Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners

Jessica J. M. Monaghan, Tobias Goehring, and Xin YangFederico BolnerShangqiguo Wang, Matthew C. M. Wright, and Stefan Bleeck

Citation: The Journal of the Acoustical Society of America **141**, 1985 (2017); doi: 10.1121/1.4977197

View online: http://dx.doi.org/10.1121/1.4977197

View Table of Contents: http://asa.scitation.org/toc/jas/141/3

Published by the Acoustical Society of America

Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners[a)]

Jessica J. M. Monaghan,[b)] Tobias Goehring, and Xin Yang
Institute of Sound and Vibration Research, University of Southampton, Southampton, United Kingdom

Federico Bolner[c)]
ExpORL, Katholieke Universiteit Leuven, Leuven, Belgium

Shangqiguo Wang, Matthew C. M. Wright, and Stefan Bleeck
Institute of Sound and Vibration Research, University of Southampton, Southampton, United Kingdom

(Received 7 August 2016; revised 23 December 2016; accepted 9 February 2017; published online 22 March 2017)

Machine-learning based approaches to speech enhancement have recently shown great promise for improving speech intelligibility for hearing-impaired listeners. Here, the performance of three machine-learning algorithms and one classical algorithm, Wiener filtering, was compared. Two algorithms based on neural networks were examined, one using a previously reported feature set and one using a feature set derived from an auditory model. The third machine-learning approach was a dictionary-based sparse-coding algorithm. Speech intelligibility and quality scores were obtained for participants with mild-to-moderate hearing impairments listening to sentences in speech-shaped noise and multi-talker babble following processing with the algorithms. Intelligibility and quality scores were significantly improved by each of the three machine-learning approaches, but not by the classical approach. The largest improvements for both speech intelligibility and quality were found by implementing a neural network using the feature set based on auditory modeling. Furthermore, neural network based techniques appeared more promising than dictionary-based, sparse coding in terms of performance and ease of implementation.
 [http://dx.doi.org/10.1121/1.4977197]

[GCS] Pages: 1985–1998

I. INTRODUCTION

Individuals with hearing impairment often have difficulty recognizing speech in background noise. In a UK survey of individuals fitted with a hearing aid (HA), a quarter of those who reported never wearing their aids indicated "lack of benefit in noisy situations" as their reason for not doing so (Kochkin, 2000). Together with the finding that HA users are more tolerant of background noise than are hearing-impaired (HI) individuals who do not choose to use aids (Nabelek *et al.*, 2006), this suggests that solving the problem of background noise could allow more HI people to benefit from HAs. One means of reducing the detrimental effect of noise on recognizing speech is to employ speech-enhancement algorithms (sometimes referred to as noise-reduction algorithms) to improve intelligibility. Single-channel speech-enhancement algorithms operate on the input from a single microphone and are therefore ideally suited to being incorporated into HA processing.

Traditional approaches to single-channel speech enhancement have demonstrated only limited success. For some noise conditions, the "auditory masked threshold noise suppression" technique (Tsoukalas *et al.*, 1997) increased recognition for both HI and NH listeners (Arehart *et al.*, 2003), whilst a sparse-code shrinkage algorithm tested in our laboratory improved speech intelligibility in speech-shaped noise (Sang *et al.*, 2014) and quality in speech-shaped and babble noise (Sang *et al.*, 2015) for HI listeners. Other studies reported no benefit to word recognition for HI listeners with single-channel enhancement algorithms, but did report an increase in listener preference (e.g., Bentler *et al.*, 2008; Zakis *et al.*, 2009; Luts *et al.*, 2010), including increased acceptable background noise level for HI listeners (Mueller *et al.*, 2006; Fredelake *et al.*, 2012). Overall, any improvements in intelligibility have been small and limited to stationary noise types (Loizou and Kim, 2011).

Recently, machine-learning approaches have shown great promise for improving speech intelligibility both for hearing-impaired and normal-hearing listeners (e.g., Healy *et al.*, 2013; Healy *et al.*, 2015; Bolner *et al.*, 2016) as well as for cochlear implant users (Goehring *et al.*, 2016). Rather than calculating a gain function based on estimates of the speech and noise statistics from the incoming signal—the classical approach—machine-learning approaches incorporate prior knowledge of patterns of speech and noise to estimate the optimal gain function to be applied to the incoming signal. Gaussian mixture models have been used to improve speech intelligibility for normal hearing listeners (Kim *et al.*, 2009) and for cochlear implant users (Hu and Loizou, 2010). Healy *et al.* (2013) demonstrated large improvements in speech intelligibility scores for both NH and HI listeners

[a)]Portions of this work were presented at the 24th European Signal Processing Conference (EUSIPCO 2016).
[b)]Current address: The Australian Hearing Hub, Sydney, Australia. Electronic mail: jessica.monaghan@gmail.com.
[c)]Also at Cochlear Technology Centre, Mechelen, Belgium.

using a deep neural network (DNN) algorithm. The main limitations to these approaches, however, were the very large classification systems (256 mixtures/128 sub-band networks), and the specificity of the training set required to achieve this level of performance. These studies used the same noise recordings for both the training and testing stages of the algorithm. A match between training and testing data is likely to overestimate the performance of the algorithm in unseen test conditions. May and Dau (2014) showed that the use of novel noise realizations for testing yielded a substantial decrease in estimation performance with a GMM-based system, such as the one used by Kim *et al.* (2009). More recently, it has been shown for both NH and HI listeners, that DNN-based algorithms can generalize well to novel realizations of the same noise type (Healy *et al.*, 2015; Bolner *et al.*, 2016; Goehring *et al.*, 2016) or to completely novel types of noise (Chen *et al.*, 2016). Chen *et al.* (2016) showed that a DNN-based algorithm that used cochleagram features similar to the first part of the AIM features in this study can generalize to novel types of noise for a speaker-dependent system. This was a promising result and was based on simple spectral features indicating that further reductions in terms of algorithm complexity may be achieved by omitting complex feature extraction stages and using large-scale training.

For potential application in hearing devices such as HAs, algorithms must fulfill the requirements of low computational complexity due to the restricted capacities of HAs in terms of memory and computational power (Löllmann and Vary, 2009) and low processing delay due to the perceptual requirements of HA users (Stone and Moore, 1999). For feedforward neural network algorithms with fully connected layers, the number of units in the consecutive layers defines the computational complexity of the algorithm (each unit in a given layer is connected to each unit in the next layer via a weight parameter). While it is unclear what the current limits for applications in hearing devices in terms of memory and computational complexity are, neural network (NN) algorithms with millions of parameters are unlikely to be implementable on current hearing devices. This motivates a decrease in the size and complexity of the NNs to allow for potential real-time operation on mobile devices (Bolner *et al.*, 2016; Goehring *et al.*, 2016).

Another machine-learning technique, dictionary-based sparse coding, has been used successfully in image denoising (Elad and Aharon, 2006), but there have been few applications to speech enhancement. This approach is attractive from a biomimetic perspective; evidence suggests a sparse representation of ecologically relevant features in both the auditory (DeWeese *et al.*, 2003; Lewicki, 2002) and visual (Olshausen and Field, 1996) systems. Following the training stage, in which the algorithm learns a "dictionary" of typical speech features from many examples of clean speech segments, an estimate of the clean signal is reconstructed using a linear combination of relatively few of these dictionary components (i.e., the representation is "sparse"). If the noise is sufficiently dissimilar to speech, it will not have a sparse representation over the dictionary; a linear combination of hundreds of dictionary elements would be required to accurately reconstruct a segment of noise, compared with the handful required to accurately reconstruct a speech segment. The sparse reconstruction, therefore, will preserve more of the speech energy than the noise energy. A neural analogy for the dictionary would be a very large set of neurons, each responding to one specific speech component. If these neurons continued to respond to these particular speech components, even in a noisy background, this would provide robustness to coding speech in noise, since familiar, speech-like elements would be better represented than unfamiliar, noise-like components. In the current study, we tested a new dictionary based sparse-coding approach to see if it can improve speech intelligibility.

Here, we first assess the performance of neural networks with greatly reduced complexity that have shown promising results in our previous study with NH listeners (Bolner *et al.*, 2016) to determine whether it is still possible to obtain improvements in speech intelligibility for HI listeners with more practically feasible algorithms than other studies used (e.g., Healy *et al.*, 2015; Chen *et al.*, 2016). We also assess the performance of a novel machine learning algorithm known as sparse coding, and compare it with both a classical approach, Wiener filtering, and with DNNs. Third, we determine the performance of the DNN approach when it derives its input from an auditory model, comparing its performance with that of an algorithm employing the standard spectrum-based feature vectors of previous studies. Finally, as well as speech recognition scores, we compare the performance of these three approaches in terms of their sound-quality ratings. Each of the four algorithms was assessed in both stationary (speech-shaped) noise and multi-talker babble noise conditions and at signal-to-noise ratios (SNRs) of 0 and +4 dB.

