Machine learning-based sound coding for cochlear implants

Bernhard U. Seeber & Esteban Bullón Tarrasó

Audio Information Processing, Technical University of Munich, email: {seeber, esteban.bullon-tarraso}@tum.de

Abstract

Sound coding strategies for cochlear implants (CIs) translate the incoming sound signal into parameters for the electrical pulse pattern delivered by the implant. Without specific noise reduction, the common envelope coding in CIs, like in the continuous interleaved sampling strategy (CIS), leads to problems with sound localization and listening in situations with noise and reverberation. Hence, noise reduction and de-reverberation stages clean the sound signal prior to pulse processing, together achieving good performance in many listening situations. Statistical approaches using deep neural networks (DNN) achieve impressive results for de-reverberation, noise reduction and segregation of sound sources from a background of other sources. These networks, like Conv-TasNet, contain stages for filtering and amplitude extraction similar to the processing in CIs. We investigated if a DNN could be trained to replace the complete sound coding strategy of a CI. We compare the amplitude stimulation sequences for each electrode computed by a new stochastic approach and compare them to those of the classic CIS strategy. Results show that a DNN is capable of replacing existing CI coding strategies like CIS, while having the potential to include preprocessing stages for source segregation at no extra computational cost.

Introduction

Cochlear implants (CI) are the most widespread sensorineural prosthesis and are the main solution for patients who suffer from profound deafness [1-3]. Continuous Interleaved Sampling (CIS) is an envelope-based coding strategy which is commonly used as CI processing algorithm. It maps the amplitude envelope of the sound filtered in frequency bands to individual envelopes, achieving high levels of speech recognition [4].

Machine learning has already been used in the field of cochlear implants and has demonstrated its potential in signal processing. One approach is using neural networks (NN) to optimize the envelope-to-current transfer function, i.e. to individually adjust threshold (THR) values and compression characteristics [5]. The THR and maximum comfortable level (MLC) can also be calculated by a neural network approach [6]. Another approach uses NN for sound source segregation [7] and for increasing the performance of CI users in noisy environments. For instance, several studies achieved automatic segregation of the speech signals from noise by using neural networks. Besides, a different NN greatly improved the signal-to-noise ratio from compared to the original CI processing strategies [8].

In this paper, we propose replacing the complete CIS strategy algorithm by a DNN, Conv-TasNet [9]. The DNN receives a single-channel input audio signal and the target is the 16-channel output of the CIS algorithm to be delivered by the CI. With this idea, our intention is to go one step further on the application of deep learning on cochlear implants, which has already shown great results, based on the speech segregation and noise suppression capabilities of Conv-TasNet [10][11].

Results show that the DNN is able to replicate the output of CIS. It has been capable of bandpass filtering the input signal, introducing a logarithmic compression function and mapping the signal to the correct current amplitudes within a simulated listener's dynamic range. This opens the door to new possibilities. For instance, the NN could be trained to support binaural hearing enhancement by the use of interaural time differences (ITDs), which it is impossible for the traditional CIS algorithm [12], or include environmental sounds and noise segregation.

Methods

Training material

The LibriSpeech corpus was used as the training, validation and test dataset. LibriSpeech contains up to 1000 hours of read English speech, sampled at 16 kHz [13]. In this study, 15.000 audio files of 3 seconds each extracted from the corpus were included in the training, which corresponds to 12.5 hours of read speech. Among all the files included, there are more than 251 different speakers (125 female speakers).

Groundtruth generation

The neural network was trained with the simulated output of the CIS strategy from the mentioned audio files as groundtruth of the model. First, the audio signal is divided into 16 frequency channels using a series of bandpass filters with logarithmically spaced cut-off frequencies from 200 to 8000 Hz. The bandpass filters consist of 6th-order type-II Chebyshev filters with a stopband attenuation of 30 dB down. The bandpass-filtered signals are rectified by full-wave rectification and low-pass filtered to extract the envelope. A type-II Chebyshev filter of 8th-order and 30 dB stopband attenuation are used, with a cut-off frequency of 200 Hz. The last stage is the mapping from the audio dynamic range to the CI output dynamic range. A logarithmic function, as equation , was used to compress an input dynamic range of 45 dB to 10 dB.

