SPEECH ENHANCEMENT BASED ON NEURAL NETWORKS APPLIED TO COCHLEAR IMPLANT CODING STRATEGIES

Federico Bolner[1,2], Tobias Goehring[3*], Jessica Monaghan[3], Bas van Dijk[2], Jan Wouters[1], and Stefan Bleeck[3]*

[*]these authors contributed equally to this work
[1]ExpORL, KU Leuven, Leuven, Belgium
[2]Cochlear Technology Centre, Mechelen, Belgium
[3]ISVR, University of Southampton, Southampton, UK
fbolner@cochlear.com, t.goehring@soton.ac.uk

ABSTRACT

Traditionally, algorithms that attempt to significantly improve speech intelligibility in noise for cochlear implant (CI) users have met with limited success, particularly in the presence of a fluctuating masker. In the present study, a speech enhancement algorithm integrating an artificial neural network (NN) into CI coding strategies is proposed. The algorithm decomposes the noisy input signal into time-frequency units, extracts a set of auditory-inspired features and feeds them to the NN to produce an estimation of which CI channels contain more perceptually important information (higher signal-to-noise ratio, SNR). This estimate is then used accordingly to retain a subset of channels for electrical stimulation, as in traditional n-of-m coding strategies. The proposed algorithm was tested with 10 normal-hearing participants listening to CI noise-vocoder simulations against a conventional Wiener filter based enhancement algorithm. Significant improvements in speech intelligibility in stationary and fluctuating noise were found over both unprocessed and Wiener filter processed conditions.

Index Terms— Cochlear implants, noise reduction, speech enhancement, neural networks, machine learning

1. INTRODUCTION

State-of-the-art cochlear implants (CI) allow for near-to-normal speech understanding in quiet acoustic conditions, however environmental noises represent one of the main challenges for CI users' speech understanding in everyday life [1]. Several speech enhancement algorithms for cochlear implants have been proposed to alleviate this problem.

Single-channel speech enhancement techniques have been successfully applied to cochlear implant sound coding and demonstrated to improve speech intelligibility in certain acoustical environments. These algorithms rely on statistical assumptions about the background noise (e.g. stationarity) for the estimation of the SNR in order to modify the spectral content of the signal. Maximum benefits of around 2.5 dB in speech reception threshold (SRT) were demonstrated in stationary noise, but the benefit is much reduced when the interfering noise is non-stationary, as in the case of competing talkers [2,3].

More recent approaches, such as supervised speech separation techniques, have been reported to improve speech intelligibility also in fluctuating background noise [4,5]. These algorithms make use of a binary classifier trained on the task of estimating the ideal binary mask (IBM). The concept of the IBM is based on retaining speech-dominant time-frequency (T-F) units while discarding masker-dominant T-F units with lower SNR by the application of an SNR threshold, the local criterion (LC) [6,7]. A demonstration of intelligibility improvement for CI users by a monaural algorithm has been provided by Hu *et al.* [4]. The authors used a Gaussian mixture model-based classifier to decide whether each CI channel was dominated by speech or by noise. Only speech-dominated channels were retained for electrical stimulation, resulting in large improvements in speech intelligibility in babble, train and exhibition hall noise. More recently, deep neural networks have been applied to the task of the IBM estimation and have shown significant improvements in speech intelligibility for normal-hearing (NH) and hearing-impaired (HI) listeners [5].

These studies represent a promising direction for improving speech enhancement algorithms, but are yet limited to a specific set of acoustic scenarios used during the training stage of the algorithm and depend on the choice of a LC for the IBM estimation. In contrast, the ideal Wiener filter (IWF) (also known as ideal ratio mask, IRM) applies a gradual weight to each T-F unit according to its local SNR and does not depend on the choice of a LC [8,9]. Listening tests conducted with NH listeners have shown that the IWF is less sensitive to estimation errors, leads to higher intelligibility scores in low SNR conditions, and is preferred in terms of perceived quality compared with the IBM [10].

The present study aims to investigate the potential improvements in speech intelligibility of a NN-based speech enhancement algorithm applied to CI coding strategies, hereafter referred as NNSE. The algorithm uses the IWF target for the training of the NN. In contrast to previous studies, we evaluated the performance on unseen noise realizations while reducing the complexity of the algorithm.

