

Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants

Tobias Goehring, Mahmoud Keshavarzi, Robert P. Carlyon, and Brian C. J. Moore

Citation: [The Journal of the Acoustical Society of America](#) **146**, 705 (2019); doi: 10.1121/1.5119226

View online: <https://doi.org/10.1121/1.5119226>

View Table of Contents: <https://asa.scitation.org/toc/jas/146/1>

Published by the [Acoustical Society of America](#)

ARTICLES YOU MAY BE INTERESTED IN

[On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise](#)

The Journal of the Acoustical Society of America **146**, 693 (2019); <https://doi.org/10.1121/1.5119240>

[Cognitive factors contribute to speech perception in cochlear-implant users and age-matched normal-hearing listeners under vocoded conditions](#)

The Journal of the Acoustical Society of America **146**, 195 (2019); <https://doi.org/10.1121/1.5116009>

[Source characterization of full-scale tactical jet noise from phased-array measurements](#)

The Journal of the Acoustical Society of America **146**, 665 (2019); <https://doi.org/10.1121/1.5118239>

[Reconfigurable topological insulator for elastic waves](#)

The Journal of the Acoustical Society of America **146**, 773 (2019); <https://doi.org/10.1121/1.5114920>

[Unified wave field retrieval and imaging method for inhomogeneous non-reciprocal media](#)

The Journal of the Acoustical Society of America **146**, 810 (2019); <https://doi.org/10.1121/1.5114912>

[Characteristics and microgeographic variation of whistles from the vocal repertoire of beluga whales \(*Delphinapterus leucas*\) from the White Sea](#)

The Journal of the Acoustical Society of America **146**, 681 (2019); <https://doi.org/10.1121/1.5119249>



Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants

Tobias Goehring,^{1,a)} Mahmoud Keshavarzi,² Robert P. Carlyon,¹ and Brian C. J. Moore²

¹Medical Research Council Cognition and Brain Sciences Unit, University of Cambridge, 15 Chaucer Road, Cambridge CB2 7EF, United Kingdom

²Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, United Kingdom

(Received 23 April 2019; revised 17 June 2019; accepted 8 July 2019; published online 31 July 2019)

Speech-in-noise perception is a major problem for users of cochlear implants (CIs), especially with non-stationary background noise. Noise-reduction algorithms have produced benefits but relied on *a priori* information about the target speaker and/or background noise. A recurrent neural network (RNN) algorithm was developed for enhancing speech in non-stationary noise and its benefits were evaluated for speech perception, using both objective measures and experiments with CI simulations and CI users. The RNN was trained using speech from many talkers mixed with multi-talker or traffic noise recordings. Its performance was evaluated using speech from an unseen talker mixed with different noise recordings of the same class, either babble or traffic noise. Objective measures indicated benefits of using a recurrent over a feed-forward architecture, and predicted better speech intelligibility with than without the processing. The experimental results showed significantly improved intelligibility of speech in babble noise but not in traffic noise. CI subjects rated the processed stimuli as significantly better in terms of speech distortions, noise intrusiveness, and overall quality than unprocessed stimuli for both babble and traffic noise. These results extend previous findings for CI users to mostly unseen acoustic conditions with non-stationary noise. © 2019 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1121/1.5119226>

[ICB]

Pages: 705–718

I. INTRODUCTION

Despite great advances in cochlear implant (CI) technology and the benefits that these provide, users of CIs still encounter difficulties understanding speech in noisy environments, especially with non-stationary backgrounds such as competing speech or traffic. CI users struggle more than normal-hearing (NH) listeners in these conditions, partly due to a decreased ability to make use of temporal fluctuations in the background noise (Stickney *et al.*, 2004; Cullington and Zeng, 2008). Furthermore, the spectral resolution that a CI can deliver is limited by the use of a small number of electrodes whose outputs interact due to current spread (Carlyon *et al.*, 2007; Oxenham and Kreft, 2014). In addition, CI listeners have poor sensitivity to the temporal fine structure of sounds (Moore and Carlyon, 2005), and this may limit their ability to perceptually segregate speech from interfering sounds. As a result, CI users rely strongly on slowly varying temporal-envelope information, and this makes them especially susceptible to the effects of modulated, or non-stationary, interfering noise (Cullington and Zeng, 2008; Fu *et al.*, 1998). Previous studies have shown improved speech intelligibility (SI) for speech in fluctuating noise using directional algorithms, but these depend on the assumption that the target speech and masking noise are spatially separated (Wouters and Vanden Berghe, 2001; Hersbach *et al.*, 2012).

In addition, such algorithms usually require the user to face the target talker, which is not always possible. Here, we describe and evaluate a single-microphone algorithm that operates without spatial information and can be applied in conjunction with directional algorithms in CI speech processors (Hersbach *et al.*, 2012).

Conventional single-microphone speech enhancement algorithms, such as those used in current CIs, are based on statistical signal processing methods that include spectral subtraction and wiener filtering (Boll, 1979; Scalart and Filho, 1996). These have been shown to improve the intelligibility of speech in stationary noise for CI users (Loizou *et al.*, 2005; Dawson *et al.*, 2011; Mauger *et al.*, 2012) and NH listeners using CI simulations (Bolner *et al.*, 2016; Lai *et al.*, 2018). Data-based algorithms using machine-learning (ML) techniques, such as deep neural networks (DNNs) or Gaussian mixture models (GMMs), were successful for speech in non-stationary, multi-talker babble and achieved significant SI improvements for NH (Kim *et al.*, 2009; Bentsen *et al.*, 2018), hearing-impaired (HI; Healy *et al.*, 2013; Healy *et al.*, 2015; Healy *et al.*, 2019; Chen *et al.*, 2016; Monaghan *et al.*, 2017; Bramsløw *et al.*, 2018), and CI listeners (Hu and Loizou, 2010; Goehring *et al.*, 2017; Lai *et al.*, 2018). Improvements of more recent approaches over earlier ones have been mainly driven by two factors: the use of more powerful DNN-based regression systems instead of classification systems, and the use of a ratio mask instead of a binary mask as the training target (Madhu *et al.*, 2013;

^{a)}Electronic mail: Tobias.Goehring@mrc-cbu.cam.ac.uk

Bentsen *et al.*, 2018). However, all of these algorithms made use of some *a priori* information about the target speech and/or interfering noise by using the same target speaker (Lai *et al.*, 2018; Chen *et al.*, 2016), background noise (Goehring *et al.*, 2017), or both (Kim *et al.*, 2009; Hu and Loizou, 2010; Healy *et al.*, 2013; Healy *et al.*, 2015; Healy *et al.*, 2019; Goehring *et al.*, 2017; Lai *et al.*, 2018; Bramsløw *et al.*, 2018; Bentsen *et al.*, 2018) for the training and testing of the algorithm.

While the results of these studies are promising, in practice the application to CI speech processors requires an algorithm to generalize to acoustic conditions that were not presented during the training. Unfortunately, performance has been found to drop substantially for unseen testing data evaluated with objective intelligibility predictors (May and Dau, 2014; Chen and Wang, 2017) and for a speaker-independent over a speaker-dependent system evaluated with CI users (Goehring *et al.*, 2017). Recent computational studies provide evidence that the generalization performance of DNNs to unseen speakers or background noise can be improved by using recurrent neural network (RNN) architectures (Weninger *et al.*, 2015; Chen and Wang, 2017; Kolbæk *et al.*, 2017). These differ from feed-forward architectures by using recurrent connections, as well as feedback and gate elements, to add temporal memory to the network (Graves *et al.*, 2013). One of the most successful RNN architectures is the “long short-term memory” (LSTM) RNN architecture that uses four gates to accumulate information about past input and state data, and learns to manage this information over time (Hochreiter and Schmidhuber, 1997; LeCun *et al.*, 2015). RNN-LSTM algorithms have shown improved generalization using objective measures, but have not been evaluated in listening studies with CI users. However, similar types of LSTM-RNNs have recently been shown to provide benefits for speech-in-noise perception for HI listeners (Bramsløw *et al.*, 2018; Keshavarzi *et al.*, 2018; Keshavarzi *et al.*, 2019; Healy *et al.*, 2019), and they represent a promising way for improving performance for CI users in conditions with non-stationary noise that was not included in the training data.

In addition to the requirement for generalization to unseen conditions, a constraint for the practical use of ML-based algorithms in CI devices is a processing delay below about 10–20 ms, to avoid subjective disturbance during speech production and limit audio-visual asynchrony (Stone and Moore, 1999; Goehring *et al.*, 2018; Bramsløw *et al.*, 2018). Most of the studies described above used non-causal signal processing by providing future frames to the input of the neural network (for example, Healy *et al.*, 2013; Healy *et al.*, 2015; Healy *et al.*, 2019; Chen *et al.*, 2016). This could not be done in a hearing device due to the excessive delay it would introduce. Other studies have used causal signal processing without future frames to stay within the tolerable range of delays (Bolner *et al.*, 2016; Monaghan *et al.*, 2017; Goehring *et al.*, 2017; Bramsløw *et al.*, 2018).

Another constraint is that current CI devices have limited computational power and memory. Furthermore, the speech processor of CI devices is worn behind the ear of the user, and therefore is limited in terms of battery power.

While this may improve in the future, the use of highly complex ML architectures with millions of parameters and extensive acoustic feature-extraction methods is unlikely to yield a practical solution for next-generation CI devices. We focussed on using a real-time-feasible, low-complexity architecture with a small number of layers and processing units in conjunction with simple acoustic features similar to those extracted by CI speech processors to facilitate the practical application of the algorithm in future CI devices.

