Development of a Sound Coding Strategy based on a Deep Recurrent Neural Network for Monaural Source Separation in Cochlear Implants

Waldo Nogueira[1], Tom Gajęcki[2], Benjamin Krüger[1], Jordi Janer[2], Andreas Büchner[1]

1 Dept. of Otolaryngology and Hearing4all, Medical University Hannover, 30625, Hannover, Germany
2 Universitat Pompeu Fabra, Music Technology Group, Barcelona, Spain
Email: `nogueiravazquez.waldo@mh-hannover.de`
Web: `https://www.mh-hannover.de/`

Abstract

The aim of this study is to investigate whether a source separation algorithm based on a deep recurrent neural network (DRNN) can provide a speech perception benefit for cochlear implant users when speech signals are mixed with another competing voice.

The DRNN is based on an existing architecture that is used in combination with an extra masking layer for optimization. The approach has been evaluated using the HSM sentence test (male voice) mixed with a competing voice (female voice) for a monaural speech separation task. Two DRNNs with two levels of complexity have been used. The algorithms have been evaluated in 8 normal hearing listeners using a Vocoder and in 3 CI users. Both DRNNs show a large and significant improvement in speech intelligibility using Vocoded speech. Preliminary results in 3 CI users seem to confirm the improvement observed using Vocoded simulations.

1 Introduction

A cochlear implant (CI) is an electronic device that is surgically implanted into the inner ear and can restore the sense of hearing of a profoundly deaf person. CI users need significantly higher signal-to-noise ratios (SNRs) to achieve the same speech intelligibility as normal-hearing listeners [1]. For this reason, speech enhancement techniques have emerged to improve the SNR in noisy acoustic conditions. Although many successful single [2] and multichannel noise reduction algorithms exist [3] and all implants have some sound coding strategies implemented in their processors, noise reduction remains one of the big challenges of the acoustic processing in CIs. All algorithms and techniques have a good performance when the noise is coherent. However, their performance is reduced when the CI user is in a noisy environment with many incoherent noise sources, in reverberant rooms or in the presence of more interfering speech sources [3].

For example, Beamforming algorithms are spatial filters able to enhance speech from a target direction in the presence of interfering speech sources from different directions [4]. Their implementation however requires several microphones. If only one microphone is available, source separation algorithms can be used to solve the issue of non-stationary noises such as speech interference. Source separation algorithms have been applied to separate musical instruments and to separate speech from interferences. Several approaches have been proposed to address the monaural source separation problem. The widely used non-negative matrix factorization (NMF) [5] or, more recently deep recurrent neural networks (DRNNs) [6].

Current results from source separation algorithms are not able to outperform human capabilities. For CI users, given their large limitations in perceiving spectral and temporal characteristics of sound, the potential artifacts introduced by monaural source separation algorithms may not be perceived and therefore, these algorithms are promising to improve speech performance in the presence of a competing voice. So far, no evaluation of source separation algorithms has been performed in CIs users.

In this work, we study a state-of-the-art DRNN in the context of CIs. We propose to evaluate the performance in normal hearing listeners and CI users. For the evaluation in normal hearing (NH) listeners we use a Vocoder to simulate performance of CI users. The Vocoder reduces spectral information into a limited number of channels to reduce speech intelligibility to a similar degree as in CI users [10]. It needs to be emphasized that no attempt is made to match the degree of smearing to specific CI subjects and also that the sound produced by the Vocoder does not correspond to the sound a CI user perceives.

The goal of this manuscript is first to propose a CI sound coding strategy architecture incorporating a DRNN. The second goal is to show whether a DRNN can improve speech intelligibility for CI users. The third goal of this manuscript is to show whether the quality of the separation by the DRNN is affected when the complexity and latency is reduced to satisfy the needs of a CI speech processor.

The organization of the manuscript is as follows: Section 2 presents the methods section giving a summary of the DRNN used and its implementation. Section 3 presents the evaluation of the DRNNs using objective measures, and subjective speech intelligibility tests in normal hearing listeners and CI users. We conclude the manuscript in Section 4.

2 Methods

The source separation algorithm has been incorporated in a CI sound coding strategy as shown (Figure 1).

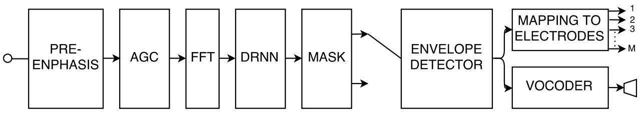

Figure 1: Sound coding strategy incorporating a DRNN.