2. ALGORITHM DESCRIPTION

The integration of NNSE into a typical CI signal path is shown in Figure 1. We used the Advanced Combination Encoder (ACE™), an n-of-m speech coding strategy, where the input signal is decomposed into $m = 22$ frequency channels from which the envelope information is extracted. Maxima selection then retains only a subset of n channels with the largest amplitudes (maxima) for electrical stimulation in each stimulation cycle. In this study, we chose a typical value of eight maxima.

The proposed algorithm consists of two main components: feature extraction and gain estimation. The integration of the NNSE into CI processing does not require a reconstruction stage, since the energy in frequency channels is directly used to determine the electrode output. Noisy input signals were first downsampled to 16 kHz and divided into 20-ms frames with 10-ms overlap, from which a set of features was extracted and passed to an artificial neural network trained on the task of estimating the IWF gain over 63 frequency channels of a gammatone filterbank with centre frequencies ranging from 50 to 8000 Hz. The IWF gain was calculated as:

$$G_{k,n} = \frac{\xi_{k,n}}{1+\xi_{k,n}}, \quad (1)$$

where $\xi_{k,n}$ is the (true) SNR of the k-th frequency channel and n-th frame. The estimated gains were then remapped to the 22 CI channels, smoothed (exponential smoothing with a time constant $\tau = 12$ ms) and applied to the noisy envelopes before ACE maxima selection. This has the main effect of attenuating masker-dominated channels, ultimately affecting maxima selection so that target-dominated channels are more likely to be selected for electrical stimulation.

2.1. Feature extraction

In contrast to previous studies that employed sub-band feature sets [4,5], we extracted features from the broadband signal. The feature set was extracted from each 20-ms long frame and consisted of two widely used speech recognition features - the Mel-Frequency Cepstral Coefficients (MFCC) and the Relative Spectral Transform Perceptual Linear Prediction (RASTA-PLP) feature - concatenated with the Gammatone log-energies (GTE) features. Our experimental results indicated that this combination led to higher estimation accuracy than the individual features alone.

To compute the MFCC and RASTA-PLP features, we applied a Hanning window to the input frame to then derive

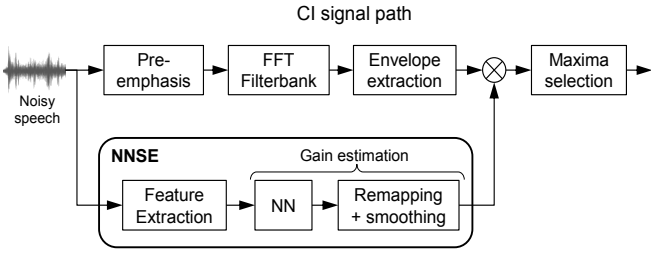

Figure 1 - System block diagram of the proposed speech enhancement strategy (NNSE) integrated in ACE.

the power spectrum using short-time Fourier transform. For MFCC, the spectrum was converted into Mel scale, followed by log-compression and discrete cosine transform to obtain 31 cepstral coefficients. For RASTA-PLP, the power spectrum was instead warped to the Bark scale, log compressed, filtered by the RASTA filter (which emphasizes the modulation frequencies relevant to human speech), and expanded again by an exponential function. Finally, a 12-th order linear prediction model analysis was performed on this filtered spectrum to derive 13 RASTA-PLP features. To extract GTE features, we passed each input signal frame through the same 63-channel gammatone filterbank used to compute the IWF target gains. The energy of each sub-band signal was log-compressed to obtain 63 GTE features. Since speech typically exhibits highly structured spectro-temporal patterns, we added contextual temporal-information in the form of the 107 features of the previous frame, for a total of 214 features for each time-frame.

2.2. Artificial neural network training

In this study, we used a feed-forward NN with two hidden layers of 100 and 50 units. We found that the use of two hidden layers increased the estimation accuracy, while additional layers did not provide further benefit. The number of units in the input and output layers is given by the dimensionality of the input feature set and the output gains (214 and 63-D, respectively). Both hidden layers used a saturating linear transfer function (linear between 0 and 1, but saturating at these values outside that range), whereas the output layer used a linear transfer function.

We trained the network to estimate the IWF gain mask using the resilient back-propagation algorithm, with the mean squared error performance function and weight decay regularization to avoid overfitting. The network was trained with a total of 80 sentences (eight lists) from the IEEE database (male talker) [11]. The interfering maskers included speech shaped noise (SSN) with the same long-term spectrum as the target speech, and BABBLE noise (4 male and 4 female talkers from the TIMIT corpus). Each noise recording was 26-seconds long. We used 18-seconds long segments of each masker for the training, while the remaining 8 seconds were left for testing. The sentences were mixed with random parts of both noises at 8 SNR

levels, in 2 dB increments from -6 to 8 dB. The training of one NN took around 15 minutes on a newer generation personal computer. A different network was trained for each masker tested (but not for each SNR condition).