II. SPEECH ENHANCEMENT ALGORITHMS

A. Wiener filtering

Wiener filtering was one of the first noise reduction algorithms to be developed (Lim and Oppenheim, 1979), and has been implemented in commercial HAs. In order to obtain the noisy speech spectrum, a short-time Fourier transform (STFT) is performed. The clean speech spectrum X is then estimated as $\hat{X}$ using the following equation:

$$\hat{X}_k = \sqrt{\frac{\xi_k}{1+\xi_k}} Y_k, \tag{1}$$

where ξ is the *a priori* SNR, Y is the noisy signal magnitude and k indexes the Fourier components. The estimate of the clean signal (used to calculate the *a priori* SNR) is derived by minimizing the difference between the clean and enhanced complex speech spectra, taking into account the phase spectra. The Wiener filter is the optimal estimator of the clean speech spectrum when the speech and noise signals are independent Gaussian processes. Scalart and Filho (1996) reported that using an estimate of the *a priori* (rather than *a posteriori*) SNR in (1) would give superior enhancement. Their method was employed in the current study. The

noise magnitude spectrum was estimated on a frame-by-frame basis using the algorithm of Gerkmann and Hendriks (2011) to estimate the *a priori* SNR and calculate the gain function for each frame and frequency component.

Hu and Loizou (2007) tested a number of single-channel speech enhancement algorithms and found that the Wiener filtering algorithm described by Scalart and Filho (1996) was the only algorithm that enhanced speech recognition for NH listeners, although this improvement was evident in only one condition (automobile noise at 5 dB SNR). Levitt *et al.* (1992) found that consonant recognition in non-stationary cafeteria babble noise was significantly increased for half of HI listeners but significantly reduced for half of NH listeners when a Wiener filter was applied. In that study, the gain of the filter was calculated by assuming knowledge of the consonant and noise spectra. This indicated that Wiener filtering can be beneficial for some HI listeners if the filter gain is approximated accurately enough. Luts *et al.* (2010) tested a Wiener filter algorithm that estimated the noise and speech spectral densities from the signal (in a different way from that employed in the current study) but found no improvements in the recognition of speech in babble noise by NH and HI listeners. Nevertheless, listeners in that study preferred the enhanced speech over the unprocessed condition. Wiener filtering was included in the current study to determine whether a traditional single-channel speech enhancement algorithm could provide improvements in speech intelligibility or quality in the conditions tested.

B. Neural networks

The next two algorithms to be tested also employed the Wiener gain function to estimate the clean speech signal. The principal difference between these algorithms and the classical Wiener filtering algorithm described in the previous chapter was the use of a more sophisticated approach to estimate the Wiener filter gain, namely, the use of an artificial NN algorithm.

The NN algorithm consisted of two parts: a front-end that extracted acoustic features from the noisy input signal and a back-end that employed a multi-layer feedforward neural network to estimate the ideal Wiener filter gain [see Eq. (1)] in each frequency channel. The estimated gain was used to enhance the noise-corrupted input signal by applying it to the noisy envelopes after the signal had been passed through a 63 channel gammatone filter bank ranging from 50 to 8000 Hz (Patterson *et al.*, 1987; Hohmann, 2002). A schematic of the NN algorithm is shown in Fig. 1.

The first processing stage for these algorithms was to split the input signal ($f_s = 16$ kHz) into 20-ms long time-frames with 10 ms overlap. Then, two sets of acoustic features were extracted from the broadband signal of each input frame: a comparison feature set (NN_COMP) similar to those used in previous studies (Healy *et al.*, 2013; Healy *et al.*, 2015) and a novel auditory-model based feature set (NN_AIM). Both feature sets comprised several sub-features that were concatenated per timeframe and directly fed to the input layer of the NN. This yielded two distinct NN algorithms: NN_COMP and NN_AIM.

The second processing stage was performed by the NN, which consisted of an input layer with a number of units determined by the dimensionality of the feature set, two hidden layers with 100 and 50 units using saturating linear transfer functions and an output layer with linear activations. The output layer had a dimensionality of 63 given by the number of gammatone frequency channels used for calculating the target Wiener filter gain function. The output layer activations of the NN were taken as the estimated Wiener filter gains and applied to the noisy envelopes. The NN was trained using the resilient backpropagation algorithm (Riedmiller and Braun, 1993) to minimize the mean squared error between the estimated and ideal Wiener filter gain in each gammatone frequency channel. The NN was trained in full-batch mode over 500 epochs using weight decay regularization of 0.5 to avoid overfitting. The learning rate was set to 0.01 and weights were updated using increment and decrement factors of 1.2 and 0.5, respectively. These hyper parameters were chosen based on our previous study (Bolner *et al.*, 2016), which yielded improvements in speech perception in noise by NH listeners.

In total 80 sentences (eight lists) from the IEEE database (Rothauser *et al.*, 1969) spoken by a male talker were mixed at 5 SNRs (−2, 0, 2, 4, and 6 dB) to amount to 400 training utterances per noise condition. The training data sets were the same as used for the sparse coding algorithm. A single NN was trained per noise type incorporating all five SNR conditions. As mentioned above, the ideal Wiener filter gain was taken as target signal to be estimated by the NN for each training utterance in each gammatone channel. The target data were calculated using the ground-truth speech and noise signals at the given SNR.

One of the goals of the current study was to assess the performance of more real-time feasible NNs for speech enhancement. Kim *et al.* (2009) and Healy *et al.* (2013) used sub-band classifiers that employed two GMMs or NNs for each frequency channel, yielding large classification

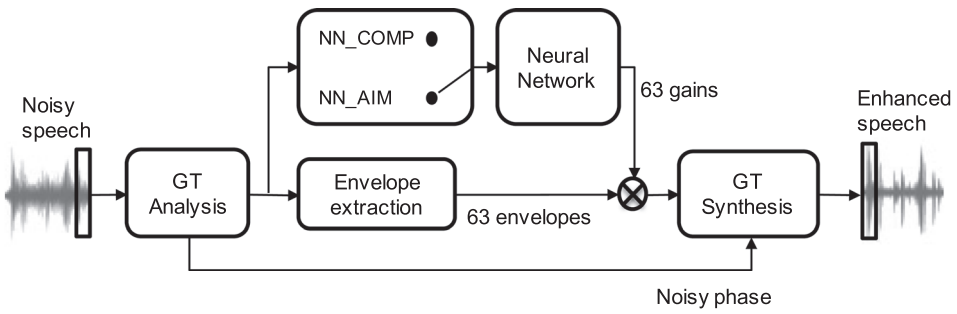

FIG. 1. Schematic of processing for the neural network algorithms.

systems. Healy *et al.* (2014) and Healy *et al.* (2015), and Bolner *et al.* (2016) used a broadband approach that employed a single NN to estimate the target gains for all frequency channels collectively. This approach yielded a large decrease in NN parameters and computational complexity (a 43-fold increase in processing speed was reported by Healy *et al.*, 2014). The memory requirements and the number of calculations performed by the NN per timeframe are determined by the number of NN parameters, consisting of the weight and bias values of the units in the hidden and output layers. In this study, the auditory-model based NN comprised 39 800 parameters, which is a 100-fold or 500-fold decrease in parameters compared with Healy *et al.* (2015) or Chen *et al.* (2016), respectively. Another aspect of real-time processing is the algorithmic processing delay, which is limited to a few milliseconds by the perceptual requirements of users of hearing aids (Stone and Moore, 1999). As reported by Healy *et al.* (2015), the inclusion of future timeframes has to be avoided for real-time processing applications such as HAs. In contrast to 2 future frames in Healy *et al.* (2015) and 11 future frames in Chen *et al.* (2016), the current study used no future frames for the processing to assess a more real-time feasible approach.

1. Comparison feature set

The comparison feature set NN_COMP was generated based on the same set of features used in Healy *et al.* (2013) and Healy *et al.* (2014; "complementary features"). To generate the feature set, the amplitude modulation spectrum (AMS; Tchorz and Kollmeier, 2003), relative-spectral transform and perceptual linear prediction coefficients (RASTA-PLP; Hermansky and Morgan, 1994), and mel-frequency cepstral coefficients (MFCC) were extracted from each 20-ms long timeframe of the noisy speech mixture (broadband features were computed as described in Healy *et al.*, 2014). The concatenated features had a dimensionality of 445 per timeframe (AMS $[25 \times 15]$ + RASTA-PLP $[3 \times 13]$ + MFCC[31]). NN_COMP was extracted from the current timeframe and concatenated with delta (differences between features in consecutive frames) and delta-delta features for RASTA-PLP only (as described in Healy *et al.*, 2014).

2. Auditory feature set

The proposed feature set NN_AIM was extracted using the auditory image model (AIM; Patterson *et al.*, 1995; Bleeck *et al.*, 2004). AIM is a time-domain functional model of auditory processing. It generates a stream of two-dimensional sound representations, referred to as "auditory images," for an acoustic input signal. AIM produces a more stable representation for periodic parts of the input sound, such as for vowels and voiced sounds in speech and tones in music signals, than for non-periodic sounds. The model consists of a cascade of processing stages that simulate peripheral auditory processing, such as pre-cochlear processing, basilar membrane motion (BMM) and the transduction process in the cochlea. Further stages of AIM are intended to model more central auditory processing stages, such as neural activity patterns in the auditory nerve and cochlear nucleus and temporal integration and source size normalization in higher auditory processing stages (finally yielding the size-shape transformed auditory image; SSI). The SSI output of AIM is based on the size covariant processing of the auditory system (Smith *et al.*, 2005; von Kriegstein *et al.*, 2007) and produces the same pattern for vowels spoken by speakers with different glottal pulse rates or vocal tract lengths. The processing of AIM has been reported to improve the SNR of voiced speech and to yield improved performance in automatic speech recognition experiments (Irino and Patterson, 2002; Monaghan *et al.*, 2008; Müller and Mertins, 2012).

