Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction

Mahmoud Keshavarzi, Tobias Goehring, Richard E. Turner, et al.

Citation: The Journal of the Acoustical Society of America **145**, 1493 (2019); doi: 10.1121/1.5094765
View online: https://doi.org/10.1121/1.5094765
View Table of Contents: https://asa.scitation.org/toc/jas/145/3
Published by the Acoustical Society of America

Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction

Mahmoud Keshavarzi[a)]
Department of Psychology, University of Cambridge, Cambridge, United Kingdom

Tobias Goehring
MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, United Kingdom

Richard E. Turner
Department of Engineering, University of Cambridge, Cambridge, United Kingdom

Brian C. J. Moore
Department of Psychology, University of Cambridge, Cambridge, United Kingdom

(Received 15 November 2018; revised 9 February 2019; accepted 1 March 2019; published online 25 March 2019)

The effects on speech intelligibility and sound quality of two noise-reduction algorithms were compared: a deep recurrent neural network (RNN) and spectral subtraction (SS). The RNN was trained using sentences spoken by a large number of talkers with a variety of accents, presented in babble. Different talkers were used for testing. Participants with mild-to-moderate hearing loss were tested. Stimuli were given frequency-dependent linear amplification to compensate for the individual hearing losses. A paired-comparison procedure was used to compare all possible combinations of three conditions. The conditions were: speech in babble with no processing (NP) or processed using the RNN or SS. In each trial, the same sentence was played twice using two different conditions. The participants indicated which one was better and by how much in terms of speech intelligibility and (in separate blocks) sound quality. Processing using the RNN was significantly preferred over NP and over SS processing for both subjective intelligibility and sound quality, although the magnitude of the preferences was small. SS processing was not significantly preferred over NP for either subjective intelligibility or sound quality. Objective computational measures of speech intelligibility predicted better intelligibility for RNN than for SS or NP. © *2019 Acoustical Society of America*.
https://doi.org/10.1121/1.5094765

[JGB] Pages: 1493–1503

I. INTRODUCTION

A major complaint of people with sensorineural hearing loss is difficulty in understanding speech in the presence of background sounds (Plomp, 1978; Moore, 2007). The difficulty relative to people with normal hearing is especially pronounced for background sounds containing one or more competing talkers (Festen and Plomp, 1990; Peters *et al.*, 1998). While directional microphones and beamformers in hearing aids can improve the ability to understand speech in the presence of spatially distributed interfering sounds (Hawkins and Yacullo, 1984; Bentler *et al.*, 2008; Launer *et al.*, 2016; Picou and Ricketts, 2017), the benefits in everyday life have been found to be modest (Picou *et al.*, 2014), partly because it is not always possible or practical to "point the beam" at the target talker. This paper compared the effects of babble reduction produced using two algorithms that do not depend on spatial separation of the target and background sounds. The two algorithms were spectral subtraction (SS) and processing using a deep recurrent neural network (RNN). These two types of algorithms are described next.

Most hearing aids incorporate some form of noise-reduction algorithm that operates independently of the direction of the sound sources. Some of these algorithms are based on SS (Hamacher *et al.*, 2005), which has four steps: (1) estimation of the short-term spectrum of the noisy speech and of the noise alone; (2) subtraction of the estimated noise spectrum from the spectrum of the speech + noise; (3) reconstruction of the speech from the resulting spectrum, using the original noisy phases; (4) repetition of this process for a series of overlapping time frames. SS can be applied to the broadband input signal. Alternatively, the noisy signal can be split into multiple frequency bands or channels, and the noise spectrum can be estimated independently in each channel. In some simplified systems that are used in hearing aids, the gain in each channel is progressively reduced as the estimated noise level in that channel increases.

The most error-prone stage in SS is estimation of the noise spectrum. There are two main approaches to this, and both are based on the assumption that the noise is stationary or nearly so (Kates, 1987). Hence, these approaches are ineffective when the background consists of one or very few talkers, but they can improve the signal-to-background ratio (at the potential cost of the introduction of artifacts) when the background is a reasonably steady noise (e.g., car noise)

a)Electronic mail: mahmoud.keshavarzi.ir@ieee.org

or multiple talkers (babble). It is not clear, however, how many talkers are required for a babble to be steady enough for reliable estimation of its spectrum.

The first approach focuses on voice-activity detection (VAD), which is the process of discriminating between time periods when speech is present and when it is absent. The discrimination is usually based on features such as short-time energy and the pattern of zero-crossings of the input signal (Loizou, 2007). The estimates of noise power and spectrum are updated only during non-speech segments, and it is assumed that the noise power and spectrum remain the same during speech segments. The second approach estimates the noise spectrum continuously and does not need to detect voice activity. This approach is based on the assumption that speech contains time-frequency (TF) segments where the energy is very low, so the output of those segments is dominated by the noise (Loizou, 2007). Hence, the noise spectrum in a given frequency region can be estimated during the level minima in that frequency region.

Several researchers (Levitt *et al.*, 1993; Jamieson *et al.*, 1995; Arehart *et al.*, 2003; Alcántara *et al.*, 2003; Natarajan *et al.*, 2005; Hu and Loizou, 2007a; Brons *et al.*, 2012) have investigated both subjective and objective effects of SS on intelligibility, listening effort and sound quality. Generally, little or no improvement in speech intelligibility has been found, although some studies showed improvements in sound quality for steady noise backgrounds.

There are two major limitations of SS. The first is that accuracy in estimating the background spectrum is poor when the background is not stationary, for example when it is babble or traffic noise. Although many hearing aids include some form of scene classification (Launer *et al.*, 2016; Moore *et al.*, 2016), the categories are usually rather coarse, for example, speech in quiet, speech in noise, music, or noise alone. Hence, in practice, SS is applied to most types of noise background, and it might have deleterious effects for backgrounds such as babble. The second limitation is that the speech is reconstructed using the original noisy phase, and this generates processing artifacts such as musical noise (Loizou, 2007). Processing schemes that jointly enhance the magnitude and the phase have been developed (Krawczyk and Gerkmann, 2014), but these are complex and have not, to our knowledge, been implemented in hearing aids.

An alternative approach to SS is to employ supervised machine-learning (ML) techniques. Techniques of this sort have been widely used and artificial neural networks (ANN) with three or more layers have led to significant progress in many supervised ML tasks (Hinton *et al.*, 2006). ANN models consist of a multi-layer structure that is used to transform n-dimensional input data into m-dimensional output data representations with arbitrary degrees of abstraction. Each layer of the model consists of a number of nodes whose parameters (e.g., connection weights and bias values) are fine-tuned during a training procedure, for example in a supervised fashion with pre-labelled output data. Especially for complex, non-linear tasks, such as single-channel speech segregation, ANNs represent a promising technique for predicting the speech-dominant parts of a speech + noise signal. We consider here ANNs with two architectures: feed-forward deep neural networks (DNN) and recurrent neural networks (RNN). While DNNs process input data exclusively in a feed-forward manner to calculate output data representations, RNNs make use of recursive connections between layers that allow for the build-up of an internal temporal memory. Hence, RNNs make use of the inter-dependence of data samples across time (Graves *et al.*, 2013; Lipton *et al.*, 2015) and have achieved the best results for the detection and recognition of temporal patterns in speech signals and time-series data in general. One highly successful RNN architecture, the long short-term memory (LSTM) model, was proposed by Hochreiter and Schmidhuber (1997) to model the long-range dependencies of temporal sequences in a more accurate way than with conventional RNNs (Sak *et al.*, 2014). The LSTM model was therefore used in this study.