We used a RNN-based algorithm to process speech in noise and assessed its benefits in terms of speech perception with CIs in two listening experiments. The main research question for both experiments was whether a RNN can generalize to an unseen speaker and noise condition over a range of signal-to-noise ratios (SNRs) that are relevant for CI users. Initially, two objective SI prediction methods were used to optimize and evaluate the RNN. The first experiment evaluated performance of the RNN for speech in babble using CI vocoder simulations presented to NH listeners (Oxenham and Kreft, 2014; Grange *et al.*, 2017; Fletcher *et al.*, 2018). Two simulated amounts of current spread were used to simulate CI users with electrodes positioned close to or far from the stimulated neural elements in an attempt to model the variability that characterizes the CI population, and evaluate its effects on the benefits of RNN processing over no processing. The second experiment measured CI users’ speech-in-noise performance for two realistic noise scenarios, multi-talker babble and traffic noise. In addition, subjective speech quality ratings were collected to determine if CI users preferred the RNN processing over no processing. For both SI and quality comparisons with CI users, an ideal noise-reduction condition was included for which the speech and background noise were available separately, to evaluate the theoretical upper limit of benefits that could be obtained with the algorithm.

II. ALGORITHM DESCRIPTION

A. Signal processing and RNN architecture

The RNN-based single-microphone algorithm is illustrated schematically in Fig. 1. The input signal was the unprocessed (UN) speech in noise that was obtained by adding the speech to the noise:

$$x(t) = s(t) + n(t), \quad (1)$$

where t is time, x is the speech in noise, s is the clean speech, and n is the noise. The input signal was segmented into 20-ms frames with 10-ms overlap between successive frames, giving 320 samples per frame at a sampling rate of 16 kHz.

Acoustic features were extracted from each frame by calculating the energy of a fast Fourier transform (FFT)-based gammatone filterbank (Patterson *et al.*, 1995) consisting of 64 channels equally spaced on the equivalent rectangular bandwidth (ERB)_N-number scale (Glasberg and Moore, 1990) with center frequencies from 50 to 8000 Hz. The gammatone features were obtained using Hanning-windowed frames. We chose these simple features because of the low computational requirements and based on a comparison study where gammatone features were only slightly

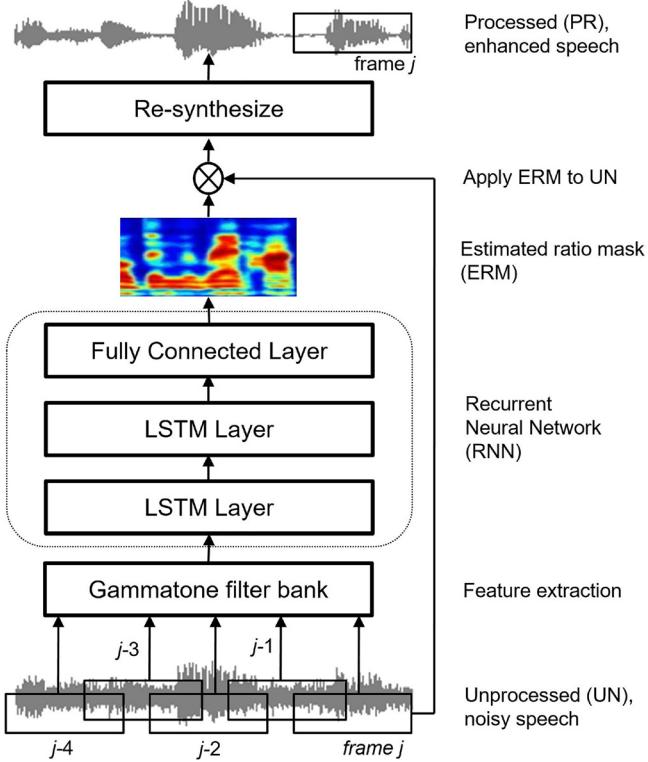


FIG. 1. (Color online) Schematic diagram of the RNN algorithm and signal processing framework.

inferior to a computationally much more complex feature set (Chen *et al.*, 2014). The acoustic features were scaled to have zero mean and unit variance by subtracting the mean and then dividing by the standard deviation calculated across the whole set of training data. The target data for training the RNN were the ideal ratio masks (IRMs) that were calculated by passing the speech and noise signals separately through the 64-channel gammatone filterbank and calculating the wiener gain in the time-frequency (T-F) domain for each frame j and frequency channel m ,

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

where $S(j,m)$ and $N(j,m)$ are the magnitudes of $s(t)$ and $n(t)$ in the m th channel of frame j , respectively. The soft gain function applied by the IRM was chosen here over the ideal binary mask (IBM) because it generally leads to better speech quality and intelligibility (Madhu *et al.*, 2013). The IRM also provides more precise information about the local SNR in each T-F segment than the IBM during the training of the algorithm. It has the additional advantage that no threshold criterion has to be chosen or adapted, in contrast to the IBM.

The RNN consisted of an input layer, two hidden LSTM layers with 128 units, followed by a fully connected layer with 64 sigmoidal units as the output layer. The LSTM processed a five-timestep input wherein each timestep was related to acoustic features extracted from a single frame of the input signal (noisy speech); steps 1, 2, 3, 4, and 5 corresponded to successive frames $j-4$, $j-3$, $j-2$, $j-1$, and j ,

respectively. We selected this architecture based on previous studies using HI listeners (Keshavarzi *et al.*, 2018; Keshavarzi *et al.*, 2019). The RNN estimated the IRM for frame j as its output (estimated ratio mask, ERM).

The ML-frameworks TFlearn and Tensorflow were used to construct, train, and test the RNN (Abadi *et al.*, 2016; Tang, 2016). The “adam” algorithm (Kingma and Ba, 2014), a method for stochastic optimization, was used as the training algorithm with the goal of minimizing the mean square error (MSE) between the ERM and IRM. The learning rate was set to 0.001, the batch size was 1024, and otherwise the default settings were used for adam, as specified by TFlearn. An early stopping criterion was used to choose the best-performing model for a validation dataset that consisted of about one-third of the testing data. Performance for the validation dataset did not improve significantly after one presentation of the full training dataset (an epoch). Instead, performance decreased with more than two epochs, as indicated by an increased MSE between the ERM and IRM when testing at SNRs of 0, 5, and 10 dB (the MSE increased by up to 30% for ten epochs vs one epoch of training). This behaviour indicated that the RNN was overfitting the training data, which could not be avoided when using dropout regularization with a proportion of 20% (Srivastava *et al.*, 2014). It seems likely that, because of the large mismatch between training and validation data (different speaker, noise recording and partly SNR), multiple presentations of the same training data would not improve performance on the validation data. Therefore, we chose to perform only one epoch of training to avoid overfitting the training data and maximise performance for the validation data. One epoch of training comprised 3185 parameter updates with gradients computed over batches of 1024 frames each (about 2 utterances per batch), but took only a few minutes on a modern laptop computer. Performance was found to be very similar for several RNN models that were trained on a single epoch each, confirming the robustness and efficiency of the adam algorithm. This approach also serves as a proof-of-concept for a system that could be quickly re-trained in practice to adapt to a new acoustic environment. This could, for example, be performed on a mobile device.

After the network had been trained, the ERM and IRM were used to process the noisy speech in each frame (by element-wise multiplication in the T-F domain) so as to attenuate T-F segments with low SNR while maintaining segments with high SNR. To avoid extreme changes in gain and preserve an awareness of the acoustic environment, the applied gain was limited to values in the range from 0.1 to 1 for both the ERM and IRM,

$$\begin{aligned} Y_{IRM}(j,m) &= \max(IRM(j,m), 0.1)X(j,m), \\ Y_{PR}(j,m) &= \max(ERM(j,m), 0.1)X(j,m), \end{aligned} \quad (3)$$

where $Y_{IRM}(j,m)$ and $Y_{PR}(j,m)$ (PR indicates conditions processed with the ERM) are the magnitudes for the m th channel and frame j of the speech in noise after weighting with the IRMs and ERMs, respectively. For both Y_{IRM} and Y_{PR} , the modified magnitudes from the processed frames were

combined with the noisy phases of the speech-in-noise signal $x(t)$ to obtain the output signals $y_{\text{IRM}}(t)$ and $y_{\text{PR}}(t)$, using the overlap-add operation and Hanning windowing. The output signals were presented acoustically to allow similar testing conditions for NH listeners and CI users. All stimuli were equalized to have the same root-mean-square (RMS) level after the processing.

B. Training and testing data

The speech data used for training the RNN consisted of sentences taken from CSTR VCTK, a British-English multi-speaker corpus with a variety of accents (available online from the University of Edinburgh; [Veaux et al., 2016](#)). We used 100 sentences from each of 80 speakers (40 female) to obtain a speech training dataset of 8000 sentences in total. The multi-talker babble used for training consisted of 25 real-world recordings of various multi-talker babbles (recorded from cafeterias, canteens, cafes, and shopping malls) obtained from Freesound Datasets ([Fonseca et al., 2017](#)). Recordings ranged in length from 5 to 81 s and were concatenated to form the training babble, giving a total duration of about 17 min. Traffic noise training data were generated using 25 real-world recordings of various traffic noises (recorded on motorways and public streets with cars passing by), also obtained from Freesound Datasets, and with a total duration of 8.5 min. The speech-in-noise data used for training were then generated by mixing the speech data (VCTK) with random cuts of either the babble or traffic noise at 5 dB SNR to obtain two separate training datasets, one for babble and one for traffic, each with 8000 utterances and a length of about 9 h. This SNR was chosen to represent a challenging condition in which CI users struggle to understand speech in babble.

