DNN for Multiple Speaker Detection and Localization

FuYanjie 2021.9.24

Overview

1. Introduction
2. Proposed Method
3. Experiment
4. Conclusion

Author's homepage: https://idiap.ch/~whe/

Paper's link: https://arxiv.org/pdf/1711.11565.pdf

Introduction

Challenges

1. Noisy environments;
2. Multiple simultaneous speakers;
3. Short and low-energy utterances;
4. Obstacles such as robot body blocking sound direct path.

Introduction

Motivation

Compared to Conventional Methods:

Conventional signal processing algorithms are derived with assumptions, many of which do not hold well under the above-mentioned conditions.

NNs can learn the mapping from the localizaition cues to the direction-of-arrival without making strong assumptions.

Introduction

Motivation

Compared to Existing NN-based SSL Methods:

1. Do not address the problem of multiple sound sources;
2. Cannot detect and localize multiple voices in real multi-party HRI scenarios simultaneously.
3. Formulate the problem as the classification of an audio input into one "class" label associated with a location, and optimizing the posterior probability of such labels. Such posterior probability encoding cannot be easily extended to multiple sound source situations.

Introduction

Contributions

TABLE I: Comparison of our methods with existing NN-based SSL approaches

Approach	Number of Sources	Input Size	Input Feature	Output Coding
Datum et al. [5]	1	-	IPD and ITD per freq.	Gaussian-shaped function
Xiao et al. [8]	1	Utterance	GCC-PHAT coefficients	Posterior probability
Takeda et al. [9]	0 or 1	200ms	MUSIC eigenvectors	Posterior probability
Yalta et al. [10]	0 or 1	200ms	Power spectrogram	Posterior probability
Ma et al. [7]	Known multiple	Utterance	CCF and ILD per freq.	Posterior probability
Takeda et al. [11]	0, 1, 2	200ms	MUSIC eigenvectors	Marginal posterior probability
Ours	Unknown multiple	170ms	GCC-PHAT and GCCFB	Likelihood-based coding

Our methods can cope with short input, overlapping speech, an unknown number of sources and strong ego-noise.

Proposed Method

A. Input Features

time frame: 170ms number of sources: N number of microphones: M

STFT of input signal: $X_i(\omega), i = 1, \ldots, M$, the mic index: i, the discrete freq: ω.

Two types of features based on GCC-PHAT at frame level:

- GCC-PHAT coefficients:

The GCC-PHAT between channel i And j Is formulated as:

$$g_{ij}(\tau) = \sum_{\omega} \mathcal{R}\left(\frac{X_i(\omega) X_j(\omega)^*}{|X_i(\omega) X_j(\omega)^*|} e^{j\omega\tau} \right)$$

Proposed Method

$\tau(\in [-25, 25])$ is the discrete delay (use the center 51 delays), $(\cdot)^*$ denotes the complex conjugation, $\mathcal{R}(\cdot)$ denotes the real part of a complex number.

- GCC-PHAT on Mel-scale filter bank:

The GCC-PHAT is not optimal for TDOA estimation of multiple source signals since it equally sums over all freq bins disregarding the "sparsity" of speech signals in the TF domain. Thus propose to use GCC-PHAT on Mel-scale filter bank (GCCFB).

$$g_{ij}(f,\tau) = \frac{\sum_{\omega\in\Omega_f} \mathcal{R}\left(H_f(\omega)\frac{X_i(\omega)X_j(\omega)^*}{|X_i(\omega)X_j(\omega)^*|}e^{j\omega\tau}\right)}{\sum_{\omega\in\Omega_f} H_f(\omega)}$$

f is the filter index, H_f is the transfer function of the f-th Mel-scaled triangular filter.
40 Mel-scale filters covering the frequencies from 100 to 8000 Hz.

