Speech Enhancement Based on Neural Networks Improves Speech Intelligibility in Noise for Cochlear Implant Users

Tobias Goehring[a]*, Federico Bolner[b,c]*, Jessica J. M. Monaghan[a], Bas van Dijk[c], Andrzej Zarowski[d], and Stefan Bleeck[a]

*these authors contributed equally to this work

[a]ISVR, University of Southampton, University Rd, Southampton, SO17 1BJ, United Kingdom

[b]ExpORL, KU Leuven, O&N II Herestraat 49, 3000 Leuven, Belgium

[c]Cochlear Technology Centre, Schaliënhoevedreef 20 I, 2800 Mechelen, Belgium

[d]European Institute for ORL-HNS, Sint Augustinus Hospital, Oosterveldlaan 24, 2610 Wilrijk, Belgium

Abstract

Speech understanding in noisy environments is still one of the major challenges for cochlear implant (CI) users in everyday life. We evaluated a speech enhancement algorithm based on neural networks (NNSE) for improving speech intelligibility in noise for CI users. The algorithm decomposes the noisy speech signal into time-frequency units, extracts a set of auditory-inspired features and feeds them to the neural network to produce an estimation of which frequency channels contain more perceptually important information (higher signal-to-noise ratio, SNR). This estimate is used to attenuate noise-dominated and retain speech-dominated CI channels for electrical stimulation, as in traditional *n*-of-*m* CI coding strategies. The proposed algorithm was evaluated by measuring the speech-in-noise performance of 14 CI users using three types of background noise. Two NNSE algorithms were compared: a speaker-dependent algorithm, that was trained on the target speaker used for testing, and a speaker-independent algorithm, that was trained on different speakers. Significant improvements in the intelligibility of speech in stationary and fluctuating noises were found relative to the unprocessed condition for the speaker-dependent algorithm in all noise types and for the speaker-independent algorithm in 2 out of 3 noise types. The NNSE algorithms used noise-specific neural networks that generalized to novel segments of the same noise type and worked over a range of SNRs. The proposed algorithm has the potential to improve the intelligibility of speech in noise for CI users while meeting the requirements of low computational complexity and processing delay for application in CI devices.

Keywords

Cochlear implants, noise reduction, speech enhancement, machine learning, neural networks

1. INTRODUCTION

A cochlear implant (CI) is an auditory prosthesis that provides a sensation of hearing for listeners with severe to profound sensorineural hearing loss. State-of-the-art CI devices allow many users to achieve near-to-normal speech understanding in quiet acoustic conditions (Fetterman and Domico, 2002; Zeng *et al.*, 2008). However, background noises such as environmental sounds or competing talkers negatively affect CI users' speech understanding. The decrease in performance can be measured with the speech reception threshold (SRT), which is defined as the signal-to-noise ratio (SNR) at which 50% of the speech is intelligible. CI users typically have SRTs that are 10 to 25 dB higher (worse) than those of normal hearing (NH) listeners (Spriet *et al.*, 2007; Wouters and Van den Berghe, 2001). It has been reported that CI recipients can take less advantage of temporal gaps or slow amplitude fluctuations on an otherwise stationary noise masker compared with NH listeners in terms of speech intelligibility (Cullington and Zeng, 2008; Stickney *et al.*, 2004; Zeng *et al.*, 2008, Oxenham and Kreft, 2014). This process is known as release from masking (Miller and Licklider, 1950). Since the spectral information conveyed by a CI is reduced to a small number of effective spectral channels (Friesen *et al.*, 2001), CI users rely strongly on temporal information (in the form of envelope modulations) and thus are more susceptible to modulated masking noise than NH listeners (Cullington and Zeng, 2008; Fu *et al.*, 2013). Most likely, a combination of reduced spectral resolution and increased modulation interference accounts for the decrease in speech understanding performance observed for CI users compared with NH listeners and with NH listeners tested with CI simulations (Cullington and Zeng, 2008; Jin *et al.*, 2013, Oxenham and Kreft, 2014).

Speech enhancement (SE) algorithms have been proposed to alleviate this problem by attenuating the noise component of the noisy mixture to increase the intelligibility and perceived quality of the speech component (Loizou, 2013). SE algorithms can be divided into algorithms that make use of two or more microphones to exploit the spatial properties of target and noise sources, and algorithms that make use of a single microphone (or the output signal of a multi-microphone algorithm). Multi-microphone algorithms have been shown to deliver large benefits in SRT scores when the target signal and the interfering noise source are spatially separated (Mauger and Warren, 2014; Spriet *et al.*, 2007; Wouters and Van den Berghe, 2001). However, in everyday listening situations, these requirements might not always be fulfilled, and single-microphone algorithms are still of interest for numerous applications, such as hearing devices, where the number of microphones is usually limited to two and the two microphones are on the same side of the head.

Single-microphone SE algorithms are based on the assumption that improving the global SNR of noisy speech will lead to improved speech intelligibility (SI) (French and Steinberg, 1947). With such algorithms, the signal is converted into the spectral domain (e.g. by Fourier analysis or filter bank processing) and a filter is applied to retain the signal in frequency channels with high SNR and attenuate the signal in frequency channels with low SNR, leading to an increased global SNR. Numerous algorithms have been proposed to estimate the SNR in each frequency channel (Gerkmann and Hendriks, 2012; Martin, 2001). This estimate is used to calculate a gain function to determine the attenuation of noise-dominated channels. SE algorithms mainly differ in the SNR estimation methods and the gain functions used for noise suppression (e.g. spectral subtraction or parametric Wiener filter, Boll, 1979; Lim and Oppenheim, 1979). In the ideal case (i.e. when the speech and noise components are known), these algorithms can lead to highly increased intelligibility, close to that for noise-free speech for NH listeners (Madhu *et al.*, 2013) and CI users (Koning *et al.*, 2015; Mauger *et al.*, 2012a; Qazi *et al.*, 2013). Similarly, extensive studies on the SI benefits of time-frequency masking with the ideal binary mask (IBM) support the potential of SNR-based suppression criteria for improving the intelligibility of speech in noise (Anzalone *et al.*, 2006; Brungart *et al.*, 2006; Hu and Loizou, 2008; Wang *et al.*, 2009).

In a real system, where only the mixture of speech and noise is available, SNR estimation errors may lead to speech distortions, introduction of musical noise or insufficient noise suppression. In challenging acoustic environments these artefacts greatly reduce and often completely undo the speech intelligibility benefits observed in the ideal case for NH and hearing-impaired (HI) listeners (Brons *et al.*, 2012; Chen and Loizou, 2012; Loizou, 2013). For CI users, where a decrease in SI performance is typically observed at higher SNRs than for NH and HI listeners, improvements in SI have been reported with several SE algorithms based on noise-estimation techniques (Dawson *et al.*, 2011; Hu *et al.*, 2007; Mauger *et al.*, 2012b; Ye *et al.*, 2013). This success may be due to the better performance (reduced estimation errors) of the algorithms for higher SNRs. In addition, Mauger *et al.* (2012a; 2012b) reported that CI users generally preferred a more aggressive gain function than the standard Wiener gain function, suggesting that CI users might be more resistant to speech removal distortions (type-II errors) and less resistant to noise addition errors (type-I) (also reported by Qazi *et al.*, 2013). For CI users, maximum benefits of about 2 dB in SRT were found for speech in stationary noise, but the benefit was much reduced when the interfering noise was non-stationary, as in the case of competing talkers (Dawson *et al.*, 2011; Mauger *et al.*, 2012b).

A recent approach to SE algorithms employs supervised machine learning to estimate the gain function (by using either classification or regression methods), instead of using conventional SNR estimation techniques (Tchorz and Kollmeier, 2003). Using a similar approach, algorithms have been trained on labelled datasets to approximate the IBM. These have been reported to provide remarkably large SI improvements for NH listeners (Kim *et al.*, 2009), HI-listeners (Healy *et al.*, 2013, 2014) and CI users (Hu and Loizou, 2010) for speech in both stationary and non-stationary noise, even at low SNRs. However, these algorithms were trained and tested on datasets using the same speaker, background noise and SNRs. This approach is likely to lead to overfitting of the training data and strongly limits generalization performance to acoustic conditions different from the ones used during training (May and Dau, 2014). Recently, it has been shown, for both NH and HI listeners, that incorporating more exemplars of the noise recordings in the training stage leads to algorithms that generalize well to novel realizations of the same noise type (Bolner *et al.*, 2016; Healy *et al.*, 2015) or to completely novel types of noise (Chen *et al.*, 2016). These studies indicate that generalization to novel noise conditions is possible when the training datasets incorporate higher degrees of variability. Furthermore, the use of a "soft" gain mask (often called *ideal ratio mask*, IRM) inspired by the Wiener filter gain function (Lim and Oppenheim, 1979) avoids the need to choose an appropriate SNR-dependent classification threshold in IBM-based processing, and can lead to a regression model that worked over a range of SNRs (Bolner *et al.*, 2016) or generalized to untrained SNRs (Chen *et al.*, 2016).

The results from the studies described above are promising. However, generalization to novel, unseen speakers was not tested (Bolner *et al.*, 2016; Chen *et al.*, 2016, Healy *et al.*, 2015). In real-world situations, in the context of SE for hearing devices, an algorithm should work well with any target speaker and meet the requirements of limited computational complexity and short processing delay (Stone and Moore, 2005). The algorithms proposed by Chen *et al.* (2016) and Healy *et al.* (2015) include non-causal information (future frames) in the processing and therefor introduce considerable processing delays (>20 ms). As described by Healy *et al.* (2015), the use of future frames has to be avoided for applications using real-time processing, such as hearing aids and CIs.

In this study, we tested whether an SE algorithm using neural networks (NNSE) can improve the SRTs of CI users for speech in stationary and non-stationary background noises. We address the important aspect of generalization performance to a novel speaker by comparing two identical systems that were trained on either the same or different speakers from the one used during testing. This study used noise-specific networks that were tested on novel segments of the same noise type (similar to Healy *et al.*, 2015). The algorithm complexity and processing delay were chosen to yield a real-time feasible architecture with low latency for potential application in CIs. We employed an aggressive gain function as preferred by CI users (Mauger *et al.*, 2012a, 2012b; Qazi *et al.*, 2013) and integrated the SE algorithm into the coding strategy of a CI to evaluate the performance of the algorithm. The algorithm was designed to work over a range of SNRs (Chen *et al.*, 2016; Bolner *et al.*, 2016) relevant to CI users and to process stimuli adaptively using online processing.

2. ALGORITHM DESCRIPTION

The NNSE algorithm, was integrated within an implementation of the Advanced Combination Encoder (ACE™) CI speech processing strategy (Seligman and McDermott, 1995). Figure 1 shows a block diagram of the algorithm.