Several ML-based approaches using DNNs to segregate speech from background sounds have shown improved perception of speech in noise for normal-hearing and hearing-impaired listeners and users of cochlear implants (Healy *et al.*, 2013; 2015; Chen *et al.*, 2016; Goehring *et al.*, 2017; Monaghan *et al.*, 2017; Bramsløw *et al.*, 2018). In the last few years, the use of RNNs for segregation of speech from noise has led to improvements over DNN-based methods, in both estimation accuracy and generalization performance (Weninger *et al.*, 2015; Huang *et al.*, 2015; Chen and Wang, 2017; Kolbæk *et al.*, 2017). ML-based approaches, in particular RNNs, appear to be good candidates for reducing babble noise, given their success in reducing non-stationary noise maskers at low signal-to-noise ratios. However, to our knowledge, they have not yet been evaluated with this application in mind. While RNN-based methods have been shown to give better performance than DNN-based methods when using computational measures of intelligibility (Chen and Wang, 2017), the benefits for the perception of speech in babble by human listeners remain unclear.

This paper investigated the effects of babble reduction produced using two algorithms. The first was an ML algorithm based on a deep (multi-layer) LSTM model. For brevity, this is referred to hereafter simply as the RNN. The RNN was first trained to predict the ideal ratio mask (IRM, a soft-gain function based on the ideal Wiener filter in the TF domain), using recordings of speech that were combined with babble. The clean (babble-free) speech was used to estimate the target IRM. The trained model was then used to process the speech in babble, so as to attenuate TF segments with low speech-to-babble ratio (SBR) while retaining segments with high SBR. The second algorithm used multi-band SS, as described by Kamath and Loizou (2002). This algorithm was included as a comparison condition because similar algorithms have been used in hearing aids and because, as described earlier, SS processing in hearing aids is likely to be applied to many types of background, not just steady noise. Both algorithms were frame based, using 5-ms frames, chosen here because for application in hearing aids a low time delay is required (Stone and Moore, 1999, 2002) and this requires short frames. A control condition used speech in babble with no processing (NP). Each condition

was compared pairwise with the two other conditions to assess subjective preferences for intelligibility and sound quality, using participants with hearing loss. Subjective judgments of sound quality were used since signal processing that leads to poor sound quality is unlikely to be accepted by users of hearing aids. Subjective intelligibility was assessed to speed testing time and because it has been shown to be highly correlated with objective intelligibility (Cienkowski and Speaks, 2000).

II. METHOD

A. Participants

Eight native British English-speaking participants with hearing loss took part in the experiment. The number of participants was chosen to be sufficient to reveal small-to-moderate effect sizes, based on previous studies using similar methods to assess preferences (Moore and Sek, 2013; 2016; Keshavarzi *et al.*, 2018). Audiometric thresholds were measured for audiometric frequencies from 0.25 to 8 kHz, using a Grason-Stadler GSI-61 audiometer (Eden Prairie, MN) and Telephonics TDH50 headphones (Huntington, NY). Only the better-hearing ear of each participant was tested (based on the average threshold across 0.5–4 kHz). The sex, age, and audiometric thresholds for the test ears of the hearing-impaired participants are shown in Table I. The participants were chosen to be representative of the general population of people with mild-to-moderate hearing loss, and most had greater hearing loss at high than at low frequencies. Seven participants were regular users of hearing aids, but the hearing aids were removed for testing. The experiment lasted about 1 h for each participant, and participants were paid for taking part as well as receiving reimbursement for travel expenses. Ethical approval was granted by the National Research Ethics Service, East of England (approval 06/Q0108/101).

B. Speech and babble materials

The speech stimuli used in this study were selected from a British English multi-speaker corpus named CSTR VCTK (Centre for Speech Technology Voice Cloning Toolkit), developed by researchers at the University of Edinburgh and available at http://homepages.inf.ed.ac.uk/jyamagis/release/VCTK-Corpus.tar.gz. The speakers had a variety of accents.

TABLE I. Sex, age and audiometric thresholds (in dB HL) of the test ears of the hearing-impaired participants.

		Frequency, kHz							
Sex	Age, years	0.125	0.5	1	2	3	4	6	8
Male	73	10	20	25	40	50	55	50	70
Female	82	40	50	55	50	55	55	50	75
Female	46	5	20	30	45	45	50	30	30
Male	71	15	20	20	10	30	45	50	60
Female	56	0	10	10	45	60	70	85	75
Male	62	15	30	20	40	55	50	55	55
Male	77	10	5	5	25	45	70	65	65
Female	72	10	10	10	35	45	50	40	40

This set of recordings was chosen to ensure that the RNN would generalize to a wide range of talkers. The recordings were made using a 96-kHz sampling rate and 24-bit resolution. They were subsequently converted by the creators of the corpus to 48-kHz sampling with 16-bit resolution. For this study, the sentences were further down-sampled to 16 kHz. We used 1600 sentences from 80 speakers (40 female and 40 male) for training the RNN and 300 sentences from six other speakers (three female and three male) for evaluating the performance of the RNN and SS algorithms using objective measures. For subjective evaluations of speech intelligibility and sound quality, we used eight sentences (two from each of two female and two male talkers) randomly chosen from the 300 used for the objective measures.

The babble was taken from a recording made by Auditec St Louis and consisted of a mixture of speech from 20 American-English speakers (eight male and 12 female). For the subjective evaluations, SBRs of 0 and 5 dB were used. For training of the RNN and for the objective evaluations, SBRs of −5, 0, 5, and 10 dB were used. The SBRs were based on the long-term average of the speech and babble levels. Independent 2-min samples of the babble were used for training and testing.

C. Machine-learning algorithm and RNN

As described in the introduction, the LSTM model variant of the RNN proposed by Hochreiter and Schmidhuber (1997) was used here. As illustrated in Fig. 1, the RNN consisted of an input layer, three LSTM layers with 128 units each, and a fully connected output layer with 64 units. This architecture is typical of RNN-based speech-segregation systems, with the modelling of increasingly abstract representations of the input features and their temporal dependencies within the three LSTM layers and a recombination of the information for the final output estimation in the fully connected layer. The RNN used fewer units in each of the hidden layers than previously proposed RNN architectures (Chen and Wang, 2017) to reduce its computational complexity and memory requirements for potential applications in practice. The signal was segmented into frames with a duration of 5 ms (80 samples) and an overlap of 50% (40 samples) between successive frames. The RNN processed a five-time-step input and each step corresponded to features extracted from one frame of speech; steps 1, 2, 3, 4, and 5 corresponded to successive frames j-4, j-3, j-2, j-1, and j, respectively. We used only four past frames and the current frame as input to keep the RNN processing causal, to minimize the delay and memory requirements for the processing, and to keep the size of the RNN small while allowing for a temporal window at the input layer comparable to the frame length used in other studies (often around 20–25 ms). The choice of five frames was also based on results from a previous study (Keshavarzi *et al.*, 2018), in which we used four input frames. The RNN took acoustic features as its inputs and predicted the IRM based on the ideal Wiener filter in the TF domain (Wang *et al.*, 2014; Delfarah and Wang, 2017; Healy *et al.*, 2017).

