LOW-LATENCY SOUND SOURCE SEPARATION USING DEEP NEURAL NETWORKS

Gaurav Naithani[1], Giambattista Parascandolo[1], Tom Barker[1]
Niels Henrik Pontoppidan[2], and Tuomas Virtanen[1]

[1]Tampere University of Technology, Department of Signal Processing, Tampere, Finland
Email:{gaurav.naithani, giambattista.parascandolo, thomas.barker, tuomas.virtanen}@tut.fi

[2]Eriksholm Research Centre, Oticon A/S, Snekkersten, Denmark
Email:{npon}@eriksholm.com

ABSTRACT

Sound source separation at low-latency requires that each incoming frame of audio data be processed at very low delay, and outputted as soon as possible. For practical purposes involving human listeners, a 20 ms algorithmic delay is the uppermost limit which is comfortable to the listener. In this paper, we propose a low-latency (algorithmic delay $\leq$ 20 ms) deep neural network (DNN) based source separation method. The proposed method takes advantage of an extended past context, outputting soft time-frequency masking filters which are then applied to incoming audio frames to give better separation performance as compared to NMF baseline. Acoustic mixtures from five pairs of speakers from CMU Arctic database [1] were used for the experiments. At least 1 dB average improvement in source to distortion ratios (SDR) was observed in our DNN-based system over a low-latency NMF baseline for different processing and analysis frame lengths. The effect of incorporating previous temporal context into DNN inputs yielded significant improvements in SDR for short processing frame lengths.

Index Terms— Source separation, Deep neural networks, Low-latency

1. INTRODUCTION

Sound source separation aims to recover the individual sounds within a mixture composed of sounds originating from several sources. Some of the most common uses for such a technology are in speech recognition [2], music transcription [3], speech de-noising, [4] and hearing aid applications [5]. Although all of these applications can employ source separation in online processing, hearing aid use is perhaps unique in the strictness of the constraint for very low processing latency, since significant listener discomfort can arise as audio delays exceed 20 ms, and it has even been shown that delays as low as 3 ms are detectable [6]. With this application considered, there is a strong motivation for developing source separation approaches which are able to improve the quality of sound for very low frame lengths.

There are two popular approaches to source separation: compositional model based approaches [7], such as non-negative matrix factorisation (NMF) or the somewhat equivalent probabilistic latent component analysis (PCLA) and deep neural network (DNN) based methods. The compositional methods decompose complex acoustic mixtures into a linear mixture of simpler sub-units or components, based on inherent structure. Deep neural networks, on the other hand, are essentially non-linear models capable of learning complex non linear input-output mappings, with the relationship between the two being embedded in the weights stored within the hidden layers. DNN-based techniques are becoming widespread in their application to source separation problems and have been found to perform better than compositional model based approaches, e.g., in [8], [9].

For low-latency source separation, a supervised, dictionary-based approach was proposed in [10], where short mask frames were generated based on the factorisation of longer past context data to predict weights for a separation filter in the difficult scenario of single-channel speech separation. A similar approach can be used for the formation of training and input vectors in a DNN-based separation, which provides greater opportunities for non-linearities present in data to be modelled.

This work specifically addresses the use of DNN-based separation in scenarios where low processing latency is important, e.g., in hearing-aid applications. Here we use a DNN that takes spectral feature vector inputs derived from the the acoustic mixture signals as inputs and predicts time-frequency mask filters corresponding to the constituent sources. We propose that including an extended previous temporal context in DNN input, which leads to improvement in separation performance for very short processing frame lengths and latencies. We also study the effect of the duration of this incorporated temporal context on separation performance and compare the results with a NMF baseline.

The paper is organized as follows: Section 2 describes the proposed method. It is followed by Section 3 which describes

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

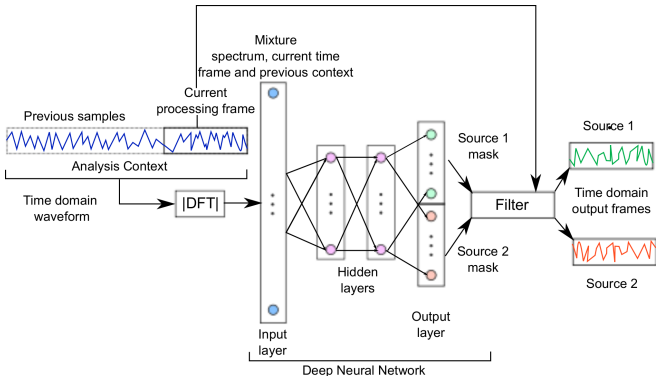

Fig. 1. A schematic illustration of the proposed source separation approach. Feature vectors are formed from greater temporal context than only the current frame for predicting mask filters

the acoustic material used, metrics used for evaluation, and experimental settings and results. Section 4 concludes the paper with a discussion and some insights about the work.