PLACEHOLDER - Figure 1

2.1 Reference strategy

A research ACE strategy implementation served as the reference strategy. The noisy speech signal was downsampled to 16 kHz, passed through a pre-emphasis filter, and sent through an automatic gain control (AGC). The AGC compressed the acoustic dynamic range such that it could be conveyed into the smaller electrical dynamic range of a CI recipient (with an attack time of 5 ms, a release time of 75 ms, a compression threshold of 73 dB SPL and compression limiting above that level). Next, a filter bank based on a Fast Fourier Transform (FFT) was applied to the compressed signal. The FFT was performed on Hanning-windowed 8-ms long input blocks, with an overlap of 7 ms. The magnitude of the complex FFT output was used to provide an estimate of the envelope for each of the M frequency channels (typically, M=22). Each channel was then allocated to one electrode. Maxima selection was applied to retain the subset of N channels with the largest envelope magnitudes (with N<M set by an audiologist during the fitting of the subject's CI processor). A loudness growth function (LGF) instantaneously mapped the envelope for each channel to the subject's dynamic range between the threshold level (THL) and maximum comfortable loudness level (MCL) for electrical stimulation (using the THL and MCL parameters from the subject's CI processor). Finally, the electrodes corresponding to the selected channels were stimulated sequentially and one cycle of stimulation was completed. The number of cycles per second is called the channel stimulation rate, and the total stimulation rate is N times the channel stimulation rate.

2.2 Speech enhancement algorithm

CI processing directly transforms the envelope of the frequency channels to an electrical output, and it does not require a reconstruction stage. We chose to integrate the NNSE directly into the CI signal path rather than performing preprocessing of the noisy signal. This avoids an unnecessary synthesis stage, which would introduce additional noise and increase the complexity and delay of the system. The NNSE algorithm consisted of two main components: feature extraction and neural network (NN) regression.

After downsampling to 16 kHz, the noisy speech signal was divided into 20-ms long segments with 50% overlap. Feature extraction was performed on each segment of the noisy signal, and the output was fed to the NN. The trained NN (the training is described below) was used to estimate the Wiener gain over 31 frequency channels equally spaced on the equivalent rectangular bandwidth (ERB_N-number, Glasberg and Moore, 1990) scale with centre frequencies ranging from 50 to 8000 Hz. Since the frequency channels assigned to the electrodes varied across subjects, the estimated gains were mapped to each subject's specific filter bank configuration. Exponential smoothing (with a time constant of 12 ms) was performed before applying the gain to the corresponding noisy envelope in the ACE signal path. The main effect of the gain application was the attenuation of noise-dominated channels. This occurred before the ACE channel selection (see Fig. 1). Therefore, speech-dominated channels were more likely to be selected for stimulation. Unlike most SE algorithms (Loizou, 2013), the algorithm does not require a voice activity detector or the estimation of noise statistics. The NNSE was designed so that it could be run in real time, with an algorithmic delay of 10 ms.

An example of an electrodogram of a Dutch sentence (*"Het verhaal is heel spannend"*) from the LIST corpus processed by the ACE coding strategy with 11 maxima is shown in Fig. 2. An electrodogram represents the stimulation pattern across electrodes (y-axis) over time (x-axes). The height of each vertical bar reflects the normalised amplitude of a single stimulation pulse.
The top panel represents the electrodogram of the clean sentence, in which the boundaries between words are clearly visible. For the second panel, the speech was corrupted by babble noise (SNR = 5 dB). The resulting stimulation sequence changed significantly: periods of silence were filled with noise, envelopes were distorted, and not all of the channels containing speech were selected. The third and fourth panels represent the conditions with NNSE processing using speaker-independent and speaker-dependent training, respectively. The processing steered channel selection to pick the channels containing speech, thus partially restoring information that was masked by the noise (Qazi *et al.*, 2013).

PLACEHOLDER - Figure 2

2.2.1 Feature extraction

Feature extraction was performed on each 20-ms segment, or frame, at a rate of 100 Hz. Each frame was passed through a Gammatone filter bank consisting of 31 channels equally spaced on the ERB_N-number scale with centre frequencies ranging from 50 to 8000 Hz (Hohmann, 2002). Then, the energy of each channel was log-compressed to obtain 31 Gammatone Frequency Energy features ($GFEN_n$, with *n* denoting the frame number). From the $GFEN_n$, two additional features were extracted: 26 Gammatone Frequency Cepstral Coefficients ($GFCC_n$) and 13 Gammatone Frequency Perceptual Linear Prediction Cepstral Coefficients ($GPLP_n$). The $GFCC_n$ features were obtained by performing the discrete cosine transform (DCT) on $GFEN_n$ for frequencies above 200 Hz (and excluding the DC component of the DCT). The $GPLP_n$ features were obtained by filtering $GFEN_n$ with the relative spectral transform (RASTA, Hermansky and Morgan, 1994) filter, which emphasises the modulation frequencies relevant to human speech, and performing a 12-th order linear prediction model analysis on the output (perceptual linear prediction, PLP).

The 31 $GFEN_n$, 26 $GFCC_n$ and 13 $GFPLP_n$ features were concatenated to form a 70-dimensional feature vector F_n. Our pilot results (Bolner *et al.*, 2016) indicated that this combination led to higher estimation accuracy than the individual features alone. Note that F_n was derived exclusively from the ERB_N-number spaced spectrum of the signal ($GFEN_n$). Evaluation with several objective measures (difference between hit and false alarm rates, HIT-FA, Kim *et al.*, 2009; short-time objective intelligibility measure, STOI, Taal *et al.*, 2011; normalized covariance metric, NCM, Holube and Kollmeier, 1996; Ma *et al.*, 2009) indicated that this choice had no detrimental effects on the estimation accuracy of the algorithm compared with the use of the more conventional MFCC (using the Mel-scale) and RASTA-PLP (using the Bark scale), and it avoided two additional filtering stages. Finally, F_n was concatenated with the features extracted from the preceding frame F_{n-1} to provide additional temporal information. The resulting 140-dimensional feature vector $[F_n, F_{n-1}]$ was fed to the NN to estimate the Wiener gain for the current frame *n*. Note that the NN estimated the Wiener gain using information related to the current and past frames only. This feature set allowed relatively low complexity and low delay making the proposed algorithm suitable for real-time processing, in contrast to most recent speech segregation studies (Chen *et al.*, 2016; Healy *et al.*, 2013, 2015).

2.2.2 Neural network regression: architecture and training procedure

A parametric Wiener gain mask (Lim and Oppenheim, 1979), the IRM, was used as the training target for the supervised training process. The ideal ratio mask is defined as follows:

$$G(f,n) = \left(\frac{SNR(f,n)}{SNR(f,n)+1}\right)^{\beta},$$

where $SNR(f,n)$ denotes the SNR in frame *n* and Gammatone frequency channel *f*. The parameter β controls the slope of the gain function $G(f,n)$. We experimented with different values of β and found $\beta = 1$ to be a good compromise between noise removal and speech distortion when the mask was applied to noisy speech. This choice was also supported by the finding that CI users generally prefer a relatively aggressive gain function (Mauger *et al.*, 2012a, 2012b) as opposed to the square-root Wiener mask ($\beta = 0.5$) used in previous studies with HI listeners (Chen *et al.*, 2016; Healy *et al.*, 2015).

The neural network consisted of an input layer, defined by the feature vector, 2 hidden layers of 75 units using a saturating-linear activation function (which resembled a piecewise linearised sigmoidal function) and 31 linear output units. Resilient backpropagation (Riedmiller and Braun, 1993) was used for training the NN in full-batch mode over 500 epochs with a learning rate of 0.01 and weight increment and decrement factors of 1.2 and 0.5, respectively. The cost function was the mean squared error (MSE) between the true and estimated Wiener gain using a weight-decay regularisation of 0.5 to avoid overfitting.

The parameters of the algorithm were chosen based on a previous study of Bolner *et al.* (2016), who observed significant improvements in speech intelligibility in noise for NH listeners using CI vocoder simulations with a supervised NN-based SE algorithm. The biggest difference between the two algorithm configurations was a reduced number of neural network parameters (node weights and biases), mainly deriving from the use of a Gammatone filter bank with 31 channels both for the feature extraction stage and Wiener gain estimation, as opposed to 63 channels used by Bolner *et al.* (2016). The Nucleus implants tested in this study maximally use 22 spectral channels, and thus 31 channels seemed a good compromise between algorithm complexity and SE performance for CI application. The 31 estimated Wiener gains were mapped to the 22 CI channels before application to the envelopes. The configuration used in the current study allowed a reduction in the algorithm complexity while maintaining comparable performance in terms of estimation accuracy and with respect to several speech intelligibility objective metrics, such as HIT-FA (between estimated and ideal ratio masks), NCM and STOI (using vocoded simulations of the enhanced and noise-free reference signals, Chen and Loizou, 2011).

The algorithm made use of feed-forward neural networks that were trained using the true Wiener gain along with the features extracted from the noisy speech. Rather than performing large-scale training with thousands of noises (as done by Chen *et al.*, 2016), the networks were noise-specific, i.e. each network was trained for a particular listening situation (similar to Hu *et al.*, 2010). This made it possible to take advantage of the learning of the distinctive spectro-temporal characteristics of each noise while limiting the NN size.

The speech materials used to train the NNSE were LISTm (sentences of equal difficulty with 2-7 keywords, equal number of syllables and key words per list, male Flemish talker, Jansen *et al.*, 2014), LISTf (similar structure to LISTm, but partially different sentences than LISTm, female Flemish talker, Van Wieringen and Wouters, 2008), NVA (lists of 10 bisyllabic words, male Flemish talker, Wouters *et al.*, 1994), and GRID (simple and syntactically identical phrases of 6 words, 18 male and 16 female English talkers, Cooke *et al.*, 2006). Three types of noise were used: steady speech weighted noise (SWN), single-speaker-modulated speech-weighted noise (ICRA), and 20-talker babble (BABBLE). The SWN had the same long-term spectrum as the sentences of the LISTm corpus (Jansen *et al.*, 2014). The modulated speech-weighted noise was the ICRA5-250 (Dreschler *et al.*, 2001) that was generated by sending English male speech through a 3-channel filter bank, randomly reversing the sign of each sample in each channel (with a probability of 0.5), filtering it again with the same filter bank, randomizing the phase in the frequency domain and applying the standard long-term average speech spectral shape of male speech. The ICRA5-250 noise has maximum silent gaps of 250 ms and may contain some intelligible fragments, at least for English native speakers, as reported by Dreschler *et al.* (2001). The BABBLE signal was recorded at Auditec St. Louis and consisted of a mixture of 20 English competing talkers (8 male, 12 female). The three types of masking noise have different degrees of temporal fluctuation (increasing from SWN to BABBLE to ICRA) and thus introduce varying amounts of modulation masking (Dau *et al.*, 1997).