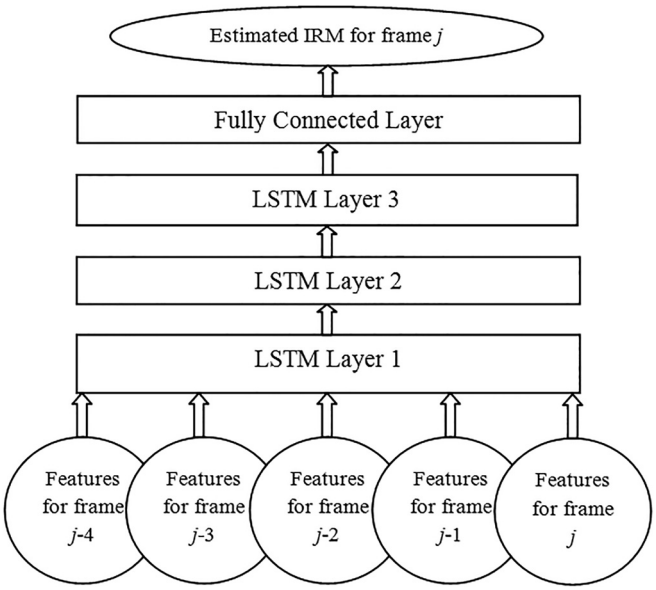

FIG. 1. Schematic diagram of the LSTM network used to estimate the IRM.

The speech in babble was considered as

$$x(t) = s(t) + v(t), \tag{1}$$

where t is time, x is the speech in babble, s is the clean speech, and v is the babble, respectively. For training, SBRs of −5, 0, +5, and + 10 dB were used, and these were intermixed in random order during training. The IRM for the jth frame and mth channel was defined as

$$IRM(j,m) = \sqrt{\frac{S^2(j,m)}{S^2(j,m) + V^2(j,m)}}, \tag{2}$$

where $S(j, m)$ and $V(j, m)$ represent the magnitudes of $s(t)$ and $v(t)$ in the mth channel of frame j, respectively (Delfarah and Wang, 2017). Essentially, this equation means that the gain applied to each channel and frame increased from 0 to 1 as the ratio of the speech signal power to the total power in that channel and frame increased as a function of the local SBR. The IRM has been widely used in previous noise-reduction studies to attenuate speech-absent TF units while preserving speech-dominant TF units. Note that the babble-free speech was used to obtain $S(j,m)$ for calculating the IRM training data, but this is not available in practice and needs to be estimated by the RNN.

The features used to train the RNN were the energy in each frame at the output of a 64-channel gammatone filter bank (Patterson *et al.*, 1995) with filter center frequencies uniformly spaced on the ERB_N-number scale (Glasberg and Moore, 1990) and spanning the range 50–8000 Hz. Similar features have been used previously in DNN- and RNN-based noise reduction (Chen *et al.*, 2016; Chen and Wang, 2017). The gammatone features were calculated using a fast Fourier transform with 5-ms Hanning windowed frames and 50% overlap. No smoothing of the spectrum was applied.

The ML frameworks TFlearn and Tensorflow, which are freely available, were used to construct, train, and test the RNN (Abadi *et al.*, 2016; Tang, 2016). The resilient back-propagation algorithm "RMSprop" (Riedmiller and Braun, 1993) was used as the optimizer function in the training algorithm, with the goal of minimizing the mean square error. The learning rate started at 0.001 and was decreased by a factor of 0.999 in each training run (a run was based on using all of the training data once). The batch size was 100 and 20 training runs were performed. After the RNN had been trained, the estimate of the IRM for each frame was used to process the speech plus babble for that frame in the TF domain so as to attenuate TF segments with low SBR while maintaining the level of segments with high SBR. The modified magnitudes from the processed frames were combined with the original noisy phases and the output signals were constructed using the overlap-add operation.

D. Spectral subtraction

The multi-channel SS algorithm described by Kamath and Loizou (2002) was used here. We used the implementation from the freely available MATLAB code provided with the book by Loizou (2007). The algorithm divides the speech spectrum into N non-overlapping frequency channels, and SS is performed independently in each channel. Here, N was set to four. The parameters of the processing used here were similar to those used by Kamath and Loizou (2002) except for the frame duration, which was chosen to be the same as for the RNN, i.e., 5 ms, whereas Kamath and Loizou used 20-ms frames. To assess whether the use of shorter frames would be likely to have deleterious effects, we used three objective intelligibility metrics to predict the intelligibility of speech in babble for SBRs from −5 to 10 dB. As is described in more detail later, all three metrics predicted slightly better intelligibility for the 5-ms frames than for the 20-ms frames, so it seems unlikely that the use of shorter frames here had deleterious effects.

The processing was frame-based, using the overlap-add procedure (Allen, 1977). Each frame used a 5-ms hamming window and there was a 2.5-ms overlap between frames. Each frame was zero-padded at the start and end to give 128 samples and a fast Fourier transform (FFT) was used to determine the short-term spectrum. The short-term spectrum was smoothed by taking a weighted average of preceding and following frames, as defined by

$$\bar{X}_j(k) = \sum_{i=-M}^{M} W_i X_{j-i}(k), \tag{3}$$

where j is the frame index, k is the frequency index, X_{j-i} is the spectrum of the speech-plus-babble for frame j-i, $\bar{X}_j$ is the smoothed short-term spectrum of the speech-plus-babble, $M = 2$, and $W = (0.09,0.25,0.32,0.25,0.09)$ defines the weight applied to frames $-M$ to M.

The babble power spectrum was estimated during periods when the SBR in a given TF segment dropped below a certain threshold, based on the method described by Hu and Loizou (2007b). The parameters of the processing were the same as described by Hu and Loizou (2007b), except for the duration of the frames. The estimate of the babble spectrum

was updated when the following condition for frame j was satisfied:

$$\mathrm{E}\left\{\left(\frac{|\bar{X}_j(k)|^2}{|\hat{V}_{j-1}(k)|^2}\right)-\log_{10}\left(\frac{|\bar{X}_j(k)|^2}{|\hat{V}_{j-1}(k)|^2}\right)-1\right\}\leq 0.45, \quad (4)$$

where E is the expected value operator, and $\hat{V}_{j-1}(k)$ is the estimate of the babble spectrum for the previous frame. Equation (4) is a discrete approximation of the Itakura–Saito distance (Itakura and Saito, 1968) as a measure of the difference between the speech-plus-babble spectrum, $\bar{X}_j(k)$, and the estimate of the babble spectrum for the previous frame, $\hat{V}_{j-1}(k)$. The decision rule of Eq. (4) is based on a log-likelihood ratio and the assumption that speech is absent in the current frame when the rule is satisfied. The running estimate of the babble spectrum was then calculated as

$$\hat{V}_j(k)=\sqrt{0.9|\hat{V}_{j-1}(k)|^2+0.1|\bar{X}_j(k)|^2}. \quad (5)$$

This meant that the estimate of the babble spectrum in any given frame had only a small influence on the running estimate.