$$y_{i,c} = A_c \cdot \log\left(\frac{x_{i,c}}{x_{rms,c}}\right) + B_c \qquad (1)$$

where $x_{i,c}$ is the envelope at instant i of channel c, $y_{i,c}$ is the compressed output of the CIS algorithm at instant i of channel c, and A_c and B_c are the individual parameters of the equation for each of the channels to obtain the mentioned dynamic range, measured in µA. Arbitrary values of 100 µA and 316.23 µA for the threshold and maximum comfortable levels, respectively, were used.

Neural Network

The model is similar to Conv-TasNet [9], with a few adjustments on the Neural Network (NN) due to the particularities of this study:

1. A Rectified Linear Unit (ReLU) is inserted at the end of the NN to force only-positive values, as it happens on CI delivered currents.
2. The output of the separation block considers only the output of the last 1-D Conv Block instead of the sum of all the blocks' output.
3. The sigmoid function at the end of the separation block is removed because the restriction of [0, 1] masks no longer exists (i.e., the sum of all channels is not the original input audio).

Conv-TasNet operated in causal mode, as it is expected in a CI. Besides, it works sample wise, with no temporal downsampling of the input audios. Hence error values are computed on a sample-by-sample basis; there is no downsampling to the lower per-electrode stimulation rate.

Training

The dataset was split into 60% (9.000 sample files, 7.5h of audio) for the training partition and 20% (3.000 sample file, 2.5 h of audio) each for the validation and test partitions. The model was trained using back propagation and ADAM optimizer, with a learning rate starting at 0.005 and being reduced every three consecutive epochs with no decrease of the validation loss. The loss function was the Mean Squared Error (MSE). The training stopped once the validation loss was not reduced for six consecutive epochs.

Results

The training finished with an MSE of 1.81 µA^2 for the validation partition. For a channel wise analysis, the RMS error was computed as equation (2) for each channel independently, only considering values above T-level.

$$RMS\ error = \frac{1}{N}\sum_{n=0}^{N}\sqrt{\frac{1}{T}\sum_{t=0}^{T}\left(x_{t,n} - \tilde{x}_{t,n}\right)^2} \qquad (2)$$

The error is lower in the lower frequency channels (i.e., the lower frequencies) and increases for higher channels. This is expected, as there is less noise-like behaviour at lower frequencies and better predictability across time. Overall, the error of every channel is very low, with a minimum mean error of 0.41 µA for channel 2 and a maximum mean error of 1.34 µA for channel 16. The mean error across all channels is 0.98 µA.

Nevertheless, as the causality of the model may affect the error of the output compared to the groundtruth, a preliminary analysis of the mean absolute error across time for the whole test partition was performed. Results are shown in Figure 1.

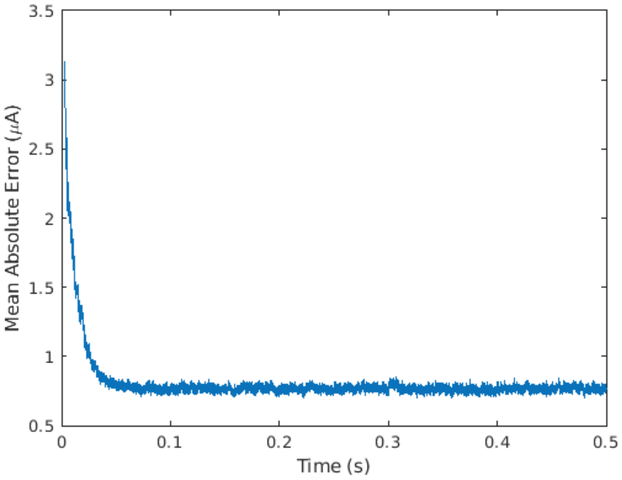

Figure 1: Mean absolute error in µA for each timepoint for all the outputs of the test partition.