2. PROPOSED NEURAL NETWORK BASED SOURCE SEPARATION

In a general spectrally-based source separation approach with DNNs, spectral features derived from an acoustic mixture of data are used as input vectors to the DNN. Source-separation time-frequency masks are then predicted at the output. These mask filters are applied to the mixture spectrum to obtain individual source spectra for reconstruction of separated source estimates. In our approach, we divide an incoming time-domain signal into blocks for processing. To ensure low latency, processing is performed on short blocks referred here as the processing frame. The latency is determined by the length of this frame, since all samples must be buffered before a discrete Fourier transform (DFT) can be applied to obtain the spectral representation. We propose that a larger previous temporal context can be utilized to generate network inputs corresponding to the current processing frame. This extended temporal context is referred here as the analysis frame. Hence, spectral features derived from analysis frame are then fed to DNN input predict the source separation masks for the processing frame. This process is shown in Figure 1.

2.1. Input features

Spectral features corresponding to the current analysis frame are generated via a short-time Fourier transform (STFT). We use a window length equal to the processing frame length and 50 % overlap in this work. As the analysis frame is longer than processing frame, this produces a set of feature vectors which are then concatenated to give a longer analysis feature vector for each processing frame. We will now elaborate upon the process of generating targets for neural network training, i.e., the time-frequency masks corresponding to the constituent sources.

2.2. Mask outputs

The proposed supervised speech separation approach aims to estimate a suitable time-frequency mask, which when applied, can improve separation and intelligibility of speech signals in the mixture, similarly to in [11]. The mask used in our approach is a soft time-frequency mask and is defined as,

$$M(t,f) = \frac{|S_1(t,f)|}{|S_1(t,f)| + |S_2(t,f)|} \tag{1}$$

where t is a an index for a particular processing frame and f is the bin index from the discrete Fourier transform (DFT). S_1 and S_2 are STFT feature vectors of corresponding constituent speech signals. Mask values are bounded between $[0, 1]$ ensuring numerical stability and forming a suitable choice as an output target for training of neural networks through back-propagation.

For each processing frame, target outputs for training DNNs are obtained in accordance with Equation 1. While training, the DNN network weights are tuned by presenting it with features corresponding to analysis frames and targets corresponding to processing frames. The aim is to capture the relevant features from the training data to produce the correct mask output estimates on unseen examples.

A single frame of mask output, and hence the minimal latency of the processing in a spectral constant-frame-based processing strategy is limited by the length of the DFT used. In order to keep the algorithmic latency of the system low, the processing frames and applied masks need to be kept accordingly short, since the DFT can not be computed before all samples required have been buffered.

2.3. Source reconstruction

The complex STFT spectra of separated sources, S_{est1} and S_{est2} from mixture $Y(t,f)$ are obtained using the estimated mask, $M(t,f)$ in the following manner,

$$\begin{aligned} S_{est1} &= M(t,f) * Y(t,f) \quad \text{and} \\ S_{est2} &= (1 - M(t,f)) * Y(t,f) \end{aligned} \tag{2}$$

where $*$ denotes element-wise multiplication. Time-domain source estimates from these complex spectra are reconstructed on-line by applying inverse discrete Fourier transform (IDFT) and overlap-add processing. Note that phase of the mixture spectra is being used for source reconstruction.

Equation 2 involves multiplication in frequency domain, and hence it should be ensured that the applied masks incorporate zero padding to avoid circular convolution. Hence, prior to applying STFT during feature extraction, time domain signals are zero-padded.

Table 1. Comparison of NMF and DNN separation metrics for 5 and 10 ms processing frame lengths.