During training, 4-minute long recordings of the three noises were mixed with two speech material training sets:

- Single talker (ST), containing 10 lists from the LISTm corpus (total of 8 minutes)
- Multiple talker (MT), containing 6 lists from the LISTf corpus, 4 lists from the NVA corpus and 120 sentences from the GRID corpus (total of 15 minutes).

In both cases, the sentences were mixed with random segments of the noise at 7 SNRs, from –6 to +6 dB in steps of 2 dB. This, in turn, produced two networks for each noise type, one trained on a single talker (LISTm) and the other trained on multiple talkers.

3. MATERIALS AND METHODS

3.1 Software/Hardware

The research ACE strategy and NNSE algorithm were developed in MATLAB (The MathWorks, Natick, Massachusetts). Stimuli were processed through a computer implementing the ACE strategy (with/without NNSE) and directly presented to the implant user. Electrical stimulation was delivered via the Cochlear NIC3 interface connected to an L34 experimental processor. The system delivered radio frequency output to the coil that transmitted stimulus data to the subject's implant.

3.2 Subjects

A group of 14 CI users, all native Dutch speakers and implanted with a Cochlear Nucleus® CI, participated. The study protocol was approved by the Commissie Medische Ethiek GZA Ziekenhuizen (Antwerp) ethics committee, and subjects gave their informed consent to participate in the study. Subjects were not paid, but travel expenses were reimbursed. This study was conducted according to the guidelines for Good Clinical Practice (GCP), ISO14155-2011 (International Standard for Clinical Investigations of medical devices for human subjects) and the Declaration of Helsinki (2013).
The mean age of the group at the start of the study was 61 years, ranging from 23 to 81 years. Only one ear of each subject was tested. If the subject had a hearing aid (HA) or CI on the contralateral side, it was turned off during the testing. The mean duration of implant use was 9.8 years at the start of the study, with a range from 1.2 to 13.6 years. All subjects were users of the ACE strategy. Demographic data for the subjects can be found in Table 1.

PLACEHOLDER - Table 1

Prior to the speech in noise test, the subjects' existing CI program parameters were transferred from their own sound processor to the control computer. Subjects informally reported that they did not perceive a difference between the daily program on their sound processor and the stimulation delivered via the ACE strategy on the test system. Details of each subject's CI parameters, such as stimulation rate, number of maxima, number of total active channels, THL and MCL, and dynamic range are presented in Table 2.

PLACEHOLDER - Table 2

3.3 Stimuli and processing conditions

Sentences from the LISTm corpus (Jansen *et al.*, 2014) were used as the target speech material. The LISTm corpus consists of 38 lists, with 10 sentences for each list, produced by a male Flemish talker. The number of keywords per sentence ranged from 2 to 7, with an average and median of 3. Since 10 lists of the corpus were used during the training stage of the algorithm, only the remaining 28 lists were employed for the listening test.

The maskers were 20-s long novel realizations of SWN, ICRA5-250 and BABBLE, from which a random segment was extracted and mixed with the target speech for each sentence. This was done in order to test the algorithm on sentences and noise segments that were not previously processed by the NNs.

The three processing conditions were:

- UN: unprocessed condition, i.e. ACE.
- NNSE-ST: processed condition with the NNSE algorithm, using the networks trained on the single-talker data. Note that in this case the algorithm was tested on the same speaker as the one used during the training stage (LISTm).
- NNSE-MT: processed condition with the NNSE algorithm, using the networks trained using multiple talkers data, which did not include the target speaker.

The NNSE-MT condition was included to assess the performance of the NNSE in more realistic and challenging conditions when the target speaker was unknown to the system, in contrast to recent SE studies (Bolner *et al.*, 2016; Chen *et al.*, 2016; Healy *et al.*, 2013, 2015; Hu and Loizou, 2010).

3.4 Study protocol

The study used a repeated measures, single-subject design in which each subject served as his/her own control. This approach made it possible to accommodate the heterogeneity that usually characterizes the CI population. At the beginning of the session, each subject was allowed to choose his/her preferred volume. Sentences from one list of the corpus (from the training set) were presented in quiet and in noise (SWN between 0 and 5 dB SNR) until the subject was satisfied with the volume. The chosen volume setting was then fixed for the rest of the testing.

The SRT was measured using an adaptive procedure for 9 conditions [3 maskers (SWN, ICRA, BABBLE) x 3 processing conditions (UN, NNSE-ST, NNSE-MT)] by an audiologist in a sound-treated room. Both subject and audiologist were blind as to which processing condition was being tested.

An SRT was measured using one list (10 sentences) randomly selected from the speech corpus. The speech level was held constant at 65 dB SPL while the noise level was adjusted according to the subject's response to each sentence in steps of 2 dB, in a one-down, one-up procedure to target the 50% correct point. After determining the level of the (hypothetical) 11th item, the SRT was calculated as the mean of the last 6 SNRs. A response was counted as correct when all the keywords in the sentence were correctly identified. Errors for non-keywords were not taken into account, but incomplete keywords or minor variations of verb tenses of keywords were penalised (van Wieringen and Wouters, 2008).

Each of the 9 conditions was tested 3 times, counterbalancing the order in which the conditions were tested for each subject. The order in which the noise and processing conditions were tested was counterbalanced across 12 subjects, and the order for the remaining two subjects was allocated randomly. The final SRT for each condition was obtained by averaging the three SRT values. At the end of the testing, subjects resumed the use of their own sound processor.

3.5 Evaluation

Prior to clinical testing, an objective analysis of the performance of each processing condition was performed. Electrodograms were computed at different SNRs, and were compared with a reference electrodogram in terms of type I and type II error rates. Although this method has not been widely used in the literature, it represents a useful way to compare noise reduction performance for CIs (Mauger *et al.*, 2012b).

In an electrodogram, stimuli have normalized values between 0 and 1, representing the electrical perception range between threshold and comfort level in each frame and frequency channel. The reference electrodogram was generated by processing speech in quiet with ACE (without NNSE), and provided the "ideal" outcome of noise reduction.

Error rates were computed as the stimulus amplitude difference of the reference electrodogram (REF-E) and the comparison electrodogram (COM-E), with the method proposed by Mauger *et al.* When the COM-E contained a stimulus (channel-frame) that was lower in amplitude than the corresponding stimulus in the REF-E, a type II error was computed as the stimuli amplitude difference. For example, if the COM-E had a stimulus amplitude of 0.3 and the REF-E had a stimulus of 0.5, this was considered as a type II error of value 0.2. A full type II error (value = 1) occured when no stimulus (amplitude = 0) was present in the COM-E, while the REF-E contained a stimulus with amplitude = 1. In a similar manner, a type I error occurred when the COM-E contained a stimulus of higher amplitude than for the REF-E. The type I error was computed as the difference of the stimulus amplitudes. For example, if the COM-E had a stimulus amplitude of 0.3 and the REF-E had a stimulus amplitude of 0, this was considered as a type I error of value 0.3. A type I error can be viewed as a noise addition error, while a type II error can be viewed as a speech removal error.

Type I and type II errors were summed across all channels and frames and divided by the total number of possible errors to obtain the type I and type II error rates. Error rates for processing condition were computed as the average error rates calculated over 20 sentences at –5, 0, 5, and 10 dB SNR, with 11 selected channels (ACE maxima selection). This was done so as to have the same number of possible errors for both error types and to avoid introducing a bias towards either of the two.

PLACEHOLDER - Figure 3

Results of the objective analysis are displayed in Figure 3. For SWN, UN gave type I error rates from 36% to 66%, and type II error rates ranging from 9% to 15% (SNR = -5 and 10 dB, respectively). The NNSE conditions gave similar error rates, with greatly reduced type I error rates ($\leq$ 6% and $\leq$ 17%, at –5 and 10 dB SNR, respectively), at the expense of slightly higher type II error rates ($\leq$ 14% and $\leq$ 20%, at –5 and 10 dB SNR, respectively).

For ICRA, UN gave type I error rates from 20% to 42%, and type II error rates from 4% to 10% (SNR = -5 and 10 dB, respectively). Again, both NNSE conditions gave greatly reduced type I error rates at the expense of higher type II error rates. Type I errors ranged from 7% to 17% for NNSE-MT, and from 6% to 14% for NNSE-ST, at –5 and 10 dB SNR, respectively, while type II error rates ranged from 7% to 12% for NNSE-MT, and from 11% to 15% for NNSE-ST (at –5 and 10 dB SNR, respectively).

For BABBLE, UN gave type I error rates from 37% to 66%, and type II error rates from 9% to 15% (SNR = -5 and 10 dB, respectively), in line with what was found for SWN. Also for BABBLE, both NNSE conditions gave reduced type I error rates but higher type II error rates compared to the UN condition. Type I errors ranged from 9% to 30% for NNSE-MT, and from 5% to 20% for NNSE-ST, at –5 and 10 dB SNR, respectively. Type II error rates ranged from 14% to 18% for NNSE-MT, and from 22% to 25% for NNSE-ST.

In conclusion, both NNSE algorithms greatly reduced the noise, but also introduced some speech removal distortions. This effect was more pronounced for NNSE-ST than for NNSE-MT for the modulated noises (ICRA and BABBLE), while the performance of the two NNSE strategies was comparable for SWN. Both NNSE-MT and NNSE-ST reduced the total error compared to UN for all noises and SNRs. These results suggested that an improvement in speech perception might be achieved and supported the clinical speech performance testing of CI users.

4. RESULTS

The group mean SRTs for all processing conditions are shown in Fig. 4 and individual SRTs and their changes relative to those for the unprocessed condition (UN) are shown in Fig. 5. The data in all conditions were normally distributed, as tested with the Kolmogorov-Smirnov (using Lilliefors significance correction) and the Shapiro-Wilk tests. The SRTs used in statistical analyses were the average of the 3 SRTs obtained for each processing condition and noise type. Performance with UN was poorer (higher SRT) than with the processed conditions for all three noises. Group mean SRTs for speech in UN increased from 2.8 dB in SWN, to 5.1 dB in ICRA, and up to 6.7 dB in BABBLE. For all three noise types, lower mean SRTs were obtained with NNSE-MT and NNSE-ST than with UN. NNSE-ST achieved the lowest SRTs for all three noise conditions with an advantage of about 1 to 1.5 dB SRT over NNSE-MT.

