Deep Neural Networks for Multiple Speaker Detection and Localization

1Idiap Research Institute, Switzerland  2EPFL, Switzerland
Weipeng He1,2, Petr Motlicek1, Jean-Marc Odobez1,2

Introduction

Task: Sound source localization in real HRI scenarios
unknown number of multiple speakers

Contributions:
• Novel deep learning-based multiple sound source localization method.
• Likelihood-based output encoding handles an arbitrary number of sources.
• Investigation of three network architectures based on different motivations.
• Study sub-band cross correlation information as an input feature for better localization cues in speech mixtures.
• Collected and released a benchmark dataset of real recordings.

Approach

Different network architectures:
• Multi-layer Perceptron (MLP-GCC)
• Basic structure (baseline)
• Convolutional neural network (CNN-GCCFB)
• Convolution to reduce number of parameters
• Two-stage network (TSNN-GCCFB):
  • Considers the sparsity of speech signal in time-frequency points
  • First predict on sub-bands
  • Then aggregate early predictions across all frequencies.
  • Training also in 2 steps: (1) Pretrain first subnet (2) End-to-end

Experiments

Data:
• 24 hours of real recordings of Pepper.
• Up to two simultaneous speakers.

Loudspeakers
(16h train / 8h test)

Human talkers
(4 min test)

Results:

1. Proposed methods have better overall performance
2. More significant with overlapping sources
3. Two-stage network performs the best

Baseline methods:
• SRP-PHAT, MVDR, MUSIC

Evaluation:
• Number of sources is unknown => detection problem
• Prediction is correct if error < 5°
• Compute precision vs recall

Conclusion

Baseline methods:
• >90% recall and precision.
• Significantly better than popular spatial spectrum methods.

Resources
• Database: https://www.idiap.ch/dataset/sslr
• Video: https://youtu.be/_4EwuvIE_pU