Evaluation of the benefit of neural network based speech separation algorithms with hearing impaired listeners

Gaurav Naithani[1] *, Tom Barker*[1]*, Giambattista Parascandalo*[1†]
Lars Bramsløw[2]*, Niels Henrik Pontoppidan*[2]*, and Tuomas Virtanen*[1]

[1]Tampere University of Technology, Finland
[2]Eriksholm Research Centre, Oticon A/S, Denmark

[1]{gaurav.naithani, thomas.barker, giambattista.parascandolo, tuomas.virtanen}@tut.fi
[2]{labw, npon}@eriksholm.com

Abstract

Source separation is a useful technology for improving the benefit from hearing aids. However, most of the existing approaches to evaluating source separation rely on computational methods, and do not consider the effect of the algorithm on the end user. We seek to address this mismatch by quantifying the benefit of two state-of-the-art deep neural network (DNN) based source separation techniques, in terms of actual speech intelligibility benefits evaluated via subjective listening tests with 15 hearing impaired (HI) listeners, as well as more established computational metrics by which most source separation algorithms are currently compared. We present here our proposed source separation approach which is a novel application of the 'Convolutional Recurrent Neural Network' (CRNN) deep learning architecture, and compare it with feedforward deep neural network (FDNN) approach. We evaluate these approaches on two talker mixtures from Danish hearing in noise test (HINT) database. We are particularly interested in speech separation in this work as the hearing-impaired listeners have problems understanding speech in the presence of one or more competing voices.

Index Terms: source separation, deep neural networks, low latency, hearing aids

1. Introduction

Source separation is an important technology for improving hearing aid performance, and recently, large advances in this domain have been achieved using a range of techniques using 'deep neural networks (DNN)' – whereby mapping of an input to a target output is realised through learning complicated non-linear relationships which are captured within the network parameters. These approaches achieve state-of-the-art performance even at very low latency, which is critical for hearing aids [1]. It has been postulated (e.g., in [2]) that delays larger than 10 ms are objectionable to hearing impaired (HI) listeners. The algorithmic delay of the DNN based approach used in our work is 8 ms. This low-latency performance is therefore one of the critical design features when considering source separation for hearing aids.

Alongside low-latency performance, another primary goal of a hearing-aid algorithm is to improve speech intelligibility, yet most of the current evaluation methods do not address this need with respect to hearing-impaired listeners. Typically, source separation algorithms, and the literature which reports them, focuses primarily on the performance of the algorithms in terms of separated source energy (e.g., source to distortion ratio (SDR) [3]), predicted perceptual quality (PEASS [4]), or predicted intelligibility. The existing predicted intelligibility metrics such as short term objective intelligibility (STOI [5]) and extended short term intelligibility (ESTOI [6]) are based on models of normal hearing and tested on normal hearing listeners, so they may not be accurate predictors of algorithm performance for use in hearing aids.

Overall, current trends for developing and evaluating source separation as a general technology do not adequately consider the needs of its use specifically for hearing aids. We therefore seek to address this in both development of low-latency source separation and evaluation strategy and present our findings to date here.

2. DNN for source separation

We use the time-frequency masking paradigm of source separation whereby a DNN is used to predict time-frequency mask corresponding to the target speaker. The input features are short-time Fourier transform (STFT) coefficients and output is soft ratio mask defined as the ratio of magnitude spectrum of the target speaker and sum of magnitude spectra of constituent sources in the acoustic mixture (e.g., in [7, 8]). The predicted time-frequency mask is multiplied with mixture spectrum to yield the target speaker spectrum.

We investigate convolutional recurrent neural network (CRNN) for source separation, originally proposed in [7]. The motivation of using this architecture is to combine the feature extraction property of convolutional layers from the input, i.e., time-frequency representation of the acoustic mixture in our case, and the ability of recurrent layers (with long short term memory (LSTM) units [9]) to model long term temporal dependencies. We compare this architecture to a feedforward DNN architecture similar to the one used in [10]. Table 1 shows the hyperparameters used for the two architectures. Note that in case of FDNN, frames spanning previous temporal context of 32 ms is fed to the input for estimation of the current frame, as was done in [10]. For more details on hyperparameter selection for the two architectures, please refer [7]. Output neurons for both topologies use sigmoid activations while hidden units for CRNN are rectified linear units and FDNN are sigmoid units. Max pooling is used after each convolutional layer in CRNN but only along frequency axis. Dropout regularization of 0.4 is used. For training DNNs Keras library [11] is used.

3. Evaluation

The dataset used for training and evaluation of neural networks is an extended version of the Danish hearing in noise test (HINT) dataset developed by [12]. The extended version consists of three male and three female speakers, each of them

* The authors wish to thank CSC-IT Centre of Science Ltd., Finland, for providing computational resources used in experiments reported in this paper.

† The author is currently with Max Planck Institute for Intelligent Systems.

