Deep Neural Networks for Multiple Speaker Detection and Localization

Weipeng He[1,2], Petr Motlicek[1] and Jean-Marc Odobez[1,2]

Abstract—We propose to use neural networks for simultaneous detection and localization of multiple sound sources in human-robot interaction. In contrast to conventional signal processing techniques, neural network-based sound source localization methods require fewer strong assumptions about the environment. Previous neural network-based methods have been focusing on localizing a single sound source, which do not extend to multiple sources in terms of detection and localization. In this paper, we thus propose a likelihood-based encoding of the network output, which naturally allows the detection of an arbitrary number of sources. In addition, we investigate the use of sub-band cross-correlation information as features for better localization in sound mixtures, as well as three different network architectures based on different motivations. Experiments on real data recorded from a robot show that our proposed methods significantly outperform the popular spatial spectrum-based approaches.

I. INTRODUCTION

A. Motivation

Sound source localization (SSL) and speaker detection are crucial components in multi-party human-robot interaction (HRI), where the robot needs to precisely detect where and who the speaker is and responds appropriately (Fig. 1). In addition, robust output from SSL is essential for further HRI analysis (e.g. speech recognition, speaker identification, etc.) which provides a reliable source of information to be combined with other modalities towards improved HRI. Although SSL has been studied for decades, it is still a challenging topic in real HRI applications, due to the following conditions:

- Noisy environments and strong robot ego-noise;
- Multiple simultaneous speakers;
- Short and low-energy utterances, as responses to questions or non-verbal feedback;
- Obstacles such as robot body blocking sound direct path.

Traditionally, SSL is considered a signal processing problem. The solutions are analytically derived with assumptions about the signal, noise and environment [1–3]. However, many of the assumptions do not hold well under the above-mentioned conditions, which may severely impact their performance. Alternatively, researchers have recently adopted machine learning approaches with neural networks (NN). Indeed, with a sufficient amount of data, the NNs can in principle learn the mapping from the localization cues to

Fig. 1: The robot, Pepper, used in our experiments and a typical HRI scenario where the robot interacts with multiple persons.

the direction-of-arrival (DOA) without making strong assumptions. Surprisingly, most of the learning-based methods do not address the problem of multiple sound sources and in particular, the simultaneous detection and localization of multiple voices in real multi-party HRI scenarios have not been well studied.

B. Existing Neural Network-based SSL Methods

Although the earliest attempts of using neural networks for SSL date back to the 1990s [4, 5], it was not until recently that researchers started to pay more attention to such learning-based approaches. With the large increase of computational power and advances in deep neural networks (DNN), several methods were shown to achieve promising single SSL performance [6–10]. Nevertheless, most of these methods aim at detecting only one source, focusing the research on the localization accuracy. In particular, they 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. Unfortunately, such posterior probability encoding cannot be easily extended to multiple sound source situations.

Neural networks that are trained for localizing a single source can be applied for multi-source localization by pooling the network outputs (i.e. posterior probabilities) over multiple time frames [7]. However, this method requires a known number of sources and a long period of time as input for pooling. Such limitations make it not practical for real applications.

Localization of two sources is addressed in [11], which encodes the output as two marginal posterior probability vectors. However, an ad-hoc location-based ordering is introduced to decide the source-to-vector assignment, rendering the posteriors dependent on each other and the encoding somewhat ambiguous. That is, the same source may need to be predicted as the first source if alone, or as the second one if another signal with a preceding label is present.

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 eigenspectrum	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

C. Contributions

This paper investigates NN-based SSL methods applied to real HRI scenarios (Fig. 1). In contrast to previous studies (Table I), the methods are required to cope with short input, overlapping speech, an unknown number of sources and strong ego-noise. We emphasize their application in real conditions by testing the methods with recorded data from the robot Pepper[1].

In this paper, we propose three NN architectures for multiple SSL based on different motivations. The NNs adopt a likelihood-based output encoding that can handle an arbitrary number of sources. And, we investigate the employment of sub-band cross correlation information as an input feature for better localization cues in speech mixtures. The experiments indicate that the proposed methods significantly outperform the baseline methods.

Furthermore, we collect and release a benchmark dataset[2] of real recordings for developing and evaluating learning-based SSL in HRI.