3. EVALUATION

3.1. Test material

The target speech material used for the testing included the remaining 64 lists from the IEEE corpus (male talker) after the neural network training, while the maskers consisted of random parts of 8-second long segments of the noise recordings (SSN and BABBLE) reserved for testing (see Section 2.2).

The noisy mixtures were processed off-line with the Nucleus MATLAB Toolbox ACE implementation in three conditions: unprocessed noisy speech (UN), signal enhanced with the proposed algorithm (NNSE), and enhanced with a Wiener-filter type algorithm (WFSE). In the WFSE condition, noisy mixtures were pre-processed with a Wiener filter based on *a priori* SNR estimation [12] using the unbiased MMSE-based noise power estimator [13] prior to CI processing. This condition was added in order to compare NNSE with a state of the art speech enhancement algorithm. A broadband correction gain was applied after the enhancement algorithms to restore the level of the speech component to its original level. Finally, after the ACE maxima selection, processed envelopes were passed through a noise-band vocoder (with the same cutoff frequencies as the ACE filterbank) to simulate CI processing and obtain the test stimuli used for the objective evaluation and the listening experiment.

3.2. Objective evaluation: procedure and results

To compare the accuracy of the neural network estimation with previous studies, we converted the estimated and IWF gain masks into binary masks using an LC set 6 dB lower than the overall mixture SNR. We then calculated the average correctly classified T-F units (HIT) and false alarm (FA, equivalent to type-I error) percentage rates [4].

Additionally, we computed two speech intelligibility measures of the processed vocoded speech (using clean vocoded speech as reference): the short-time objective intelligibility measure (STOI) [14], which has been developed for T-F weighted noisy speech, and the normalized covariance metric (NCM), which is closely related to CI processing and has been successfully applied to vocoded signals previously [15,16]. Accuracy rates and intelligibility scores were computed over 100 sentences and are shown in Table 1 and Table 2, respectively.

The proposed algorithm reached high performance in terms of HIT-FA rate, which has been shown to correlate with speech intelligibility. Intelligibility measures predicted higher scores for the NNSE compared with unprocessed speech (UN) and - to a lower extent - over the WFSE.

Table 1 – HIT-FA and FA rates (expressed in percent) obtained with NNSE for the four noise conditions.

	SSN		BABBLE	
	0 dB	5 dB	5 dB	10 dB
HIT-FA	74.40	76.18	69.65	67.38
FA	7.75	3.17	8.75	3.20

Table 2 - STOI and NCM scores for the four noise conditions and the three processing conditions.

	SSN		BABBLE	
	0 dB	5 dB	5 dB	10 dB
		STOI		
UN	0.54	0.63	0.58	0.65
WFSE	0.60	0.69	0.59	0.67
NNSE	0.64	0.70	0.64	0.70
		NCM		
UN	0.50	0.63	0.42	0.57
WFSE	0.60	0.69	0.44	0.59
NNSE	0.64	0.69	0.58	0.66

3.3. Listening experiment: procedure and results

Ten normal-hearing native English speakers (six males and four females, with an average age of 29 years) participated in this study. The test material is described in Section 3.1.

Testing began with a short training to acclimatize the subject to the vocoded stimuli, consisting of one list of clean speech followed by one list at 10 dB SNR for each masker type. The listening test involved a sentence recognition task (five keywords per sentence) of vocoded speech in four noise conditions: in SSN at 0 and 5 dB SNR, and in BABBLE noise at 5 and 10 dB SNR. The SNR levels were chosen to avoid floor and ceiling effects. Subjects were presented with two lists for each of the 12 conditions [3 processing strategies (UN, WFSE and NNSE) × 2 SNRs × 2 maskers]. The presentation order of processing strategy and SNR level was randomised for each subject. Stimuli were presented diotically over closed circumaural headphones (Sennheiser HD380 pro) at 65 dB SPL.