A two-way analysis of variance (ANOVA) with repeated measures was conducted with factors processing condition (UN, NNSE-ST and NNSE-MT) and noise type (SWN, ICRA, and BABBLE). There were significant main effects of processing condition [$F(2,26) = 31.83$, $p < 0.001$], noise type [$F(2,26) = 37.63$, $p < 0.001$] and a significant interaction [$F(4,54) = 13.73$, $p < 0.001$].

Further statistical analysis was conducted separately for each noise type to compare the 3 processing conditions.

For SWN noise, Mauchly's test showed no violation of sphericity and a one-way repeated measures ANOVA indicated a significant effect of processing condition [$F(2,12) = 8.165$, $p = 0.006$]. *Post hoc* pairwise comparisons using Bonferroni correction revealed significant differences between UN and both NNSE-MT ($p = 0.019$) and NNSE-ST ($p = 0.003$), but not between NNSE-MT and NNSE-ST ($p = 0.10$), with improvements in SRT scores relative to those for UN of 1.4 and 2.3 dB for NNSE-MT and NNSE-ST, respectively. Apart from three subjects for NNSE-MT and one subject for NNSE-ST, subjects benefitted from the processing with both NNSE algorithms for speech in SWN.

For ICRA noise, Mauchly's test showed no violation of sphericity and a one-way repeated measures ANOVA indicated a significant effect of processing condition [$F(2,12) = 28.13$, $p < 0.001$]. *Post hoc* pairwise comparisons using Bonferroni correction revealed significant differences between UN and both NNSE-MT ($p < 0.001$) and NNSE-ST ($p < 0.001$) but not between NNSE-MT and NNSE-ST (p

= 0.67), with improvements in SRT scores relative to those for UN of 5.4 and 6.4 dB for NNSE-MT and NNSE-ST, respectively. Apart from subject 14, all subjects benefitted from the processing with both NNSE algorithms for speech in ICRA. For some subjects, there were improvements in SRT scores of more than 10 dB.

For BABBLE noise, Mauchly's test showed a violation of sphericity ($p = 0.023$) and a one-way repeated measures ANOVA using the Greenhouse-Geisser correction indicated a significant effect of processing condition [$F(1.364,32.727) = 7.45$, $p = 0.009$]. *Post hoc* pairwise comparisons using Bonferroni correction revealed significant differences between UN and NNSE-ST ($p < 0.001$) and between NNSE-MT and NNSE-ST ($p = 0.035$). A significant improvement in SRT scores relative to UN was observed only for NNSE-ST. Apart from subject 4, all subjects benefitted from NNSE-ST for speech in BABBLE. For NNSE-MT, 8 out of the 14 subjects showed SRT improvements relative to UN of 1.5-3 dB. However, the rest of the subjects performed either the same or more poorly with NNSE-MT than with UN.

PLACEHOLDER - Figure 4

PLACEHOLDER - Figure 5

5. DISCUSSION

Significant improvements in speech intelligibility for CI subjects were produced by NNSE for the three background noises over a range of SNRs. To accomodate the large variability among CI users, algorithm performance was evaluated using an adaptive procedure measuring SRT scores, in contrast to previous studies that tested at fixed SNRs. The magnitude of the improvements in SRT ranged from 1.4 dB for speech in SWN with NNSE-MT up to 6.4 dB for speech in ICRA with NNSE-ST. Apart from NNSE-MT with BABBLE, significant improvements were found for NNSE relative to UN in all conditions.

For SWN, improvements tended to be larger for NNSE-ST than for NNSE-MT (2.3 / 1.4 dB SRT), but this difference was not statistically significant. There was also a non-significant difference of 1 dB between NNSE-MT and NNSE-ST for ICRA (SRTs of 5.4 and 6.4 dB, respectively) but there was a significant difference of 1.6 dB for BABBLE (SRTs of 0.4 and 2.0 dB, respectively). The advantage of NNSE-ST over NNSE-MT was expected due to the mismatch between training and testing sets for NNSE-MT. Nevertheless, NNSE-MT led to significant improvements relative to UN for speech in SWN and ICRA despite the mismatch in speakers. NNSE-MT failed to give significant improvements relative to UN for BABBLE. For this noise condition, competing speakers might be wrongly detected as the target speaker and not attenuated adequately. Especially for lower SNRs, where the spectral energy of the target speaker was less dominant, NNSE-MT performed worse than NNSE-ST (it should be noted, that the training data were increased by nearly a factor of 2 for NNSE-MT, to increase its robustness to unseen speakers). The latter can use *a priori* information about the target speaker's spectral characteristics.

For ICRA, the improvements produced by NNSE (ST and MT) relative to UN were remarkable (about 5 to 6 dB) and were about 3 times larger than for the other two noise conditions. The average SRT for UN was comparable for ICRA and BABBLE. The processing produced a much larger improvement relative to UN for ICRA than for BABBLE. The ICRA noise employed in this study had much stronger spectro-temporal modulations (obtained from one male talker) than the BABBLE noise (20 talkers), leading to more and larger time-frequency (T-F) regions with a positive SNR. We speculate that the NNSE algorithm exploits these positive-SNR T-F regions in the feature space to predict adjacent or even more distant spectro-temporal patterns of the target speech signal. This would enable the algorithm to extrapolate its prediction over potentially masked T-F regions with lower SNR in the corresponding time frame (similar to the mechanism often called "glimpsing" or listening in the dips by human listeners). The algorithm was presented with numerous examples and variations of potential masking patterns during training and thus learned typical spectral patterns of the speech. This constitutes a potential benefit of machine learning algorithms in conjunction with acoustic broadband features over traditional signal processing schemes that operate independently on separate frequency channels.

The machine learning based algorithm proposed by Hu *et al.* (2010) showed large improvements in percentage correct scores for speech in three different non-stationary noise backgrounds for CI listeners. A direct comparison between the performance of their system and NNSE is difficult because we used an adaptive procedure in contrast to testing at fixed SNRs, and we used different speech materials and background noises. Hu *et al.* showed large improvements with an IBM-based processing scheme, but their system was trained on the same speaker, noise realizations and SNRs as used for testing. May *et al.* (2014) showed that the use of novel noise realizations for testing led to a substantial decrease in estimation performance with a Gaussian Mixture Model (GMM) based system, such as the one used by Hu *et al*. Recently, Healy *et al.* (2015) and Bolner *et al.* (2016) have shown that neural network based regression systems can achieve high estimation performance with novel realizations of the same noise type. Both studies tested at fixed SNRs and used acoustic stimuli to test normal hearing and hearing-impaired listeners' speech understanding in noise. Bolner *et al.* tested NH listeners using CI vocoder simulations and reported an improvement of 18% in percentage correct scores for speech in BABBLE at an SNR of 5 dB. This improvement can be compared to the 2-dB improvement in SRT for NNSE-ST, since the two algorithms used the same speaker for training and testing. Jansen *et al.* (2013) reported that, for CI users, an improvement in SRT scores of about 1 dB corresponds to an improvement in percentage correct scores of 18.7% with the LISTm corpus and SWN. This suggests that CI users benefitted more from NNSE processing than the NH listeners with

CI simulations for speech in BABBLE. For SWN at 5 dB SNR, Bolner *et al.* measured an improvement relative to UN of 27%, whereas in this study an improvement of 2.3 dB was achieved by NNSE-ST. Again, this suggests larger benefits for CI users than for NH listeners, but less so than for BABBLE.

Other studies of single-microphone noise reduction for CI users showed consistent improvements in understanding of speech in stationary noise such as SWN (Dawson *et al.*, 2011; Hu *et al.*, 2007; Mauger *et al.*, 2012; Ye *et al.*, 2013). However, the improvements were usually smaller with non-stationary noise and only a few studies achieved significant improvements for both stationary and non-stationary noise (Dawson *et al.*, 2011). Machine-learning based algorithms like NNSE have the potential to overcome this challenge and achieve consistent improvements in both stationary and non-stationary noises, as indicated by the performance of NNSE with BABBLE and ICRA.

Several architectures for machine learning based noise reduction have been proposed in the last few years. In the studies of Kim *et al.* (2009) and Hu and Loizou (2010), GMM classifiers were used, which recently have been surpassed by artificial neural networks with several hidden layers (*deep neural network*, DNN) (Chen *et al.*, 2016; Healy *et al.*, 2013, 2015). Similar to the architecture of the previous GMM-based classification systems, where the SNR of each frequency channel is predicted independently, Healy *et al.* (2013) used two successive stages of multiple-subband DNNs (one DNN for each of the 64 frequency channels) resulting in a very large classification system. Healy *et al.* (2014) reduced the complexity of the DNN by a factor of 43 by using a single DNN for the prediction of the SNR of all frequency channels simultaneously. They used a DNN with 3 hidden layers, each composed of 1024 rectified linear units, and changed the feature extraction process to broadband features (being extracted across all frequency channels simultaneously) resulting in a greatly reduced number of features (64 times smaller) and an input layer dimensionality of just 259. However, this DNN system still had nearly 2.5 million tunable parameters. In the most recent studies on DNN-based speech separation, the complexity was increased again to DNNs with nearly 4 million (Healy *et al.*, 2015) and more than 20 million tunable parameters (Chen *et al.*, 2016). Recent advances in computational power through the use of supercomputers and graphics processing units (GPUs) made it possible to train and execute such complex algorithms in reasonable amounts of time. However, the application of such complex algorithms to hearing devices with strongly limited computational and memory resources is not feasible at present. In contrast, the NNSE algorithm uses a smaller number of relatively simple features combined with a much smaller NN regression system consisting of 2 hidden layers with 75 units each. This NN system has 18,631 tunable parameters, 2/3 of those used by Bolner *et al.* (2016). NNSE employs 200 times fewer parameters than the system used by Healy *et al.* (2015) and has a 1000-fold smaller system complexity than the system used by Chen *et al.* (2016).

Real-time processing requires a processing delay of less than 20-30 ms to ensure perceived audio-visual synchrony and acceptance by users of hearing devices (Stone and Moore, 2005). Besides the computational complexity aspect, which may become less relevant with the steady increase in computational power, the algorithm architectures used in many studies make use of non-causal processing involving the analysis of “future” frames (e.g. from feature sets using 2 future frames used by Healy *et al.*, 2015, up to 11 future frames used by Chen *et al.*, 2016). Generally, algorithms need to work in a causal way to be implementable in hearing devices that meet the perceptual requirements of potential end-users. The NNSE algorithm proposed in this study satisfies this requirement by using only the past and the current frames.