It should be noted that a more modern VAD method has been proposed for estimating the spectrum of the background (Gerkmann and Hendriks, 2013). However, this newer method has been used in studies on normal-hearing listeners with cochlear-implant simulations (Bolner *et al.*, 2016) and with hearing-impaired listeners (Monaghan *et al.*, 2017) and it did not provide significant improvements in speech intelligibility in babble. Also, it did not improve perceived speech quality for 17 hearing-impaired listeners. Therefore, it seems that there is no clear advantage in using this newer method over the method used here.

The four channels were constructed from the appropriate bins of the smoothed spectrum so as to cover the frequency ranges 0–2, 2–4, 4–6, and 6–8 kHz. The estimate of the clean-speech spectrum in the ith channel, $\hat{S}_i$, in a given frame was obtained by subtracting a scaled version of the estimate of the babble spectrum from the combined speech-plus-babble spectrum:

$$|\hat{S}_i(k)|^2=|\bar{X}_i(k)|^2-\alpha_i\delta_i|\hat{V}_i(k)|^2, \quad b_i\leq k\leq e_i, \quad (6)$$

where $\bar{X}_i$ is the smoothed signal spectrum for the ith frequency channel, $\hat{V}_i$ is the estimated babble power spectrum, α_i and δ_i are parameters controlling the amount of babble reduction in the ith frequency channel, and b_i and e_i are the beginning and ending frequency bins of the ith frequency channel. The values of α_i and δ_i varied across channels, as described below.

Equation (6) can sometimes result in negative values. When this happened, the estimated power spectrum in the ith frequency channel was replaced by an attenuated version of the spectrum of noisy speech in that channel using the following equation:

$$|\hat{S}_i(k)|^2=\begin{cases}|\hat{S}_i(k)|^2 & |\hat{S}_i(k)|^2>0\\ \beta|X_i(k)|^2 & otherwise,\end{cases} \quad (7)$$

where the spectral floor parameter β was set to 0.002, corresponding to an attenuation of 27 dB.

Parameter α_i is called the oversubtraction factor for the ith channel. It was a function of the SBR in the ith frequency channel, SBR_i:

$$\alpha_i=\begin{cases}5 & SBR_i<-5\\ 4-\frac{3}{20}SBR_i & -5\leq SBR_i\leq 20\\ 1 & SBR_i>20,\end{cases} \quad (8)$$

where SBR_i is

$$SBR_i(\mathrm{dB})=10\log_{10}\left(\frac{\sum_{b_i}^{e_i}|\hat{X}_i(k)|^2}{\sum_{b_i}^{e_i}|\hat{V}_i(k)|^2}\right). \quad (9)$$

Hence, the amount of babble reduction increased as SBR_i decreased.

Parameter δ_i was a function of frequency:

$$\delta_i=\begin{cases}1 & f_i\leq 1\,\mathrm{kHz}\\ 2 & 1\,\mathrm{kHz}<f_i\leq\frac{F_s}{2}-2\,\mathrm{kHz}\\ 1.5 & f_i>\frac{F_s}{2}-2\,\mathrm{kHz},\end{cases} \quad (10)$$

where F_S is the sampling rate (16 kHz here). Hence, the amount of babble reduction was greater for frequencies between 1 and 6 kHz than for lower or higher frequencies.

To reconstruct the signal for a given frame, the processed spectrum in each channel [Eq. (6)] was combined across channels and an inverse FFT was applied. This was repeated for each frame and the overlap-and-add method was utilized to obtain the babble-reduced speech.

E. Equipment and conditions

The participants were seated in a soundproof room and wore Sennheiser HD580 headphones (Wedemark, Germany) connected to the sound card of a computer (with 24-bit resolution and a sampling rate of 16 000 Hz). Three conditions were used: NP, RNN, and SS. For condition NP, the root-mean-square level of the speech (excluding the noise and before frequency-dependent amplification) was 65 dB sound pressure level (SPL). The overall level when the babble was added to the speech was 68 dB SPL for SBR = 0 dB and 66.2 dB SPL for SBR = 5 dB. The SS and RNN processing would have reduced the level of the babble markedly and reduced the level of the speech slightly. To allow for this, the overall level of the processed mixture of speech-plus-babble was set to 65 dB SPL.

The stimuli for each condition and each subject were subjected to linear frequency-dependent amplification according to the "Cambridge formula" (Moore and Glasberg, 1998) to ensure that the speech was audible over a wide frequency range. This was done independently for each participant, based on the audiogram of the test ear, using a 513-tap finite impulse response filter implemented using the fir2 function in Matlab.

F. Procedure

The three conditions were compared in terms of subjective speech intelligibility and sound quality, using the paired-comparison procedure described by Moore and Sek (2013) and Keshavarzi *et al.* (2018). There were three types of paired comparisons: RNN vs SS, RNN vs NP, and SS vs NP. The two sounds to be compared were presented in succession in random order with a 1-s silent interval between them.

Preferences in terms of speech intelligibility were assessed first. The instructions on the computer screen were as follows: "On each trial you will hear the same sentence twice in succession. Please decide whether the first or second sentence is easier to understand and by how much, by using the mouse to position the slider on the screen."

Next, preferences in terms of sound quality were assessed. The instructions on the screen were: "On each trial you will hear the same sentence twice in succession. Please decide whether the first or second sentence has a better sound quality and by how much, by using the mouse to position the slider on the screen."

Each pair of sounds was presented once on a given trial. Participants used a mouse to select the position of a slider on the screen along a continuum, which was labeled "1 much better," "1 moderately better," "1 slightly better," "equal," "2 slightly better," "2 moderately better," and "2 much better." Any point along the slider could be chosen, so the ratings were continuous rather than categorical. Within a given block of trials, a given pair of conditions (e.g., SS and NP) was presented in both orders for both SBRs (5 and 0 dB) and for each of the eight sentences, yielding 32 trials in a block. Each participant was tested in six blocks, corresponding to three pairs of comparisons (RNN vs SS, RNN vs NP, and SS vs NP) and two types of judgment (intelligibility and sound quality).

Preference scores for each participant and each pair of conditions were computed as described by Moore and Sek (2013). Briefly, the extreme positions of the slider were assigned arbitrary values of −3 and +3. Regardless of whether condition X or condition Y was presented first, if X was preferred the slider position was coded as a negative number and if Y was preferred the slider position was coded as a positive number. The overall score for a given SBR and a given comparison was based on the average of the scores for the two orders for that comparison and SBR for each participant. Therefore preference scores had to fall in the range −3 to +3.

III. RESULTS

A. Objective evaluation of the accuracy of the RNN

To evaluate the accuracy of the RNN in estimating the IRM from the unprocessed speech in babble, the estimated SBR for each TF segment was compared with that for the IRM determined using the speech alone and the babble alone. This was done using 300 sentences produced by six talkers who had not been used for training the RNN. An arbitrary threshold of 0 dB SBR was used to convert the IRM estimated by the RNN into a binary mask in order to calculate the hit rate (the percentage of correctly classified TF units with SBR $\geq$ 0 dB) and the false alarm rate (the percentage of TF units with SBR $<$ 0 dB that were incorrectly classified as having SBR $\geq$ 0 dB) (Kim *et al.*, 2009; Goehring *et al.*, 2017). The hit and false-alarm rates were used to calculate the detectability index d′ (Green and Swets, 1974) (see Table II).