For the first evaluation based on objective measures, the speech-in-noise data used for testing the algorithm in babble were generated from the Bamford-Kowal-Bench (BKB) sentences (English, spoken by a male talker; [Bench et al., 1979](#)) mixed with different multi-talker babble recordings at SNRs of 0, 5, and 10 dB. Six babbles with 2, 4, 8, 16, 32, and 64 talkers were generated to evaluate the objective measures (Sec. II3), using sentences from the TIMIT corpus ([Garofolo et al., 1993](#)). Each babble had equal numbers of male and female talkers and a duration of 1 min. These multi-talker babbles were filtered to have the same long-term spectrum as the BKB sentences.

For the test stimuli in the listening experiments, the 20-talker babble from Auditec (St. Louis, MO) was used, as in previous publications (e.g., [Goehring et al., 2017](#)). For the second listening experiment, we also used a traffic noise recording (“Traffic02”) obtained from MusicRadar, available online.¹ The dataset used for testing the RNN algorithm in the listening experiments consisted of 270 sentences (18 lists) from the BKB corpus mixed with either the 20-talker babble or the traffic noise at SNRs between –10 and 20 dB (in 2-dB steps). We generated a second set for evaluation with the objective measures with these stimuli at SNRs of 0, 5, and 10 dB. It should be noted that the RNN was evaluated using a range of SNRs, both higher and lower than used for training. Furthermore, all speech and noise recordings used

for the objective measures and listening experiments were not part of the training data, and there were two separate conditions for training and testing two RNNs: one for babble and one for traffic.

C. RNN performance evaluation using objective measures

As a preliminary evaluation and to quantitatively compare the performance of the RNN to that for previous studies, the RNN was evaluated using two objective SI measures, the short-time objective intelligibility metric (STOI; [Taal et al., 2011](#)), and the normalized-covariance metric (NCM; [Holube and Kollmeier, 1996](#)), using utterances from the two objective-measure datasets. Both NCM and STOI are intrusive SI prediction methods that use the clean speech signal as reference for the speech signal under test. The NCM applies a filter bank to both signals, extracts the temporal envelope for each filter channel, and calculates a weighted sum over the normalized covariance (linear correlation) between the envelopes of the reference and the test signals in each filter bank channel. The STOI follows a similar method but calculates the mean of the linear correlation coefficients between the filter bank envelopes of the signals in 384-ms long time frames. NCM and STOI have been used in previous studies for predicting the effects on SI of speech enhancement algorithms based on T-F masks. The first evaluation compared the predicted SI produced by the RNN algorithm for speech-in-babble noise for conditions with different numbers of competing talkers in the babble. Twenty BKB sentences from the testing data were mixed with random cuts of the 6 artificially generated multi-talker babbles with between 2 and 64 talkers (2T–64T) and the 20-talker babble. Each babble was mixed at SNRs of 0, 5, and 10 dB. Note that the 20T babble was not filtered to have the same long-term spectrum as the BKB sentences, but was used in its original form, as for the listening experiments.

The results for the speech-in-noise processed with the RNN algorithm (2T–64T) are shown in Fig. 2, together with the mean scores (across babble types) for the UN and ideal (IRM) conditions. The RNN processing improved the NCM scores over those for condition UN for babble with two or more talkers and improved the STOI scores for babble with four or more talkers. For condition UN, the NCM metric predicted an increase in SI with increasing number of talkers (from 0.45 for 2T to 0.61 for 64T at 0 dB SNR), whereas the STOI metric predicted SI to vary only slightly with the number of talkers (not shown). The improvement in predicted SI produced by processing with the RNN increased with increasing number of talkers. Both the STOI and NCM predicted slightly smaller improvements for the 20T babble (from Auditec, St. Louis, MO) than for the other babbles, especially at 0 dB SNR. Overall, these results indicate that the RNN processing generalized well over babbles with 8–64 competing talkers.

The second performance evaluation compared the feed-forward DNN architecture as used in [Goehring et al. \(2017\)](#) and the RNN architecture used here. The number of hidden units and layers of the DNN were made to be similar to those

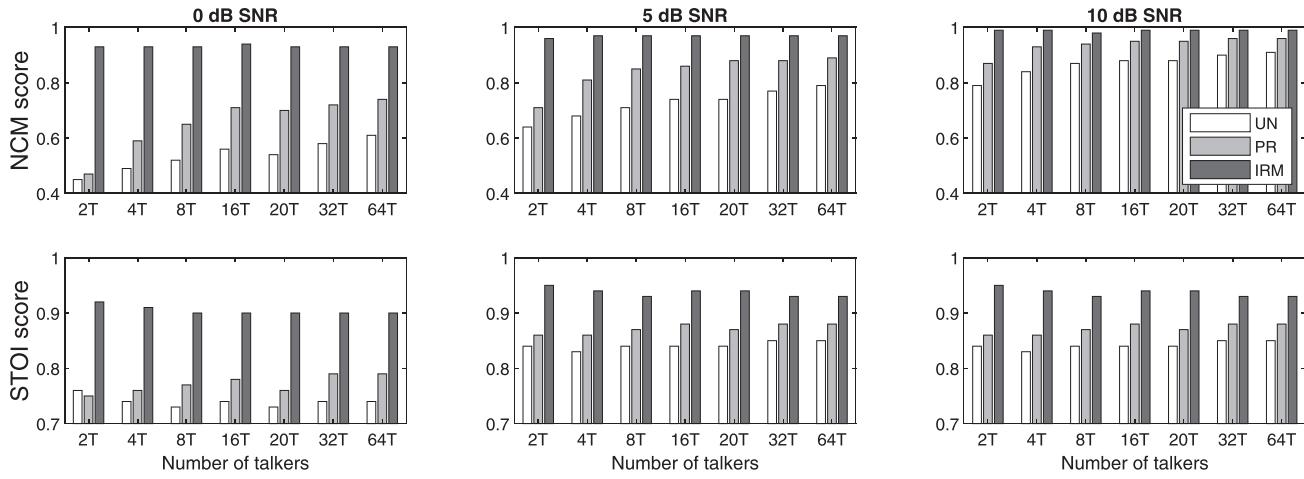


FIG. 2. NCM and STOI scores for seven multi-talker babbles using 2–64 different talkers (2T–64T) and at 0, 5, and 10 dB SNR. UN, PR, and IRM scores are shown for each noise condition.

for the RNN and the same feature set was used. The training data and training procedure were the same as for the RNN. The results for speech in the 20T babble are shown in Fig. 3. The NCM metric predicted larger improvements in SI for the RNN than for the DNN for all three SNRs, while the STOI metric predicted larger improvements for the SNRs of 0 and 10 dB with similar outcomes for the SNR of 5 dB. On average, the relative improvements predicted by STOI and NCM were 38% for the DNN and 46% for the RNN, indicating an advantage of the RNN of about 8 percentage points. It should be noted that the RNN provided the largest benefit

over the DNN of about 15 percentage points on average for the SNR of 10 dB, which represents a condition that is challenging for many CI users (Boyle *et al.*, 2013; Goehring *et al.*, 2017; Croghan and Smith, 2018).

Several measures of the accuracy of the ERM were also calculated, including the MSE, the classification score (HIT-FA score calculated as hit rate, HIT, minus false-alarm rate, FA; Kim *et al.*, 2009; Goehring *et al.*, 2017), and the NCM and STOI scores for the RNN-processed signals used for the listening experiments. The results are shown in Table I for both babble (20T) and traffic noise and for three SNRs, 0, 5, and 10 dB. Scores are shown for the RNN trained using the same class of noise (RNN-B for babble and RNN-T for traffic), and the RNN trained on babble but tested with traffic noise and vice versa. Based on the NCM and STOI scores for condition UN, babble was predicted to lead to lower SI than the traffic noise by an amount equivalent to a change in

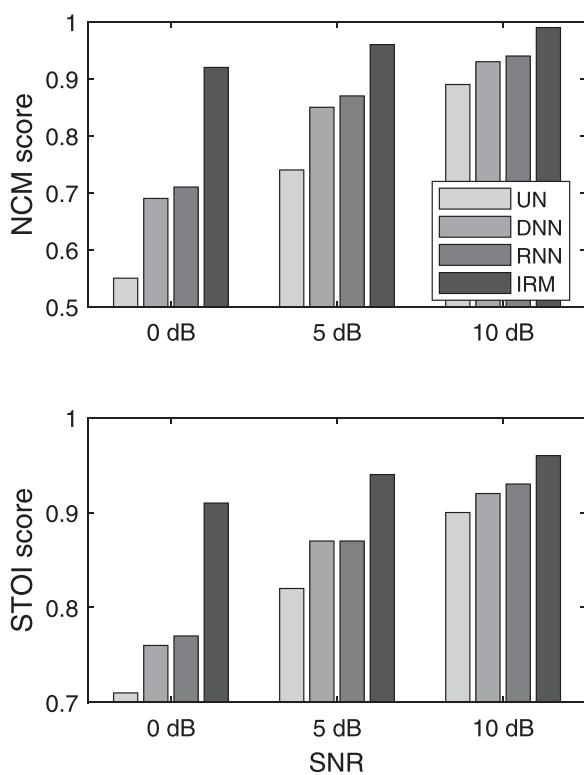


FIG. 3. STOI and NCM scores for speech in the 20T babble at 0, 5, and 10 dB SNR for conditions UN, DNN, RNN, and IRM.

TABLE I. Objective measure scores: HIT-FA alarm rates (with FA scores in brackets), MSE between ERM and IRM, and NCM and STOI scores for the RNN algorithms used in the listening experiment, RNN-B and RNN-T, and UN and IRM in both test noise conditions (20-talker babble and traffic noises) and three SNRs. Results are shown both for matched-noise (RNN-B in babble, RNN-T in traffic) and unmatched-noise (RNN-B in traffic, RNN-T in babble) conditions between training and testing.

