Speaker aware neural network for speaker extraction from overlapping speech

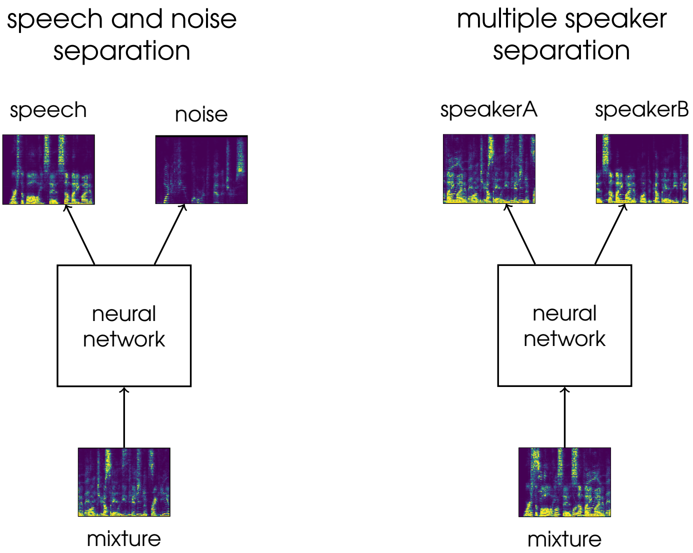
Katerina Zmolikova, Marc Delcroix

Brno University of Technology, Faculty of Information Technology
NTT Communication Science Laboratories, Kyoto

BRNO UNIVERSITY OF TECHNOLOGY FACULTY OF INFORMATION TECHNOLOGY

NTT

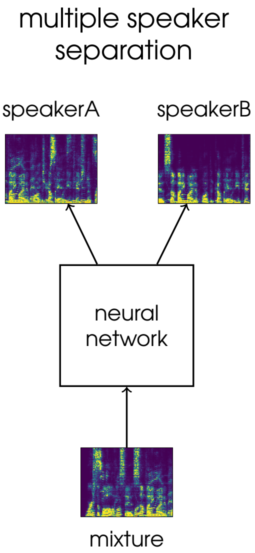
NN based speech separation



NN based speech separation

Issues

1. dependency on number of speakers
2. label permutation
3. speaker tracing



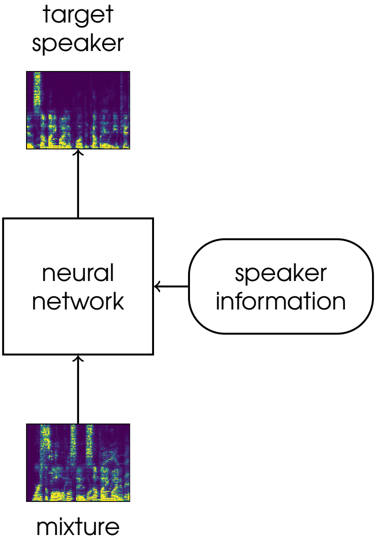
NN based speech separation

Popular approaches

1. **permutation invariant training**
   permutes outputs of the network during training
2. **deep clustering**
   projects T-F points to embedding space

	PIT	DC
# of speakers	✓	✓
label permutation	✓	✓
speaker tracing	**?**	**?**

Speaker aware neural network



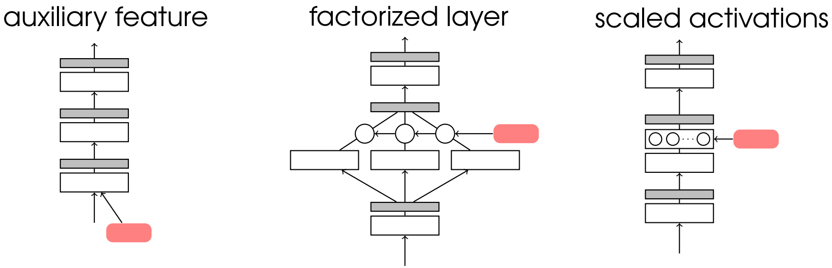
Speaker information

- informs the network about target speaker
- extracted from an adaptation utterance

Solves issues

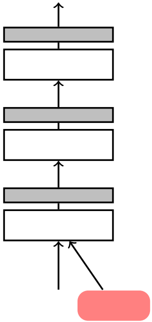
1. independent of number of speakers
2. no label permutation
3. tracks the speaker