2.1 Sound Coding Strategy

An audio signal containing a target speech and some other interference is captured by the CI microphone. The signal from the microphone is digitized with a sampling frequency (FS) of around 16 kHz and sent through an adaptive gain control (AGC). Next, a filter bank implemented as a fast Fourier transform (FFT) is applied to the compressed signal. A typical buffer size for the FFT is 128 or 256 samples weighted by an analysis window such as the Hanning window. Next a DRNN is used to separate the target speech signal from the interfering speech signal. The output of the DRNN is the spectrum of the target signal and the interferer signal. Next, a mask is applied so that the sum of both predicted signals is equal to the original mixture. After that, an estimation of the desired envelope is calculated for each spectral band of the target speech signal. The envelopes are obtained by computing the magnitude of the complex FFT bins. Each band is allocated to one electrode after non linear compression (mapping). For each frame of the audio signal M channels with the highest amplitudes are stimulated. Some sound coding strategies perform a selection of the bands for stimulation right after the envelope detector, the so called NofM strategies [7]. Typical values for M and N are 22 and 8 respectively.

In the proposed architecture the DRNN is incorporated right after the FFT. By doing so it is possible to use the same DRNN for CI users having different number of electrodes activated or different processing stages after the FFT such as noise reduction algorithms, multichannel compressors or NofM selection algorithms.

2.2 The Deep Recurrent Neural Network

In this section we summarize the implementation of the DRNN proposed by [6]. The DRNN learns the optimal hidden representations to reconstruct the target spectrum by applying a generated soft mask to the original source mixture. The general architecture is based on Figure 2.

As mentioned before, the incoming sound from the microphone is segmented into frames and transformed into the frequency domain using the FFT by the CI sound processor. Each spectral frame of the spectrum is denoted by $\boldsymbol{X}_{\mathrm{n}}$. At frame n, the training input $\boldsymbol{X}_{\mathrm{n}}$ of the network is the concatenation of spectral features.

The output predictions, $\boldsymbol{Y}_{1_{\mathrm{n}}}$ and $\boldsymbol{Y}_{2_{\mathrm{n}}}$ of the network are the spectra of different sources. In a DRNN, the l^{th} hidden layer, $l > 1$, is calculated based on the current input X_{n} and the hidden activation from the previous time step $h^l(\boldsymbol{X}_{n-1})$,

$$h^l(\boldsymbol{X}_n) = \sigma\left(\boldsymbol{W}^l h^{l-1}(\boldsymbol{X}_n) + \boldsymbol{b}^l + \boldsymbol{W}_{rec}{}^l h^l(\boldsymbol{X}_{n-1})\right),$$

where $\boldsymbol{W}^l$ and $\boldsymbol{W}_{rec}{}^l$ are the weight matrices for the feed forward and the recurrent connections, and $\boldsymbol{b}^l$ is the bias vector.

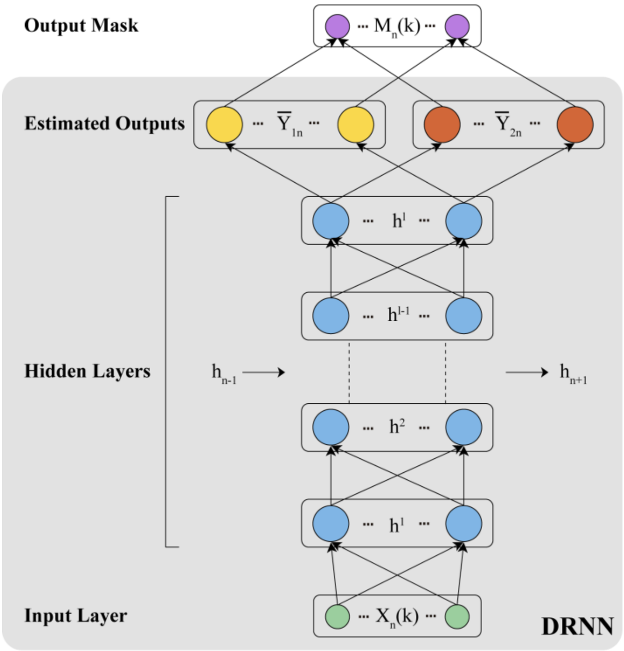

Figure 2: Scheme of a recurrent neural network adapted from [3].