Metric	SNR	Tested with babble noise				Tested with traffic noise			
		UN	RNN-B	RNN-T	IRM	UN	RNN-T	RNN-B	IRM
HIT-FA (FA)	0	65 (18)	30 (53)			74 (18)	71 (14)		
	5	78 (9)	46 (42)			80 (10)	77 (9)		
	10	82 (3)	62 (25)			84 (6)	79 (4)		
MSE	0	0.079	0.230			0.064	0.061		
	5	0.039	0.170			0.037	0.041		
	10	0.028	0.100			0.028	0.036		
STOI	0	0.71	0.77	0.74	0.91	0.82	0.86	0.85	0.94
	5	0.82	0.87	0.86	0.94	0.90	0.92	0.91	0.96
	10	0.90	0.93	0.92	0.96	0.94	0.95	0.95	0.98
NCM	0	0.55	0.71	0.63	0.92	0.72	0.81	0.79	0.96
	5	0.74	0.87	0.82	0.96	0.85	0.91	0.90	0.98
	10	0.89	0.95	0.94	0.99	0.93	0.96	0.95	0.99

SNR of about 5 dB. As expected, the RNN models that were trained on a specific type of noise performed best for a noise of that type. For cross-testing, RNN-B performed well with traffic noise, with only slight decreases in estimation accuracy and NCM and STOI values compared to RNN-T. However, the scores for HIT-FA, MSE, and NCM for speech in babble processed with RNN-T were all substantially worse than for babble processed with RNN-B. This suggests that training the RNN using a more difficult noise type (babble) can lead to good generalization to an easier noise type (traffic), but the converse is not the case. In general, the objective measures indicated good estimation performance in terms of HIT-FA scores with acceptable levels of FA (<20%; [Hu and Loizou, 2010](#)) and large improvements for conditions RNN-B and RNN-T over condition UP, as predicted by NCM and STOI. The RNN processing often led to at least 50% of the improvement that the IRM achieved.

III. LISTENING EXPERIMENT 1: CI SIMULATIONS

A. Subjects

Ten native speakers of British English (five female, with an average age of 35 yr and a range of 20–61 yr) with self-reported normal hearing were tested. The subjects were blind as to which condition was being presented and unaware of the goal of the experiment until after testing was complete. Subjects were not used to listening to CI simulations based on vocoder processing. The study was part of a larger research program that was approved by the National Research Ethics committee for the East of England. Before commencing, subjects gave their informed consent and were informed that they could withdraw from the study at any point.

B. CI simulation and listening procedure

All stimuli were processed using the SPIRAL vocoder to simulate CI processing ([Grange et al., 2017](#)). SPIRAL decouples the analysis and carrier stages of the vocoder processing and combines a continuous mixture of envelopes from the analysis filters with a large number of carrier tones to simulate current spread and/or neural degeneration along the cochlea. It has been argued that the SPIRAL vocoder provides a more accurate simulation of the perceptual effects of current spread on speech perception than traditional noise-band or tone vocoders ([Shannon et al., 1995](#); [Oxenham and Kreft, 2014](#)), and resulting speech scores more accurately match those obtained from CI listeners ([Grange et al., 2017; Fletcher et al., 2018](#)). We used 16 analysis filter bands within SPIRAL to represent the 16 electrode channels in CIs from Advanced Bionics (AB, Valencia, CA), and used two current decay slopes of –8 and –16 dB/oct to simulate the effects of current spread observed with typical CIs ([Oxenham and Kreft, 2014](#)). The evaluation stimuli (each at SNRs from –10 to 20 dB) were processed with the SPIRAL vocoder using a sampling rate of 16 kHz and presented to the left ear of the subjects using Sennheiser HD650 circumaural headphones (Sennheiser, Wedemark, Germany) connected to a Roland Quad-Capture external soundcard (Roland, Hamamatsu,

Japan). The setup was calibrated with a KEMAR Artificial Head (GRAS, Holte, Denmark) and HP3561A Signal Analyzer (Hewlett-Packard, Palo Alto, CA) to give a presentation level of 65 dB sound pressure level (SPL), using a noise stimulus with the same long-term spectrum as the target speech. Testing was performed in a sound-attenuating room.

To let the subjects acclimatize to the CI simulation, the test started with the presentation of ten practice sentences in quiet, ten sentences in babble and UN, and ten sentences in babble processed with the RNN algorithm (PR) at 10 dB SNR with the text presented on a screen (and equally split between current spread settings of –8 and –16 dB/oct). Next, a one-up, one-down adaptive procedure ([MacLeod and Summerfield, 1990](#)) was used to measure the speech reception threshold (SRT) at which 50% of the sentences in babble were understood correctly. A trial was deemed correct if all three keywords in that sentence were correctly repeated by the subject. The starting SNR was –4 dB, which was chosen to give low intelligibility, and the step size was 2 dB. The first sentence from a randomly chosen list was repeated until it was correctly understood before the remaining 14 sentences from that list were presented in random order. The average SNR used with the last ten sentences was taken as the SRT for that run. If the adaptive procedure called for a SNR below –10 dB, the SNR was kept at –10 dB, but the adaptive track continued (this was never the case for conditions UN and PR). There were two processing conditions (UN, PR) and two current spread simulations (–8, –16 dB/oct), giving four conditions in total. Three runs were performed for each condition, giving 12 runs in total. The order of the 12 runs was randomized for each subject. Note that only the 20T babble was used, as the objective measures predicted this to be more difficult than the traffic-noise condition.

C. Results

Figure 4 shows individual results for the ten subjects and the group average for conditions UN and PR and both simulated current spread values. As expected, the SRTs were lower (better) for the –16 dB/oct condition than for the –8 dB/oct condition by 4.7 dB for condition UN and 6.2 dB for condition PR. For the simulated current spread of –16 dB/oct, the average SRT was 7.3 dB for condition UN and 4.4 dB for condition PR. All ten subjects showed lower SRTs for PR than for UN, the difference ranging from 1.5 to 4.5 dB. For the simulated current spread of –8 dB/oct, the average SRT was 12 dB for condition UN and 10.6 dB for condition PR. All subjects but one showed better speech reception for condition PR than for condition UN, the difference ranging from –1.0 to 2.8 dB. A two-way, repeated-measures analysis of variance (ANOVA) was conducted with factors processing condition (UN and PR) and simulated current spread (–8 dB/Oct and –16 dB/Oct). There were significant effects of processing condition [$F(1,9)=43.6, p < 0.001$], simulated current spread [$F(1,9)=93.8, p < 0.001$], and a significant interaction [$F(1,9)=5.9, p = 0.022$]. Post hoc tests with Bonferroni correction for each of the two simulated current spread settings showed significant differences between

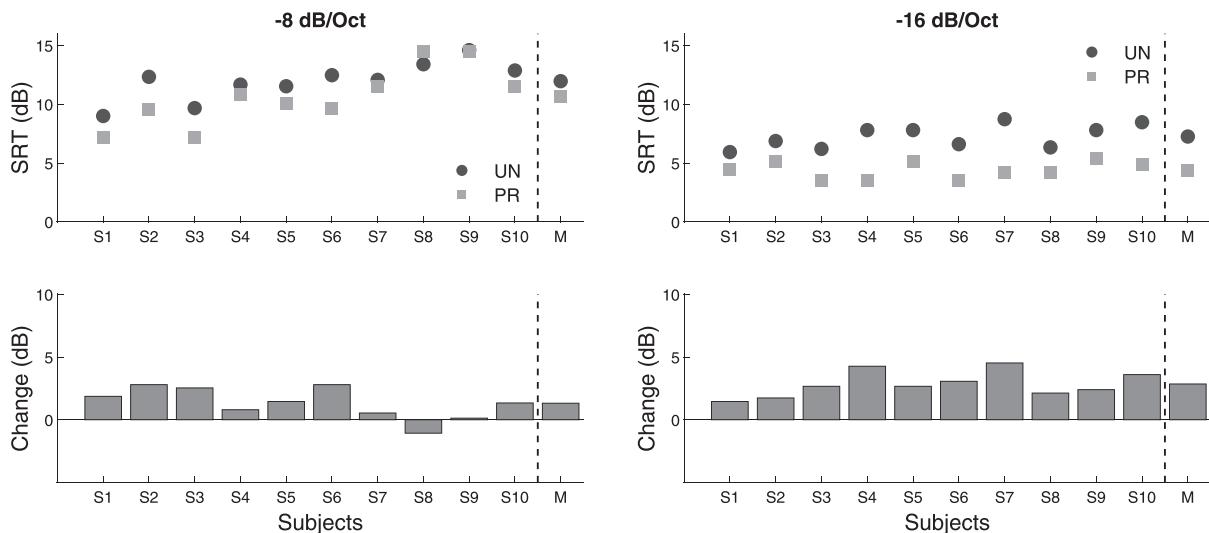


FIG. 4. (Top) Individual and group mean (M) SRTs for the NH subjects listening to CI simulations for conditions UN and PR and the two simulated current spread settings of -8 dB/Oct and -16 dB/Oct . The background was 20T babble. (Bottom) The difference in SRT between conditions UN and PR.

UN and PR for both -8 dB/Oct [$t(9)=3.3, p=0.009$] and -16 dB/Oct [$t(9)=8.8, p<0.001$] current spreads.