The NN_AIM feature set combined the output of two processing stages of AIM: the BMM and SSI. The two features were concatenated to obtain a dimensionality of each feature vector of 315, consisting of 63 BMM features and 252 SSI features. The BMM features were obtained by calculating the logarithm of the envelope power of a linear gammatone filterbank with 63 frequency channels (Hohmann, 2002) and represented predominantly spectral information of the current timeframe. The SSI features were obtained by calculating a two-dimensional discrete cosine transform (DCT) of the SSI output of AIM. The DCT was performed for de-correlation and a reduction of the dimensionality of the SSI. Before the DCT was performed, each SSI channel was downsampled to 400 Hz to reduce the temporal resolution of the data. After performing the DCT using the downsampled signal, only the 2nd to 22nd coefficients were used for the NN_AIM feature set. The first coefficient was omitted since it is related to the overall energy of the SSI and more susceptible to noise degradation and the higher order coefficients above the 22nd were found to be numerically close to zero. The SSI represented both spectral and temporal information of the current timeframe in form of enhanced periodicity information and increased SNR for voiced components of speech signals. The NN_AIM feature set was extracted using only the current timeframe.

3. Objective measures

Two "objective measures"—computationally derived scores intended to predict how well humans will recognize a given sample of noisy or enhanced speech—were used to optimize the performance of the NN and sparse coding algorithms: the Short Time Objective Intelligibility (STOI; Taal *et al.*, 2011) and the Normalized Covariance Metric (NCM; Holube and Kollmeier, 1996). Additionally, for the NN algorithms, hit–false alarms (HIT-FA) and false alarm (FA) rates were determined and used for optimization (Kim *et al.*, 2009). These measures required the estimated gain function to be converted into a binary mask. The hit rate was defined as the percentage of speech-dominated time-frequency bins correctly classified by the binary mask, and the false-alarm rate was defined as the percentage of noise-dominated time-frequency bins incorrectly classified as speech-dominated. During optimization of the algorithms, their performance was assessed using objective measure scores from two sentence lists that were not part of the training set or test set (the set used for the human testing). After testing with the human

listeners had taken place objective measures were also applied to the sentences in the test set to determine the correlation between these measures and the human performance (see Sec. IV D).

C. Sparse coding

The fourth algorithm was a novel speech enhancement algorithm based on dictionary-based sparse coding (Elad and Aharon, 2006). This algorithm was also a machine-learning algorithm but involved a different approach from the NN based algorithms. Rather than estimating a gain function to be applied to the noisy signal (as with the other algorithms tested here), an estimate of the clean filter bank outputs was produced directly.

The algorithm requires a "dictionary" of typical elements of speech, known as "atoms." This dictionary is learned from many frames of clean speech during the training stage. The dictionary is typically over-complete, i.e., the number of atoms in the dictionary is greater than the length of the atoms. Any speech signal can then be approximated by a linear combination of just a few atoms from the dictionary, i.e., it is a "sparse" representation. Because stationary noise is unstructured, and therefore cannot be predicted, it cannot be sparsely represented. Therefore, for noisy speech, the speech signal can more easily be approximated in the form of a sparse code than can the noise, leading to de-noising.

For a noisy speech frame, $\mathbf{y}$, consisting of noise $\mathbf{n}$, and clean speech $\mathbf{x}$:

$$\mathbf{y} = \mathbf{x} + \mathbf{n}. \tag{2}$$

It is assumed that the clean speech can be represented as

$$\mathbf{x} = \mathbf{D}\boldsymbol{\alpha}, \tag{3}$$

where the matrix $\mathbf{D}$ is a dictionary and $\boldsymbol{\alpha}$ is a sparse coefficient vector (i.e., most entries are zero). The estimate of the clean speech is then given by

$$\hat{\mathbf{x}} = \mathbf{D}\hat{\boldsymbol{\alpha}}, \tag{4}$$

where

$$\hat{\boldsymbol{\alpha}} = \min_{\boldsymbol{\alpha}} \|\boldsymbol{\alpha}\|_0, \tag{5}$$

such that

$$\|\mathbf{y} - \mathbf{D}\hat{\boldsymbol{\alpha}}\|_2^2 < \varepsilon. \tag{6}$$

The zero norm of $\boldsymbol{\alpha}$, $\|\boldsymbol{\alpha}\|_0$, is the number of non-zero elements in $\boldsymbol{\alpha}$ and ε is the desired error, which is chosen to be approximately equal to the estimated noise power.

The signals were processed with the same sampling rate, frame length and using the same gammatone filter as in the NN algorithm except that 30 channels were used rather than 63 channels. Increasing the number of channels for the sparse-coding algorithm to 63 did not improve performance (as assessed using the objective measures described in Sec. II B 3) but greatly increased processing time. Each channel of the filter output was normalized to have a root mean square (RMS) amplitude of 1. In the training stage, the dictionary was trained on eight sentence lists from the same speaker and corpus as used in the testing stage. The K-singular value decomposition (KSVD) algorithm (Aharon *et al.*, 2006) was used to train a dictionary for each channel. The Orthogonal Matching Pursuit (OMP) algorithm (Pati *et al.*, 1993) was modified so that the selection of atoms was optimized across all frequency channels. The first stage of the original OMP finds the atom from the dictionary that gives the highest correlation with the noisy signal. Rather than selecting atoms independently for each frequency channel and dictionary, the atom that gave the highest correlation over each frequency channel and dictionary was selected. The corresponding atom from each dictionary was chosen for the other frequency channels. This was intended to capture across frequency correlations in the speech. Five atoms per frame were chosen as the optimal number for training the dictionary during pilot tests.

For the testing stage the average noise power in each channel was estimated using the approach of Gerkmann and Hendriks (2011) and used to define the desired error in each channel (ε). In the denoising stage the least angle regression algorithm (LARS) algorithm (Efron *et al.*, 2004) was used rather than OMP because it was found to give superior performance in terms of both objective measures. For each frame, atoms were selected until the sum over channels of the RMS difference between the noisy and sparse signals was less than the sum of the desired errors for each channel.

A separate approach was used for the babble noise, as proposed by Sigg *et al.* (2012) because in this case the noise and speech are more similar and so the representation of the noise might also be sparse over the speech dictionary. Therefore, in the training stage, in addition to the speech dictionary, a noise dictionary was trained using an example of the babble noise (distinct from the noise segment used in testing). As for the speech dictionary, the KSVD with modified OMP was used to train the noise dictionary. To reduce the similarity of the noise and speech dictionaries (and thus the probability of speech components being misclassified as noise), any atom with a correlation greater than 0.95 was removed from the noise dictionary. In the testing stage the noise and speech dictionaries were concatenated and the LARS algorithm was used to find a fit to the noisy speech. Atoms selected from the noisy dictionary were discarded, and only atoms from the speech dictionary were used to reconstruct the signal. The values of the free parameters were optimized using objective measure scores from a sentence list not used in the training or testing sets.

A similar approach proposed by Sigg *et al.* (2012) operated instead on the STFT domain. Rather than using the reconstructed speech signal directly (as in the current study) Sigg *et al.* used it to estimate the speech and noise magnitude to calculate the Wiener filter gain [see Eq. (1)] which was then applied to the original noisy speech. An objective measure of speech quality (the cepstral distance) showed improvement for their approach relative to multiband spectral subtraction (Kamath and Loizou, 2002) and a vector

quantization based approach, but tests with human listeners were not performed.

III. METHODS

Seventeen native speakers of British English (seven female, median age 65 years, IQR 13 years) with mild to moderate sensorineural hearing loss were recruited. Volunteers were recruited using advertisements at the University of Southampton and the Southampton local community, such as churches and libraries. A screening process was performed and participants were excluded from the study if they failed the screening. As part of the screening, otoscopy and tympanometry were performed to check for normal ear-canal anatomy and normal middle ear function. A questionnaire was given to exclude any recent ear surgery, otalgia, tinnitus and hyperacusis and pure tone audiometry was performed. Participants with unilateral or conductive hearing loss were excluded. All participants were experienced hearing aid users (>1 year use). The audiograms of each of the participants are shown in Fig. 2. The mean pure tone average (PTA) measured at 0.5, 1, and 2 kHz was 31.4 dB hearing level (HL).

Speech recognition in each condition was assessed as the percentage of keywords identified correctly in IEEE (Rothauser *et al.*, 1969) sentences spoken by a British male speaker. Two types of noise were tested: speech-shaped noise (SSN) and multi-talker babble noise. In the SSN conditions, a noise generated to have the same long-term average spectrum as the IEEE sentences was used. For the babble noise conditions, the noise was constructed by mixing different sentences from eight speakers (four male and four female) taken from the TIMIT corpus (Garofolo *et al.*, 1993). Both the SSN and the babble noise were 26 s in duration, 18 s of which was used to train the algorithms and the remaining 8 s of which was used in the testing stage. A