Proposed Method

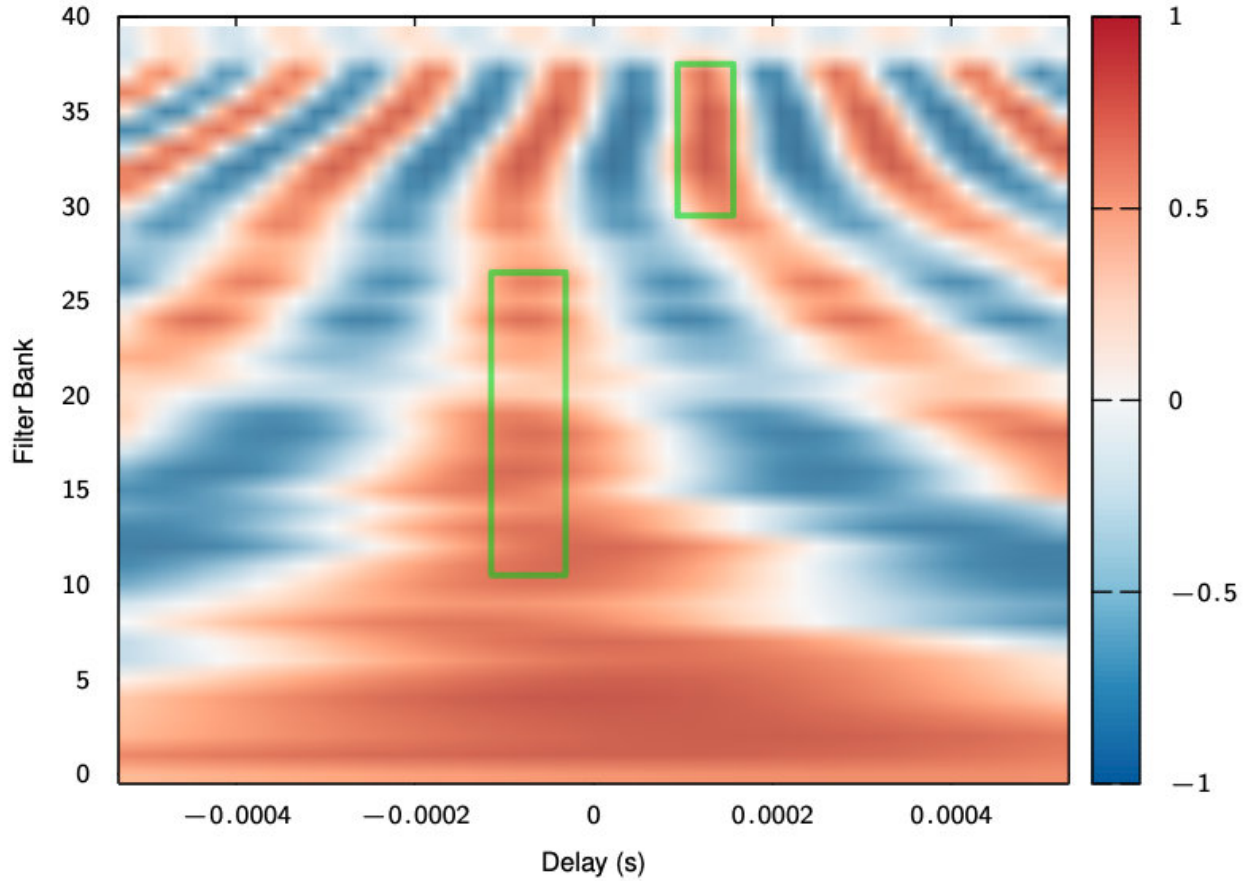


Fig. 2: Example of GCCFB extracted from a frame with two overlapping sound sources.

Proposed Method

B. Likelihood-based Output Coding

Encoding: the output (the likelihood of a sound source being in each direction) is encoded into a vector $\{o_i\}$ of 360 values (分别对应 θ_i). The values are defined as the maximum of Gaussian-like functions centered around the true DOAs:

$$o_i = \begin{cases} \max_{j=1}^{N} \left\{ e^{-d\left(\theta_i, \theta_j^{(s)}\right)^2 / \sigma^2} \right\} & \text{if } N > 0 \\ 0 & \text{otherwise} \end{cases}$$

其中 $\theta_j^{(s)}$ 是第 j 个声源DOA的真实值，σ 是高斯分布的标准差（尺度参数），$d(\cdot,\cdot)$ 表示 angular distance

Proposed Method

B. Likelihood-based Output Coding

Posterior probability coding is constrained as a probability distribution (the output layer is normalized by a softmax function). It can be all zero when there is no sound source, or contains N peaks when there are N sources.