D. Discussion

The results were consistent with the predictions of the objective measures and showed substantial mean benefits in SRT of between 1.4 and 2.9 dB for speech in 20T babble. There were significant effects of the simulated current spread, with higher SRTs (worse performance) for the -8 dB/oct spread and a larger benefit of the RNN processing for the -16 dB/oct spread. While the former effect was expected due to the greater spectral smearing produced by the -8 dB/oct spread, the latter effect was somewhat surprising, as it may indicate that the RNN processing would be less beneficial for CI listeners with lower spectral resolution. However, the most likely explanation for the reduced benefit of the RNN processing with the greater current spread is the fact that with this spread some listeners struggled to understand the speech even without babble. This explanation is supported by the observation that the two subjects who performed worst in condition UN (S8,S9) also received the smallest benefit (S9) or even a degradation of performance with PR (S8). In contrast, the two subjects with the best performance in condition UN (S1 and S3) showed substantial benefits in SRT of 1.9 and 2.5 dB. It is likely that the simulated spread of -8 dB/oct is more suitable for simulating CI users who struggle with speech understanding in quiet than for simulating CI users who mainly struggle when noise is present. For the simulated current spread of -16 dB/oct , the average SRT for condition UN was 7.3 dB (ranging from 5.9 to 8.7 dB), which is consistent with SRTs obtained with well-performing CI users (e.g., 6.7 dB for the same 20T babble in [Goehring et al., 2017](#); 7.9 dB for a 4T babble in [Croghan and Smith, 2018](#)). Our SRTs are also consistent with those of [Grange et al. \(2017\)](#), who reported that for speech-shaped noise a current spread setting of -16 dB/oct yielded SRTs with SPIRAL that matched those found for CI users.

IV. LISTENING EXPERIMENT 2: CI USERS

A. Subjects

Ten post- or peri-lingually deafened native speakers of British English were tested (six female, mean age of 65 years with a range from 49 to 74 years). Subjects were unilaterally implanted users of an AB HiRes 90K CI with a minimum of 3 years of experience with their device (mean duration of implant use of 5.5 years). During testing, the subjects listened only with their implanted ear. If a subject usually wore a hearing aid in the other ear, it was taken off during the experiment. Prior to the experiment, the most recent clinical map was obtained for each subject (usage experience with the current maps ranged from 10 months to 2 years). Demographic and device information for the subjects is given in Table II.

The study was part of a larger research program that was approved by the National Research Ethics committee for the East of England. Before commencing, subjects gave their informed consent and were informed that they could withdraw from the study at any point. Subjects were paid for taking part and reimbursed for travel expenses.

B. Technical setup and study design

The acoustic stimuli were presented via a Harmony speech processor (AB, Valencia, CA) that was battery powered and worn by the subject during the listening tests. The stimuli were delivered to the subject using an external USB soundcard (Roland UA-55 Quad-Capture USB, Hamamatsu, Japan) that was connected to the auxiliary (AUX) input port of the processor with an audio cable provided by AB, and with the input from the microphone disabled. The use of a clinical AB speech processor for this part of the experiment ensured that the stimuli did not exceed limitations in output current and comfortable listening levels, as specified in the individual clinical map of the subject. The most recent clinical map of the subjects was used, and adaptive pre-processing functions were switched off (e.g., adaptive noise reduction). Most subjects used a AB HiRes Optima-S

TABLE II. Subject demographics: sex, age (years), etiology of deafness, duration since implanted (years), duration of deafness (years), device type, electrode type, coding strategy, pulse width (μ s), and implanted ear. (n.a., not available.)

Subject	Identifier	Sex	Age	Etiology of deafness	Duration implanted	Duration of deafness	Device	Electrode	Strategy	Pulse width	Implanted ear
S1	AB25	f	65	Sinus infection/post-ling.	3	34	Naida Q90	HiFocus MS	HiRes Optima-S	18	R
S2	AB6	f	70	Unknown/peri-lingually	6	65	Naida Q70	HiFocus 1J	HiRes Optima-S	35	R
S3	AB20	m	73	Unknown/post-lingually	3	45	Naida Q90	HiFocus MS	HiRes Optima-S	29.6	R
S4	AB2	f	59	Possible ototoxicity/post-lingually	3	58	Naida Q70	HiFocus 1J	HiRes Optima-S	31.4	L
S5	AB5	m	76	Otosclerosis/post-lingually	9	27	Harmony 90K	HiFocus 1J	HiRes-S w/ Fidelity 120	18	L
S6	AB23	f	57	Enlarged vestibular aqueduct/ post-lingually	3	58	Naida Q90	HiFocus MS	HiRes Optima-S	23.3	R
S7	AB24	f	49	Unknown/post-lingually	3	4	Naida Q90	HiFocus MS	HiRes Optima-S	36.8	L
S8	AB3	m	72	Otosclerosis/post-lingually progression	11	36	Naida Q70	HiFocus 1J	HiRes Optima-S	29.6	L
S9	AB26	f	57	Unknown/post-lingually	5	21	Naida Q70	HiFocus MS	HiRes Optima-S	22.4	L
S10	AB19	m	74	Unknown	3	n.a.	Naida Q90	HiFocus MS	HiRes Optima-S	30.5	L

strategy but S5 used a AB HiRes-S Fidelity 120 strategy. Subjects were allowed to take breaks when required, and the whole testing procedure took about 2.5 h.

Initially, the input to the speech processor was adjusted to the most comfortable level using a randomly chosen sentence in quiet. The level was then kept constant. An adaptive procedure similar to that for experiment 1 was used to measure the SRT. There were three processing conditions (UN, PR, IRM) and two noise conditions (babble and traffic noise), giving six conditions in total. The two noise conditions were tested in two separate blocks whose order was counterbalanced across subjects. Three runs were performed for each condition. The order of the nine runs per block was randomized for each subject.

After the SI measurements were completed, a subjective quality rating procedure was used in accordance with ITU-T P.835 (Hu and Loizou, 2008). Subjects were asked to rate the stimuli in terms of speech distortions (SDs), background noise intrusiveness (NI), and overall speech quality (OQ). Subjects used a graphical user interface (GUI; programmed in MATLAB, MathWorks, Natick, MA) that allowed them to play a sentence in noise by clicking on one of three cursors (numbered 1–3), one for each processing condition (UN, PR, IRM). The task was to place the three cursors on continuous scales arranged horizontally in the GUI window (with labels left and right: for SD, “not distorted” to “very distorted”; for NI, “not intrusive” to “very intrusive”; for OQ, “bad quality” to “excellent quality”). For each trial, with a given sentence in noise, the subject had to position each of the three cursors in each of the three types of scale, giving nine judgments per trial. For every trial, the initial locations of the cursors within the scales were chosen randomly and the scales were assigned to a range of arbitrary units from 0 to 100, with higher scores reflecting better ratings. In total, each subject completed 20 trials, based on 20 sentences drawn from the BKB corpus and mixed with either babble or traffic noise (10 sentences per noise, equally split between SNRs of 10 and 4 dB). The subjects were blind as to which condition was being presented and which condition was associated with each cursor.

C. Results

Figure 5 shows box plots of the SRTs for the three processing conditions for speech in babble (left) and traffic noise (right). Overall performance was best for condition IRM, with SRTs of -8.0 and -8.6 dB (close to the minimum of -10 dB) in babble and traffic noise, respectively, and worst for condition UN, with SRTs of 7.9 and 2.8 dB, respectively. The RNN algorithm (PR) led to improvements in SRTs relative to condition UN by 3.4 and 2 dB for babble and traffic noise, respectively.

A two-way, repeated-measures ANOVA was conducted with factors processing condition (UN, PR, and IRM) and noise type (babble, traffic). There were significant effects of processing condition [$F(2,18) = 273.2, p < 0.001$] and noise type [$F(1,9) = 53.3, p < 0.001$] and a significant interaction [$F(2,18) = 14.6, p < 0.001$]. Mauchly’s test showed no violation of sphericity for any of these effects. Bonferroni-corrected *post hoc* tests revealed highly significant differences between all three pairs of processing conditions (UN vs PR, $p = 0.006$; UN vs IRM, $p < 0.001$; PR vs IRM, $p < 0.001$).

The performance of the RNN algorithm was assessed further by comparing the SRTs for conditions UN and PR without including the IRM condition. The individual SRTs for conditions UN and PR are shown in Fig. 6. For the

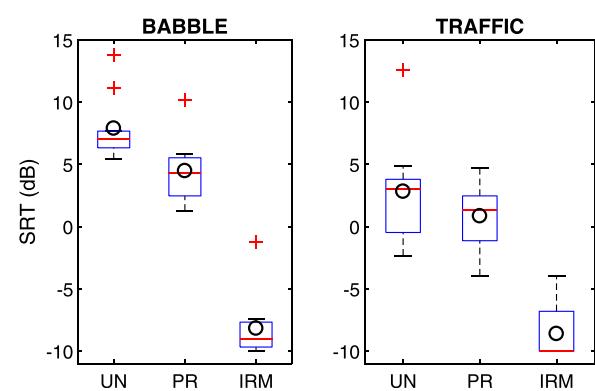


FIG. 5. (Color online) CI group mean SRTs (circles) and box plots for conditions UN, PR, and IRM for speech in babble and traffic noise.

