Low-Latency Sound Source Separation Using Deep Neural Networks

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

TAMPERE UNIVERSITY OF TECHNOLOGY

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

EriksholmResearchCentre PART OF OTICON

Overview

The monoaural source separation method is applied to two talker mixtures using feedforward deep neural networks (DNN) with no prior information other than the identity of speakers. The proposed approach is focused at low algorithmic delay applications, e.g., hearing aids. Around 1-2 dB improvement in source to distortion ration (SDR) compared to non negative matrix factorization baseline are achieved.

- Low- algorithmic delay is paramount for real time applications. For example, in hearing aids even delays ≈ 20 ms results in listener discomfort.
- DNNs models the source separation task as non linear regression between input (mixture spectrum) and output (constituent source spectra or intermediate time-frequency masks).
- DNNs are better equipped to handle this task in comparison to compositional model based approaches, e.g., non negative matrix factorization (NMF).

Method

- Spectral short-time Fourier transform (STFT) features derived from two talker acoustic mixtures are used as DNN input to estimate time-frequency masks corresponding to individual speakers.
- Algorithmic latency as low as 5 ms have been achieved.

Time-frequency masking

- Soft time-frequency masks are used:

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

where S_1 and S_2 are spectral features of corresponding constituent sources.

Source reconstruction

- Individual source spectra are calculated from estimated DNN output M_{est} , as,

$$\boldsymbol{S_{est1}} = \boldsymbol{M_{est}(t,f) * Y(t,f)}$$

and

$$\boldsymbol{S_{est2}} = \boldsymbol{\left(1 - M_{est}(t,f)\right) * Y(t,f)}$$

where $Y(t,f)$ is the mixture spectrum.

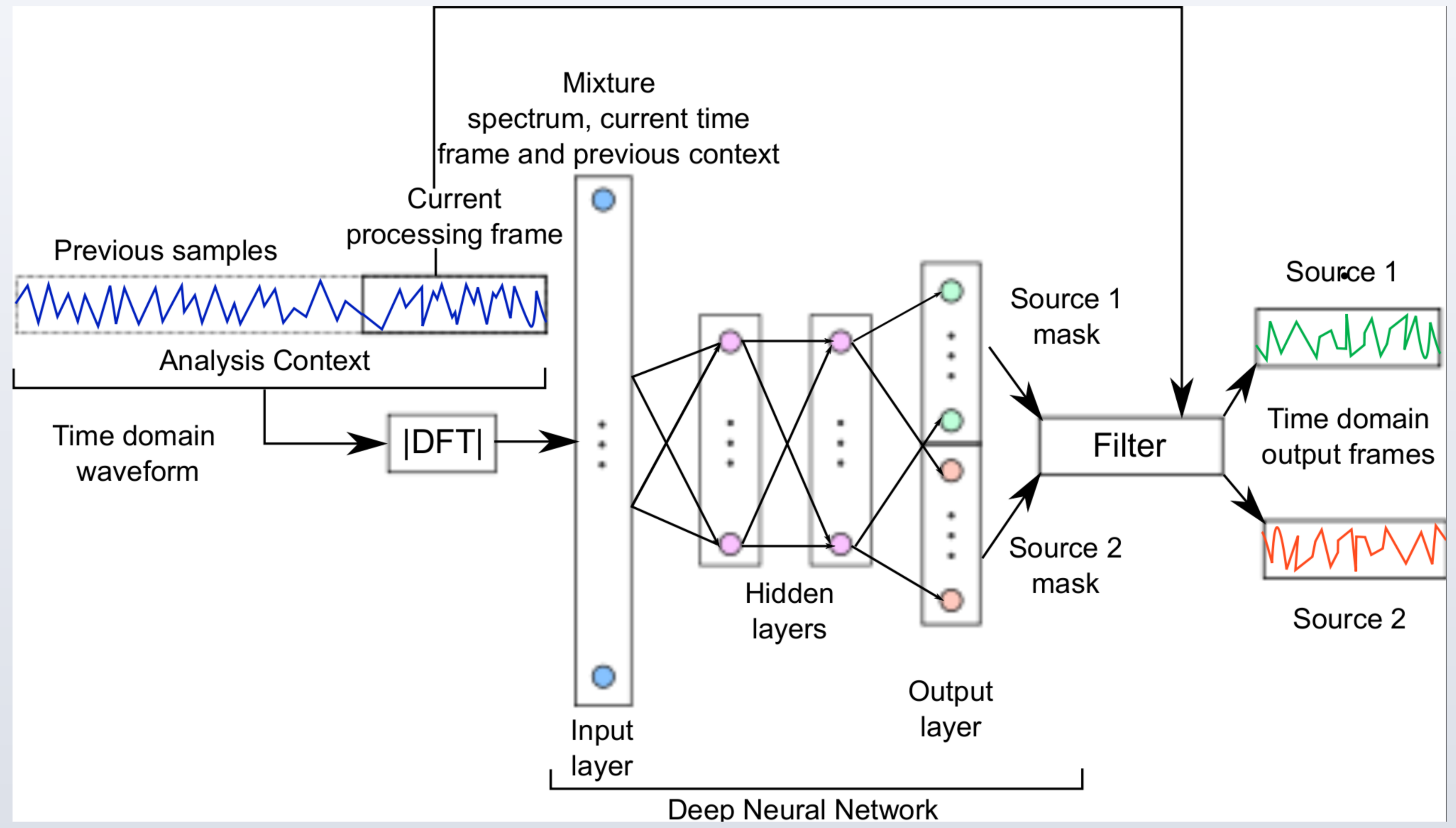

Fig.1.. Proposed DNN based sound source separation. approach.