II. PROPOSED METHOD

In this section, we describe our proposed NN models for multiple SSL. We consider the localization of sounds in the azimuth direction in individual time frames, which are 170ms long. We denote the number of sources by N and the number of microphones by M. The input signal is represented by the short time Fourier transforms (STFT): $X_i(\omega), i = 1, \ldots, M$, where i is the microphone index and ω is the frequency in the discrete domain. We omit the time index for clarity, because none of the methods described below exploit context information or temporal relations.

A. Input Features

The generalized cross-correlation with phase transform (GCC-PHAT) [1] is the most popular method for estimating the time difference of arrival (TDOA) between microphones, which is an important clue for SSL. Here, we use two types of features based on GCC-PHAT at frame level.

GCC-PHAT coefficients: The first type of input feature is represented by the center GCC-PHAT values of all $M(M-1)/2$ microphone pairs as used in [8]. The GCC-PHAT between channel i and j is formulated as:

$$g_{ij}(\tau) = \sum_{\omega} \mathcal{R}\left(\frac{X_i(\omega)X_j(\omega)^*}{|X_i(\omega)X_j(\omega)^*|}e^{j\omega\tau}\right), \quad (1)$$

where τ is the delay in the discrete domain, $(\cdot)^*$ denotes the complex conjugation, and $\mathcal{R}(\cdot)$ denotes the real part of a complex number. The peak in GCC-PHAT is used to estimate the TDOA. However, under real condition, the GCC-PHAT is corrupted by noise and reverberation. Therefore, we use the full GCC-PHAT function as the input feature instead of a single estimation of the TDOA. In our experiments, we use the center 51 delays ($\tau \in [-25, 25]$).

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 frequency bins disregarding the "sparsity" of speech signals in the time-frequency (TF) domain and the randomly distributed noise which may be stronger than the signal in some TF bins. To preserve delay information on each frequency band and to allow sub-band analysis, we propose to use GCC-PHAT on mel-scale filter bank (GCCFB). Hence, the second type of input feature is formulated as:

$$g_{ij}(f,\tau) = \frac{\sum_{\omega \in \Omega_f} \mathcal{R}\left(H_f(\omega)\frac{X_i(\omega)X_j(\omega)^*}{|X_i(\omega)X_j(\omega)^*|}e^{j\omega\tau}\right)}{\sum_{\omega \in \Omega_f} H_f(\omega)}, \quad (2)$$

where f is the filter index, H_f is the transfer function of the f-th mel-scaled triangular filter, and Ω_f is the support of H_f. Fig. 2 shows an example of the GCCFB of a frame where two speech signals overlap. Each row corresponds to the GCC-PHAT in an individual frequency band. The frequency-based decomposition allows the estimation of the TDOAs by looking into local areas rather than across all frequency bins. In the example, the areas marked by the green rectangles correspond to two separate sources with different delays and which produce high cross-correlation values in different frequency bands (and hence, for different filter indices). In the experiments, we use 40 mel-scale filters covering the frequencies from 100 to 8000 Hz.

B. Likelihood-based Output Coding

Encoding: We design the multiple SSL output coding as the likelihood of a sound source being in each direction. Specifically, the output is encoded into a vector $\{o_i\}$ of 360 values, each of which is associated with an individual azimuth direction θ_i. The values are defined as the maximum of Gaussian-like functions centered around the true DOAs:

$$o_i = \begin{cases} \max_{j=1}^{N}\left\{e^{-d(\theta_i, \theta_j^{(s)})^2/\sigma^2}\right\} & \text{if } N > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$
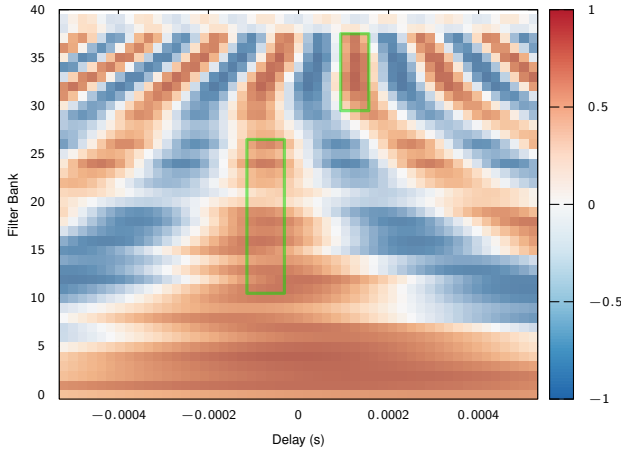
Fig. 2: Example of GCCFB extracted from a frame with two overlapping sound sources.
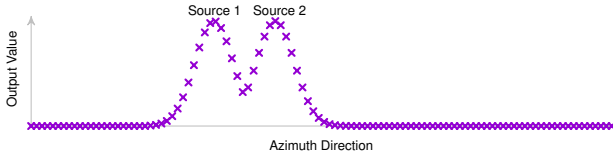