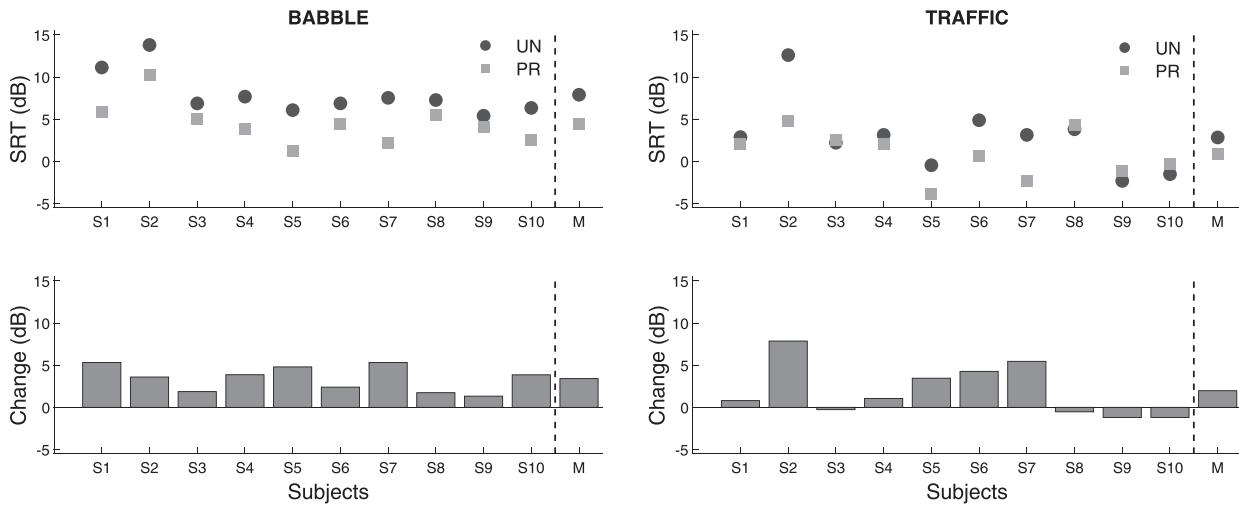


FIG. 6. (Top) Individual and group mean (M) SRTs for the CI subjects and conditions UN and PR for babble and traffic noise. (Bottom) The difference in SRTs between conditions UN and PR.

babble, all subjects performed better with PR than with UN with a mean benefit of 3.4 dB. For the traffic noise, results were mixed, with six subjects showing benefits with PR and four subjects showing worse SRTs. A two-way, repeated-measures ANOVA was conducted with factors processing condition (UN and PR) and noise type (babble and traffic). There were significant effects of processing condition [$F(1,9) = 72.5, p = 0.002$] and noise type [$F(1,9) = 86.8, p < 0.001$] but no significant interaction [$F(1,9) = 2.6, p = 0.144$]. *Post hoc* tests with Bonferroni correction for each of the two noise types showed a significant difference between conditions UN and PR for babble [$t(9) = 7.2, p < 0.001$] but not for traffic [$t(9) = 1.9, p = 0.077$].

The results of the subjective rating procedure are shown in Fig. 7 for each processing condition and noise type. Mean scores were higher for condition PR than for condition UN for all conditions, with improvements ranging from 17 to 50 units for babble and 12 to 33 units for traffic noise. The improvements were larger for NI than for SD. The benefits for OQ were intermediate. Condition IRM was always rated highest, with improvements over UN from 23 to 55 units for babble and 21 to 59 units for traffic noise. The magnitude of the improvements for IRM over UN was similar across the different types of ratings. For both PR and IRM, there were smaller benefits in terms of SD at 4 dB SNR, due to better ratings for condition UN.

A four-way, repeated-measures ANOVA was conducted with factors rating scale (SD, NI, and OQ), SNR (4 and 10 dB), processing condition (UN, PR, and IRM), and noise type (babble and traffic). To reduce the effects of the bounded range of the rating scores, for statistical analysis the scores were transformed using the rationalized arcsine transform (RAU; Studebaker, 1985). Following this transform, the scores for each condition were approximately normally distributed. There were significant effects of SNR [$F(1,9) = 24.9, p = 0.001$], processing condition [$F(1,1,10,1) = 35.5, p < 0.001$, using the Greenhouse-Geisser correction for a violation of sphericity] and noise type [$F(1,9) = 45.7, p < 0.001$] and significant interactions between rating scale and processing condition

[$F(1,9,17,9) = 8.0, p = 0.004$] and between SNR and processing condition [$F(1,5,13,7) = 16.5, p < 0.001$]. No further effects were significant. For the main effect of processing condition, *post hoc* tests with Bonferroni correction showed significant differences between conditions UN and PR ($p = 0.002$), UN and IRM ($p < 0.001$), and PR and IRM ($p = 0.001$).

D. Discussion

The results for CI subjects showed significant improvements in SRTs with the RNN processing over condition UN for the babble but not for the traffic noise. SRTs improved with the RNN processing for all CI subjects for the babble noise, but only for six out of ten subjects for the traffic noise. SRTs were generally higher for the speech in babble than for the speech in traffic noise, with a mean difference of 5.1 dB for the UN stimuli. This may partly explain the observed difference in outcomes, as the RNN algorithm is likely to introduce more estimation errors at lower SNRs. Furthermore, the traffic noise was highly non-stationary with very slow modulations of amplitude (e.g., the sound of a car or bus passing by), and this led to strongly time-varying masking of the speech. The local SNR was likely to be strongly negative for the more masked parts of the speech, resulting in large estimation errors of the RNN algorithm and therefore no benefits or even some degradation of SI for those parts. This effect may have been exacerbated by the high SNR of 5 dB used for training of the RNN algorithm. This was chosen beforehand based on typical performance with the babble background, but it was less appropriate for the easier traffic noise background.

The subjective ratings showed that, relative to condition UN, the RNN processing gave significant benefits in terms of less SD, less intrusiveness of the background noise, and better OQ for both babble and traffic noise. These benefits were larger for the babble background than for the traffic noise background, consistent with the SRTs. While there were substantial improvements of between 12 and 55 units for PR over UN, the IRM condition was rated best in all comparisons, reflecting the limited accuracy of the ERM.

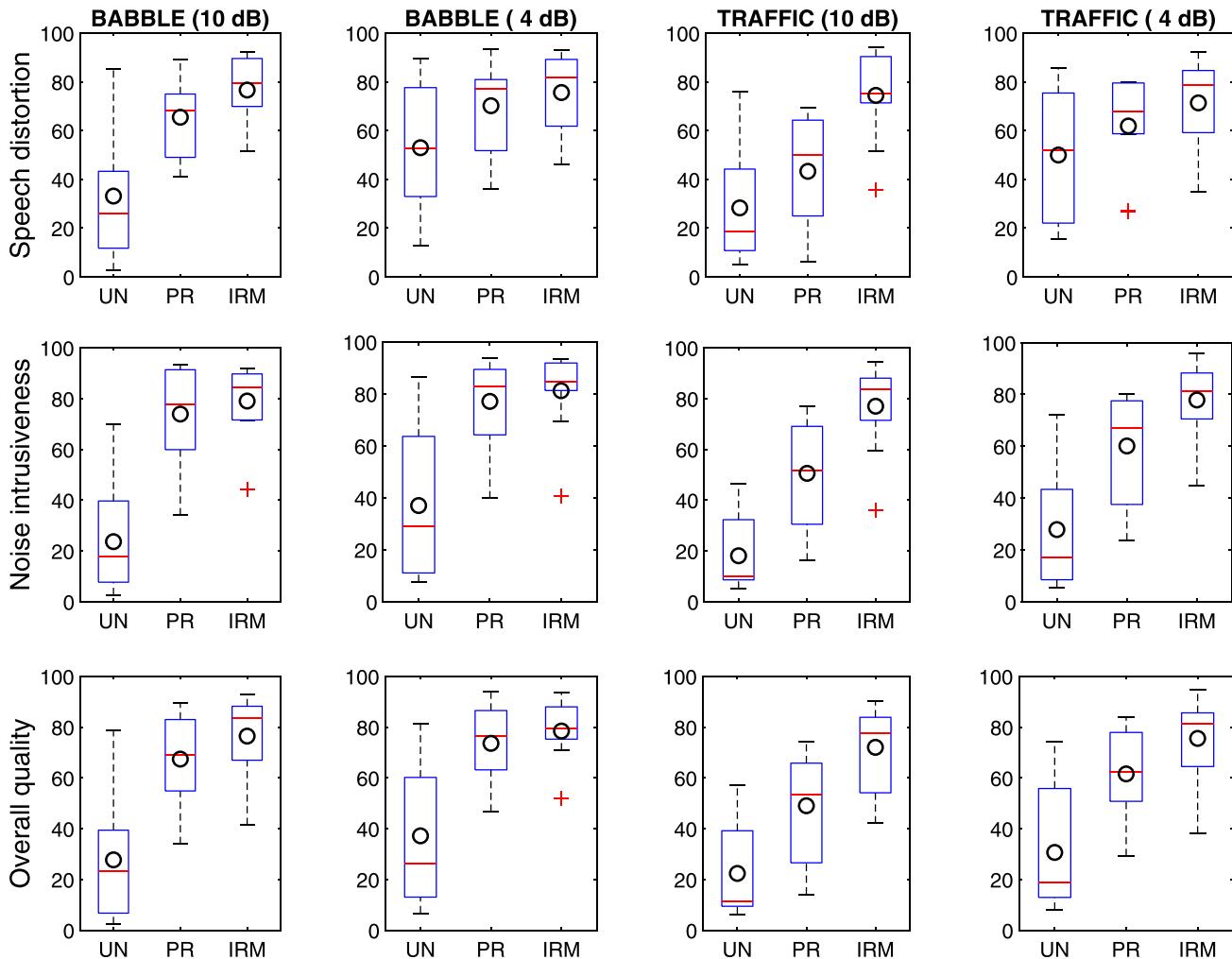


FIG. 7. (Color online) Subjective ratings shown as boxplots and mean scores (circles) for SD, NI, and OQ for conditions UN, PR, and IRM, and SNRs of 4 and 10 dB. The backgrounds were babble (columns 1 and 2) and traffic (columns 3 and 4). Lower scores indicate more negative ratings (e.g., “very distorted” and “bad quality”).