As expected, the d′ values decreased with decreasing SBR. Nevertheless, the d′ values were above 1 for all SBRs, indicating reasonably good classification accuracy.

B. Objective evaluation of speech intelligibility and quality

Three objective metrics for estimating speech intelligibility were computed for the stimuli used in conditions NP, RNN, and SS. In addition, the metrics were computed for stimuli processed using the original frame size of 20 ms for the SS processing method of Kamath and Loizou (2002). The metrics were the normalized covariance metric (NCM) (Ma *et al.*, 2009), the short-time objective intelligibility (STOI) measure (Taal *et al.*, 2011), and a method called sEPSM^{cor} that combines the auditory processing front end of the multi-resolution speech-based envelope power spectrum model (Jørgensen *et al.*, 2013) with a correlation-based stage similar to that for the STOI (Relano-Iborra *et al.*, 2016). The sEPSM^{cor} method has been shown to give accurate predictions of the intelligibility of speech that has been processed using SS with a variety of parameter values (Relano-Iborra *et al.*, 2016). In addition, we used a method developed by Kates and Arehart (2014), the Hearing Aid Speech Quality Index, version 2 (HASQI2), to predict the quality of the processed speech signals. All metrics cover the range from 0 to 1, where 1 indicates perfect intelligibility (or quality) and 0 indicates zero intelligibility (or quality). The HASQI requires specification of the audiogram of the listener, and takes into account the effects of hearing loss, when present. For the present purpose, the mean audiometric thresholds of the participants were used. The outcomes are shown in Fig. 2.

Both the intelligibility and quality metrics predicted a worsening with decreasing SBR, as would be expected. All three intelligibility metrics predicted slightly higher intelligibility for condition RNN than for condition NP, for all SBRs. Intelligibility for condition SS with the 5-ms frame size used here was predicted to be similar to or slightly lower than for condition NP and lower than for condition RNN. Intelligibility for condition SS with the 20-ms frame size used by Kamath and Loizou (2002) was predicted to be

TABLE II. Hit and false alarm rates (percent) and d′ values for the estimated IRM.

SBR (dB)	Hit rate, %	False alarm rate, %	d′
10	91.7	16.0	2.37
5	85.9	17. 9	1.99
0	69.9	17.1	1.47
−5	55.6	16.4	1.11

similar to or slightly worse than for the 5-ms frame size. Except for the lowest SBR, the speech quality predicted using the HASQI was best for condition NP, slightly worse for condition RNN, worse still for condition SS with the 5-ms frame size and worst of all for the 20-ms frame size. For the lowest SBR (not used for testing the participants) the HASQI predicted very slightly better performance for condition RNN than for condition NP. The predictions of the intelligibility metrics were broadly consistent with the results presented in Sec. III C, while the results for the HASQI quality metric were not fully consistent with the results.

C. Preferences scores for speech intelligibility

To assess whether the preference scores for a given comparison and a given SBR were significantly different from zero (which would indicate a significant preference for one condition relative to another at that SBR), the scores for each participant were first averaged across the eight sentences and two orders of presentation used for evaluation. Wilcoxon non-parametric tests were used to assess whether the mean of the eight resulting scores (one for each participant) was significantly different from zero (using two-tailed tests). This was done separately for each pair of conditions and each SBR. Since the number of participants was small, the W statistic was used. Since we had specific hypotheses, namely that RNN would produce a benefit relative to NP while SS would not, no correction for multiple comparisons was applied.

Figure 3 shows box plots of the preference scores for speech intelligibility for each SBR and each pair of conditions. For each pair, if the score fell above 0, then the first condition in the pair was preferred; otherwise the second condition was preferred. For the pair RNN vs SS (panel a), the mean and median preference scores were slightly positive for both SBRs, indicating preferences for the RNN. The Wilcoxon test was significant for both SBRs ($W=1$, $p<0.05$ for SBR $=0$ dB and $W=0$, $p<0.05$ for SBR $=5$ dB). For the pair RNN vs NP (panel b), the mean and median preference scores were all positive, favoring the RNN. The Wilcoxon test was significant for both SBRs ($W=1$, $p<0.05$ for SBR $=0$ dB and $W=0$, $p<0.05$ for SBR $=5$ dB). For the pair SS vs NP (panel c), the mean and median preference scores were close to 0 and the mean preference scores were not significantly different from 0 ($W=17$, $p>0.05$ for SBR $=0$ dB and $W=7$, $p>0.05$ for SBR $=5$ dB). In summary, RNN was significantly preferred over both SS and NP, but SS was not significantly preferred over NP.

D. Preferences scores for quality

Figure 4 shows box plots of the preference scores for sound quality. For the pair RNN vs SS (panel a), the mean and median preference scores were positive for both SBRs, favoring the RNN, and the Wilcoxon test was significant for both SBRs ($W=0$, $p<0.05$ for SBR $=0$ dB and $W=0$, $p<0.05$ for SBR $=5$ dB). For the pair RNN vs NP (panel b), the mean and median preference scores were all positive, favoring the RNN. The Wilcoxon test was significant for both SBRs ($W=1$, $p<0.05$ for SBR $=0$ dB and $W=0$, $p<0.05$ for SBR $=5$ dB). For the pair SS vs NP (panel c), the mean and median preference scores were close to 0 and the mean and median preference scores were not significantly different from 0 ($W=15$, $p>0.05$ for SBR $=0$ dB and $W=15$, $p>0.05$ for SBR $=5$ dB). In summary, RNN was preferred over both SS and NP, and SS was not significantly preferred over NP.

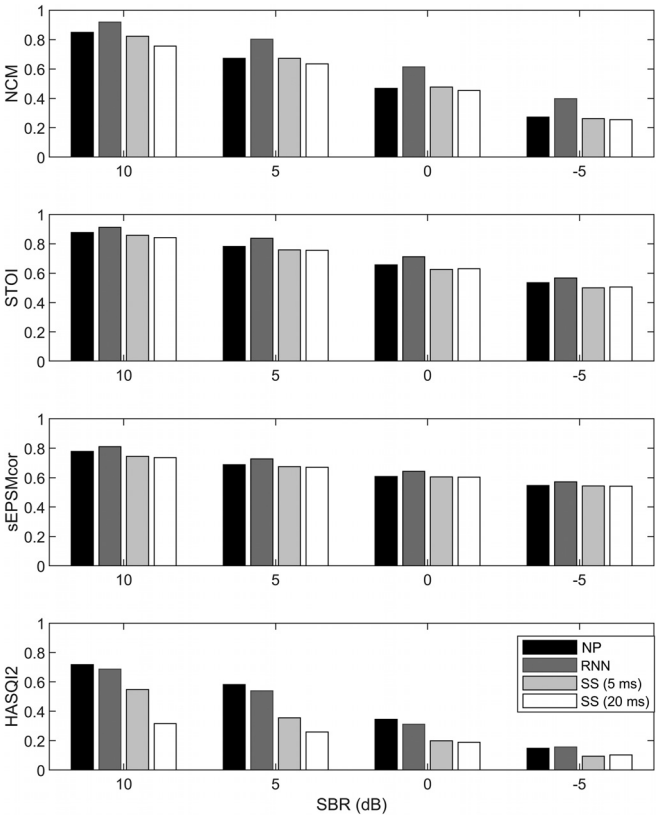

FIG. 2. NCM, STOI, sEPSM^{cor}, and HASQI2 values for conditions RNN, SS, and NP for each SBR.