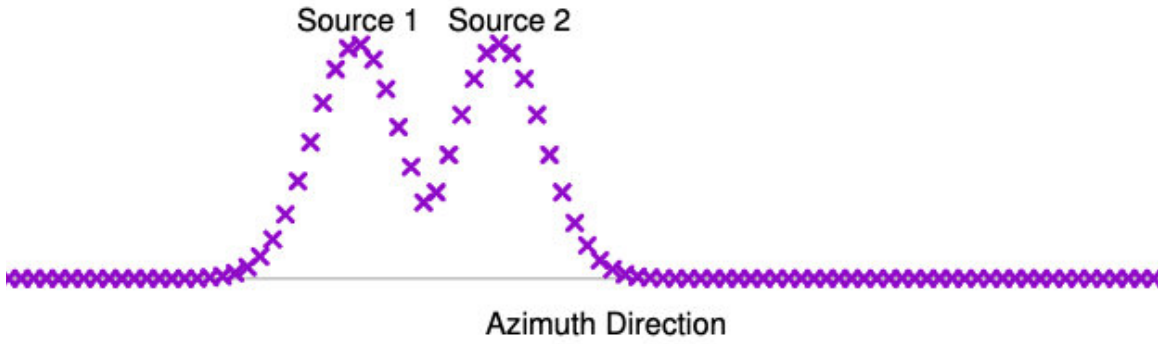


Fig. 3: Output coding for multiple sources.

Proposed Method

B. Likelihood-based Output Coding

Decoding: During the test phase, we decode the output by finding the peaks that are above a given threshold ξ :

$$\text{Prediction} = \left\{ \theta_i : o_i > \xi \quad \text{and} \quad o_i = \max_{d(\theta_j, \theta_i) < \sigma_n} o_j \right\}$$

with σ_n being the neighborhood distance. We choose $\sigma = \sigma_n = 8^\circ$ for the experiments.

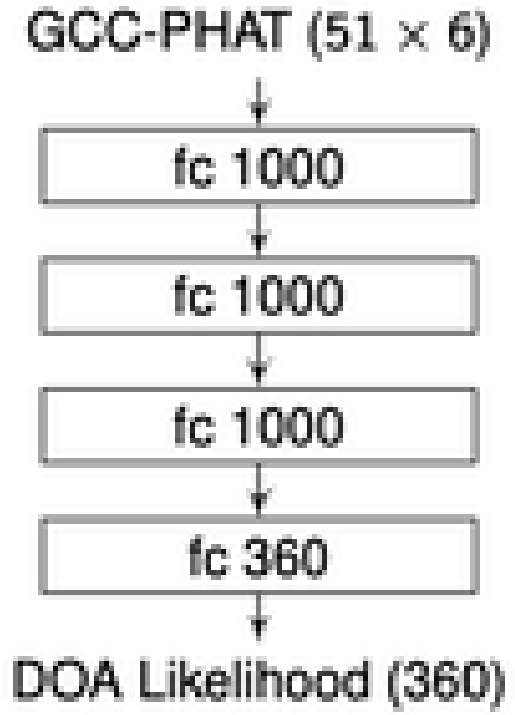
Proposed Method

C. 3 different Neural Network Architectures

MLP-GCC (Multilayer perceptron with GCC-PHAT)

Three hidden fully-connected layers with ReLU activation function and Batch Normalization

The last fully-connected layer with sigmoid activation function



(a) MLP-GCC

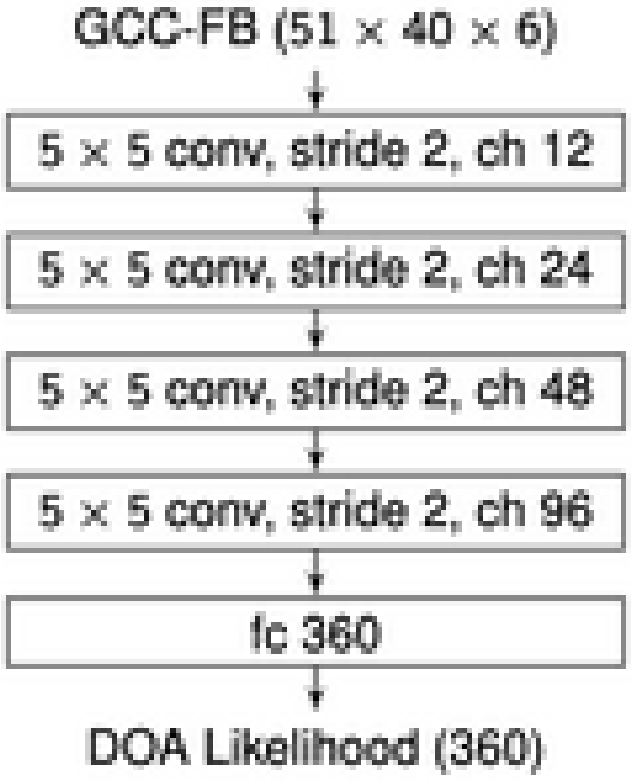
Proposed Method

C. 3 different Neural Network Architectures

CNN-GCCFB (Convolutional neural network with GCCFB)

FC NNs are not suitable for high-dimensional input features (such as GCCFB) (introduces a large amount of parameters; prone to overfitting).