An important aspect of SE algorithms is their ability to generalize to unseen acoustic conditions. NNSE was designed to satisfy several generalization requirements. Firstly, multiple SNRs were used for training, yielding an algorithm that worked over a range of SNRs. This was assessed by using an error rate analysis where NNSE gave decreased total error rates relative to the unprocessed condition for all noise types and SNRs (and even for an untrained SNR of 10 dB). Secondly, novel realizations of a specific type of background noise were used for evaluation. NNSE performed well in these more challenging conditions (as it was also shown by Bolner *et al.*, 2016, and Healy *et al.*, 2015). Thirdly, NNSE-MT was tested using a novel speaker and substantial improvements were found for two out of three noise types. However, generalization to unseen types of noise was not assessed with the current study that used noise-specific training and testing. A future goal is to design a system that works in completely novel noise conditions, but still meets the constraints on delay and computational power of CI processors.

Kim and Loizou (2010) reported that a GMM classifier using amplitude modulation spectrum (AMS) features for estimating the IBM, that was trained on a large number of noise types, failed to achieve satisfactory performance with unseen noises (low classification rates). This was the case even when a speaker-dependent classifier was used. Instead of employing large-scale training to improve generalization, they proposed incrementally adapting the system to new noises. May and Dau (2014) have shown that a GMM-based classifier trained on AMS features tended to overfit the training data more when they increased the dimensionality of the feature space and the complexity of the classifier. The authors observed a larger decrease in classification performance when the algorithm was tested on novel segments of the same noise type for the more complex classifier and feature combinations than for the less complex ones (no evaluation on unseen noise types was performed). They proposed addressing the problem of overfitting with the use of a less complex classification system and a lower dimensionality of the feature space. Chen *et al.* (2016) used large-scale training with thousands of background noises in combination with a powerful DNN system and showed that generalization to unseen noises could be achieved when speaker-dependent models were used. This is a promising result and suggests that DNN-based systems might improve generalization to unseen noises compared to the GMM-based systems that were used in previous studies (Kim and Loizou, 2010; May and Dau, 2014).

GMM-based systems have been used mostly in combination with AMS features (Kim *et al.*, 2009; Kim and Loizou, 2010; Hu and Loizou, 2010; May and Dau, 2014). Chen *et al.* (2014), showed that Gammatone-based features performed better than other features (including AMS) in terms of classification accuracy and HIT-FA rates with a DNN-based system. During the optimization of NNSE, we found similar results, confirming an advantage of Gammatone-based energy features over AMS features. We combined the processing paradigms of Gammatone-based RASTA-PLP features (that incorporate temporal aspects of speech such as modulations), and GFCC features (that perform a de-correlation of the spectral information), with log-compressed Gammatone-energy features in order to increase the robustness to noise and changes in speaker characteristics.

We performed a pilot experiment to evaluate the performance of the NNSE algorithm with unseen types of noise. We used 12 real-world recordings from different noisy environments (various recordings from a stadium, several restaurants and cafeterias, a classroom, a train, city and highway traffic situations; all obtained from freesound.org) and combined 20-s long segments of each recording to form a multi-noise recording with a total length of 4 minutes (the same length as employed for the noise-specific NNSE). The NNSE algorithm was trained on the multi-noise recording using the same procedure as for the listening experiment, and its performance to the noises employed for the training of the noise-specific NNSE was assessed objectively using the NCM speech intelligibility predictor. The NCM scores are shown in Fig. 6 for the single- and multi-talker NNSE algorithm for both noise-specific and noise-independent training (the NCM scores were calculated using 20 sentences from the LISTm corpus).

PLACEHOLDER - Figure 6

For SWN and BABBLE, there was a small decrease in performance with the noise-independent algorithm compared to the noise-specific algorithm for NNSE-ST, and a larger decrease in performance with the noise-independent algorithm compared to the noise-specific algorithm for NNSE-MT. Interestingly, large improvements in NCM scores for both NNSE-ST and NNSE-MT were achieved with the noise-independent algorithms relative to UN. This is promising, because NCM was proven useful for predicting intelligibility outcomes for vocoded stimuli in our pilot study using CI simulations (Bolner *et al.*, 2016) and for CI users (Chen and Loizou, 2011), but it remains unclear if the predicted improvements relative to UN will occur for CI users. For ICRA, the performance of the noise-independent algorithm was much reduced in comparison to that for the noise-specific algorithm for NNSE-ST, and the predicted performance of the noise-independent algorithm equaled that for UN for NNSE-MT (it should be noted that the noise-independent algorithm did not impair intelligibility relative to UN). We speculate that the difference in predicted performance between noise conditions depends on the degree of similarity of the spectro-temporal characteristics between the training and testing noise types. The NCM scores indicate that both the speaker-dependent and the speaker-independent NNSE algorithms generalize better to unseen noise types for cases when the spectro-temporal modulation patterns are somewhat similar between the training and testing noises (as was the case for SWN and BABBLE) than when the training and testing noises contain different spectro-temporal modulation patterns (in the case of ICRA). Instead of using multi-noise training to increase algorithm performance in unseen noise types, a noise-specific algorithm could be combined with an environmental classifier to provide *a priori* knowledge about the noise type (Hazrati *et al.*, 2014; May and Dau, 2013), while retaining the advantages of high SE performance in combination with low processing delay and potentially reduced computational complexity compared to a "one-for-all" large-scale algorithm.

6. CONCLUSIONS

A speech enhancement algorithm based on neural networks (NNSE) intended to improve the perception of speech in noise was evaluated using 14 CI users. Significant improvements, ranging from 1.4 to 6.4 dB in SRT, were achieved with noise-specific neural networks using stationary and non-stationary background noise. The architecture and low processing delay of the NNSE algorithm make it suitable for application in hearing devices. While NNSE was evaluated using a noise-specific approach, several aspects of generalization to unseen acoustic conditions were addressed, most importantly performance with a speaker not used during the training stage. Even though improvements in SRT scores were about 1 to 1.5 dB lower than for the speaker-dependent algorithm, substantial and statistically significant improvements were found for 2 out of 3 noise conditions for the speaker-independent NNSE algorithm. The benefits in CI users' speech in noise understanding are promising and provide motivation for further investigations of this approach. Future development in the rapidly growing field of machine learning can be expected to improve the estimation accuracy and generalization performance to unseen conditions.

ACKNOWLEDGEMENTS

The authors would like to thank all of the subjects who participated in the experiments. We are also very thankful to Prof. Brian Moore for his helpful comments and to Taike Bruyneel, Birgit Philips and Anke Plasmans for performing the listening experiments.

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. PITN-GA-2012-317521 (ITN ICanHear) and from EPSRC grant no. EP/K020501/1.

REFERENCES

Anzalone, M.C., Calandruccio, L., Doherty, K.A., Carney, L.H., 2006. Determination of the potential benefit of time-frequency gain manipulation. Ear Hear. 27, 480–492. doi:10.1097/01.aud.0000233891.86809.df

Boll, S.F., 1979. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Trans. Acoust. 27, 113–120. doi:10.1109/TASSP.1979.1163209

Bolner, F., Goehring, T., Monaghan, J., Van Dijk, B., Wouters, J., 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 (ICASSP), Shanghai. IEEE, pp. 6520–6524. doi:10.1109/ICASSP.2016.7472933

Brons, I., Houben, R., Dreschler, W.A., 2012. Perceptual effects of noise reduction by time-frequency masking of noisy speech. J. Acoust. Soc. Am. 132, 2690–9. doi:10.1121/1.4747006

Brungart, D.S., Chang, P.S., Simpson, B.D., Wang, D., 2006. Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation. J. Acoust. Soc. Am. 120, 4007. doi:10.1121/1.2363929

Chen, F., Loizou, P.C., 2012. Impact of SNR and gain-function over- and under-estimation on speech intelligibility. Speech Commun. 54, 272–281. doi:10.1016/j.specom.2011.09.002

Chen, F., Loizou, P.C., 2011. Predicting the intelligibility of vocoded. Ear Hear. 32, 331–338. doi:10.1097/AUD.0b013e33181ff3535

Chen, J., Wang, Y., Yoho, S.E., Wang, D., 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. doi:10.1121/1.4948445

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120, 2421–2424. doi:10.1121/1.2229005

Cullington, H.E., 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, 450–61. doi:10.1121/1.2805617

Dau, T., Kollmeier, B., Kohlrausch, A., 1997. Modeling auditory processing of amplitude modulation .1. Detection and masking with narrow-band carriers. J. Acoust. Soc. Am. 102, 2892–2905. doi:10.1121/1.420344

Dawson, P.W., Mauger, S.J., Hersbach, A.A., 2011. Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus® cochlear implant recipients. Ear Hear. 32, 382–390. doi:10.1097/AUD.0b013e318201c200

Dreschler, W.A., Verschuure, H., Ludvigsen, C., Westermann, S., 2001. ICRA Noises: Artificial Noise Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment. Audiology 40, 148–157. doi:10.3109/00206090109073110

Fetterman, B.L., Domico, E.H., 2002. Speech recognition in background noise of cochlear implant patients. Otolaryngol. Head. Neck Surg. 126, 257–63. doi:10.1067/mhn.2002.123044

French, N.R., Steinberg, J.C., 1947. Factors Governing the Intelligibility of Speech Sounds. J. Acoust. Soc. Am. 19(1), 90–119. doi:10.1121/1.1916407

Friesen, L.M., Shannon, R. V, Baskent, D., Wang, X., 2001. Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. J. Acoust. Soc. Am. 110, 1150–1163. doi:10.1121/1.1381538

Fu, Q.J., Shannon, R. V, Wang, X., 2013. Effects of noise and spectral resolution on vowel and consonant recognition : Acoustic and electric hearing. J. Acoust. Soc. Am. 104, 3586–3596. doi:10.1121/1.423941

Gerkmann, T., Hendriks, R.C., 2012. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans. Audio, Speech Lang. Process. 20, 1383–1393. doi:10.1109/TASL.2011.2180896

Gibak Kim, Loizou, P.C., 2010. Improving Speech Intelligibility in Noise Using Environment-Optimized Algorithms. IEEE Trans. Audio. Speech. Lang. Processing 18, 2080–2090. doi:10.1109/TASL.2010.2041116

Glasberg, B.R., Moore, B.C.., 1990. Derivation of auditory filter shapes from notched-noise data. Hear. Res. 47, 103–138. doi:10.1016/0378-5955(90)90170-T

Hazrati, O., Sadjadi, S.O., Hansen, J.H.L., 2014. Robust and efficient environment detection for adaptive speech enhancement in cochlear implants. 2014 IEEE Int. Conf. Acoust. Speech Signal Process. 900–904. doi:10.1109/ICASSP.2014.6853727

Healy, E.W., Yoho, S.E., Wang, Y., Wang, D., 2013. An algorithm to improve speech recognition in noise for hearing-impaired listeners. J. Acoust. Soc. Am. 134, 3029–38. doi:10.1121/1.4820893

Healy, E.W., Yoho, S.E., Wang, Y., Apoux, F., 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. doi:10.1121/1.4901712

Healy, E.W., Yoho, S.E., Chen, J., Wang, Y., 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. doi:10.1121/1.4929493

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Trans. Speech Audio Process. 2, 587–589.

