Development of a Sound Coding Strategy based on a Deep recurrent Neural Network for Monoaural Source Separation in Cochlear Implants.

Waldo Nogueira[1], Tom Gajecki[1,2], Benjamin Krüger[1], Jordi Janer[2], Andreas Büchner[1]
1. Department of Otolaryngology, Medical University Hannover and Hearing4all, Germany
2. Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain

INTRODUCTION

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.

CI users need significantly higher distortion ratios to achieve the same speech intelligibility as normal-hearing listeners. For this reason, speech enhancement techniques have emerged to improve the SNR in noisy acoustic conditions.

Goals

- Investigate whether a DRNN can improve speech intelligibility in noise for CI users.
- Study 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.

METHOD

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

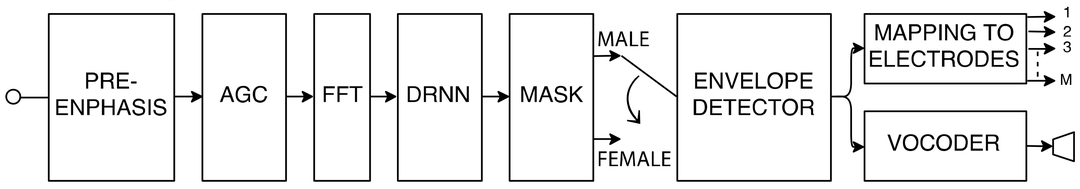

Figure 1: Sound coding strategy incorporating a DRNN.

The proposed approaches were evaluated for monaural speech separation using the HSM sentence test. Two different DRNNs were implemented.

DRNN 1

- 1 Hidden recurrent layer
- 16 Hidden units.

DRNN 2

- 3 Hidden layers
- 1000 Hidden units
- Only one recurrent layer (the third one).

The temporal connection of the recurrence was set to two for both networks.

RESULTS

Objective Evaluation.

Although the complexity of the DRNN2 was much higher than the DRNN1 the performance of both networks was similar. Figure 3 presents the evaluation comparing the two DRNNs.

Evaluation in NH listeners

The described algorithms were evaluated in a group of NH listeners using the Vocoder. Figure 4 presents the individual and averaged speech performance scores in % of correct words.

Evaluation in CI users

Three CI users participated in the evaluation of the DRNNs. The three study participants were bilateral CI users, only the best CI side was tested. Figure 5 presents the speech intelligibility scores obtained by the 3 CI users.

Optimization for CI

The length of the FFT needs to be reduced for this purpose. 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.

CONCLUSION

Preliminary tests in CI users 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.

MATERIAL

The Deep Recurrent Network

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.

Training

An Intel Xeon CPU E5-1620@3.5 GHz with 16 GB RAM and a NVIDIA Tesla K40 was used to train the models.

Evaluation

An M-Audio mobile Pre sound card connected to a Genelec 8240A Loudspeaker.

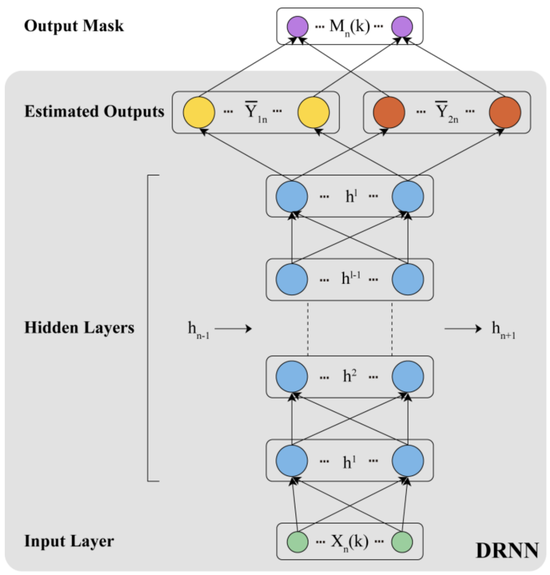

Figure 2: Scheme of a recurrent neural network.

- Magnitude Spectrum
- 1024-point short time Fourier transform (STFT)
- 32 ms Hamming window with 50% overlap

The models were implemented in Matlab and the training of each model in its standard configuration took around 7.5 hours using a discriminative cost function together with the mean squared error. The model is optimized by back-propagating the gradients through time with respect to the training objectives. In our case, 1200 iterations and where used in order to obtain the optimum minima. The limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm is used to train the models from a random initialization. Once the model was trained the whole HSM sentence test 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.

Evaluation

The speech test was presented using a loudspeaker at a 1 m distance from the study participant conducted in a sound treated room at a presentation level of 60 dB(A) SPL. 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.

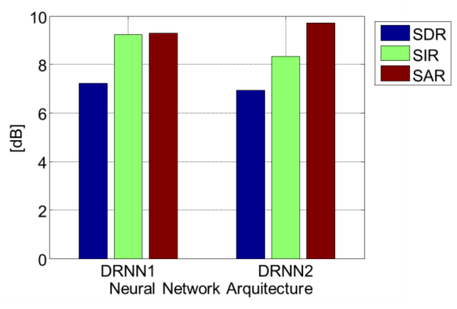

Figure 3: Effect of DRNN architecture on objective measures (SDR, SIR and SAR).

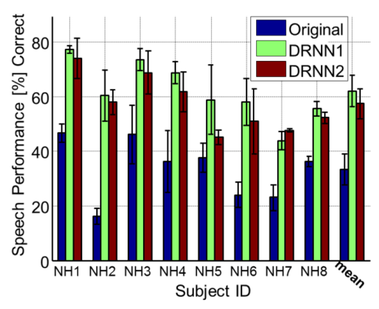

Figure 4: Speech intelligibility scores using the HSM sentence test with a competing female voice using a Vocoder. The SSIR was 0 dB for all participants.

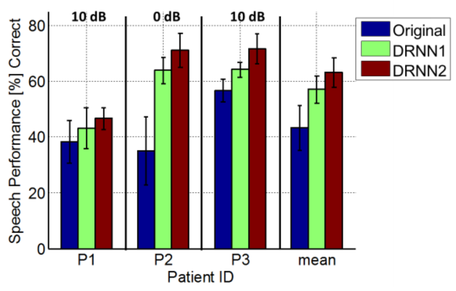

Figure 5: Speech intelligibility scores using the HSM sentence test mixed with HSM sentences uttered by a female voice. The target voice was the male voice. The SSIR was 0 dB or 10 dB as indicated in the labels on top of the bars.

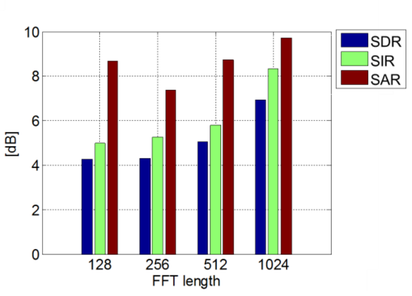

Figure 6: Effect of FFT length on objective measures performance.

REFERENCES

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.
II. 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.
III. 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.
IV. 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.