Four convolutional layers (with ReLU activation and BN) and a FC layer at the output (with sigmoid activation)



(b) CNN-GCCFB

Proposed Method

C. 3 different Neural Network Architectures

TSNN-GCCFB (Two-stage neural network with GCCFB)

- do analysis or implicit DOA estimation in each freq band before such info is aggregated into a broadband prediction
- Features with the same delay on different microphone pairs do not correspond to each other locally. Instead, feature extraction or filters should take the whole delay axis into account.

Training scheme: First, we train the Subnet 1 in the first stage using the DOA likelihood as the desired latent feature. During the second step, both stages are trained in an end-to-end manner.

Proposed Method

C. 3 different Neural Network Architectures

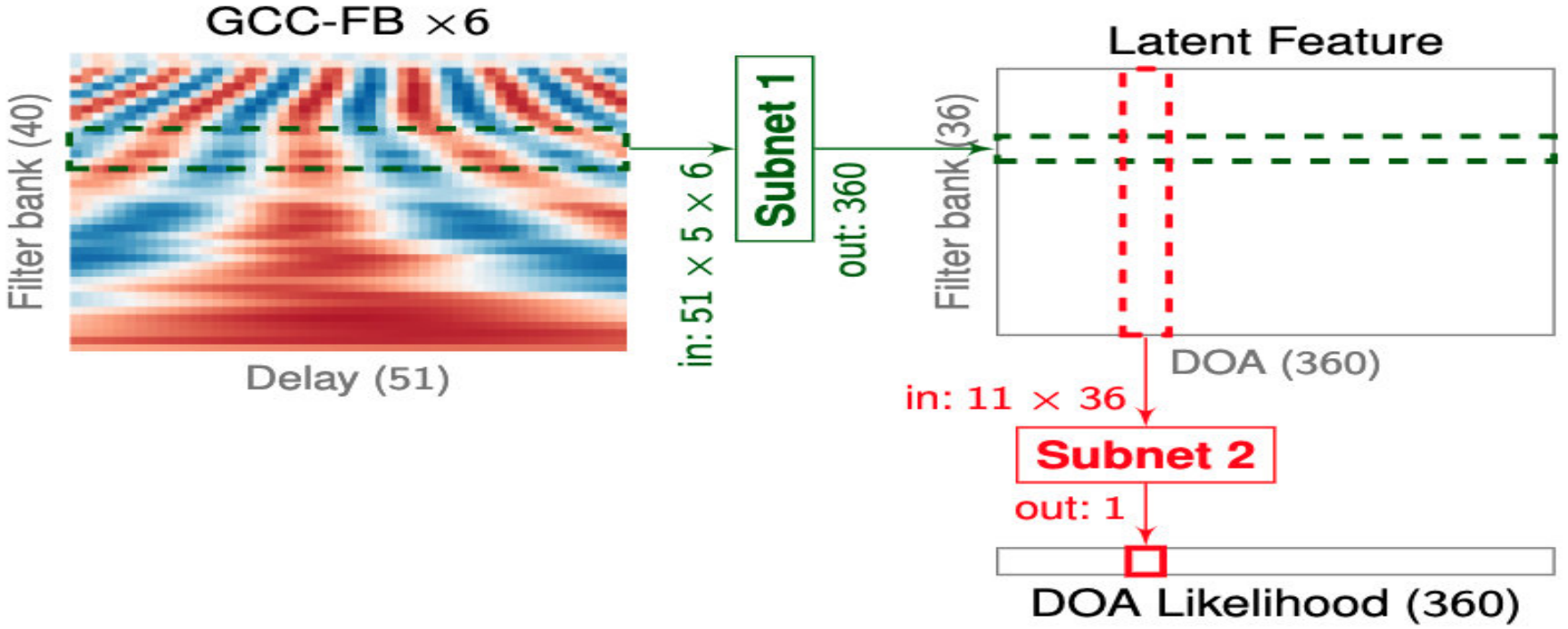


Fig. 5: NN architecture of two-stage neural network with GCCFB as input. The first and second stages are marked as green and red, respectively.