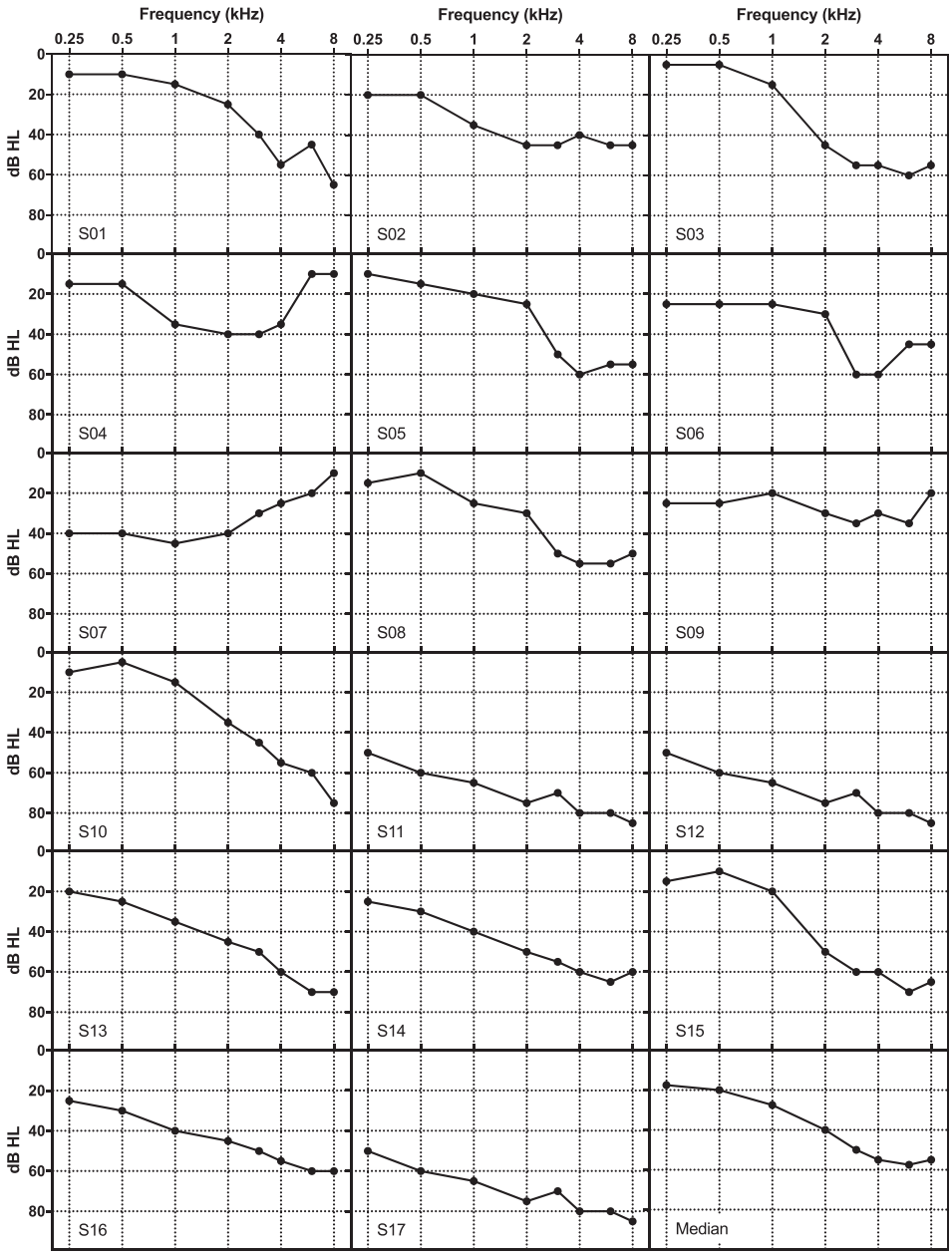

FIG. 2. Audiogram of ear tested for each participant and the median audiogram over all ears tested.

segment of noise was selected at random from the test noise and added to the sentences at SNRs of 0, and +4 dB. Fixed SNRs were used rather than using an adaptive procedure to find the speech reception threshold (SRT; the SNR for which the speech recognition score is 50%), because the goal was to better compare the performance of the algorithms at a specified SNR. Two sentence lists, each comprising ten sentences, were used for each condition. A different, and random, order of conditions was used for each participant, and a Latin square was employed so that the same list would be used in the same condition as seldom as possible. Participants practised the procedure with a different sentence list from the ones tested, with an SNR of +10 dB and no speech enhancement applied.

Custom MATLAB software was used to process and present the stimuli. Pre-processed sentences were loaded using a laptop computer and presented to the participant, who was seated in a quiet room in the clinic, through an RME Babyface soundcard over headphones (HD 380 Pro, Sennheiser, Wedemark, Germany). A finite impulse response headphone filter was designed in MATLAB so that the headphones produced a flat frequency response at the ear reference point (measured using a Brüel & Kjær type 4153 artificial ear with the standard cone YJ0304 above the adapter plate for circumaural headphones, type DB 0843). The spectrum and overall sound level were measured using a Brüel & Kjær type 2250 sound level meter. The stimuli were presented monaurally to the participant's better ear, which was the left ear for nine of the participants. In order to compensate partially for each participant's hearing loss, a linear hearing-loss dependent gain was applied at each audiometric frequency according to the NAL-R prescription formula (Byrne and Dillon, 1986). There was one experimental session lasting approximately 2 hours. Participants were able to have rest breaks if they felt fatigued.

Noisy sentences were generated by setting the level of the clean speech to 65 dB sound pressure level (SPL) and adding noise scaled to give an SNR of 0 or +4 dB. Before amplification was applied, the level of the stimulus (the speech and noise mixture) was approximately 68 dB SPL in the unenhanced 0 dB SNR condition and 66 dB SPL in the +4 dB SNR condition. These noisy sentences were processed by each of the four enhancement algorithms and the corresponding enhanced sentences stored. Because the maximum gain that can be applied in enhancement is unity, enhancement will generally result in attenuation of the speech energy as well as the noise. This may reduce the audibility of the speech and render speech enhancement less effective. So that the level of the speech was unchanged between the enhanced and unenhanced conditions, "shadow-filtering" was used as described in Fredelake *et al.* (2012) for the Wiener filter and Neural Network conditions: the attenuation applied to the speech signal was determined by multiplying the clean speech by the same gain function that was applied to the noisy speech and measuring the corresponding reduction in RMS level relative to the original speech. In the case of the sparse-coding algorithm, there was no gain function applied, so instead the reconstructed signal was set to 65 dB SPL, the level of the clean speech.

The experimenter scored each sentence list and condition using a graphical user interface (GUI) without knowledge of which condition was being presented. After each sentence was presented, the participant was asked to repeat what they had heard as accurately as possible. Using the scoring GUI, the experimenter recorded how many of the keywords the participant had identified correctly. After each sentence list, the participant was asked to rate the perceived quality of the speech ("How would you rate the quality of the speech?"). A paper sheet was provided on which the participant indicated the rating on a scale from 0 to 7 (with labels at 0, "bad"; 4, "fair"; and 7, "excellent"). For finer resolution, there were ten subdivisions for each of its seven values.The data are available at http://dx.doi.org/10.5258/SOTON/D0020.

IV. RESULTS

A. Speech intelligibility

Speech intelligibility in speech shaped and multi-talker babble noise at SNRs of 0 and +4 dB was determined for four speech enhancement algorithms and compared with the corresponding unenhanced conditions. Figure 3 shows the group-mean percentage of key words correctly recognized in

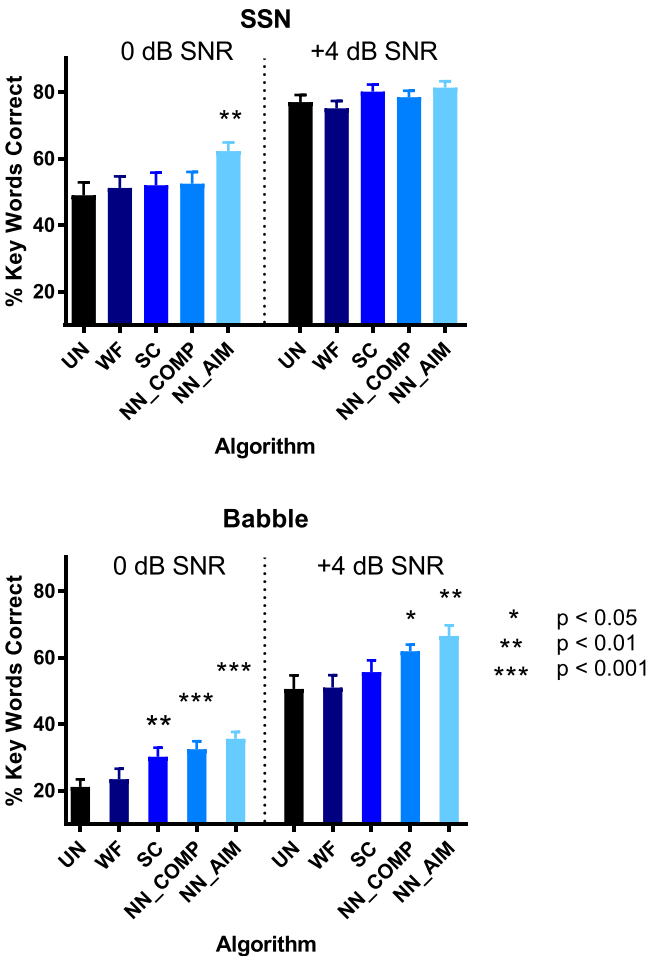

FIG. 3. (Color online) Group-mean percentage of key words correctly recognised for each algorithm in the speech-shaped noise (upper panel) and babble noise (lower panel) conditions. Error bars show standard errors of the mean. Asterisks indicate conditions for which the enhanced scores were significantly different from the unenhanced condition.