Percentage correct word scores are shown in Figure 2. For SSN, analysis of variance with repeated measures indicated significant effects of both SNR level [$F(1,9) = 686.2$, $p < 0.001$] and processing condition [$F(2,18) = 47.3$, $p < 0.001$], and a significant interaction between the two [$F(2,18) = 5.9$, $p = 0.011$]. For BABBLE, significant effects of both SNR level [$F(1,9) = 265.3$, $p < 0.001$] and processing condition [$F(2,18) = 70.2$, $p < 0.001$] were found.

Bonferroni corrected *post hoc* tests were run to assess the statistical significance between conditions. Results show significant improvements of the proposed NNSE algorithm over both UN and WFSE in all noise types and SNR levels (*p*-values are shown in Figure 2).

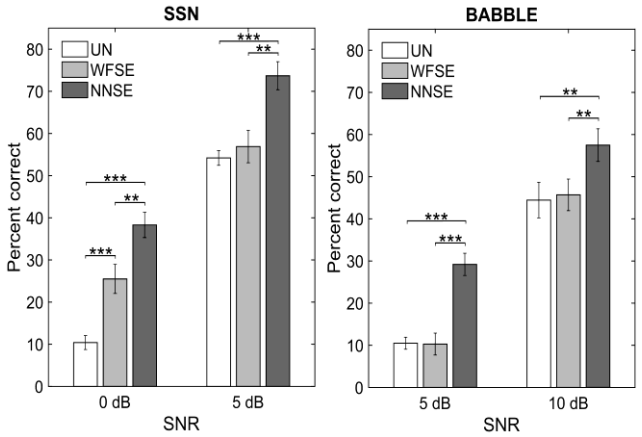

Figure 2 - Mean speech intelligibility scores of 10 NH subjects in noisy speech (UN), Wiener-filter-based speech enhancement (WFSE), and the proposed algorithm (NNSE) in SSN and BABBLE noise. Error bars represent the standard error of the mean; (**)$p \leq 0.01$, (***)$p \leq 0.001$.

4. DISCUSSION AND CONCLUSION

Significant improvements in intelligibility were observed with the proposed neural-network based speech enhancement strategy (NNSE). These improvements were consistent in all the conditions tested, both compared with unprocessed noisy (UN) and with a conventional Wiener filter based speech enhancement algorithm (WFSE). The improvements were generally more noticeable in the lower SNR level of each masker. For instance, the improvement in mean scores for NNSE reached 27% in SSN and 18% in BABBLE noise over UN ($p < 0.001$), and 11% in SSN ($p < 0.01$) and 18% in BABBLE noise ($p < 0.001$) over the WFSE. Improvements were predicted by the objective intelligibility scores, even though both STOI and NCM underestimated the performance increase and did not correctly predict higher scores in SSN at 5 dB SNR compared with the WFSE.

As we used vocoded stimuli with NH listeners, it does not allow for direct comparison of speech intelligibility improvement with previous speech separation studies. The HIT-FA rate is a popular measure used in such studies and it has been found to correlate highly with intelligibility scores obtained from NH listeners [17]. The proposed algorithm produced high HIT-FA rates, while maintaining low FA rates, which according to Kim *et al.* are necessary to obtain high levels of speech intelligibility [17]. In the same paper, the authors also showed that a conventional Wiener filter algorithm reaches much lower HIT-FA rates. This was most likely the reason behind the better performance of NNSE compared with the tested WFSE.

The NNSE algorithm provided HIT-FA rates close to those reported in recent speech separation studies [4,5], although a minor decrease in performance was expected given the more challenging conditions in which NNSE was tested. For example, Hu *et al.* obtained HIT-FA rates about 5 and 7% higher compared with NNSE, in BABBLE noise at 5 and 10 dB SNR, respectively. Although previous studies proved that supervised speech separation algorithms are promising strategies for speech enhancement, their implementation poses a major challenge in real-word applications. This is due to several reasons. Firstly, previous studies used the same noise realization for both classifier training and testing [4,5], a situation that is unlikely to occur in practice. May *et al.* have shown that the use of unseen noise realizations leads to a substantial decrease in HIT-FA rates with a Gaussian Mixture Model based system [18], such as the one employed by Hu *et al.* [4]. In the present study, the proposed algorithm was tested with unseen realisations. Even in these more difficult conditions, we found significant speech intelligibility improvements for both tested maskers and SNR levels. The generalization of the proposed algorithm to mismatched noises and speakers between the training and testing stage still needs to be investigated. This question could be addressed in several ways, for instance by enlarging the training dataset [19] or by the integration of a noise-classification system (such as [20,21]) into the proposed algorithm. Secondly, the computational power required by these algorithms scales with their complexity, reflected by the feature extraction stage and by the employed classifier. Previous studies used sub-band based feature sets and classification systems, while we opted for a smaller architecture. NNSE uses one set of features extracted from the broadband signal and one neural network (for each masker type), resulting in reduced algorithm complexity and lower risk of over-fitting the training dataset. Finally, we used the IWF as the estimation target of the network, as opposed to the IBM. For intelligibility to be maintained with the IBM, the optimal LC should change according to the overall mixture SNR (as in [5]). Compared with the IBM, the IWF does not depend on the setting of an LC, it is more robust to estimation errors and it is preferred in terms of perceived quality [9,10,22]. Thus, the IWF was expected to present a more reliable training target for speech enhancement purposes.