Informing the neural network

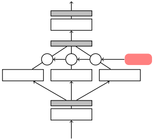


Informing the neural network

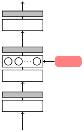
auxiliary feature



factorized layer



scaled activations

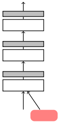


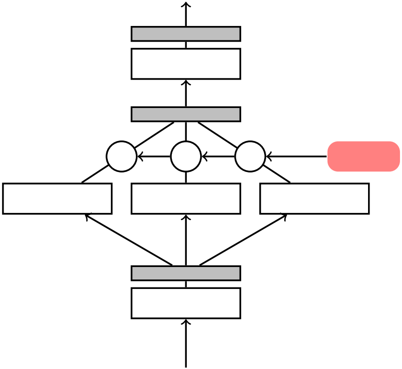
- appending speaker information as additional input
- (Saon et al. 2013; Senior et al. 2014)

Informing the neural network

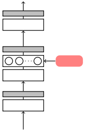
factorized layer

auxiliary feature



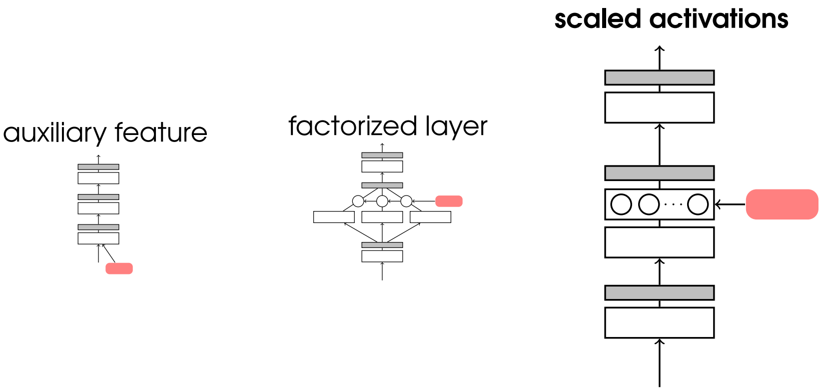


scaled activations



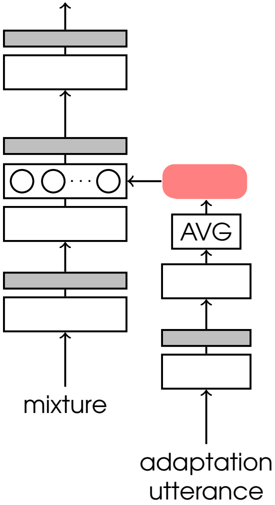
- splitting one of the layers into sublayers
- sublayers combined with weights derived from speaker info
- (Delcroix et al. 2015; Wu et al. 2015)

Informing the neural network



- activations in one layer scaled by weights derived from speaker info
- (Swietojanski et al. 2014; Samarakoon et al. 2016)

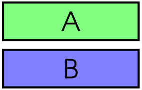
Extracting the speaker information



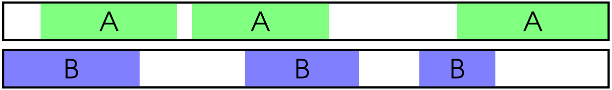
- speaker information extracted from adaptation utterance with auxiliary network
- average pooling to create utterance-wise vector from frame-wise features
- jointly trained with the main network

Datasets

- **WSJ0-2mix** (Hershey et al. 2016)
  about 10 second long fully overlapped mixtures
  based on WSJ utterances



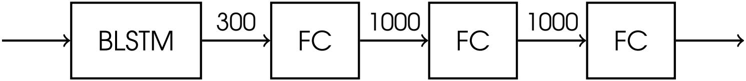
- **WSJ0-2mix-long**
  same mixing process as WSJ0-2mix
  three utterances from each speaker
  about 1 minute long mixtures



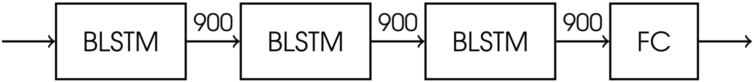
Experimental settings

Network configurations

- smaller configuration



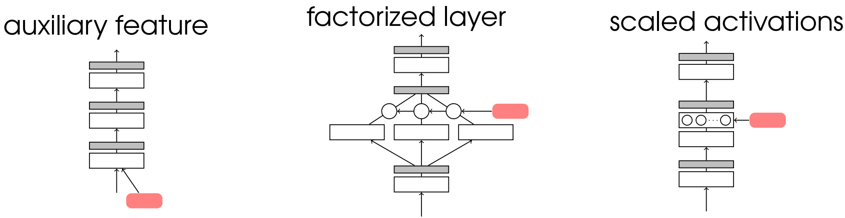
- larger configuration



- magnitude STFT as input
- predicting T-F mask, MSE objective

- WSJ0-2mix, smaller NN configuration

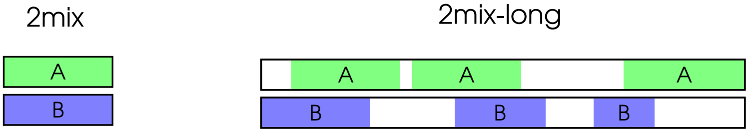
method	ΔSDR
auxliary feature	-2.2
factorized layer	6.2
scaled activations	5.7
IBM	12.8