Hohmann, V., 2002. Frequency analysis and synthesis using a Gammatone filterbank. Acta Acust. united with Acust. 88, 433–442.

Holube, I., Kollmeier, B., 1996. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. J. Acoust. Soc. Am. 100, 1703–1716. doi:10.1121/1.417354

Hu, Y., Loizou, P.C., 2010. Environment-specific noise suppression for improved speech intelligibility by cochlear implant users. J. Acoust. Soc. Am. 127, 3689–95. doi:10.1121/1.3365256

Hu, Y., Loizou, P.C., 2008. A new sound coding strategy for suppressing noise in cochlear implants. J. Acoust. Soc. Am. 124, 498–509. doi:10.1121/1.2924131

Hu, Y., Loizou, P.C., Li, N., Kasturi, K., 2007. Use of a sigmoidal-shaped function for noise attenuation in cochlear implants. J. Acoust. Soc. Am. 122, EL128-34. doi:10.1121/1.2772401

Jansen, S., Koning, R., Wouters, J., van Wieringen, A., 2014. Development and validation of the Leuven intelligibility sentence test with male speaker (LIST-m). Int. J. Audiol. 53, 55–9. doi:10.3109/14992027.2013.839886

Jin, S.H., Nie, Y., Nelson, P., 2013. Masking release and modulation interference in cochlear implant and simulation listeners. Am. J. Audiol. 22, 135–146. doi:10.1044/1059-0889(2013/12-0049)

Kim, G., Lu, Y., Hu, Y., Loizou, P.C., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. 126, 1486–94. doi:10.1121/1.3184603

Koning, R., Madhu, N., Wouters, J., 2015. Ideal Time – Frequency Masking Algorithms Lead to Different Speech Intelligibility and Quality in Normal-Hearing and Cochlear Implant Listeners 62, 331–341.

Lim, J., Oppenheim, A., 1979. Enhancement and band width compression of noisy speech. Proc. IEEE 67, 1586–1604.

Loizou, P.C., 2013. Speech Enhancement: Theory and Practice. CRC Press. Boca Raton, United States.

Ma, J., Hu, Y., Loizou, P.C., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 125, 3387–3405. doi:10.1121/1.3097493

Madhu, N., Spriet, A., Jansen, S., Koning, R., 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. Processing 21, 63–72. doi:10.1109/TASL.2012.2213248

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9, 504–512. doi:10.1109/89.928915

Mauger, S., Warren, C., 2014. Clinical evaluation of the Nucleus® 6 cochlear implant system: Performance improvements with SmartSound iQ. Int. J. Audiol. 53, 564–76. doi:10.3109/14992027.2014.895431

Mauger, S.J., Arora, K., Dawson, P.W., 2012. Cochlear implant optimized noise reduction. J. Neural Eng. 9, 65007. doi:10.1088/1741-2560/9/6/065007

Mauger, S.J., Dawson, P.W., Hersbach, A.A., 2012. Perceptually optimized gain function for cochlear implant signal-to-noise ratio based noise reduction. J. Acoust. Soc. Am. 131, 327–36. doi:10.1121/1.3665990

May, T., Dau, T., 2013. Environment-aware ideal binary mask estimation using monaural cues, in: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. doi:10.1109/WASPAA.2013.6701821

May, T., Dau, T., 2014. Requirements for the evaluation of computational speech segregation systems. J. Acoust. Soc. Am. 136, EL398. doi:10.1121/1.4901133

Miller, G.A., Licklider, J.C.R., 1950. The Intelligibility of Interrupted Speech. J. Acoust. Soc. Am. 22, 167. doi:10.1121/1.1906584

Oxenham, A.J., 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. doi:10.1177/2331216514553783

Qazi, O.U.R., van Dijk, B., Moonen, M., Wouters, J., 2013. Understanding the effect of noise on electrical stimulation sequences in cochlear implants and its impact on speech intelligibility. Hear. Res. 299, 79–87. doi:10.1016/j.heares.2013.01.018

Riedmiller, M., Braun, H., 1993. A direct adaptive method for faster backpropagation learning: The RPROP algorithm. IEEE Int. Conf. Neural Networks - Conf. Proc, San Francisco. January, 586–591. doi:10.1109/ICNN.1993.298623

Seligman P. M. and McDermott H. J., 1995. Architecture of the spectra 22 speech processor. Ann. Otol. Rhinol. Laryngol. 104, 139–141.

Spriet, A., Van Deun, L., Eftaxiadis, K., Laneau, J., Moonen, M., van Dijk, B., van Wieringen, A., Wouters, J., 2007. Speech understanding in background noise with the two-microphone adaptive beamformer BEAM in the Nucleus Freedom Cochlear Implant System. Ear Hear. 28, 62–72. doi:10.1097/01.aud.0000252470.54246.54

Stickney, G.S., Zeng, F.-G., Litovsky, R., Assmann, P., 2004. Cochlear implant speech recognition with speech maskers. J. Acoust. Soc. Am. 116, 1081–1091. doi:10.1121/1.1772399

Stone, M.A., Moore, B.C.J., 2005. Tolerable hearing-aid delays: IV. effects on subjective disturbance during speech production by hearing-impaired subjects. Ear Hear. 26, 225–35.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. IEEE Trans. Audio. Speech. Lang. Processing 19, 2125–2136. doi:10.1109/TASL.2011.2114881

Tchorz, J., Kollmeier, B., 2003. SNR estimation based on amplitude modulation analysis with applications to noise suppression. IEEE Trans. Speech Audio Process. 11, 184–192. doi:10.1109/TSA.2003.811542

Van Wieringen, A., Wouters, J., 2008. LIST and LINT: Sentences and numbers for quantifying speech understanding in severely impaired listeners for Flanders and the Netherlands. Int. J. Audiol. 47, 348–355.

doi:10.1080/14992020801895144

Wang, D., Kjems, U., Pedersen, M.S., Boldt, J.B., Lunner, T., 2009. Speech intelligibility in background noise with ideal binary time-frequency masking. J. Acoust. Soc. Am. 125, 2336–47. doi:10.1121/1.3083233

Wouters, J., Damman, W., Bosman, A.J., 1994. Vlaamse opname van woordenlijsten voor spraakaudiometrie. Logop. informatiemedium van Vlaam. Ver. voor Logop.

Wouters, J., Van den Berghe, J., 2001. Speech recognition in noise for cochlear implantees with a two microphone monaural adaptive noise reduction system. Ear Hear. 22, 420–430.

Ye, H., Deng, G., Mauger, S.J., Hersbach, A.A., Dawson, P.W., Heasman, J.M., 2013. A Wavelet-Based Noise Reduction Algorithm and Its Clinical Evaluation in Cochlear Implants. PLoS One 8. doi:10.1371/journal.pone.0075662

Zeng, F.G., Rebscher, S., Harrison, W., Sun, X., Feng, H., 2008. Cochlear implants: system design, integration, and evaluation. IEEE Rev. Biomed. Eng. doi:10.1109/RBME.2008.2008250

Figure Captions

Figure 1. Block diagram of the proposed speech enhancement algorithm integrated into the ACE signal path (including an automatic gain control, AGC, and loudness growth function, LGF). The algorithm has two components: Feature Extraction and Neural Network.

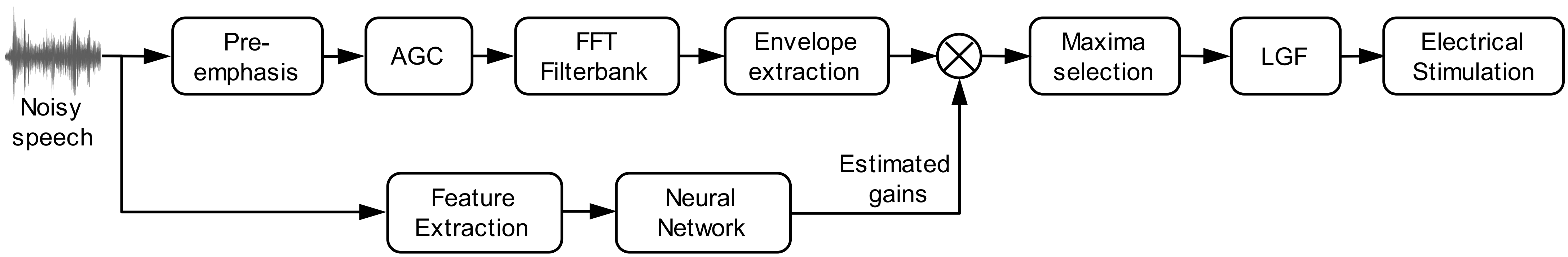
Figure 2. Electrodogram of the sentence *'Het verhaal is heel spannend'* produced by a male speaker (LISTm) at a level of 65 dB SPL. The top panel is for the noise-free signal. The second panel is for the signal with BABBLE noise (SNR = 5 dB). The third and fourth panels are for the conditions with NNSE-MT and NNSE-ST, respectively.