Improvements in subjective ratings were larger for NI than for SDs by about a factor of 2. This indicates that the RNN algorithm was successful in reducing the background noise while keeping SDs at tolerable levels. However, the RNN algorithm led to smaller benefits in terms of SDs for the lower SNR than for the higher one due to better ratings of SDs for condition UN. This may have occurred because of larger estimation errors at the lower SNR, leading to SDs with the RNN that were more comparable to those for condition UN. It may also have occurred because some CI subjects struggled to “hear out” the speech signal from the background at the lower SNR with condition UN and therefore gave ratings of less distortions of the speech than for the higher SNR.

The results for condition IRM showed large improvements of about 10–20 dB in the SRTs for both backgrounds and all subjects. This shows that—in theory—there is room for further improvements in SRT using the RNN or similar approaches via improved accuracy of the ERM. It should be noted that a maximum attenuation of 20 dB was applied for condition IRM (and for PR) and this could have limited the benefits of condition IRM at very low SNRs. This limit could be changed easily or even optimized for different acoustic

environments and/or user preferences. In addition, the processed stimuli for conditions PR and IRM were re-synthesized using the phase information from the noisy speech and this introduces distortions. This problem could be avoided if the RNN algorithm were integrated into the speech processor of a CI device and applied directly to the CI filter bank envelopes so as to avoid the re-synthesis of the signals that was done here. Even with these potential limitations in the IRM condition, all subjects reached the lowest possible SNR of –10 dB during at least one adaptive track. This further supports the IRM as a strong target for RNN training since it can provide very large improvements in SI and SQ for CI subjects.

V. GENERAL DISCUSSION

The results of experiment 2 indicate that the speech-in-babble perception of CI users was improved using the RNN algorithm. There were significant improvements of the SRTs, with improvements up to 2.9 dB for NH subjects listening to CI simulations (experiment 1) and 3.4 dB for CI subjects. The performance of the CI subjects for speech in babble was typical for the CI population, with a mean SRT

for condition UN of 7.9 dB (similar to SRTs reported by Goehring *et al.*, 2017, and Croghan and Smith, 2018). There was also a mean improvement of 2 dB in CI users' SRTs for speech in traffic noise, but this was not statistically significant, and some CI subjects performed worse with the RNN algorithm than without, by up to 1.2 dB. However, for the CI subjects, SRTs for speech in traffic noise were significantly lower than for speech-in-babble noise, by about 5 dB. Therefore, the CI users would have less need for noise reduction when the background was traffic noise.

Subjective ratings of the CI group showed significantly lower SDs, less intrusiveness of the background noise, and better overall quality for condition PR over condition UN for both babble and traffic noise. This is an interesting finding and shows that CI listeners were sensitive to changes in sound-quality characteristics due to the processing. The subjective ratings are consistent with the SRTs and indicate that CI subjects may prefer the RNN processing over no processing in terms of subjective quality.

While these results are consistent with improvements in speech reception reported in previous studies that evaluated ML-based algorithms for CI users (Hu and Loizou, 2010; Goehring *et al.*, 2017; Lai *et al.*, 2018), there were some important differences in the design that make the current findings an important confirmation of this approach and extend its practical application to more unseen acoustic conditions. Most importantly, the RNN algorithm was evaluated on a novel speaker and background noise, neither of which was included in the training data, and the algorithm was evaluated for SNRs that were different from the single SNR used for training. Despite the "unseen" nature of the talker, background and SNR, the RNN algorithm led to a significant 3.4-dB mean improvement in SRT for speech in babble for CI users. This is larger than the 2-dB improvement reported for a speaker- and noise-dependent DNN algorithm by Goehring *et al.* (2017). The greater benefit found here can be explained by the better generalization performance of RNN over DNN approaches, as shown by computational studies based on objective SI predictions (Kolbæk *et al.*, 2017; Chen and Wang, 2017), and the larger training dataset and better training algorithm than used by Goehring *et al.* (2017). Direct comparisons with the results of Hu and Loizou (2010) and Lai *et al.* (2018) are more complicated because they used different test noises and measured percentage correct scores at a fixed SNR, but they also found improvements in speech reception for babble noise using CI subjects. In addition, Hu and Loizou (2010) and Lai *et al.* (2018) used the same speaker for the training and testing datasets, which further limits the comparability of the results.

It should be noted that the RNN algorithm here was trained using a range of noises of the same type as the test noise, so the RNN can be described as an environment-specific algorithm. Many hearing aids and some CIs include some form of scene analysis to identify the acoustic environment (May and Dau, 2013; Launer *et al.*, 2016; Lai *et al.*, 2018), and in principle such an analysis could be used to determine when processing using the RNN algorithm should be activated.

Interestingly, the SRTs for the CI subjects were very similar to the SRTs for the NH subjects listening to CI simulations when using the "more focussed" current-spread setting of -16 dB/oct. Mean SRTs for condition UN were 7.9 and 7.3 dB for CI and NH subjects, respectively, while those for condition PR were 4.5 and 4.4 dB, respectively. This indicates that the vocoder simulation with the more focussed current spread setting was successful in simulating the speech-reception performance of a group of CI subjects when listening to speech in babble and in conditions UN and PR. This extends the results of Grange *et al.* (2017), who reported similar SRTs for CI subjects and NH subjects listening to stimuli processed with SPIRAL for speech in speech-shaped noise. However, it remains unknown if the SRTs would have been similar for simulated and real CI subjects for speech in traffic noise. Also, CI simulations cannot readily account for the very large individual differences in speech reception that are found for CI subjects.

The objective measures, NCM and STOI, showed that the RNN algorithm trained with the set of babble noises generalized better to traffic noise than the other way around. This could indicate that training of a RNN using noises that lead to high SRTs leads to better generalization than training with noises that lead to low SRTs, and/or it could mean that the training dataset for traffic noise did not utilize the full potential of the RNN algorithm, due to less variability in the training data. Interestingly, the NCM and STOI metrics predicted a SRT difference between babble and traffic noise for condition UN of about 5 dB, which corresponds to the difference found in the experiment with CI subjects. Consistent with the data, the NCM and STOI metrics predicted that the improvement produced by the RNN algorithm relative to condition UN would be smaller for traffic noise than for babble noise (10% smaller relative improvement). It should be noted that the NCM and STOI metrics were not designed to predict SI for CI listeners. However, the results indicate that the pattern of differences between conditions can be predicted for CI listeners to a certain degree, perhaps because the metrics are based on the temporal envelopes in different frequency bands, and these are the cues that are conveyed to CI listeners. However, the objective measures failed to predict the variability found within the CI population and overestimated the benefit of the RNN processing for speech in traffic noise.

If a CI user mainly conversed with a few specific people, the performance of the RNN algorithm could be further improved by training using speech from those specific people, as was shown by Goehring *et al.* (2017) for a DNN algorithm. Bramsløw *et al.* (2018) argued that such a system would be practical for applications in future hearing devices, where users could choose spouses, family members, and friends and use recordings of their voices to train the algorithm. This is feasible in practice because just a few minutes of recorded speech for a given speaker seems sufficient for training (Kim *et al.*, 2009; Bolner *et al.*, 2016; Goehring *et al.*, 2017; Bramsløw *et al.*, 2018). However, this approach would not ameliorate communication difficulties in situations with speakers for whom the RNN was not trained, as would be the case for many social and professional situations. These situations can have a

tremendous impact on a person's career prospects and overall well-being, and avoidance of such social interactions due to communication difficulties can lead to mental health problems such as depression or anxiety (Huber *et al.*, 2015; Choi *et al.*, 2016). For communication situations with unknown speakers, our speaker-independent approach, optimized for a specific acoustic environment, would be more suitable, especially when combined with an environmental sound classifier (May and Dau, 2013; Lai *et al.*, 2018), as mentioned above. With respect to the external validity of our test setup, CI subjects informally described the background noises as sounding realistic and similar to those in everyday environments with comments such as "lots of people talking" or "like being in a pub" for the babble and "a car or lorry going past" or "like being in traffic" for the traffic noise. This indicates that the experiment used testing stimuli that were representative of everyday listening situations encountered by CI users.

Improving the speech-in-noise performance of CI users is one of the most important challenges for research and development of future CI devices, as CI users typically spend large proportions of their daily usage time in noisy situations (Busch *et al.*, 2017). The results of this study confirm and extend the promising findings of previous studies based on ML techniques to ameliorate speech-in-noise difficulties for users of CI devices, and future implementations of this approach will hopefully be incorporated in CI devices.

VI. SUMMARY AND CONCLUSIONS

A RNN algorithm was trained to enhance speech in non-stationary babble and traffic noise and shown to provide benefits for speech perception using objective measures and two listening experiments, one with CI simulations and one with CI users. The RNN was trained using speech from many talkers mixed with real-world recordings of multi-talker babble or traffic noise and evaluated using an unknown talker and unseen noise recording of the same type as for the training noise, using a range of SNRs. The objective measures indicated small benefits of using a RNN over a DNN, and predicted that RNN processing would lead to improvements in SI. These predictions were confirmed for speech in babble by the results of the two listening experiments; mean SRTs across conditions were improved significantly by between 1.4 and 3.4 dB. Performance was comparable for the NH subjects listening to a CI simulation and for real CI subjects when a CI simulation with a current-spread setting of -16 dB/oct was used. However, for traffic noise the RNN did not give a significant benefit for the CI subjects. The CI subjects performed better overall for speech in traffic noise than for speech in babble. For traffic noise, the low SNRs in the region of the SRT meant that the RNN algorithm had to operate under conditions where there were likely to be significant errors in the ERM. This may account for the limited benefit of RNN processing for speech in traffic noise.