Experiment

C. Network Training

Adam optimizer

MSE loss

Mini-batch size: 256

10 epochs for MLP-GCC & CNN-GCCFB

4 epochs for the first stage of TSNN-GCCFB and 10 epochs for the end-to-end training

Experiment

A. Datasets

sr: 48kHz
4 mics, rectangle of 5.8 X 6.9 cm

Loudspeaker Recordings

```
├── lsp_train_106 (a large conference room)
├── lsp_train_301 (a small conference room)
├── lsp_test_106
├── lsp_test_library (a small room with shelves)

├── lsp_<*>
│   ├── audio
│   ├── gt_file
│   └── gt_frame
```

Experiment

A. Datasets

Audio Files: "RECORD_ID.wav"

File-Level Ground Truth:

- the recording ID
- the start & end time of the recording in the original ROS bag file
- list of source labels, each source label is a tuple of:
  - 3D source location
  - source audio file (a segment from the AMI corpus)
  - the start & end time of the source in the recording
  - relative volume of the source

Experiment

A. Datasets

Frame-Level Ground Truth:

- the frame ID, which start from 0. The frame of ID t contains samples between [t * HOP_SIZE, t * HOP_SIZE + WIN_SIZE).
- list of active sources (can be empty list if there is no active source). Each active source contains:
  - 3D source location
  - source type, which is always 1 (speech source).
  - speaker ID

Experiment

A. Datasets

Human Talker Recordings
```
├── human
│   ├── audio
│   ├── audio_gt
│   ├── gt_frame
│   └── video_gt
```

Audio Ground Truth:

The audio_gt directory includes the voice activity ground truth.

Video Data and Ground Truth:

The video_gt directory includes the video data and source location ground truth.

Experiment

B. Evaluation Protocol

We evaluate multiple SSL methods at frame level under two different conditions: the number of sources is known or unknown.

Known:
We select the N highest peaks of the output as the predicted DOAs and match them with ground truth DOAs one by one, and we compute the mean absolute error (MAE). Evaluate by ACC of predictions.
By saying a prediction is correct, we mean the error of the prediction is less than a given admissible error E_a.

Unknown:
detection - given ground truth sources, compute recall (the percentage of correct detection out of all ground truth sources)
localization - compute precision (the percentage of correct predictions among all predictions)

Experiment

D. Baseline Methods

Spatial spectrum-based methods:

SRP-PHAT

SRP-NONLIN: SRP-PHAT with a non-linear modification of the score

MVDR-SNR: minimum variance distortionless response beam forming with signal-to-noise ratio as score

SEVD-MUSIC: multiple signal classification, assuming spatially white noise and one signal in each bin

GEVD-MUSIC: MUSIC with generalized eigenvector decomposition, assuming noise is pre-measured and one signal in each TF bin

Experiment

E. Results

TABLE III: Performance assuming a known number of sources. $E_a = 5°$.

Dataset	Loudspeaker						Human					
Subset (# of frames)	Overall (207k)		$N = 1$ (178k)		$N = 2$ (29k)		Overall (929)		$N = 1$ (788)		$N = 2$ (141)	
	MAE (°)	ACC	MAE (°)	ACC	MAE (°)	ACC	MAE (°)	ACC	MAE (°)	ACC	MAE (°)	ACC
MLP-GCC	4.89	**0.92**	4.18	**0.94**	9.21	**0.77**	4.99	0.93	4.44	0.94	8.06	0.84
CNN-GCCFB	**4.80**	0.90	**4.11**	0.93	**9.06**	0.73	4.82	0.93	4.19	**0.96**	8.34	0.77
TSNN-GCCFB	5.41	0.91	4.64	0.93	10.10	**0.77**	**4.14**	**0.95**	3.84	**0.96**	**5.84**	**0.90**
SRP-PHAT [3]	21.51	0.78	19.00	0.82	36.95	0.50	5.39	0.88	2.62	0.93	20.90	0.56
SRP-NONLIN [18]	25.71	0.73	23.77	0.77	37.61	0.51	4.84	0.90	2.47	0.94	18.11	0.68
MVDR-SNR [18]	23.17	0.76	21.22	0.79	35.19	0.55	4.39	0.90	**2.45**	0.94	15.21	0.68
SEVD-MUSIC [2]	29.07	0.66	27.59	0.69	38.14	0.47	6.36	0.85	3.00	0.88	25.14	0.64
GEVD-MUSIC [20]	25.43	0.64	23.18	0.67	39.28	0.44	6.45	0.81	3.62	0.85	22.24	0.63