IV. DISCUSSION

The results showed significant preferences for RNN over both SS and NP for both subjective intelligibility and sound quality. In contrast, SS was not significantly preferred over NP, confirming the limited effectiveness of SS for babble backgrounds, as found in other studies (Elberling *et al.*, 1993; Hu and Loizou, 2007a). One question that arises is whether the responses to the two subjective questions (subjective intelligibility and quality) were different. To assess this, the preference scores for each participant were averaged across the two SBRs used, and the Pearson correlation between intelligibility preferences and quality preferences was calculated for each pair of conditions. For the comparison SS vs NP, the intelligibility and quality preference scores were highly correlated ($r=0.84$, $p<0.05$), reflecting the fact that participants who preferred SS for intelligibility also tended to prefer it for quality, while participants who preferred NP for intelligibility also tended to prefer it for quality. For the comparison RNN vs NP, the correlation was smaller and was not significant ($r=0.52$, $p>0.05$). For the

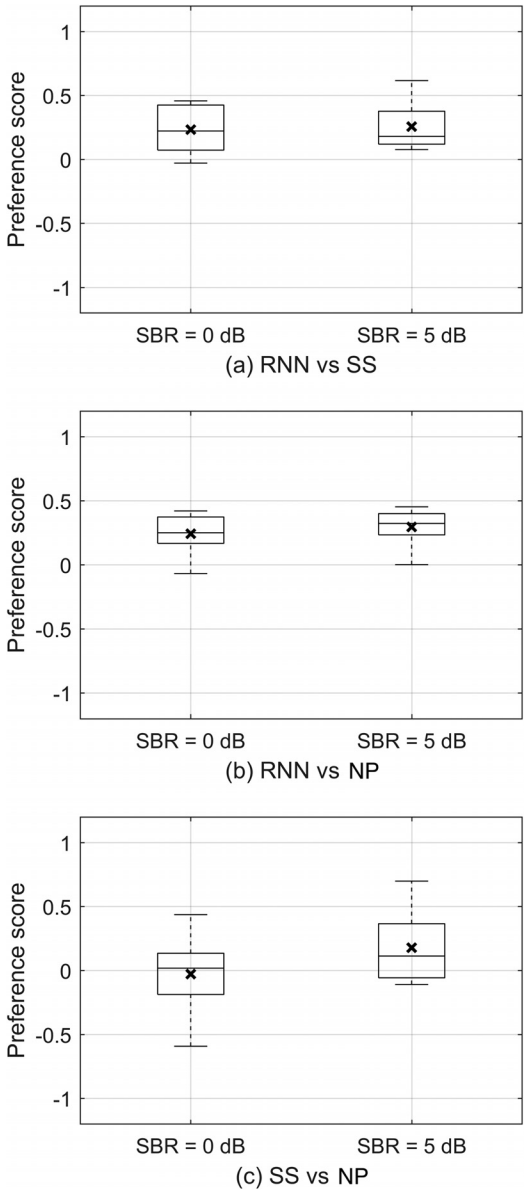

FIG. 3. Box plots of preference scores for speech intelligibility for each SBR and each pairwise comparison. The lower and upper edges of each box represent the second and third quartiles, the line inside each box shows the median value, the × shows the mean value, and the lines on either side of the rectangle show the lower and upper quartiles. Each panel shows results for a different comparison, as indicated in the key: RNN: recurrent neural network; SS: spectral subtraction; NP: no processing.

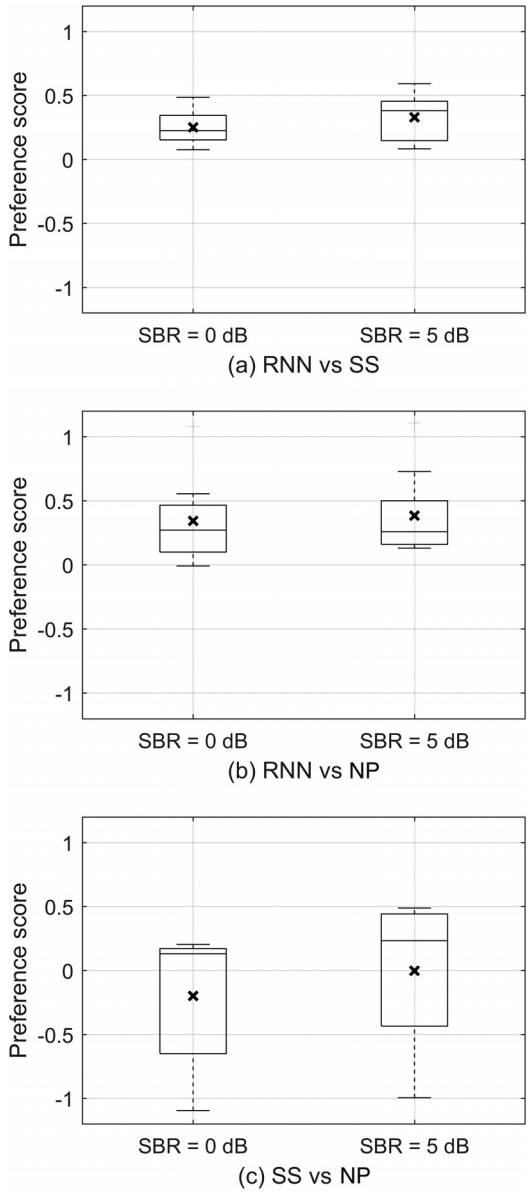

FIG. 4. As Fig. 3, but for sound quality.

comparison RNN vs SS, the correlation was even smaller and was again not significant ($r = 0.25$, $p > 0.05$). Thus, for the comparisons for which the overall preferences were significant (RNN vs NP and RNN vs SS), the judgments of quality and intelligibility were at least somewhat independent.

The preferences were mostly rather small. On a scale that went from −3 to +3, the median preferences for condition RNN over condition NP were about 0.25 to 0.4. This partly reflects the reluctance of participants to use the extremes of a rating scale (Poulton, 1979). It also probably reflects the fact that both signal-processing methods have undesired side effects. While they both resulted in a reduction of the babble, this came at the expense of some audible artifacts; participants reported hearing some "gurgling" effects and musical noise and these tended to be more pronounced for the SS than for the RNN. Some of the artifacts probably arose from the fact that the original noisy phase was used in reconstructing the signal. This would make the speech sound somewhat noisy even if the IRM were estimated perfectly. It may be possible to reduce the artifacts by setting a limit to the attenuation applied by the RNN or the SS or by limiting the speed of the gain changes across frames, as is done in the noise-reduction systems of some commercial hearing aids (Launer *et al.*, 2016). It may also be possible to reduce musical noise using processing that attempts to enhance both the magnitude and the phase (Krawczyk and Gerkmann, 2014; Williamson *et al.*, 2016). Further research is needed to assess the benefits of such modifications.