Analysis frame length	Metric	NMF 5 ms	DNN 5 ms	NMF 10 ms	DNN 10 ms
5 ms	SIR	5.4	6.9	-	-
	SAR	7.7	9.5	-	-
	SDR	2.7	4.5	-	-
10 ms	SIR	6.7	8.3	6.6	7.9
	SAR	8.0	9.4	8.1	9.6
	SDR	3.5	5.3	3.5	5.1
20 ms	SIR	7.2	8.4	7.5	8.3
	SAR	8.0	9.5	8.2	9.4
	SDR	3.9	5.5	4.2	5.3
40 ms	SIR	-	-	7.6	8.4
	SAR	-	-	8.4	9.4
	SDR	-	-	4.3	5.4

3. EVALUATION

This section describes the metrics used in evaluation, dataset used, experimental settings and finally the results obtained. For the baseline, we used NMF with 10000 basis atoms with generalized KL-divergence metric. Larger dictionaries have been shown to give better separation performance owing to their ability to better model the sources present in the mixture, e.g. in [12], [10]. The chosen NMF configuration is the best performing NMF as reported in [10], and thus serves as a good baseline for the proposed DNN-based system, as is at the limit of achievable performance with a basic dictionary-based NMF implementation.

3.1. Acoustic material

The CMU Arctic dataset [1] was used to generate acoustic mixtures for evaluating the DNN-based speech separation approach. Five pairs of speakers consisting of three male and two female speakers were chosen. The speakers were: *US-awb, US-clb, US-jmk, US-ksp*, and *US-slt*. In total, there were two male-male, two male-female, and one female-female mixture sets. For generating training data for each speaker, 32 utterances were chosen at random from the database's utterance set A. A training set of 1024 mixtures was generated for each speaker pair by summing all possible permutations. Test data was formed of utterances from CMU Arctic set B to ensure that training/validation and test sets were disjoint. It consisted of 10 utterances for each speaker and all possible permutations, i.e., 100 test mixtures were generated for each pair. In instances where one of the utterances being added was shorter than the other, the shorter utterance was zero padded. All utterances had a sampling rate of 16 kHz. The same set of acoustic mixtures which were used for training DNNs were also used for generating dictionaries for NMF baseline.

3.2. Metrics

The separation performance of the proposed approach was evaluated using BSS-EVAL toolkit [13]. It consists of three measures: Source to Interference Ratio (SIR), Source to Artifacts Ratio (SAR), and Source to Distortion ratio (SDR). SIR and SAR denote the interference and artifacts suppression in the separated speech signals, whilst SDR is a measure of the overall separation performance. While evaluating, the original time domain signals along with the corresponding separated signals were used to compute these metrics.

3.3. DNN architecture and training

For training DNNs, Keras neural networks library [14] was used. Separate DNNs were trained for all five speaker pairs. To select a reasonably performing model, different values of model hyperparameters, i.e., number of hidden layers and number of neurons were experimented with. The DNN architecture consisting of three hidden layers, each having 250 hidden neurons was finally used on the basis of its performance on validation data. A sigmoid activation function was used for neurons both in hidden as well as output layers. Mean squared error (MSE) was used as the cost function along with Adam optimization method [15]. The network was trained with learning rate of $\eta = 0.001$ along with decay rates $\beta_1 = 0.9$ and $\beta_2 = 0.999$, which are the default parameters for Adam optimizer provided in [15]. In order to reduce overfitting, dropout regularization [16] and batch normalization [17] were experimented with. Batch normalization, in addition to ensuring faster convergence, also yielded better performance on the validation set and was chosen over dropout as the regularization method. Note that batch normalization was used

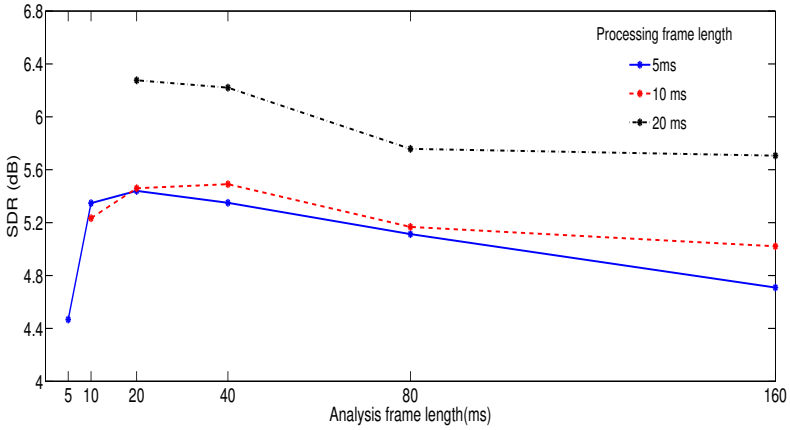

Fig. 2. DNN separation SDRs obtained with the proposed method for different analysis frame lengths.

between each hidden layer after application of sigmoid activations. Additionally, early stopping method [18] was used while DNN training, stopping the training process when no improvement in validation loss was observed continuously for 20 epochs.