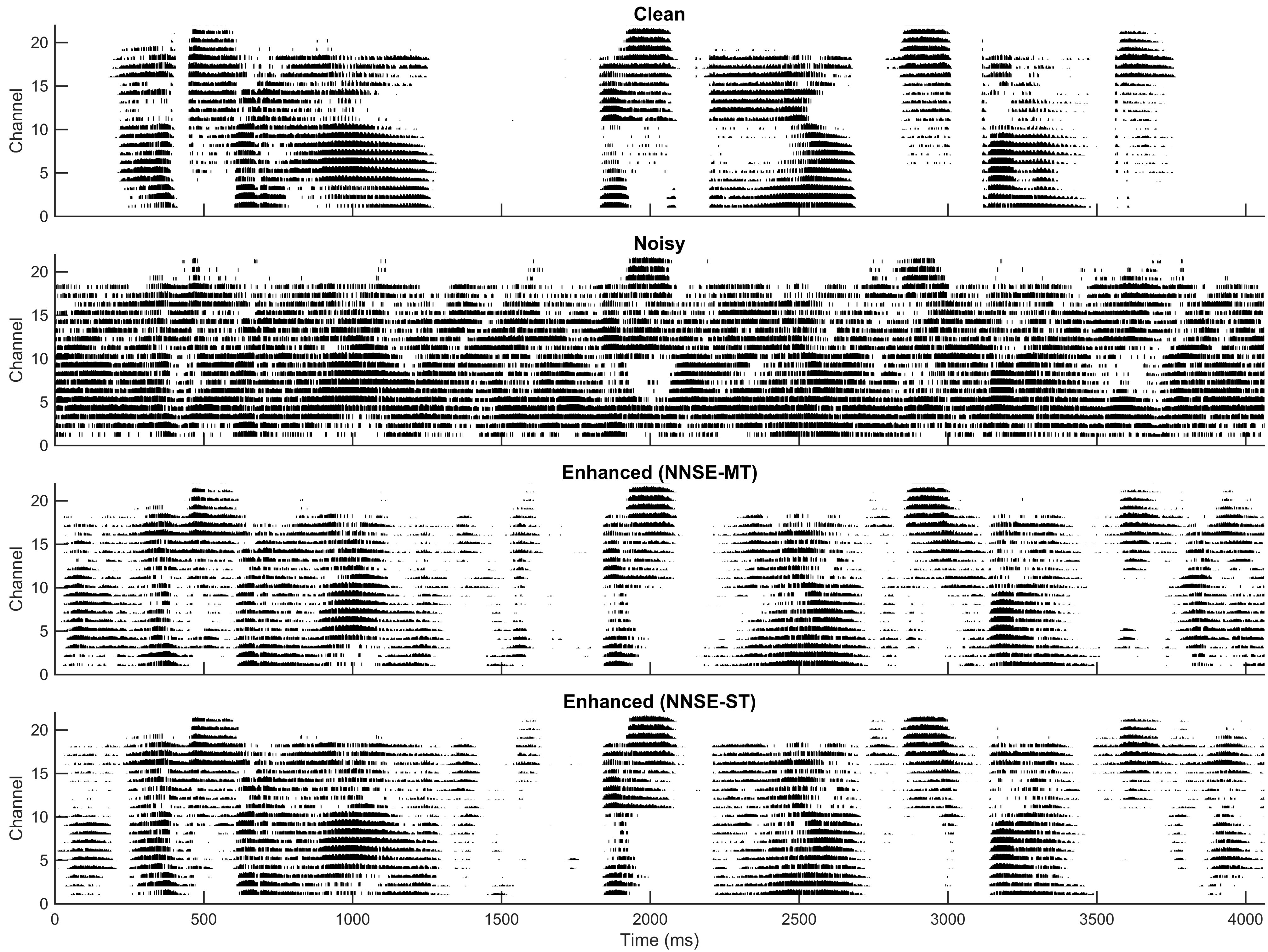
Figure 3. Error rate analysis for UN, NNSE-MT and NNSE-ST processing conditions for the three noises, at –5, 0, 5 and 10 dB SNR. Lines join error rates for the same input SNR. The target speech was LISTm sentences (not part of the training database of either of the NNSE algorithms).

Figure 4. Group mean SRTs with UN (ACE), NNSE-MT (multi-talker) and NNSE-ST (single-talker) processing for each noise type (left: SWN, center: ICRA, right: BABBLE). Error bars represent the standard error of the mean; (*) $p \leq 0.05$, (**) $p \leq 0.01$, (***) $p \leq 0.001$.

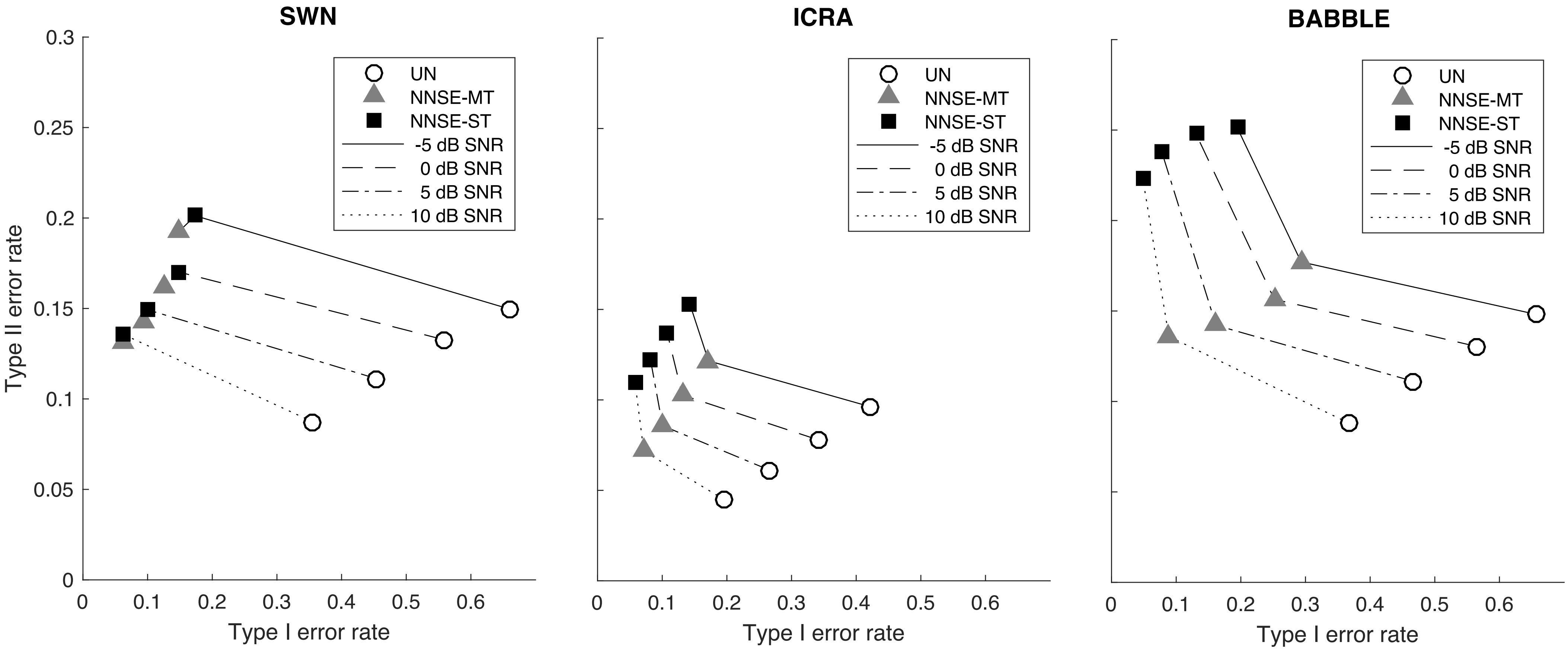
Figure 5. Top panel: Individual SRTs for UN (ACE), NNSE-MT (multi-talker) and NNSE-ST (single-talker) processing for each noise type (left: SWN, center: ICRA, right: BABBLE). Bottom panel: individual SRT change (positive is better) relative to the UN condition for NNSE-MT and NNSE-ST, for the three noises. Subjects are ordered by their performance for speech in UN (ascending SRT from left to right).

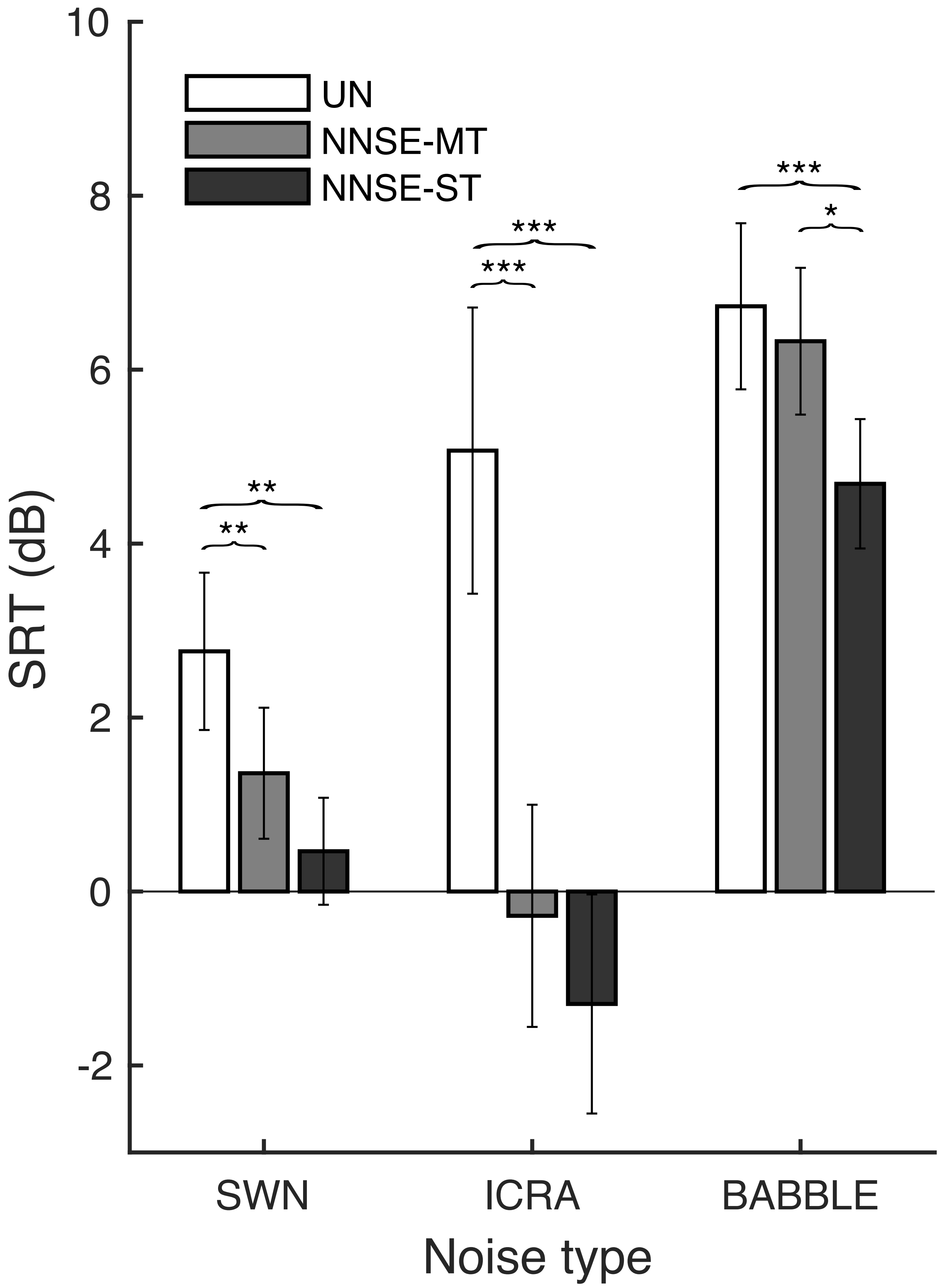
Figure 6. NCM intelligibility prediction scores for UN (ACE), MT-NI (NNSE-MT with noise-independent training), MT-NS (NNSE-MT with noise-specific training), ST-NI (NNSE-ST with noise-independent training), ST-NS (NNSE-ST with noise-specific training) and IRM (ideal ratio mask) for each noise type (left: SWN, center: ICRA, right: BABBLE).