The first hidden layer is computed as $h^1(\boldsymbol{X}_n) = \sigma(\boldsymbol{W}^1\boldsymbol{X}_n + \boldsymbol{b}^1)$. We used rectified linear unit [15]. The output layer is a linear layer and is computed as:

$$\overline{\boldsymbol{Y}}_n = \boldsymbol{W}^l h^{l-1}(\boldsymbol{X}_n) + \boldsymbol{b},$$

where $\overline{\boldsymbol{Y}}_{\mathrm{n}}$ is the concatenation of two predicted sources $\overline{\boldsymbol{Y}}_{1_{\mathrm{n}}}$ and $\overline{\boldsymbol{Y}}_{2_{\mathrm{n}}}$.

Directly training the previously mentioned networks does not have the constraint that the sum of the prediction results is equal to the original mixture. As proposed by [5] a soft time-frequency mask attached to the output layer is used to jointly optimize the network.

The DRNN was trained using a discriminative cost function [5] with a discriminative gamma factor of 0.05, together with the mean squared error (MSE) cost function.

Such discriminative cost decreases the similarity between the prediction and the targets of other sources while the MSE cost increases the similarity between the target and the prediction of the same source. The model is optimized by back-propagating the gradients through time with respect to the training objectives. In our case, 1200 iterations are used to obtain the optimum minima. The limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm [11] is used to train the models from a random initialization.

2.3 Implementation

Two different DRNNs were implemented. The first one (DRNN1) uses one single hidden layer with 16 hidden units. The second one (DRNN2) uses 3 layers with 1000 hidden units, where only the third layer is recurrent. The temporal connection of the recurrence was set to two for both networks. The input features are the magnitude spectrum of the incoming sound. The spectral representation is extracted using a 1024-point short time Fourier transform (STFT) with 50% overlap. In its standard configuration a 32 ms Hamming window with a 50% overlap was used.

The number of units in the input and output layers is given by the dimensionality of the input feature set and the output gains. In its standard configuration the number of input features is set to half the length of the FFT, therefore the number of input units is 511 and the number of output units is 1022.

An Intel Xeon CPU E5-1620@3.5 GHz with 16 GB RAM and a NVIDIA Tesla K40 was used to train the models. The models were implemented in Matlab and the training of each model in its standard configuration took around 7.5 hours. Once the model was trained the whole HSM sentence test [6] mixed with other HSM sentences uttered by the female speaker, was processed to separate the male from the female voice. For each network architecture, the whole HSM sentence test was processed.

3 Results

3.1 Objective evaluation

The proposed approaches were evaluated for monaural speech separation using the HSM sentence test. Eight HSM sentences from a male and a female speaker, respectively, were used for training. The male and female voices were mixed at a speech to speech interference ratio (SSIR) of 0 dB. An additional sentence for the male and for the female were used as the development set. All objective tests presented in the following sections were obtained processing 46 additional sentences for the male and the female speaker not included in the training or the development set.

As proposed by [5], in order to increase the variety of training samples, the male voice signals were circularly shifted in the time domain and mixed with female utterances. In total 18 minutes of male voice and female voice were used to train the DRNN.

The source separation evaluation was measured using the source to interference ratio (SIR), the source to artifacts ratio (SAR), and the source to distortion ratio (SDR), defined in the BSS-EVAL metrics [9]. The larger are the outcomes of these measures, the better the quality of the separation. Figure 3 presents the objective evaluation comparing the DRNN1 with the DRNN2. Although the complexity of the DRNN2 was much higher than the DRNN1 the performance achieved by both networks was similar.

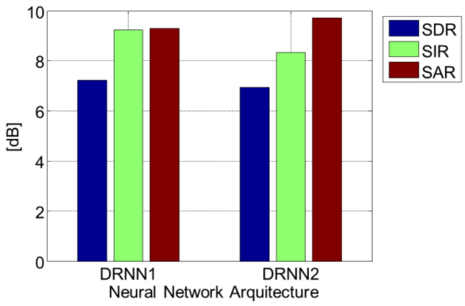

Figure 3: Effect of DRNN architecture on objective measures (SDR, SIR and SAR).