3.4. Test conditions

For each speaker pair, evaluation metrics described in Section 3.2 were calculated for the NMF baseline and DNNs. Processing frame lengths of 5, 10, and 20 ms were investigated. Longer processing frame lengths are not suitable when it comes to low-latency applications, e.g., in hearing aid systems. Additionally, for each processing frame length, the effect of incorporating past contexts was studied. Specifically, the effect of utilizing analysis frame lengths of 5, 10, 20, 40, 80, and 160 ms were investigated.

3.5. Results

The separation performance metrics were computed for the five speaker pairs, and averaged across speakers to yield the final metrics. Figure 2 shows the separation performance of DNN for different analysis frame lengths. It was observed that incorporating previous temporal context improves the separation performance for 5 ms and 10 ms processing frames. The improvement in performance was most significant for the shortest processing frame, i.e., 5 ms. As longer temporal context is incorporated, these improvements become less significant with the maximum improvement observed for analysis frames 2-4 times the processing frame length. No improvement in separation performance was observed for extended temporal context with 20 ms processing frame length. The separation performance of DNN-based method was found to be better than the NMF baseline for all values of analysis and processing frame lengths. Table 1 compares DNN and NMF performance for 5 and 10 ms processing frames for 5, 10 and 20, and 40 ms analysis frames. It can be seen that the DNN-based method consistently outperforms its NMF counterpart by at least 1.5 dB for 5 ms processing frame, and by at least 1 dB for 10 ms processing frame , in terms of SDR.

4. CONCLUSION AND DISCUSSION

In this paper, a DNN-based method for single channel source separation has been proposed for low-latency applications. We show that it gives a better separation performance than state-of-the-art low-latency NMF baseline. Incorporation of previous context has been shown to improve the performance, with the improvement being most significant for very short processing frame length, i.e., 5 ms. This observation is consistent with the findings reported in [10].

It should be noted that for large analysis frame lengths, the dimensionality of DNN input feature vectors also increases. In such situations, the DNN architecture used here might be suboptimal and increasing number of hidden neurons or hidden layers might help in improving separation performance. Moreover, increasing the amount of training data would also help in improving separation performance. This study utilized conventional feed-forward DNN architecture. Use of architectures capable of modelling temporal dependencies, e.g., Long Short Term Memory (LSTMs) [19] are expected to further improve separation performance.

5. REFERENCES

[1] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[2] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen,

"Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing (ICASSP)*, vol. 19, no. 7, pp. 2067–2080, 2011.

[3] M. D. Plumbley, S. A. Abdallah, J. P. Bello, M. E. Davies, G. Monti, and M. B. Sandler, "Automatic music transcription and audio source separation," *Cybernetics and Systems*, vol. 33, no. 6, pp. 603–627, 2002.

[4] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors.," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4029–4032.

[5] DeLiang Wang, "Time-frequency masking for speech separation and its potential for hearing aid design," *Trends in amplification*, 2008.

[6] 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.

[7] T. Virtanen, J. F. Gemmeke, B. Raj, and P. Smaragdis, "Compositional models for audio processing: Uncovering the structure of sound mixtures," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125–144, 2015.

[8] Po-Sen 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 (ICASSP)*. IEEE, 2014, pp. 1562–1566.

[9] E. M. Grais, M. U. Sen, and H. Erdogan, "Deep neural networks for single channel source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 3734–3738.

[10] T. Barker, T. Virtanen, and N. H. Pontoppidan, "Low-latency sound-source-separation using non-negative matrix factorisation with coupled analysis and synthesis dictionaries," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[11] DeLiang Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*, pp. 181–197. Springer, 2005.

[12] T Virtanen, J. F. Gemmeke, and B. Raj, "Active-set newton algorithm for overcomplete non-negative representations of audio," *IEEE Transactions on Audio, Speech, and Language Processing (ICASSP)*, vol. 21, no. 11, pp. 2277–2289, 2013.

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

[14] F. Chollet, "Keras," 2015, GitHub, Available at *https://github.com/fchollet/keras*.

[15] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[18] R. Caruana, S. Lawrence, and L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," .

[19] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling.," in *INTERSPEECH*, 2014, pp. 338–342.