each of the four noise conditions. Across all algorithms, performance improved with increasing SNR, and performance was always lower in babble noise than in SSN. In SSN, performance did not differ greatly across algorithm conditions, with the exception of the NN_AIM algorithm at 0 dB SNR. In babble noise, all of the algorithms, other than Wiener filtering, improved performance at least at one SNR, compared with the unprocessed condition. In the SSN conditions, there were significant main effects of algorithm, as determined by a repeated measures two way analysis of variance (ANOVA) [$F(4, 64) = 5.89$, $p < 0.001$], and SNR [$F(1, 16) = 126.88$, $p < 0.001$], but no significant interaction between the two [$F(4, 64) = 2.22$, $p = 0.077$]. Bonferroni-corrected planned comparisons were performed between the unprocessed and enhanced conditions at each SNR. The only significant improvement in speech recognition for SSN was at 0 dB SNR for the NN_AIM algorithm [$F(1, 16) = 17.20$, $p = 0.003$], with a mean gain in intelligibility of 13 percentage points. In the babble noise conditions, there were significant main effects of algorithm [$F(4, 64) = 17.10$, $p < 0.001$] and SNR [$F(1, 16) = 323.85$, $p < 0.001$]. The Greenhouse-Geisser correction was applied when testing the interaction between SNR and algorithm because the assumption of sphericity was violated in this case. The interaction was not significant [$F(36.27, 2.27) = 0.84$, $p = 0.45$]. Bonferroni-corrected planned comparisons were performed between the unprocessed and enhanced conditions at each SNR. The sparse-coding algorithm led to a significant improvement in speech recognition at 0 dB SNR [$F(1, 16) = 17.37$, $p = 0.003$], as did the NN_COMP [$F(1, 16) = 47.56$, $p < 0.001$], and NN_AIM [$F(1, 16) = 114.32$, $p < 0.001$]. Mean gains in intelligibility of 9, 11, and 14 percentage points were obtained for the sparse coding, NN_COMP and NN_AIM, respectively. At +4 dB SNR there were significant improvements in speech recognition scores for both NN_COMP [$F(1, 16) = 11.95$, $p = 0.013$], and NN_AIM [$F(1, 16) = 18.64$, $p = 0.002$]. Mean gains in intelligibility of 11 and 16 percentage points were obtained for the NN_COMP and NN_AIM, respectively.

B. Speech quality

Speech-quality ratings were also determined for each algorithm in each noise condition and for both 0 and +4 dB SNR, and are plotted in Fig. 4. The data were not normally distributed in the majority of conditions, so box and whisker plots are shown and non-parametric statistics were used. As for speech intelligibility, speech quality improved at the higher SNR, and the algorithms elicited greater improvements in babble noise compared with SSN. Wiener filtering was ineffectual in improving speech quality. A non-parametric Friedman's ANOVA indicated a significant effect of algorithm for the SSN at +4 dB [$\chi^2(4) = 12.27$, $p = 0.016$] and the babble noise at 0 [$\chi^2(4) = 22.01$, $p < 0.001$] and +4 dB SNR [$\chi^2(4) = 24.31$, $p < 0.001$]. Bonferroni-corrected planned comparisons were performed between the unprocessed and enhanced conditions at each SNR. The paired-samples sign test was used as the distributions were not all symmetric about the median. In the SSN conditions there was significant improvement in quality ratings for NN_AIM

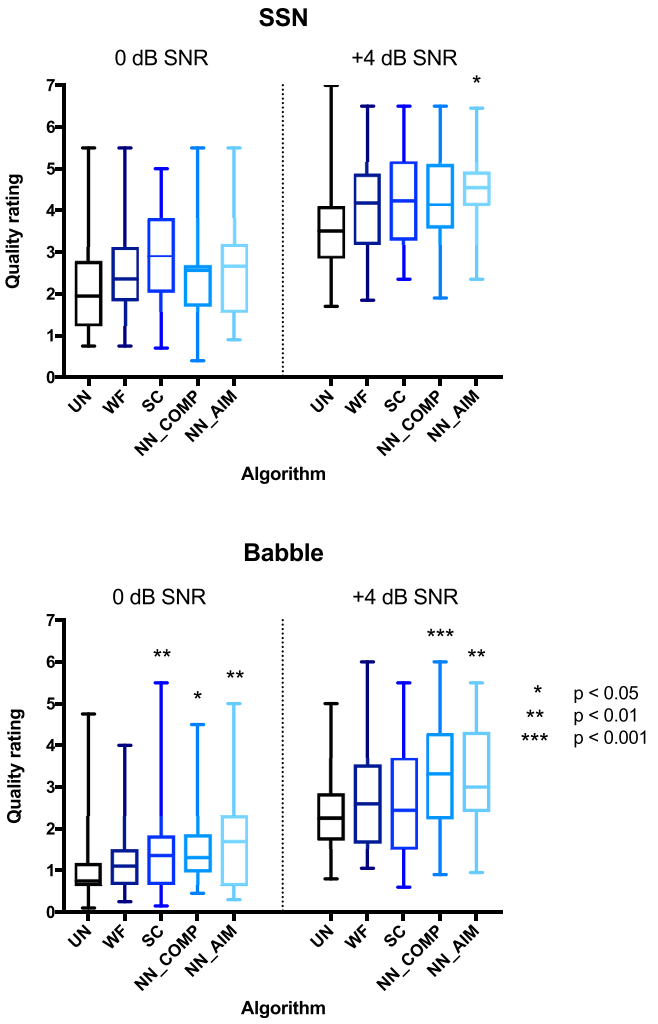

FIG. 4. (Color online) Box-and-whisker plots of speech quality ratings for each algorithm in the speech-shaped noise (upper panel) and babble noise (lower panel) conditions. Whiskers indicate the range (1.5 times the interquartile range). Asterisks indicate conditions for which the enhanced scores were significantly different from the unenhanced condition.

[$p = 0.017$] at +4 dB SNR, with a gain of 0.81 in quality rating. In the babble conditions there were significant improvements at 0 dB SNR for sparse coding [$p = 0.0021$], NN_COMP [$p = 0.017$] and NN_AIM [$p = 0.0073$], with improvements of 0.44 0.51 and 0.66, respectively. At +4 dB SNR there were significant improvements for NN_COMP [$p < 0.001$], and NN_AIM [$p = 0.0011$] of 0.98 and 0.85, respectively.

C. Correlation between speech intelligibility and quality

Overall, there was a moderately high correlation ($r = 0.605$, $p < 0.001$) between intelligibility and quality scores, pooled over all noise conditions and algorithms. This would be expected since intelligibility and quality are both influenced in a similar way by SNR and noise type. However, it is also possible that the quality ratings were biased by the fact that they were elicited following the intelligibility task. Participants may have been inclined to give high quality ratings in those conditions for which they found

TABLE I. For the Neural Network based algorithms HIT-FA and FA scores were calculated for both noise types and SNR conditions. To calculate the HIT-FA scores, the ratio masks (estimated and ideal) were converted to binary masks by applying a local SNR criterion of −5 dB.

	SSN		Babble	
% HIT-FA (% FA)	0 dB	+4 dB	0 dB	+4 dB
NN_COMP	72 (8)	75 (7)	64 (18)	65 (17)
NN_AIM	76 (7)	79 (7)	67 (18)	67 (18)

the intelligibility test less challenging. A partial correlation controlling for the effects of algorithm, SNR, noise type and participant was calculated by fitting a linear mixed model, including these factors, separately for quality rating and intelligibility. The correlation between the residuals of the two models was then calculated. A significant ($p < 0.001$) partial correlation between quality and intelligibility was found with an r value of 0.273. This indicates that there was a small influence of intelligibility on the quality ratings given by the participants, accounting for ∼7% of the variance.

D. Objective measures

In each of the four noise conditions NCM and STOI scores were calculated from all the sentences used in testing for the four algorithms and the unenhanced signals. Descriptive statistics for each condition are reported in Table I. Figure 5 shows the objective measures scores plotted as a function of the final intelligibility scores obtained from the participants in the 20 conditions tested. In the SSN conditions, correlations between speech recognition scores and the objective measures were high, with r^2 values of 0.91 for both NCM and STOI. Correlations were lower in the babble noise conditions, with r^2 values of 0.70 and 0.83 for NCM and STOI, respectively. These results confirm that both NCM and STOI are effective for predicting the intelligibility of sentences for HI listeners for stationary noise and to a lesser extent also for non-stationary noise. For the two neural-network algorithms HIT-FA scores were also calculated and are shown in Table II (see Sec. V A 1 for discussion).

V. DISCUSSION

We assessed the performance of four speech enhancement algorithms in improving speech intelligibility and speech quality in two types of interfering noise, speech-shaped noise (i.e., noise with the same long-term spectrum as speech) and eight-talker babble noise. Algorithms based on sparse coding or neural networks improved performance compared with the unprocessed signal, most notably in babble noise and for the lower of the two SNRs (0 dB) we explored. Wiener filtering—commonly applied in hearing technologies—had no effect on speech intelligibility and speech quality compared with the unprocessed signal. This suggests that machine-learning algorithms, particularly those based on neuro-mimetic principles, can improve speech-in-noise performance in challenging listening conditions with fluctuating background noise.