Due to the causality of the model, a higher error was expected for the initial timepoints of each audio file. Specifically, the error should be higher until 120 ms, which corresponds to the kernel size of the last 1-D Conv Block, as it has a dilation of 2^7 with sampling frequency of 16 kHz. However, some channels reach the steady state error at an earlier timepoint. The reason why this happens is the skip connection (i.e., the bypass summed at the end), which enables the NN to set the parameters to 0 in some channels and thus ignore long-time dependencies. Even so, as the timepoint on when the steady state error is reached varies across channels, only timepoints above 150 ms were considered in the computation of the RMS error.

Discussion

Results show that the neural network is capable of performing bandpass filtering for the 16 channels, logarithmic compression and envelope amplitude prediction for every channel of the CI with a very small error. The mean error across all channels is 0.98 µA, which is much smaller than the current quantization of CI stimulators (5-10 µA). Those preliminary results state that the model has some headroom for reducing the algorithm complexity or include additional processing states for enhancing CI performance in noise.

Conclusion

This is the first attempt known by us for replacing the whole CI processing algorithm by a neural network. The results demonstrate that deep learning has the capability of processing audio as the CIS strategy. Besides, the excellent performance shown suggests the possibility of reducing the algorithm complexity or extending the current training to include environmental sounds or noise reduction.

Acknowledgments

Work supported by the BMBF Bernstein Center for Computational Neuroscience, 01GQ1004B.

References

[1] P. C. Loizou, "Introduction to cochlear implants," ed, 1999.

[2] F. G. Zeng, "Trends in Cochlear Implants," *Trends in Amplification,* 2004, doi: 10.1177/108471380400800102.

[3] B. S. Wilson and M. F. Dorman, "Cochlear implants: A remarkable past and a brilliant future," *Hearing Research,* vol. 242, no. 1-2, pp. 3-21, 2008, doi: 10.1016/j.heares.2008.06.005.

[4] B. S. Wilson, C. C. Finley, D. T. Lawson, R. D. Wolford, D. K. Eddington, and W. M. Rabinowitz, "Better speech recognition with cochlear implants," *Nature,* 1991, doi: 10.1038/352236a0.

[5] C. H. Chang, G. T. Anderson, and P. C. Loizou, "A neural network model for optimizing vowel recognition by cochlear implant listeners," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 9, no. 1, pp. 42-48, 2001, doi: 10.1109/7333.918275.

[6] J. Torresen, A. H. Iversen, and R. Greisiger, "Data from past patients used to streamline adjustment of levels for cochlear implant for new patients," 2017, doi: 10.1109/SSCI.2016.7850063.

[7] T. Goehring, F. Bolner, J. J. M. Monaghan, B. van Dijk, A. Zarowski, and S. Bleeck, "Speech enhancement based on neural networks improves speech intelligibility in noise for cochlear implant users," *Hearing Research,* 2017, doi: 10.1016/j.heares.2016.11.012.

[8] F. Hajiaghababa, H. R. Marateb, and S. Kermani, "The design and validation of a hybrid digital-signal-processing plug-in for traditional cochlear implant speech processors," *Computer Methods and Programs in Biomedicine,* 2018, doi: 10.1016/j.cmpb.2018.03.003.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE/ACM Transactions on Audio Speech and Language Processing,* vol. 27, no. 8, pp. 1256-1266, 2019, doi: 10.1109/TASLP.2019.2915167.

[10] H. Li, K. Chen, and B. U. Seeber, "ConvTasNet-based anomalous noise separation for intelligent noise monitoring," *INTER-NOISE and NOISE-CON Congress and Conference Proceedings,* vol. 263, no. 4, pp. 2044-2051, // 2021, doi: 10.3397/IN-2021-2035.

[11] H. Li, K. Chen, and B. U. Seeber, "Auditory filterbanks benefit universal sound source separation," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings,* pp. 181-185, 2021.

[12] J. J. M. Monaghan and B. U. Seeber, "A method to enhance the use of interaural time differences for cochlear implants in reverberant environments," *The Journal of the Acoustical Society of America,* 2016, doi: 10.1121/1.4960572.

[13] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015, vol. 2015-Augus, pp. 5206-5210, doi: 10.1109/ICASSP.2015.7178964. [Online]. Available: http://www.gutenberg.org