- Features derived from a larger past temporal context used to predict time-frequency masks for current frame.
- Three layer feedforward DNN with 256 neurons in each layer is used.
- Neural network hyper parameters are chosen based on a validation set different from training and test sets.

Evaluation

Baseline

Non-negative matrix factorization (NMF) based source separation system utilizing 10000 basis atoms with generalized Kullback-Leibler divergence.

Acoustic Material

- CMU Arctic dataset A for training/validation and B for testing.
- Five speaker pairs: two male-male, two male-female and one female-female speaker pairs.
- 1024 acoustic mixtures for training, and 100 acoustic mixtures for testing for each speaker pair.

Metrics

- Source to distortion ratio (SDR), Source to Interference ratio (SIR), Source to artifact ratio (SAR)

Results

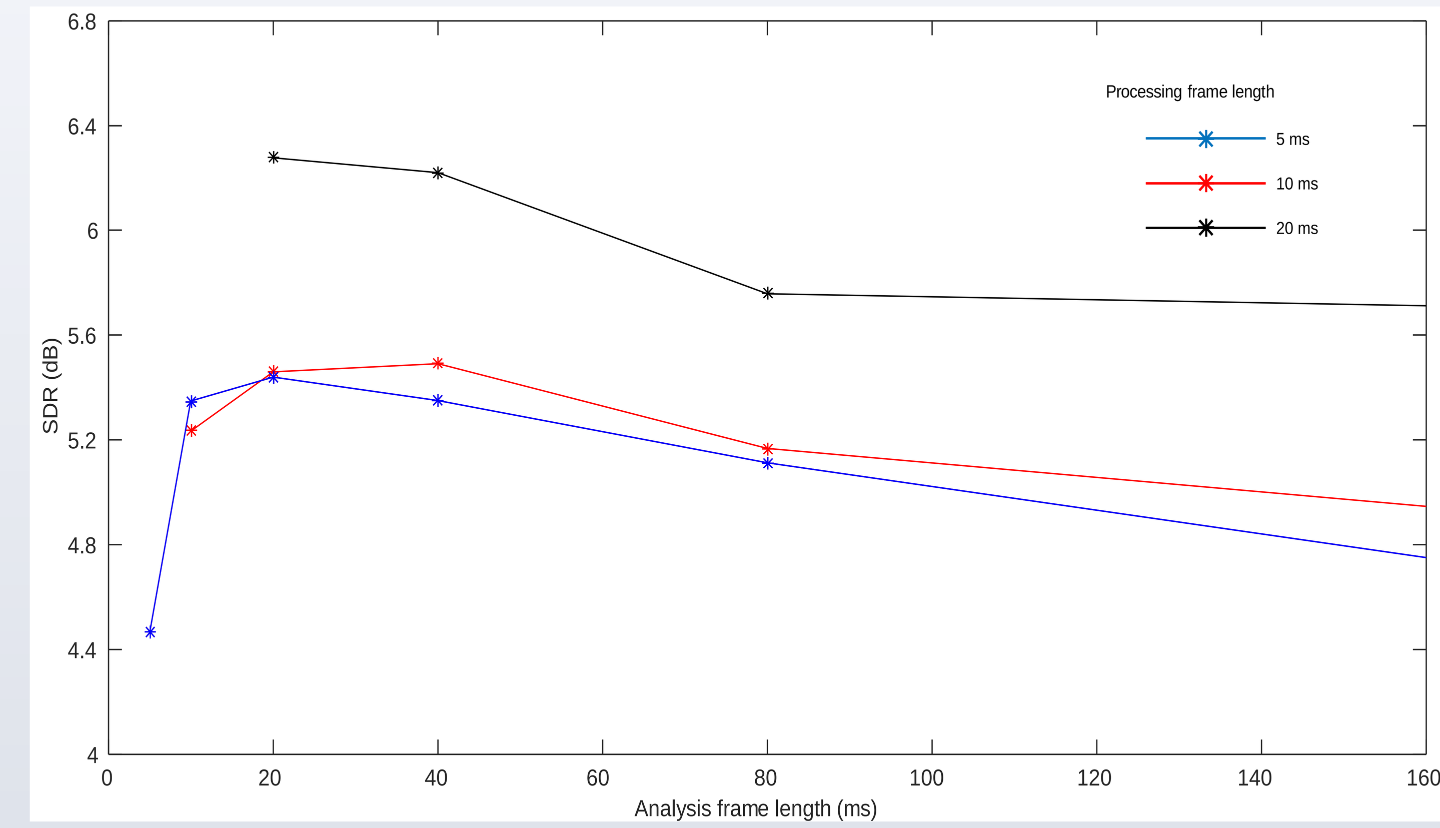

Fig.2. Variation of separation performance with analysis frame lengths for different processing frame lengths.

- Significant improvement in separation performance at low processing frame lengths.
- For larger processing frame lengths, using previous context is not of much help.
- Consistent improvement in separation performance over NMF baseline over all processing frame lengths.

for 5ms , SDR improvement over NMF is 1.8 dB.

for 10 ms , SDR improvement over NMF is 1.5 dB.

Conclusion

- A DNN based single channel source separation method for two talker mixtures has been proposed for low algorithmic delay applications.
- The effect of duration of the incorporated past temporal context on separation performance has been studied.
- The DNN based approach consistently outperforms NMF baseline for all latencies.
- Improvement in separation performance is most significant for very short processing frame lengths.

Contact details

gaurav.naithani@tut.fi