Despite the small magnitude of the preferences, RNN processing did significantly improve subjective speech

intelligibility and quality relative to NP for the speech in babble, while SS processing had no beneficial effects. The beneficial effects of RNN processing and the lack of beneficial effects for the SS processing are consistent with the objective evaluations of speech intelligibility using the three metrics. These showed little difference in predicted intelligibility between SS and NP but higher predicted intelligibility for RNN. It should be acknowledged, however, that while a relationship between subjective and objective intelligibility has been observed for speech in steady noise (Cienkowski and Speaks, 2000), it has not to our knowledge been demonstrated that such a relationship occurs for speech in babble. Hence further evaluation of the RNN based on real measures of intelligibility is desirable.

In contrast to the predictions of the HASQI, condition RNN was slightly preferred over condition NP for sound quality. The HASQI is strongly influenced by spurious amplitude modulation introduced by any signal processing, and this probably accounts for why it predicted lower quality for all conditions with signal processing relative to NP. It appears that our participants were influenced by the considerable reduction in the level of the babble produced by the RNN processing, and this outweighed the deleterious effects of spurious amplitude fluctuations introduced by the RNN processing.

The RNN processing was deliberately designed to operate with a short time delay to make it applicable in hearing aids. The frame duration was 5 ms and the frame overlap was 50%, which would lead to an inherent delay of about 7.5 ms; this corresponds to one whole frame that is needed to perform spectral analysis for that frame and half a frame that is needed for the overlap-add procedure (Allen, 1977). This is within the range that is acceptable for hearing aids (Stone and Moore, 1999; 2005; Goehring *et al.*, 2018). For implementation in a hearing aid, the RNN processing could make use of the frequency analysis that is often performed in hearing aids for dynamic range compression and directional processing, in which case the RNN processing would not increase the overall time delay produced by the hearing aid.

The RNN used here was trained using sentences from a large number of talkers and it showed good generalization to speech from the other talkers used for testing. However, it was trained and tested using only a single type of background babble. For a practical application in hearing aids or cochlear implants, the RNN would need to be trained using other types of background noises, including babble with different numbers of talkers and babble mixed with other types of noises, such as the clinking of glasses and cutlery. To make the RNN effective for many different talkers and many types of background sounds would require a very large set of training materials, and training might take a considerable time. However, this is not a problem in principle, because the training would be done offline. Also, it is not known how many units would be required in each layer of a multi-layer RNN to achieve good generalization across talkers and background types. Nevertheless, our results offer a proof of principle that an RNN can be effective in improving the subjective intelligibility and quality of speech in a babble background for speech produced by talkers who were not used for training the RNN.

V. SUMMARY AND CONCLUSIONS

Subjective intelligibility and sound quality were compared for conditions NP, SS, and RNN. The sentences used for testing were spoken by different talkers than those used for training. Eight hearing-impaired participants were tested and frequency-dependent linear amplification was provided to compensate for their hearing loss on an individual basis. RNN processing was significantly preferred over SS processing and NP for both subjective intelligibility and sound quality. SS processing was not significantly preferred over NP. Objective measures of intelligibility based on three metrics gave results consistent with the subjective evaluations: predicted intelligibility was higher for RNN than for SS or NP, and did not differ for SS and NP. However, an objective metric of sound quality, the HASQI, predicted poorer quality for condition RNN than for condition NP, except for the lowest SBR, whereas the participants rated the quality as higher for condition RNN than for condition NP.

The results provide a proof of concept that RNN processing can improve the subjective intelligibility and sound quality of speech in babble for speech produced by talkers who were not used for training. However, further work is required to assess whether an RNN can be trained to work effectively for speech from many talkers in a variety of types of background sounds.

ACKNOWLEDGMENTS

This work was supported by the Engineering and Physical Sciences Research Council (UK, Grant No. RG78536), by Action on Hearing Loss, and by the Allen Trust. We thank James Kates for providing the code for computing the HASQI. We thank the Associate Editor Joshua Bernstein and two reviewers for helpful comments on an earlier version of this paper.

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (**2016**). "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," arXiv:1603.004467.

Alcántara, J. I., Moore, B. C. J., Kühnel, V., and Launer, S. (**2003**). "Evaluation of the noise reduction system in a commercial digital hearing aid," Int. J. Audiol. **42**, 34–42.

Allen, J. B. (**1977**). "Short term spectral analysis, synthesis and modification by discrete Fourier transform," IEEE Trans. Acoust. Speech Signal Process. **25**, 235–238.

Arehart, K. H., Hansen, J. H., Gallant, S., and Kalstein, L. (**2003**). "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners," Speech Commun. **40**, 575–592.

Bentler, R., Wu, Y. H., Kettel, J., and Hurtig, R. (**2008**). "Digital noise reduction: Outcomes from laboratory and field studies," Int. J. Audiol. **47**, 447–460.