3.2 Subjective Evaluation

Speech intelligibility was measured by means of the HSM sentence test [8]. The standard test is uttered by a male voice. The sentences were mixed with other HSM sentences uttered by a female voice. 2 lists or 3 lists of 20 sentences for each condition (Original, DRNN1 and DRNN2) were presented to NH or CI users respectively. For the subjective evaluation, the output of the DRNN after applying the soft mask was used to separate the spectrogram of the target from the interferers. Next the signals were converted back into the time domain using an inverse FFT. The target signal was then processed by the CI speech processor or by a Vocoder for experimentation in NH listeners.

The speech test was presented using a loudspeaker at a 1 m distance from the study participant. An M-Audio mobile Pre sound card was used for that purpose connected to a Genelec 8240A Loudspeaker. The test was conducted in a sound treated room at a presentation level of 60 dB(A) SPL.

3.2.1 Evaluation in NH listeners

The described algorithms were evaluated in a group of NH listeners using a Vocoder. The Vocoder simulated the typical processing performed by a CI and the spread of excitation that may occur in the electrically stimulated cochlea.

Each token was digitally sampled at 16 kHz. A 128-short-time FFT was computed with a 75% overlap. Next, the FFT bins were grouped into 22 non-overlapping, logarithmically spaced bands. The envelope of each band was computed taking the square root of the total energy in the band. The output of each band was used to modulate a noise band. The noise band was generated similarly synthesized in the frequency domain [10]. The center frequency of the noise band was identical to the center frequency of the corresponding frequency band. The noise band was configured to decay at a rate of 25 dB/octave to simulate the effect of spread of excitation.

8 NH listeners participated in the evaluation. Figure 4 presents the individual and averaged speech performance scores in % of correct words.

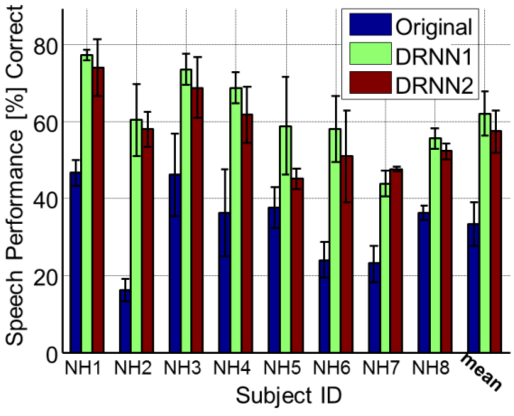

Figure 4: Speech intelligibility scores using the HSM sentence test with a competing female voice using a Vocoder. The SSIR was 0 dB for all participants.

The results show a significant improvement in speech intelligibility for the DRNN1 and DRNN2 with respect to the baseline condition. No significant difference was observed between the DRNN1 and the DRNN2 conditions.

3.2.2 Evaluation in CI users

Three CI users participated in the evaluation of the DRNNs (Table 1). The three study participants were bilateral CI users, only the best CI side was tested.

ID	Age	Duration of Deafness	Cause of Deafness	Implant Experience (in years)	Electrode type
P1	80	13.42	Sudden Hearing Loss	4.5	CI512
P2	67	33.25	Genetic	8	Sonata TI100
P3	35	0	Unknown	5.5	HiRes90k Helix

Table 1: Subject Details

Figure 5 presents the speech intelligibility scores obtained by the 3 CI users.

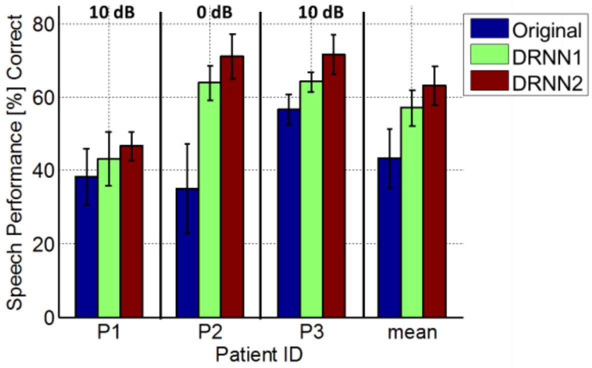

Figure 5: Speech intelligibility scores using the HSM sentence test mixed with HSM sentences uttered by a female voice. The target voice was the male voice. The SSIR was 0 dB or 10 dB as indicated in the labels on top of the bars.

The results show that subjects obtained an improvement of 50% in speech intelligibility using the DRNN with respect to the non-processed condition.