Table 1: *Hyperparameters used for the FDNN and CRNN. The pooling scheme represents max pooling operation along time and frequency axes.*

FDNN			CRNN						
hidden layers	hidden neurons	previous context	conv. layers	recurr. layers	recurr. neurons	conv. filters	pooling scheme	sequence length	conv. kernel size
4	1024	32 ms	3	1	256	256	*1 by 2*	512 ms	3×3

recorded speaking 13 lists consisting of 5 word natural sentences [13]. We use four lists for training and one list for validation. The remaining eight lists are used for testing. The test mixtures are prepared by summing the signals corresponding to the two talkers. The evaluation of the methods is based upon: 1) Computational metrics of separation, i.e., source to distortion ratio (SDR), and extended short term objective intelligibility (ESTOI), the latter being better suited to our task as interferer in our case (i.e., for two talker mixtures) is non stationary; and 2) Word recognition tests with hearing impaired listeners.

For subjective listening tests, a target-masker (TM) set up is used where one of the constituent speaker serves as the target signal. A cue is provided before the playback to indicate which of the speaker sentence the listener must reproduce. The listening test scores are percentage of correct word scores reproduced by the listener, transformed according to [14] to remove floor and ceiling effects. The study involves 15 hearing-impaired listeners with moderate to severe sloping hearing losses. In addition to the two DNN test conditions, we have two more test conditions: one where unprocessed mixture is presented (referred as *Sum*) and the other where the ground truth source is presented (referred as *Separate*). A comparison between these four test conditions is made.

4. Results and conclusions

Table 2 reports SDR and ESTOI values corresponding to FDNN and CRNN, for three speaker pairs: M1 F1, M1 M2, and F1 F2. CRNNs here showed a slightly better average ESTOI scores than FDNN. The subjective listening test, as depicted in Figure 1, showed a significant benefit of 35 % points with the DNN methods in comparison to the *Sum* condition. The difference between the two DNN modes was not found statistically significant albeit a slightly higher mean accuracy was observed for CRNN as compared to FDNN. It is interesting to observe that ESTOI metric showed similar pattern but the difference in performance between the two DNN architecture is not large enough to infer if the ESTOI metric is a good predictor of intelligibility performance for HI listeners.

The obtained results in this study show that DNN based algorithms have significant potential for improving speech intelligibility for HI listeners in tasks where a speech signal of interest is to be attended to in the presence of a masker speech signal. A more exhaustive description of listening test results will be reported in [15].

Table 2: *Performance metrics for the two DNN architectures.*

Speaker pair	FDNN		CRNN	
	SDR	ESTOI	SDR	ESTOI
M1 F1	7.42	0.77	7.44	0.79
M1 M2	5.96	0.76	6.06	0.78
F1 F2	5.40	0.71	5.56	0.72

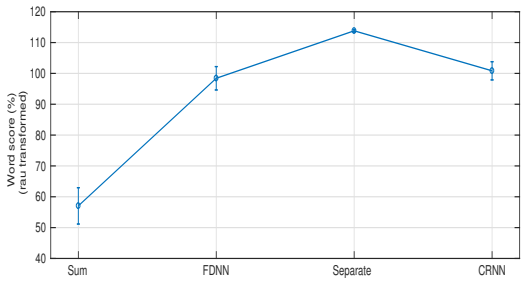

Figure 1: *Word recognition rates for the two DNN architectures for TM task. The vertical bars denote 0.95 confidence intervals.*

5. References

[1] L. Bramsløw, "Preferred signal path delay and high-pass cut-off in open fittings," *International journal of audiology*, vol. 49, no. 9, pp. 634–644, 2010.

[2] J. Agnew and J. M. Thornton, "Just noticeable and objectionable group delays in digital hearing aids," *Journal of the American Academy of Audiology*, vol. 11, no. 6, pp. 330–336, 2000.

[3] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[4] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, 2011.

[5] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.

[6] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2009–2022, 2016.

[7] G. Naithani, T. Barker, G. Parascandolo, L. Bramsløw, N. H. Pontoppidan, and T. Virtanen, "Low latency sound source separation using convolutional recurrent neural networks," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2017, (in press).

[8] P. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2014, pp. 1562–1566.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] G. Naithani, G. Parascandolo, T. Barker, N. H. Pontoppidan, and T. Virtanen, "Low-latency sound source separation using deep neural networks," in *IEEE Global Conference on Signal and Information Processing*, 2016.

[11] F. Chollet, "Keras." GitHub, 2016, available at *https://github.com/fchollet/keras*.

[12] J. B. Nielsen and T. Dau, "The Danish hearing in noise test." *International journal of audiology*, vol. 50, no. 3, pp. 202–8, 2011.

[13] L. Bramsløw, M. Vatti, R. K. Hietkamp, and N. H. Pontoppidan, "A new competing voices test paradigm to test spatial effects and algorithms in hearing aids," in *International Hearing Aid Research Conference (IHCON)*, 2016, p. 1.

[14] G. Studebaker, "A "rationalized" arcsine transform," *Journal of Speech and Hearing Research*, vol. 28, no. September, pp. 455–462, 1985.

[15] L. Bramsløw, G. Naithani, T. Barker, A. Hafez, N. H. Pontoppidan, and T. Virtanen, "Hearing impaired listeners benefit from deep neural network source separation in competing voice scenarios," manuscript in preparation for the Journal of The Acoustical Society of America.