Relative to condition UN, RNN processing led to significant improvements in subjective ratings of the CI subjects for SDs, NI, and OQ for speech in both babble and traffic noise. This indicates that subjects would prefer RNN processing over no processing. However, processing using the IRM was always rated as highest, and this IRM processing

led to improvements in SRT of 10–15 dB and significantly better speech-quality ratings than with the RNN algorithm, indicating room for further improvements in the RNN algorithm.

ACKNOWLEDGMENTS

We thank the subjects who took part in this study. This work was funded by Action on Hearing Loss (UK, Grant No. 82 and Flexi Grant No. 92). Author B.C.J.M. was supported by the Engineering and Physical Sciences Research Council (EPSRC) (UK, Grant No. RG78536). We also thank two reviewers for very helpful comments.

¹<https://www.musicradar.com/news/tech/sampleradar-286-free-real-world-fx-samples-467432> (Last viewed 20 November 2018).

- 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*.
- Bench, J., Kowal, Å., and Bamford, J. (1979). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," *Br. J. Audiol.* **13**, 108–112.
- Bentsen, T., May, T., Kressner, A. A., and Dau, T. (2018). "The benefit of combining a deep neural network architecture with ideal ratio mask estimation in computational speech segregation to improve speech intelligibility," *PLoS One* **13**, e0196924.
- Boll, S. (1979). "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust.* **27**, 113–120.
- Bolner, F., Goehring, T., Monaghan, J. J., Van Dijk, B., Wouters, J., and Bleek, 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)*, Shanghai, China, pp. 6520–6524.
- Boyle, P. J., Nunn, T. B., O'Connor, A. F., and Moore, B. C. J. (2013). "STARR: A speech test for evaluation of the effectiveness of auditory prostheses under realistic conditions," *Ear Hear.* **34**, 203–212.
- 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.
- Busch, T., Vanpoucke, F., and van Wieringen, A. (2017). "Auditory environment across the life span of cochlear implant users: Insights from data logging," *J. Speech, Lang. Hear. Res.* **60**, 1362–1377.
- Carlyon, R. P., Long, C. J., Deeks, J. M., and McKay, C. M. (2007). "Concurrent sound segregation in electric and acoustic hearing," *J. Assoc. Res. Otolaryngol.* **8**, 119–133.
- 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., and Wang, D. (2014). "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **22**, 1993–2002.
- 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.
- Choi, J. S., Betz, J., Li, L., Blake, C. R., Sung, Y. K., Contrera, K. J., and Lin, F. R. (2016). "Association of using hearing aids or cochlear implants with changes in depressive symptoms in older adults," *JAMA Otolaryngol. Neck Surg.* **142**, 652–657.
- Croghan, N. B. H., and Smith, Z. M. (2018). "Speech understanding with various maskers in cochlear-implant and simulated cochlear-implant hearing: Effects of spectral resolution and implications for masking release," *Trends Hear.* **22**, 2331216518787276.

- Cullington, H. E., and Zeng, F.-G. (2008). "Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects," *J. Acoust. Soc. Am.* **123**, 450–461.
- Dawson, P. W., Mauger, S. J., and Hersbach, A. A. (2011). "Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus® cochlear implant recipients," *Ear Hear.* **32**, 382–390.
- Fletcher, M. D., Mills, S. R., and Goehring, T. (2018). "Vibro-tactile enhancement of speech intelligibility in multi-talker noise for simulated cochlear implant listening," *Trends Hear.* **22**, 2331216518797838.
- Fonseca, E., Pons, J., Favory, X., Font, F., Bogdanov, D., Ferraro, A., Oramas, S., Porter, A., and Serra, X. (2017). "Freesound datasets: A platform for the creation of open audio datasets," in *Proceedings of the 18th ISMIR Conference*, Suzhou, China, International Society for Music Information Retrieval, Canada, pp. 486–493.
- Fu, Q.-J., Shannon, R. V., and Wang, X. (1998). "Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing," *J. Acoust. Soc. Am.* **104**, 3586–3596.
- Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM NIST speech disc 1-11," in *NASA STI/Recon Technical Report n 93*.
- 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. 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., 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.
- Grange, J. A., Culling, J. F., Harris, N. S. L., and Bergfeld, S. (2017). "Cochlear implant simulator with independent representation of the full spiral ganglion," *J. Acoust. Soc. Am.* **142**, EL484–EL489.
- Graves, A., Mohamed, A., and Hinton, G. (2013). "Speech recognition with deep recurrent neural networks," in *2013 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 6645–6649.
- Healy, E. W., Delfarah, M., Johnson, E. M., and Wang, D. (2019). "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation," *J. Acoust. Soc. Am.* **145**, 1378–1388.
- Healy, E. W., 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.
- Hersbach, A. A., Arora, K., Mauger, S. J., and Dawson, P. W. (2012). "Combining directional microphone and single-channel noise reduction algorithms: A clinical evaluation in difficult listening conditions with cochlear implant users," *Ear Hear.* **33**, 13–23.
- Hochreiter, S., and Schmidhuber, J. (1997). "Long short-term memory," *Neural Comput.* **9**, 1735–1780.
- Holube, I., and Kollmeier, B. (1996). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoust. Soc. Am.* **100**, 1703–1716.
- Hu, Y., and Loizou, P. C. (2008). "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio. Speech. Lang. Process.* **16**, 229–238.
- Hu, Y., and Loizou, P. C. (2010). "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users," *J. Acoust. Soc. Am.* **127**, 3689–3695.
- Huber, M., Burger, T., Illg, A., Kunze, S., Giourgas, A., Braun, L., Kröger, S., Nickisch, A., Rasp, G., Becker, A., and Keilmann, A. (2015). "Mental health problems in adolescents with cochlear implants: Peer problems persist after controlling for additional handicaps," *Front. Psychol.* **6**, 953.
- Keshavarzi, M., Goehring, T., Turner, R. E., and Moore, B. C. J. (2019). "Comparison of effects on subjective intelligibility and quality of speech in babble for two algorithms: A deep recurrent neural network and spectral subtraction," *J. Acoust. Soc. Am.* **145**, 1493–1503.
- 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**, 233121651877096.
- 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.
- Kingma, D. P., and Ba, J. (2014). "Adam: A method for stochastic optimization," *arXiv:1412.6980*.
- Kolbaek, 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/ACM Trans. Audio Speech Lang. Process.* **25**, 1901–1913.
- Lai, Y.-H., Tsao, Y., Lu, X., Chen, F., Su, Y.-T., Chen, K.-C., Chen, Y.-H., Chen, L.-C., Po-Hung, L. L., and Lee, C.-H. (2018). "Deep learning-based noise reduction approach to improve speech intelligibility for cochlear implant recipients," *Ear Hear.* **39**, 795–809.
- 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.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). "Deep learning," *Nature* **521**, 436.
- Loizou, P. C., Lobo, A., and Hu, Y. (2005). "Subspace algorithms for noise reduction in cochlear implants," *J. Acoust. Soc. Am.* **118**, 2791–2793.
- MacLeod, A., and Summerfield, Q. (1990). "A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use," *Br. J. Audiol.* **24**, 29–43.
- Madhu, N., Spriet, A., Koning, R., and Wouters, J. (2013). "The potential for speech intelligibility improvement using the ideal binary mask and the ideal wiener filter in single channel noise reduction systems: Application to auditory prostheses," *IEEE Trans. Audio. Speech. Lang. Process.* **21**, 63–72.
- Mauger, S. J., Dawson, P. W., and Hersbach, A. A. (2012). "Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction," *J. Acoust. Soc. Am.* **131**, 327–336.
- May, T., and Dau, T. (2013). "Environment-aware ideal binary mask estimation using monaural cues," in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4.
- May, T., and Dau, T. (2014). "Requirements for the evaluation of computational speech segregation systems," *J. Acoust. Soc. Am.* **136**, EL398–EL404.
- Monaghan, J. J. M., Goehring, T., Yang, X., Bolner, F., Wang, S., Wright, M. C. M., and Bleeck, S. (2017). "Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners," *J. Acoust. Soc. Am.* **141**, 1985–1998.
- Moore, B. C. J., and Carlyon, R. P. (2005). "Perception of pitch by people with cochlear hearing loss and by cochlear implant users," in *Pitch Perception*, edited by C. J. Plack, A. J. Oxenham, R. R. Fay, and A. N. Popper (Springer, New York), pp. 234–277.
- Oxenham, A. J., and Kreft, H. A. (2014). "Speech perception in tones and noise via cochlear implants reveals influence of spectral resolution on temporal processing," *Trends Hear.* **18**, 1–14.
- Patterson, R. D., Allerhand, M. H., and Giguere, C. (1995). "Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform," *J. Acoust. Soc. Am.* **98**, 1890–1894.
- Scalart, P., and Filho, J. V. (1996). "Speech enhancement based on *a priori* signal to noise estimation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Atlanta, GA.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.* **15**, 1929–1958.
- Stickney, G. S., Zeng, F.-G., Litovsky, R., and Assmann, P. (2004). "Cochlear implant speech recognition with speech maskers," *J. Acoust. Soc. Am.* **116**, 1081–1091.
- 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.
- 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.
- Tang, Y. (2016). “TF Learn: TensorFlow’s high-level module for distributed machine learning,” [arXiv:1612.04251](https://arxiv.org/abs/1612.04251).
- Veaux, C., Yamagishi, J., and MacDonald, K. (2016). “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit,” University of Edinburgh, Edinburgh, UK.
- 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*, pp. 91–99.
- Wouters, J., and Vanden Berghe, J. (2001). “Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system,” *Ear Hear.* **22**, 420–430.