Noisy speech
Pre-emphasis
AGC
FFT Filterbank
Envelope extraction
Maxima selection
LGF
Electrical Stimulation
Feature Extraction
Neural Network
Estimated gains

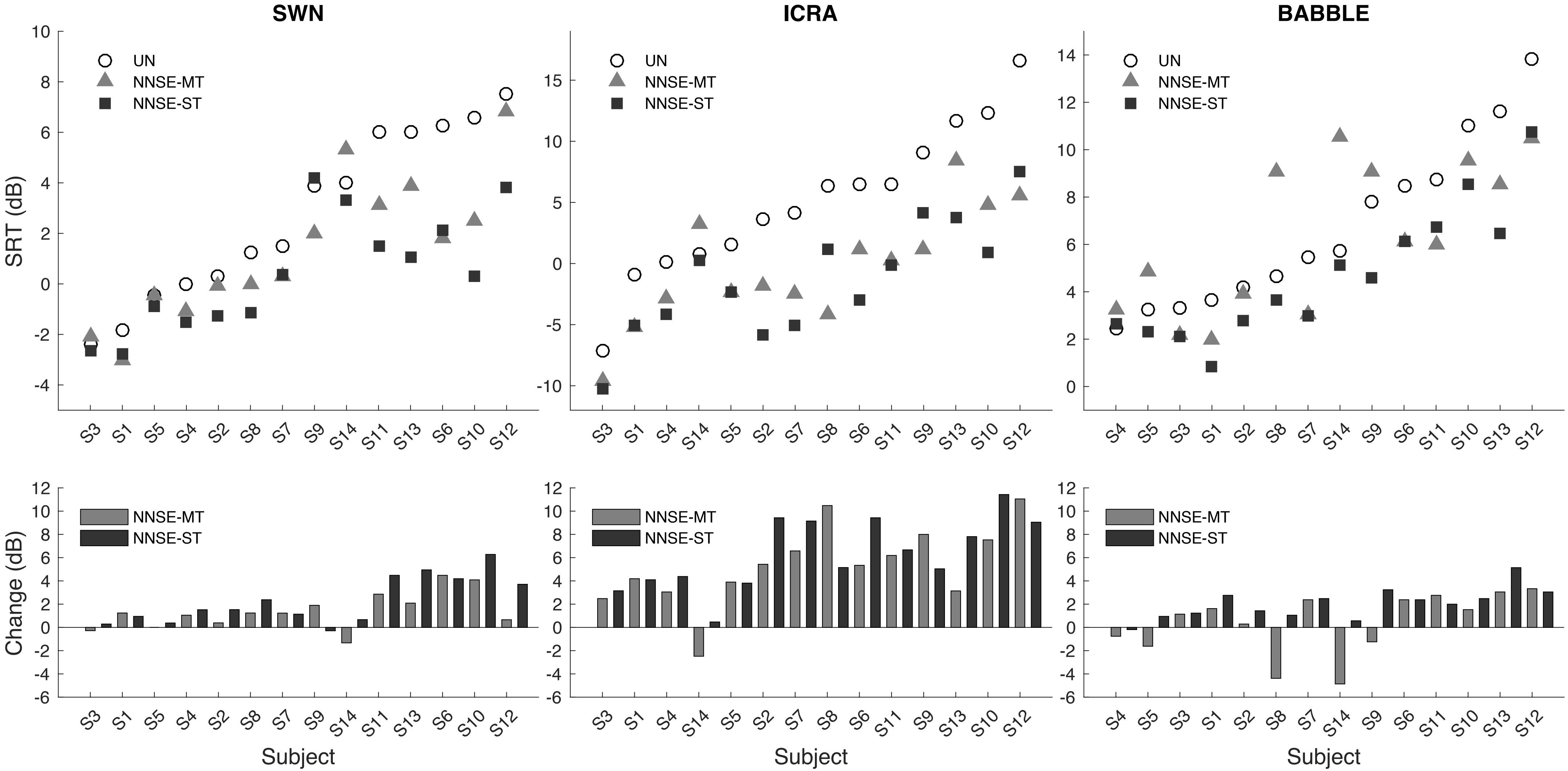


Clean
Noisy
Enhanced (NNSE-MT)
Enhanced (NNSE-ST)
Channel
Time (ms)
0
5
10
15
20
500
1000
1500
2000
2500
3000
3500
4000


SWN
ICRA
BABBLE
Type II error rate
Type I error rate
UN
NNSE-MT
NNSE-ST
-5 dB SNR
0 dB SNR
5 dB SNR
10 dB SNR



10
8
6
4
2
0
-2
SRT (dB)
UN
NNSE-MT
NNSE-ST
**
**
***
***
***
*
SWN
ICRA
BABBLE
Noise type



SWN
ICRA
BABBLE
SRT (dB)
Change (dB)
UN
NNSE-MT
NNSE-ST
Subject

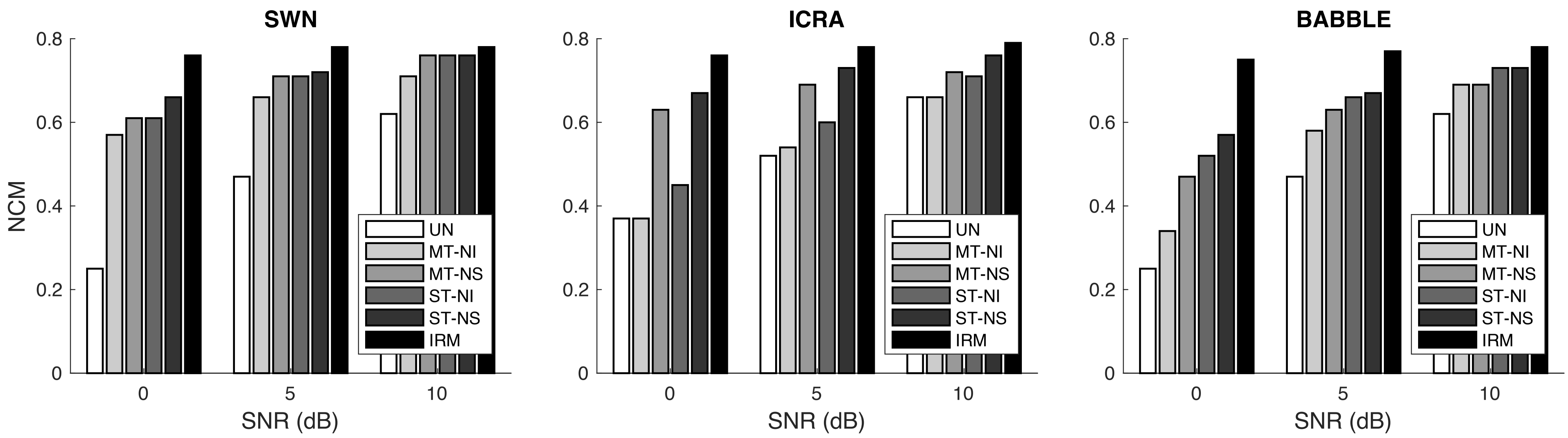

SWN
ICRA
BABBLE
NCM
0
0.2
0.4
0.6
0.8
0
5
10
SNR (dB)
UN
MT-NI
MT-NS
ST-NI
ST-NS
IRM

Table Captions

Table 1. Individual subject demographics: age (years), tested ear (left/right), duration of implant use (years), implant type, origin of hearing loss, etiology, and duration of profound hearing loss (years).

Table 2. CI parameters for each of the 14 subjects during the study: channel stimulation rate (Hz), number of maxima/number of active electrodes, THL and MCL (threshold and comfort levels in current level, CL), minimum and maximum of the dynamic range (DR, in CL).

Table 1.

Subject	Age	Tested Ear	Implant use	Implant type	Type of HL	Etiology	Duration of profound HL	Contralateral ear
01	62	R	12.6	CI24R	Progressive	Unknown	Unknown	-
02	62	L	11.3	CI24R	Progressive	Cholesteatoma	48	-
03	53	L	12.6	CI24R	Progressive	Unknown	47	-
04	68	L	8.1	CI24RE	Progressive	Meniere's Disease	17	HA
05	70	L	13.3	CI24R	Progressive	Otosclerosis	60	HA
06	69	R	10.6	CI24RE	Progressive	Meningitis and Otosclerosis	45	-
07	60	R	5.1	CI512	Sudden	Cholesteatoma	5	HA
08	35	L	11.5	CI24RE	Sudden	Meningitis	3	-
09	81	R	12.6	CI24R	Progressive	Cholesteatoma and Chronic Mastoiditis	Unknown	-
10	69	L	9.6	CI24RE	Sudden	Unknown	53	-
11	72	L	6.6	CI24RE	Progressive	Meniere's Disease	8	-
12	76	R	1.2	CI512	Progressive	Familial	5	HA
13	52	L	8.1	CI24RE	Congenital	Unknown	52	HA
14	23	R	13.6	CI24R	Congenital	Waardenburg Syndrome	1	CI24R

Table 2.

Subject	Channel stimulation rate	Pulse Width	Maxima / no. active electrodes	THL-current level		MCL-current level		DR	
UNIT	Hz	µs		Min CL	Max CL	Min CL	Max CL	Min CL	Max CL
01	900	25	14/20	105	130	150	193	39	68
02	900	25	10/19	120	135	174	184	39	60
03	900	25	14/22	108	134	165	194	47	79
04	900	25	14/22	109	176	171	200	24	62
05	900	25	14/20	113	129	159	182	42	66
06	1800	20	10/22	150	180	177	228	27	48
07	900	25	14/22	130	160	153	185	15	28
08	2400	12	10/22	111	125	195	205	75	88
09	900	25	14/20	135	152	157	175	17	28
10	900	25	8/22	78	145	108	168	10	36
11	900	25	11/22	129	171	158	203	28	32
12	900	25	12/22	98	144	132	178	32	34
13	900	25	10/21	109	151	137	190	18	73
14	900	25	14/22	120	145	186	205	50	80

Highlights

- An algorithm for improving speech understanding in noise for cochlear implant users is evaluated
- Significant improvements were found for stationary and non-stationary noise types
- It generalizes to a novel speaker and works over a range of signal-to-noise ratios
- The small algorithmic delay makes it suitable for real-time application