," Office of Health and Ecological Effects, Office of Research and Development, US EPA, Washington, DC.

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

Santurette, S., Ng, E. H. N., Jensen, J. J., and Loong, B. M. K. (**2020**). "Oticon more clinical evidence," Oticon Whitepaper.

Smeds, K., Wolters, F., and Rung, M. (**2015**). "Estimation of signal-to-noise ratios in realistic sound scenarios," J. Am. Acad. Audiol. **26**, 183–196.

Stone, M. A., and Moore, B. C. J. (**1999**). "Tolerable hearing aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses," Ear Hear. **20**, 182–192.

Stone, M. A., and Moore, B. C. J. (**2002**). "Tolerable hearing aid delays. II. Estimation of limits imposed during speech production," Ear Hear. **23**, 325–338.

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.

Stuart, A., Kalinowski, J., Rastatter, M. P., and Lynch, K. (**2002**). "Effect of delayed auditory feedback on normal speakers at two speech rates," J. Acoust. Soc. Am. **111**, 2237–2241.

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, L., and Polosukhin, I. (**2017**). "Attention is all you need," Adv. Neural Inf. Process. Syst. **30**, 1–11.

Wang, D. (**2017**). "Deep learning reinvents the hearing aid," IEEE Spectrum **54**, 32–37.

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.

Wright, A., Keller, M., Kuehnel, V., Latzel, M., and Seitz-Paquette, K. (**2024**). "Spheric speech clarity applied DNN processing to significantly improve speech understanding from any direction and reduce the listening effort," Pnonak Field Study News 1–6.

Wu, Y. H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., and Oleson, J. (**2018**). "Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss," Ear Hear. **39**, 293–304.