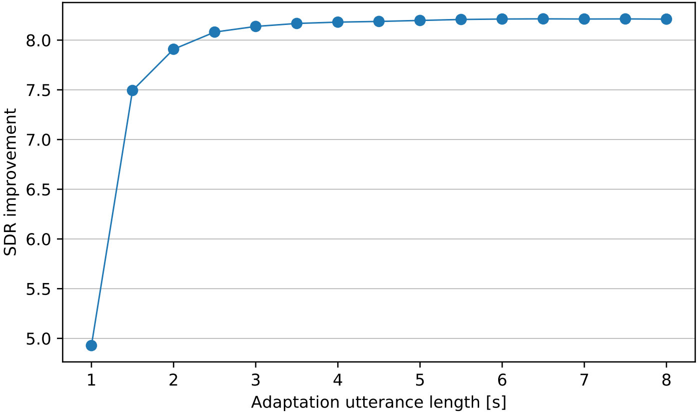
Comparison with DC and PIT

- larger NN configuration, scaled activations method

method	2mix	2mix-long
SpeakerBeam	8.2	12.2
PIT	8.2	9.9
DC	8.7	10.0
SpeakerBeam+DC	9.1	12.6
IBM	12.8	17.1

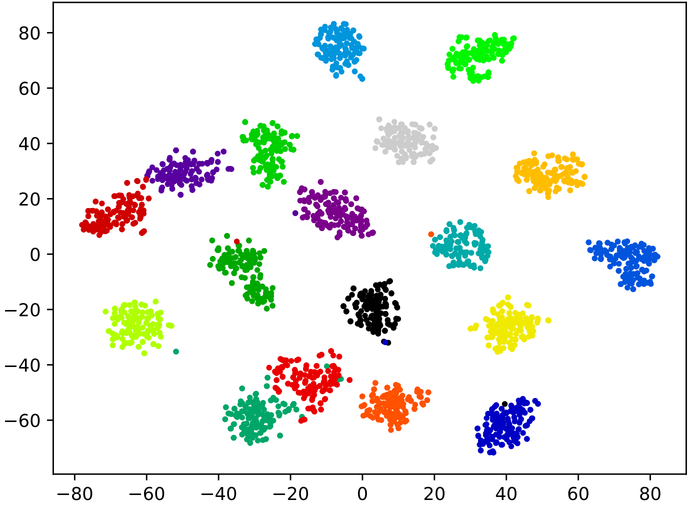


Length of adaptation data



- for 0.5 seconds, -0.9 dB SDR degradation

Extracted speaker representations



Conclusions

- Additional speaker information can help to avoid problems of NN based speech separation and do speaker tracing.
- Methods adapting parameters of entire layer work well.
- This can be combined with deep clustering to enhance its accuracy, especially on longer mixtures.

Thank you!

izmolikova@fit.vutbr.cz

- Marc Delcroix et al. "Context adaptive deep neural networks for fast acoustic model adaptation". In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. 2015, pp. 4535–4539.
- John R Hershey et al. "Deep clustering: Discriminative embeddings for segmentation and separation". In: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. 2016, pp. 31–35.
- Lahiru Samarakoon and Khe Chai Sim. "Subspace LHUC for Fast Adaptation of Deep Neural Network Acoustic Models." In: INTERSPEECH. 2016, pp. 1593–1597.
- G. Saon et al. "Speaker adaptation of neural network acoustic models using i-vectors". In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. 2013, pp. 55–59.
- A. Senior and I. Lopez-Moreno. "Improving DNN speaker independence with I-vector inputs". In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. 2014, pp. 225–229.

Pawel Swietojanski and Steve Renals. “Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models”. In: Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE. 2014, pp. 171–176.

C. Wu and M. J. F. Gales. “Multi-basis adaptive neural network for rapid adaptation in speech recognition”. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. 2015, pp. 4315–4319.