Improvement in speech intelligibility was modest in the SSN conditions, with only significant improvement apparent for NN_AIM at 0 dB SNR, for which an improvement of 13 percentage points was evident. Subjective listening indicated that the absence of any improvement in the 0-dB condition may be due to the introduction of fluctuating distortions in the signal, counteracting the beneficial effect of noise

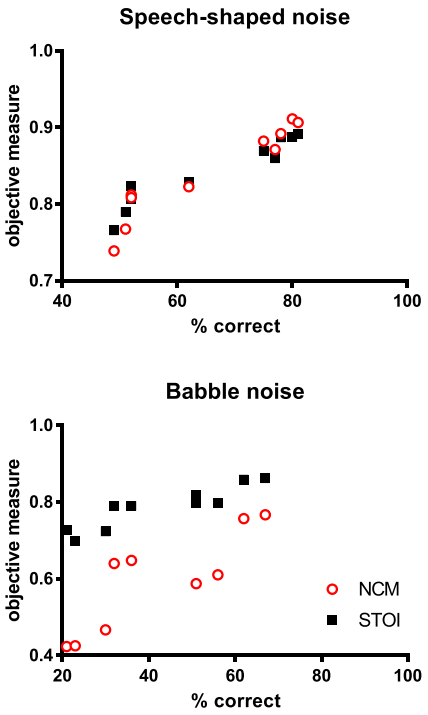

FIG. 5. (Color online) Average values of the objective measures NCM and STOI plotted as a function of the mean intelligibility scores obtained from the participants in each of the ten conditions (five algorithm conditions, two SNRs) for each noise type.

TABLE II. Mean values of NCM and STOI for the sentences used in each condition. Standard deviation are shown in brackets.

	NCM				STOI			
	SSN		Babble		SSN		Babble	
	0 dB	+4 dB	0 dB	+4 dB	0 dB	+4 dB	0 dB	+4 dB
UN	0.74(0.04)	0.87(0.03)	0.42(0.05)	0.59(0.05)	0.77(0.04)	0.86(0.03)	0.73(0.05)	0.82(0.04)
Wiener	0.77(0.04)	0.88(0.03)	0.43(0.05)	0.59(0.05)	0.79(0.04)	0.87(0.03)	0.70(0.05)	0.80(0.05)
SC	0.81(0.03)	0.91(0.03)	0.47(0.05)	0.61(0.04)	0.81(0.04)	0.89(0.03)	0.72(0.05)	0.80(0.04)
NN_COMP	0.81(0.03)	0.89(0.03)	0.64(0.05)	0.76(0.04)	0.82(0.03)	0.89(0.02)	0.79(0.04)	0.86(0.03)
NN_AIM	0.82(0.03)	0.91(0.03)	0.65(0.05)	0.77(0.04)	0.83(0.03)	0.89(0.02)	0.79(0.04)	0.86(0.03)

reduction. Interestingly, greater improvements were seen in babble noise conditions, where algorithms typically perform worse than in stationary noise conditions. In this case, significant improvements were seen for all three machine-learning algorithms at 0 dB SNR and for both neural-network-based algorithms at +4 dB SNR. A similar pattern of results was seen for the speech quality ratings, but this may have been partially due to there being a confounding effect of intelligibility (see Sec. IV C).

Figure 6 shows the group-mean improvement provided by each algorithm for each of the conditions tested, plotted in terms of gain in speech quality as a function of gain in speech intelligibility. Promisingly, almost all algorithms elicited improvements in both quality and intelligibility (albeit not significantly in many cases), although, again, this may be partially due to the confounding effect of intelligibility. Exceptions to this are the Wiener filter at +4 dB SNR for both noises, for which there was a reduction in speech intelligibility despite a small increase in speech quality ratings.

A. NN based algorithms

1. Comparison with other studies

In the current study, improvements in speech intelligibility of 14 and 16 percentage points were found for the NN_AIM in babble, at 0 dB SNR and +4 dB SNR, respectively. These results can be compared with Healy *et al.* (2015) who found improvements of 44.4 and 27.8 percentage points and with Chen *et al.* (2016) who reported 27 and 11.6 percentage points in babble noise at 0 dB SNR and +5 dB SNR, respectively. Note that participants in the current study had milder hearing losses, with a mean PTA of 31.4 dB compared with 50.5 dB in Healy *et al.* (2015) and 42.2 dB in Chen *et al.* (2016). This means that algorithm performance cannot be completely equated and that the lower improvement in SI found in the current study at 0 dB SNR may partly be explained by the milder hearing losses of the participants. However, compared to the current study, improvements were still greater for the NH listeners tested by Healy *et al.* (2015) with an increase of 21 percentage points in the babble condition at −2 dB SNR.

The decrease in performance was predicted by lower HIT-FA scores found for the algorithm in this study in comparison with the HIT-FA scores reported by Healy *et al.* (2015). This indicates a lower estimation quality of the masks in this study, especially concerning the removal of background noise (indicated by higher FA rates). However, the use of ratio masking instead of binary masking makes predictions based on HIT-FA rates less applicable. A better objective comparison of the algorithm performance may be the improvements in terms of STOI scores over the unenhanced condition. The algorithms used in this study achieved smaller improvements in STOI scores than the ones reported by Healy *et al.* (2015) and Chen *et al.* (2016), consistent with the difference in speech intelligibility improvements by the participants.

One difference between the current approach and that of other studies was the size of the neural networks and the training dataset employed. The networks in the current study had two hidden layers with 100 and 50 units (similar to the NNs used by Bolner *et al.*, 2016), whereas Healy *et al.* (2015) used four hidden layers of 1024 units each, and Chen *et al.* (2016) used five hidden layers of 2048 units each. This results in a 100-fold or 500-fold increase in the number of parameters of the networks used in Healy *et al.* (2015) and Chen *et al.* (2016), respectively, and explains partly the performance advantage of the networks used in those studies over the current study. An increase in the size of the training

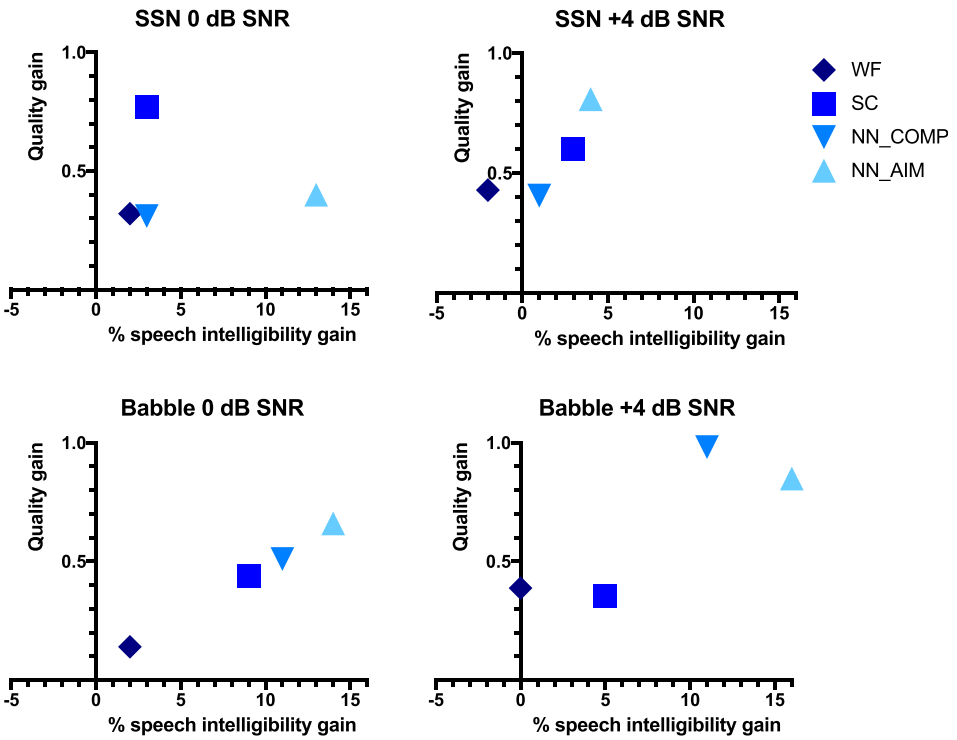

FIG. 6. (Color online) Group-mean improvement in speech quality versus improvement in speech intelligibility for the four algorithms in each noise condition.

dataset allowed the training of larger networks and reduced the risk of overfitting the training data.

Another difference between the algorithm used in this study and previous algorithms is the causality of the processing. The current study uses only the current and past frames as input signals, since this would be the case in real-time implementations. Other studies used a window of 5 (Healy *et al.*, 2013; Healy *et al.*, 2014; Healy *et al.*, 2015) or 23 (Chen *et al.*, 2016) consecutive frames centered at the current frame. In real-time implementations, the inclusion of future frames should be avoided, since this would introduce large processing delays (i.e., >20–30 ms), which most likely would not be tolerated by users of hearing aids (Stone and Moore, 1999).

The findings of this study support the results of Healy *et al.* (2013), Healy *et al.* (2014), Healy *et al.* (2015), and Chen *et al.* (2016), and demonstrate that significant (albeit more modest) improvements in speech intelligibility and quality can be provided by scaled-down neural network approaches that operate in a causal way.

2. Comparison of feature sets

In addition to assessing the speech enhancement performance of neural networks with lower complexity, a further goal of the study was to determine whether using feature vectors derived from an auditory model would improve speech enhancement relative to standard feature vectors. Two sets of feature vectors were assessed using the same model architecture: a set derived from an auditory model, "NN_AIM," and a standard feature vector set for comparison, "NN_COMP."

Although no significant difference was found between intelligibility scores or quality ratings for the two sets of feature vectors, the AIM features gave the highest scores in both dimensions in almost all conditions (see Fig. 6). Indeed it was the only algorithm tested that generated an improvement in the stationary noise conditions. This better performance overall suggests that the auditory model based features are to be preferred in terms of optimizing speech quality and intelligibility. Recently, Chen *et al.* (2016) have shown that a DNN-based algorithm, that used cochleagram features similar to the first part of the AIM features in this study, can generalize to novel types of noise when using the same target speaker. This is a promising result and motivates further investigation of auditory inspired features. For use in real-time mobile devices other considerations must be taken into account, primarily the amount of computational power required to perform the algorithms, and the ability to generate the features in real-time. Although the generation of AIM features requires considerably more computational complexity than standard spectral features or cochleagram features, they can be generated in real time by a modern PC without the use of non-causal information.