To summarise, the results obtained in this study with the proposed algorithm indicate significant improvements in speech intelligibility in noise for NH listeners using CI vocoder simulations. This motivates the investigation of the potential benefit of the proposed NNSE strategy with CI users in future studies.

5. ACKNOWLEDGEMENTS

The authors would like to thank all subjects who voluntarily participated in the 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 by EPSRC grant no. EP/K020501/1.

6. REFERENCES

[1] F. G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, "Cochlear implants: system design, integration, and evaluation.," *IEEE reviews in biomedical engineering*, vol. 1, no. dc. pp. 115–142, 2008.

[2] P. W. Dawson, S. J. Mauger, and A. a Hersbach, "Clinical evaluation of signal-to-noise ratio-based noise reduction in Nucleus® cochlear implant recipients.," *Ear Hear.*, vol. 32, no. 3, pp. 382–90, 2011.

[3] A. Buechner, M. Brendel, H. Saalfeld, L. Litvak, C. Frohne-Buechner, and T. Lenarz, "Results of a pilot study with a signal enhancement algorithm for HiRes 120 cochlear implant users.," *Otol. Neurotol.*, vol. 31, no. 9, pp. 1386–1390, 2010.

[4] Y. Hu and P. C. Loizou, "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users.," *J. Acoust. Soc. Am.*, vol. 127, no. 6, pp. 3689–95, Jun. 2010.

[5] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners.," *J. Acoust. Soc. Am.*, vol. 134, no. 4, pp. 3029–38, Oct. 2013.

[6] D. S. Brungart, P. S. Chang, B. D. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, no. 6, p. 4007, 2006.

[7] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking.," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2336–47, Apr. 2009.

[8] J. Lim and A. Oppenheim, "Enhancement and band width compression of noisy speech," *Proc. IEEE* , vol. 67, no. 12, pp. 1586–1604, 1979.

[9] Y. Wang, A. Narayanan, and D. Wang, "On Training Targets For Supervised Speech Separation," *Taslp*, vol. 22, no. 12, pp. 1849–1858, 2014.

[10] N. Madhu, a. Spriet, S. Jansen, R. Koning, and J. Wouters, "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*, vol. 21, no. 1, pp. 63–72, Jan. 2013.

[11] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AU-17, no. 3, pp. 225–246, 1969.

[12] P. Scalart and J. V Filho, "Speech enhancement based on a priori signal to noise estimation," *Acoust. Speech, Signal Process. 1996. ICASSP-96. Conf. Proceedings., 1996 IEEE Int. Conf.*, vol. 2, pp. 629–632 vol. 2, 1996.

[13] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

[15] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions.," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, 2009.

[16] F. Chen and P. C. Loizou, "Predicting the intelligibility of vocoded and wideband Mandarin Chinese.," *Ear Hear.*, vol. 129, no. 5, pp. 3281–3290, 2011.

[17] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners.," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1486–94, Sep. 2009.

[18] T. May and T. Dau, "Requirements for the evaluation of computational speech segregation systems.," *J. Acoust. Soc. Am.*, vol. 136, no. 6, p. EL398, Dec. 2014.

[19] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, 2013.

[20] T. May and T. Dau, "Environment-aware ideal binary mask estimation using monaural cues," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013.

[21] O. Hazrati, S. O. Sadjadi, and J. H. L. Hansen, "Robust and efficient environment detection for adaptive speech enhancement in cochlear implants," *2014 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 900–904, 2014.

[22] R. Koning, N. Madhu, and J. Wouters, "Ideal Time – Frequency Masking Algorithms Lead to Different Speech Intelligibility and Quality in Normal-Hearing and Cochlear Implant Listeners," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 331–341, 2015.