3.3 Optimization for Cochlear Implants

The proposed DRNN shows promising results to be integrated in a CI sound coding strategy as presented in Figure 1. However, the length of the FFT needs to be reduced for this purpose. Given the SIR, SAR and SDR values, we investigated the effect of reducing the length of the FFT on objective performance (Figure 6). Figure 6 shows that reducing the length of the FFT causes a reduction in SDR, SIR and SAR that may impact the benefits observed with a long 1024-FFT.

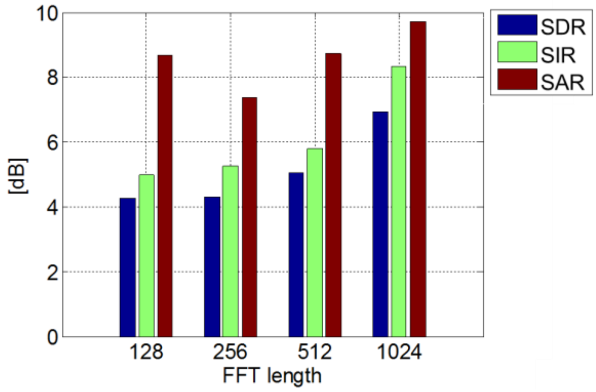

Figure 6: Effect of FFT length on objective measures performance.

4 Conclusion

In this manuscript we propose a CI sound coding strategy that integrates a DRNN to improve speech intelligibility in the presence of a competing voice. First we demonstrate that a DRNN can significantly improve speech intelligibility performance using Vocoder simulations. Preliminary tests in CI users also indicate a speech intelligibility benefit for the new sound coding strategy. Given the objective performance measures we show how to reduce the complexity and latency of the DRNN so that it can be incorporated into a CI speech processor. Additional experiments using DRNNs not trained with the same voices used for testing are necessary to show whether this technique can be generalized for a daily life application.

ACKNOWLEDGEMENTS

The authors would like to thank the subjects who have participated in the experiments. This work was supported by the DFG Cluster of Excellence EXC 1077/1 Hearing4all. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] I. Hochberg, A. Boothroyd, M. Weiss, and S. Hellman, Effects of noise and noise suppression on speech perception by CI users, Ear and hearing, vol. 13, pp. 263–271, 1992.

[2] W. Nogueira, R. Rode, A. Büchner, Spectral contrast enhancement improves speech intelligibility in noise for cochlear implants, J. Acoust. Soc. Am Feb;139(2):728, 2016

[3] W. Nogueira, M. Lopez, T. Rode, S. Doclo, A. Büchner, Individualizing a Monaural Beamformer for Cochlear Implant Users, IEEE International Conference on Acoustics, Speech and Signal Processing, Brisbane, Australia 2015.

[4] A. Buechner, K-H. Dyballa, P. Hehrmann, S. Fredelake, and Th. Lenarz, "Advanced beamformers for cochlear implant users: Acute measurement of speech perception in challenging listening conditions," PLoS ONE, vol. 9, 2014.

[5] D. D. Lee and H. S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature, vol. 401, no. 6755, pp. 788–791, 1999.

[6] P-S. Huang, M. Kim, M. Hasegawa-Johnson, Paris Smaragdis, Deep Learning for Monaural Speech Separation, IEEE International Conference on Acoustics, Speech and Signal Processing, 2014.

[7] W. Nogueira, A. Büchner, T. Lenarz, B. Edler, A psychoacoustic "NofM"-type speech coding strategy for cochlear implants. EURASIP J. Appl. Signal Processing 2005, 3044–3059, 2005.

[8] I. Hochmair-Desoyer, E. Schulz, L. Moser, and M. Schmidt, The HSM sentence test as a tool for evaluating the speech understanding in noise of cochlear implant users., The American Journal of Otology, 18, 1997.

[9] E. Vincent, R. Gribonval, and C. Fevotte, Performance measurement in blind audio source separation, IEEE Trans. Audio, Speech, and Language Processing, 14(4), pp. 1462 –1469, 2006.

[10] L. M. Litvak, A. K. Spahr, A. a. Saoji, G. Y. Fridman, Relationship between perception of spectral ripple and speech recognition in cochlear implant and vocoder listeners. J. Acoust. Soc. Am. 122, 982–91, 2007

[11] D. C. Liu, and J. Nocedal. On the limited memory BFGS method for large scale optimization. Mathematical programming 45.1-3, 503-528, 1989.