A further potential benefit of using AIM feature vectors, not assessed by the current study, is their ability to generalize to different talkers. In the current study, and in previous studies employing neural networks based speech-enhancement techniques (e.g., Healy *et al.*, 2013; Healy *et al.*, 2014; Healy *et al.*, 2015; Bolner *et al.*, 2016), networks were trained and tested on the same talkers. It was recently shown by Goehring *et al.* (2016) that a speaker-independent but noise-specific algorithm for application in cochlear implants improved speech understanding in two out of three noises for CI listeners, but performance was decreased relative to a speaker-dependent algorithm that used *a priori* information about the target speaker. Although a noise reduction system optimized for a particular talker has practical applications, widespread adoption of NN-based speech enhancement will require generalization to novel talkers and novel listening-situations. Unlike traditional feature vectors, such as MFCCs (Monaghan *et al.*, 2008), the AIM features are robust to changes in speaker-size as well as pitch, and so provide a better prospect for good performance with novel speakers.

3. Speech quality ratings

It is important for its general acceptance in hearing devices that an algorithm provides good speech quality (Kochkin, 2000). The results indicate that neural networks produced significant improvements in speech quality in all conditions for which the speech intelligibility was also significantly improved, although, as discussed above (see Sec. IV C) there was found to be a small influence of intelligibility on the quality ratings that may account for some of the improvement seen in intelligibility. One factor in the good quality ratings seen here may be the use of the Wiener filter gain, which has been shown to produce better speech quality than the binary mask (Madhu *et al.*, 2013) which is often used for neural network speech enhancement. However, speech quality scores in all conditions were similar or higher than those for traditional Wiener filtering. Since the gain function used by the neural network approaches used is identical to that used in WF, this indicates that the greater accuracy of speech and noise estimates provided by the neural network is crucial to the quality of the enhanced speech. These speech quality results support the promise of neural networks as a good candidate for speech enhancement for hearing aids. Recently, Williamson *et al.* (2016) have shown that further increases of speech quality can be achieved in comparison with the conventional IRM by estimation of complex ratio masks.

B. Dictionary-based sparse-coding algorithm

An additional goal of this study was to assess the performance of a novel dictionary-based, sparse-coding algorithm. Overall, the performance of the sparse-coding algorithm was similar to that of the NN_COMP algorithm except in babble noise at +4 dB SNR.

A disadvantage of the dictionary-based sparse-coding approach is the relative computational complexity of the de-noising stage. In the neural network approach, after the network is trained, its application in the de-noising stage is straightforward, with the same non-linear formula being applied for each frame to determine the gain. In the case of sparse coding, however, the de-noising stage still requires a sparse approximation to the noisy signal to be found, which

is more challenging to optimize. This makes dictionary-based sparse coding a less plausible candidate for a real-time noise-reduction algorithm. In contrast, image de-noising typically takes place offline and so is better suited to a sparse-coding approach. Nevertheless, it remains feasible that the brain employs mechanisms analogous to sparse coding for de-noising speech.

C. Effects of audibility

Although the use of the NAL-R gain formula in this study was intended to compensate partially for the hearing loss of the participants, it does not provide equal audibility for all listeners. Since the effect of sensation level on speech quality judgments is not well understood, difference in audibility may have influenced individual differences in speech quality ratings. Therefore, the speech intelligibility index (SII; ANSI, 1997) was calculated for each participant and condition as a measure of the audibility of the speech. SII values are shown in Table III. In the case of the enhanced conditions, the SII was calculated based on the spectra of the speech and noise after the application of the enhancement gain function and shadow filtering. The sparse coding processing did not make use of a gain function, but in order to calculate the SII a gain function was calculated based on the difference in level between the original and enhanced signals in 10-ms frames and one-third octave bands.

Additionally, the SII was calculated for the enhanced speech spectrum and the original noise spectrum, to determine whether any benefit could have been derived by changes in the level of the speech spectrum alone. For most processing and noise conditions, this resulted in a small reduction in the SII values relative to the unenhanced conditions, but there were small increases for the Wiener filtering, NN_AIM and NN_COMP in both SSN conditions (mean values of increases in SSI of 0.0157, 0.0131, and 0.0121, respectively, at 0 dB SNR and 0.00807, 0.00424, and 0.00357 at +4 dB SNR). Considering the magnitude of the increases in the SII resulting

TABLE III. Values of the SII calculated for each subject and condition.

	0 dB SNR					4 dB SNR				
Subject	UN	WF	SC	NN_COMP	NN_AIM	UN	WF	SC	NN_COMP	NN_AIM
Speech shaped noise										
1	0.412	0.540	0.565	0.593	0.591	0.526	0.635	0.658	0.666	0.663
2	0.427	0.562	0.578	0.607	0.605	0.547	0.644	0.660	0.667	0.664
3	0.390	0.502	0.510	0.539	0.537	0.487	0.575	0.592	0.598	0.595
4	0.448	0.598	0.639	0.667	0.666	0.583	0.708	0.737	0.746	0.742
5	0.393	0.514	0.535	0.561	0.559	0.501	0.602	0.624	0.632	0.629
6	0.388	0.502	0.518	0.548	0.545	0.486	0.585	0.607	0.616	0.613
7	0.426	0.569	0.615	0.641	0.640	0.554	0.680	0.717	0.728	0.725
8	0.401	0.523	0.539	0.568	0.566	0.508	0.607	0.630	0.638	0.634
9	0.452	0.598	0.637	0.664	0.663	0.583	0.709	0.742	0.752	0.748
10	0.408	0.529	0.545	0.574	0.572	0.515	0.614	0.635	0.644	0.640
11	0.280	0.343	0.326	0.350	0.348	0.333	0.348	0.341	0.345	0.344
12	0.411	0.543	0.569	0.597	0.595	0.529	0.637	0.659	0.667	0.664
13	0.388	0.501	0.519	0.545	0.543	0.488	0.585	0.601	0.608	0.605
14	0.377	0.484	0.497	0.524	0.521	0.472	0.560	0.576	0.583	0.579
15	0.369	0.471	0.480	0.507	0.504	0.458	0.544	0.560	0.567	0.564
16	0.393	0.508	0.526	0.553	0.551	0.496	0.592	0.611	0.618	0.614
17	0.360	0.466	0.482	0.505	0.503	0.454	0.541	0.557	0.564	0.561
Babble noise										
1	0.375	0.405	0.412	0.524	0.532	0.492	0.525	0.516	0.621	0.627
2	0.375	0.406	0.417	0.531	0.544	0.502	0.536	0.526	0.615	0.624
3	0.369	0.395	0.388	0.481	0.487	0.467	0.492	0.475	0.560	0.564
4	0.375	0.405	0.426	0.566	0.575	0.509	0.544	0.552	0.676	0.684
5	0.364	0.391	0.398	0.507	0.513	0.471	0.501	0.501	0.596	0.600
6	0.359	0.385	0.380	0.480	0.489	0.461	0.485	0.472	0.564	0.571
7	0.345	0.369	0.392	0.531	0.540	0.465	0.496	0.510	0.641	0.647
8	0.371	0.399	0.399	0.503	0.510	0.480	0.509	0.499	0.592	0.597
9	0.382	0.412	0.430	0.568	0.579	0.512	0.546	0.551	0.679	0.686
10	0.372	0.402	0.410	0.515	0.521	0.488	0.516	0.509	0.604	0.608
11	0.249	0.264	0.251	0.310	0.312	0.301	0.314	0.293	0.321	0.322
12	0.372	0.403	0.406	0.522	0.532	0.491	0.524	0.516	0.616	0.624
13	0.355	0.384	0.395	0.492	0.498	0.467	0.493	0.487	0.574	0.577
14	0.352	0.378	0.379	0.472	0.478	0.451	0.475	0.466	0.547	0.550
15	0.349	0.374	0.367	0.456	0.462	0.442	0.465	0.451	0.535	0.538
16	0.359	0.387	0.393	0.490	0.498	0.471	0.497	0.485	0.573	0.579
17	0.338	0.363	0.364	0.458	0.464	0.429	0.452	0.450	0.528	0.531

from the suppression of the noise spectrum (see Table III), these comparatively small increases in SII due to changes to the speech spectrum alone are unlikely to have a strong effect on the ratings of speech quality. The greatest reduction in SII occurred with sparse coding for the babble noise conditions with a mean reduction in SII of −0.0859 at 0 dB SNR and −0.0603 at +4 dB SNR. It is possible that for these conditions the performance of the sparse coding algorithm was adversely affected by a lower audibility relative to the other algorithms, although the SII of the sparse coding conditions may have been underestimated by the application of a gain function that was calculated retrospectively.

Overall, there was a significant correlation between SII and quality ratings [$r = 0.467$, $p < 0.001$]. However, a partial correlation between SII and quality rating, controlling for the effects of SNR, algorithm, and noise type, was not significant. This indicates that individual differences in audibility had no influence on the rating of speech quality. The overall correlation between SII value and intelligibility was also calculated and found to be significant [$r = 0.623$, $p < 0.001$]. The partial correlation between SII and intelligibility accounting for the effect of SNR, algorithm, and noise type was also significant ($r = 0.197$, $p < 0.001$), indicating that individual differences in audibility accounted for a very small amount of the variance.