Bolner, F., Goehring, T., Monaghan, J. J., Van Dijk, B., Wouters, J., and Bleeck, S. (**2016**). "Speech enhancement based on neural networks applied to cochlear implant coding strategies," in *IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Shanghai, China, pp. 6520–6524.
Bramsløw, L., Naithani, G., Hafez, A., Barker, T., Pontoppidan, N. H., and Virtanen, T. (**2018**). "Improving competing voices segregation for hearing impaired listeners using a low-latency deep neural network algorithm," J. Acoust. Soc. Am. **144**, 172–185.
Brons, I., Houben, R., and Dreschler, W. A. (**2012**). "Perceptual effects of noise reduction by time-frequency masking of noisy speech," J. Acoust. Soc. Am. **132**, 2690–2699.
Chen, J., and Wang, D. (**2017**). "Long short-term memory for speaker generalization in supervised speech separation," J. Acoust. Soc. Am. **141**, 4705–4714.
Chen, J., Wang, Y., Yoho, S. E., Wang, D., 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.
Cienkowski, K. M., and Speaks, C. (**2000**). "Subjective vs. objective intelligibility of sentences in listeners with hearing loss," J. Speech Lang. Hear. Res. **43**, 1205–1210.
Delfarah, M., and Wang, D. L. (**2017**). "Features for masking-based monaural speech separation in reverberant conditions," IEEE Trans. Audio, Speech Lang. Proc. **25**, 1085–1094.
Elberling, C., Ludvigsen, C., and Keidser, G. (**1993**). "The design and testing of a noise reduction algorithm based on spectral subtraction," Scand. Audiol. **Suppl. 38**, 39–49.
Festen, J. M., and Plomp, R. (**1990**). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," J. Acoust. Soc. Am. **88**, 1725–1736.
Gerkmann, T., and Hendriks, R. C. (**2013**). "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," IEEE Trans. Audio, Speech Lang. Process. **20**, 1383–1393.
Glasberg, B. R., and Moore, B. C. J. (**1990**). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.
Goehring, T., Bolner, F., Monaghan, J. J., 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., Bleeck, 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.
Graves, A., Mohamed, A. R., and Hinton, G. (**2013**). "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, pp. 6645–6649.
Green, D. M., and Swets, J. A. (**1974**). *Signal Detection Theory and Psychophysics* (Krieger, New York), p. 479.
Hamacher, V., Chalupper, J., Eggers, J., Fischer, E., Kornagel, U., Puder, H., and Rass, U. (**2005**). "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," EURASIP J. Appl. Signal Process. **18**, 2915–2929.
Hawkins, D. B., and Yacullo, W. S. (**1984**). "Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation," J. Speech Hear. Disord. **49**, 278–286.
Healy, E. W., Delfarah, M., Vasko, J. L., Carter, B. L., and Wang, D. (**2017**). "An algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker," J. Acoust. Soc. Am. **141**, 4230–4239.
Healy, E. W., Yoho, S. E., Chen, J., Wang, Y., and Wang, D. (**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.
Hinton, G. E., Osindero, S., and Teh, Y. W. (**2006**). "A fast learning algorithm for deep belief nets," Neur. Comput. **18**, 1527–1554.
Hochreiter, S., and Schmidhuber, J. (**1997**). "Long short-term memory," Neur. Comput. **9**, 1735–1780.
Hu, Y., and Loizou, P. C. (**2007a**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.
Hu, Y., and Loizou, P. C. (**2007b**). "Subjective comparison and evaluation of speech enhancement algorithms," Speech Commun. **49**, 588–601.
Huang, P. S., Kim, M., Hasegawa-Johnson, M., and Smaragdis, P. (**2015**). "Joint optimization of masks and deep recurrent neural networks for monaural source separation," IEEE Trans. Audio, Speech Lang. Proc. **23**, 2136–2147.
Itakura, F., and Saito, S. (**1968**). "Analysis synthesis telephony based on the maximum likelihood method," in *International Congress on Acoustics*, ICA, Tokyo, Japan, pp. C17–C20.
Jamieson, D. G., Brennan, R. L., and Cornelisse, L. E. (**1995**). "Evaluation of a speech enhancement strategy with normal-hearing and hearing-impaired listeners," Ear Hear. **16**, 274–286.
Jørgensen, S., Ewert, S. D., and Dau, T. (**2013**). "A multi-resolution envelope-power based model for speech intelligibility," J. Acoust. Soc. Am. **134**, 436–446.
Kamath, S., and Loizou, P. C. (**2002**). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, IEEE, Orlando, FL, pp. 1–4.
Kates, J. M. (**1987**). "The short-time articulation index," J. Rehabil. Res. Dev. **24**, 271–276.
Kates, J. M., and Arehart, K. H. (**2014**). "The hearing-aid speech quality index (HASQI) version 2," J. Audio Eng. Soc. **62**, 99–117.
Keshavarzi, M., Goehring, T., Zakis, J., Turner, R. E., and Moore, B. C. J. (**2018**). "Use of a deep recurrent neural network to reduce wind noise: Effects on judged speech intelligibility and sound quality," Trends Hear. **22**, 1–12.
Kim, G., Lu, Y., Hu, Y., and Loizou, P. C. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.
Kolbæk, M., Yu, D., Tan, Z. H., and Jensen, J. (**2017**). "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," IEEE Trans. Audio, Speech Lang. Proc. **25**, 1901–1913.
Krawczyk, M., and Gerkmann, T. (**2014**). "STFT phase reconstruction in voiced speech for an improved single-channel speech enhancement," IEEE/ACM Trans. Audio, Speech, Lang. Process. **22**, 1931–1940.
Launer, S., Zakis, J., and Moore, B. C. J. (**2016**). "Hearing aid signal processing," in *Hearing Aids*, edited by G. R. Popelka, B. C. J. Moore, A. N. Popper, and R. R. Fay (Springer, New York), pp. 93–130.
Levitt, H., Bakke, M., Kates, J., Neuman, A., Schwander, T., and Weiss, M. (**1993**). "Signal processing for hearing impairment," Scand. Audiol. **Suppl. 38**, 7–19.
Lipton, Z. C., Berkowitz, J., and Elkan, C. (**2015**). "A critical review of recurrent neural networks for sequence learning," arXiv:1506.00019v4.
Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, LA), p. 632.
Ma, J., Hu, Y., and Loizou, P. C. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**, 3387–3405.
Monaghan, J. J., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C., and Bleeck, S. (**2017**). "Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners," J. Acoust. Soc. Am. **141**, 1985–1998.
Moore, B. C. J. (**2007**). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*, 2nd ed. (Wiley, Chichester, UK), p. 332.
Moore, B. C. J., Baer, T., Ives, D. T., Marriage, J., and Salorio-Corbetto, M. (**2016**). "Effects of modified hearing-aid fittings on loudness and tone quality for different acoustic scenes," Ear Hear. **37**, 483–491.
Moore, B. C. J., and Glasberg, B. R. (**1998**). "Use of a loudness model for hearing aid fitting. I. Linear hearing aids," Br. J. Audiol. **32**, 317–335.
Moore, B. C. J., and Sek, A. (**2013**). "Comparison of the CAM2 and NAL-NL2 hearing-aid fitting methods," Ear Hear. **34**, 83–95.
Moore, B. C. J., and Sek, A. (**2016**). "Comparison of the CAM2A and NAL-NL2 hearing-aid fitting methods for participants with a wide range of hearing losses," Int. J. Audiol. **55**, 93–100.
Natarajan, A., Hansen, J. H., Arehart, K. H., and Rossi-Katz, J. (**2005**). "An auditory-masking-threshold-based noise suppression algorithm GMMSE-AMT[ERB] for listeners with sensorineural hearing loss," EURASIP J. Adv. Signal Process. **18**, 2938–2953.
Patterson, R. D., Allerhand, M. H., and Giguère, C. (**1995**). "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," J. Acoust. Soc. Am. **98**, 1890–1894.
Peters, R. W., Moore, B. C. J., and Baer, T. (**1998**). "Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people," J. Acoust. Soc. Am. **103**, 577–587.

Picou, E. M., Aspell, E., and Ricketts, T. A. (**2014**). "Potential benefits and limitations of three types of directional processing in hearing aids," Ear Hear. **35**, 339–352.
Picou, E. M., and Ricketts, T. A. (**2017**). "How directional microphones affect speech recognition, listening effort and localisation for listeners with moderate-to-severe hearing loss," Int. J. Audiol. **56**, 909–918.
Plomp, R. (**1978**). "Auditory handicap of hearing impairment and the limited benefit of hearing aids," J. Acoust. Soc. Am. **63**, 533–549.
Poulton, E. C. (**1979**). "Models for the biases in judging sensory magnitude," Psychol. Bull. **86**, 777–803.
Relano-Iborra, H., May, T., Zaar, J., Scheidiger, C., and Dau, T. (**2016**). "Predicting speech intelligibility based on a correlation metric in the envelope power spectrum domain," J. Acoust. Soc. Am. **140**, 2670–2679.
Riedmiller, M., and Braun, H. (**1993**). "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *IEEE Conference on Neural Networks*, pp. 586–591.
Sak, H., Senior, A., and Beaufays, F. (**2014**). "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech 2014*, pp. 338–342.
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.
Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio, Speech Lang. Proc. **19**, 2125–2136.
Tang, Y. (**2016**). "TF. Learn: TensorFlow's high-level module for distributed machine learning," arXiv:1612.04251.
Wang, Y., Narayanan, A., and Wang, D. (**2014**). "On training targets for supervised speech separation," IEEE Trans. Audio, Speech Lang. Proc. **22**, 1849–1858.
Weninger, F., Erdogan, H., Watanabe, S., Vincent, E., Le Roux, J., Hershey, J. R., and Schuller, B. (**2015**). "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*, Springer, Liberec, Czech Republic.
Williamson, D. S., Wang, Y., and Wang, D. (**2016**). "Complex ratio masking for monaural speech separation," IEEE/ACM Trans. Audio, Speech, Lang. Process. **24**, 483–492.