Experiment

E. Results

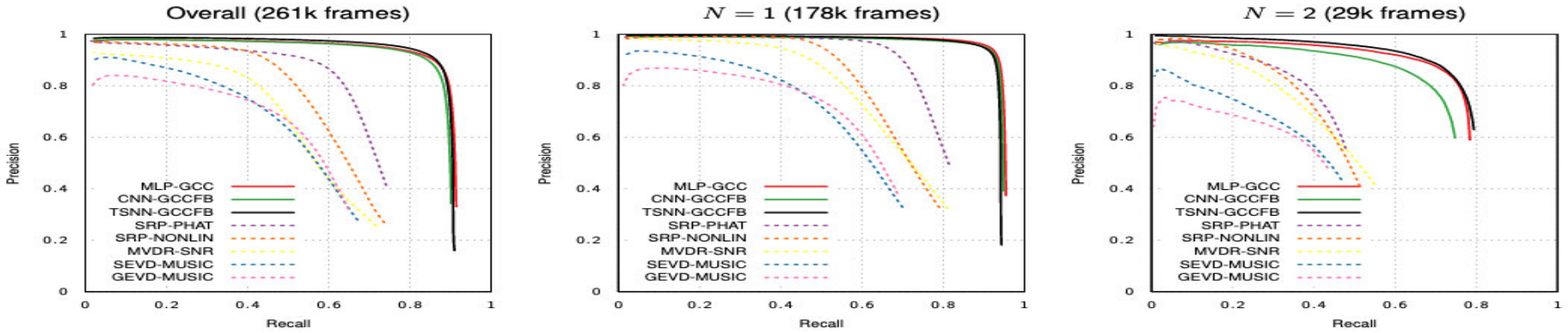


Fig. 7: Detection and localization performance on recordings with loudspeakers. $E_a = 5°$.

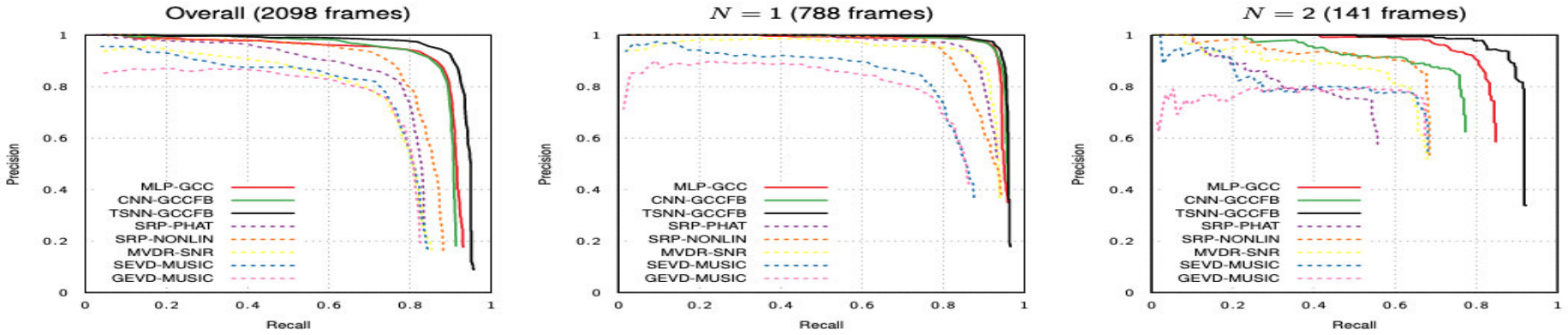


Fig. 8: Detection and localization performance on recordings with human subjects. $E_a = 5°$.

Conclusion

Limitation:

- The current study is potentially limited by the training data samples, which are not likely to cover all possible combinations of source positions, since the number of combinations grows exponentially with the number of sources.
- Will investigate the incorporation of temporal context.

Thanks for your time!