VI. CONCLUSIONS

Three speech-enhancement algorithms based on machine learning and one traditional approach to the problem of speech enhancement (Wiener filtering) were assessed in terms of speech recognition and speech quality ratings in mild-to-moderately hearing-impaired listeners. Unseen tokens of noise were used in the testing stage.

Significant increases in speech-recognition scores and quality ratings were seen for all three machine-learning approaches in at least one of the four noise conditions, although quality ratings were found to be somewhat confounded by the effect of intelligibility. In contrast, the Wiener filtering algorithm produced no significant improvement in either speech recognition or quality rating in any noise condition.

Two neural-network approaches were tested, comparing a network using standard feature-vectors with one using a novel set of features derived from the auditory model AIM. Auditory-based feature vectors performed better than standard feature vectors in terms of both speech recognition and quality in all conditions (except quality in babble noise at +4 dB SNR), although none of these differences was significant.

Although sparse coding shows some improvement in speech recognition and quality, neural networks seem preferable, both because of their higher performance (even for small networks like those used here) and because they are more efficient in the testing stage.

ACKNOWLEDGMENTS

We would like to thank Thomas Blumensath for advice concerning sparse coding. This project received funding from EPSRC Grant No. EP/K020501/1 and from the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant agreement No. PITN-GA-2012-317521 (ITN ICanHear). We also thank two anonymous reviewers for their helpful suggestions. J.J.M.M. and T.G. contributed equally to this work.

Aharon, M., Elad, M., and Bruckstein, A. (**2006**). "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," IEEE Trans. Signal Process. **54**, 4311–4322.

ANSI (**1997**). S3.5-1997, *American National Standard Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

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.

Bleeck, S., Ives, T., and Patterson, R. D. (**2004**). "Aim-mat: The auditory image model in MATLAB," Acta Acust. Acust. **90**, 781–787.

Bolner, F., Goehring, T., Monaghan, J., van Dijk, B., Wouters, J., and Bleeck, S. (**2016**). "Speech enhancement based on neural networks applied to cochlear implant coding strategies," in *2016 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 6520–6524.

Byrne, D., and Dillon, H. (**1986**). "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear Hear. **7**, 257–265.

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**(5), 2604–2612.

DeWeese, M. R., Wehr, M., and Zador, A. M. (**2003**). "Binary spiking in auditory cortex," J. Neurosci. **23**, 7940–7949.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (**2004**). "Least angle regression," Ann. Stat. **32**, 407–499.

Elad, M., and Aharon, M. (**2006**). "Image denoising via learned dictionaries and sparse representation," in *2006 IEEE Computer Society Conference on Computer Vision Pattern Recognition*, pp. 895–900.

Fredelake, S., Holube, I., Schlueter, A., and Hansen, M. (**2012**). "Measurement and prediction of the acceptable noise level for single-microphone noise reduction algorithms," Int. J. Audiol. **51**, 299–308.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (**1993**). "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM NIST speech disc 1-11," NASA STIRecon Tech. Rep.

Gerkmann, T., and Hendriks, R. C. (**2011**). "Noise power estimation based on the probability of speech presence," in *2011 IEEE Workshop on Applied Signal Processing and Audio Acoustics*, pp. 145–148.

Goehring, T., Bolner, F., Monaghan, J. J., van Dijk, B., Zarowski, A., and Bleeck, S. (**2016**). "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," Hearing Res. **344**, 183–194.

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., Apoux, F., and Wang, D. (**2014**). "Speech-cue transmission by an algorithm to increase consonant recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **136**, 3325–3336.

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (**2013**). "An algorithm to improve speech recognition in noise for hearing-impaired listeners," J. Acoust. Soc. Am. **134**, 3029–3038.

Hermansky, H., and Morgan, N. (**1994**). "RASTA processing of speech," IEEE Trans. Speech Audio Process. **2**(4), 578–589.

Hohmann, V. (**2002**). "Frequency analysis and synthesis using a Gammatone filterbank," Acta Acust. Acust. **88**, 433–442.

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. (**2007**). "A comparative intelligibility study of speech enhancement algorithms," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 561.
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.
Irino, T., and Patterson, R. D. (**2002**). "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform," Speech Commun. **36**, 181–203.
Kamath, S., and Loizou, P. (**2002**). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, p. 4164.
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.
Kochkin, S. (**2000**). "MarkeTrak V: 'Why my hearing aids are in the drawer': The consumers' perspective," Hear. J. **53**, 34–36.
Levitt, H., Bakke, M., Kates, J., Neuman, A., Schwander, T., and Weiss, M. (**1992**). "Signal processing for hearing impairment," Scand. Audiol. Suppl. **38**, 7–19.
Lewicki, M. S. (**2002**). "Efficient coding of natural sounds," Nat. Neurosci. **5**, 356–363.
Lim, J. S., and Oppenheim, A. V. (**1979**). "Enhancement and bandwidth compression of noisy speech," Proc. IEEE **67**(12), 1586–1604.
Loizou, P., and Kim, G. (**2011**). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," IEEE Trans. Audio Speech Language Process. **19**(1), 47–56.
Löllmann, H. W., and Vary, P. (**2009**). "Low delay noise reduction and dereverberation for hearing aids," EURASIP J. Adv. Signal Process. **2009**(1), 1–9.
Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Büchler, M., Dillier, N., Houben, R., Dreschler, W. A., Froehlich, M., Puder, H., Grimm, G., Hohmann, V., Leijon, A., Lombard, A., Mauler, D., and Spriet, A. (**2010**). "Multicenter evaluation of signal enhancement algorithms for hearing aids," J. Acoust. Soc. Am. **127**, 1491–1505.
Madhu, N., Spriet, A., Jansen, S., 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.
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., Feldbauer, C., Walters, T. C., and Patterson, R. D. (**2008**). "Low-dimensional, auditory feature vectors that improve vocal-tract-length normalization in automatic speech recognition," J. Acoust. Soc. Am. **123**, 3066.
Mueller, H. G., Weber, J., and Hornsby, B. W. (**2006**). "The effects of digital noise reduction on the acceptance of background noise," Trends Amplif. **10**, 83–93.
Müller, F., and Mertins, A. (**2012**). "Enhancing vocal tract length normalization with elastic registration for automatic speech recognition," in *INTERSPEECH*, pp. 1364–1367.
Nabelek, A. K., Freyaldenhoven, M. C., Tampas, J. W., Burchfield, S. B., and Muenchen, R. A. (**2006**). "Acceptable noise level as a predictor of hearing aid use," J. Am. Acad. Audiol. **17**, 626–639.
Olshausen, B. A., and Field, D. J. (**1996**). "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature **381**, 607–609.
Pati, Y. C., Rezaiifar, R., and Krishnaprasad, P. S. (**1993**). "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *IEEE 27th Asilomar Conference on Signals, Systems and Computing*, pp. 40–44.
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.
Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P. (**1987**). "An efficient auditory filterbank based on the gammatone function," in *RSRE Meeting on the IOC Speech Group Auditory Modelling*, pp. 1–33.
Riedmiller, M., and Braun, H. (**1993**). "A direct adaptive method for faster backpropagation learning: The RPROP algorithm," in *1993 IEEE International Conference on Neural Networks*, pp. 586–591.
Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., and Weinstock, M. (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.
Sang, J., Hu, H., Zheng, C., Li, G., Lutman, M. E., and Bleeck, S. (**2014**). "Evaluation of the sparse coding shrinkage noise reduction algorithm in normal hearing and hearing impaired listeners," Hear. Res. **310**, 36–47.
Sang, J., Hu, H., Zheng, C., Li, G., Lutman, M. E., and Bleeck, S. (**2015**). "Speech quality evaluation of a sparse coding shrinkage noise reduction algorithm with normal hearing and hearing impaired listeners," Hear. Res. **327**, 175–185.
Scalart, P., and Filho, J. V. (**1996**). "Speech enhancement based on a priori signal to noise estimation," in *1996 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 629–632.
Sigg, C. D., Dikk, T., and Buhmann, J. M. (**2012**). "Speech enhancement using generative dictionary learning," IEEE Trans. Audio Speech Lang. Process. **20**, 1698–1712.
Smith, D. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, T. (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**(1), 305–318.
Stone, M. A., and Moore, B. C. (**1999**). "Tolerable hearing aid delays I Estimation of limits imposed by the auditory path alone using simulated hearing losses," Ear Hear. **20**, 182–192.
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.
Tchorz, J., and Kollmeier, B. (**2003**). "SNR estimation based on amplitude modulation analysis with applications to noise suppression," IEEE Trans. Speech Audio Process. **11**(3), 184–192.
Tsoukalas, D. E., Mourjopoulos, J. N., and Kokkinakis, G. (**1997**). "Speech enhancement based on audible noise suppression," IEEE Trans. Speech Audio Process. **5**, 497–514.
von Kriegstein, K., Smith, D. R., Patterson, R. D., Ives, D. T., and Griffiths, T. D. (**2007**). "Neural representation of auditory size in the human voice and in sounds from other resonant sources," Curr. Biol. **17**, 1123–1128.
Williamson, D. S., Wang, Y., and Wang, D. L. (**2016**). "Complex ratio masking for monaural speech separation," IEEE/ACM Trans. Audio Speech Language Process. **24**, 483–492.
Zakis, J. A., Hau, J., and Blamey, P. J. (**2009**). "Environmental noise reduction configuration: Effects on preferences, satisfaction, and speech understanding," Int. J. Audiol. **48**, 853–867.