Fig. 3: Output coding for multiple sources.

where $\theta_j^{(s)}$ is the ground truth DOA of the j-th source, σ is the value to control the width of the Gaussian-like curves and $d(\cdot, \cdot)$ denotes the angular distance. The output coding resembles a spatial spectrum, which is a function that peaks at the true DOAs (Fig. 3).

Unlike posterior probability coding, the likelihood-based coding is not constrained as a probability distribution (the output layer is not 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. The coding can handle the detection of an arbitrary number of sources. In addition, the soft assignment of the output values, in contrast to the $0/1$ assignment in posterior coding, takes the correlation between adjacent directions into account allowing better generalization of the neural networks.

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

$$\text{Prediction} = \left\{ \theta_i : o_i > \xi \quad \text{and} \quad o_i = \max_{d(\theta_j, \theta_i) < \sigma_n} o_j \right\},$$
(4)

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

C. Neural Network Architectures

We investigate three different types of NN architectures for sound source localization.

MLP-GCC (Multilayer perceptron with GCC-PHAT): As illustrated in Fig. 4a, the MLP-GCC uses GCC-PHAT as input and contains three hidden layers, each of which is a fully-connected layer with a rectified linear unit (ReLU) activation function [12] and batch normalization (BN) [13]. The
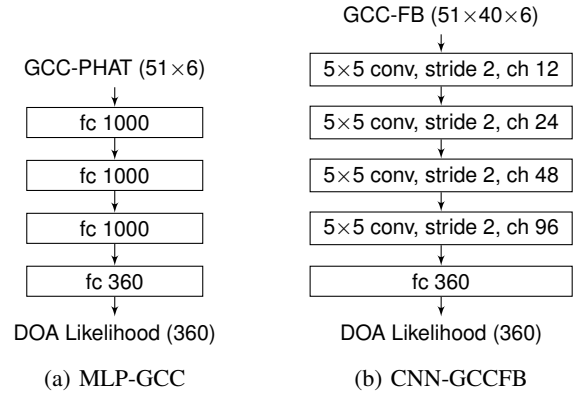

Fig. 4: Two neural network architectures for multiple SSL.

last layer is a fully connected layer with sigmoid activation function. The sigmoid function is bounded between 0 and 1, which is the range of the desired output. According to our experiments, this helps the network to converge to a better result.

CNN-GCCFB (Convolutional neural network with GC-CFB): Fully connected NNs are not suitable for high-dimensional input features (such as GCCFB) because the large dimension introduces a large amount of parameters to be trained, making the network computationally expensive and prone to overfitting. Convolutional neural networks (CNN) can learn local features with reduced amount of parameters by using weight sharing. This leads to the idea of using CNN for the input feature of GCCFB.

We use the CNN structure shown in Fig. 4b, which consists of four convolutional layers (with ReLU activation and BN) and a fully connected layer at the output (with sigmoid activation). The local features are not shift invariant since the position of the feature (the delay and frequency) is the important cue for SSL. Therefore, we do not apply any pooling after convolution. Instead, we apply the filters with a stride of 2, expecting that the network learns its own spatial downsampling.

TSNN-GCCFB (Two-stage neural network with GCCFB): The CNN-GCCFB considers the input features as images without taking their properties into account, which may not yield the best model. Thus, for the third architecture, we design the weight sharing in the network with the knowledge about the GCCFB:

- In each TF bin, there is generally only one predominant speech source, thus we can do analysis or implicit DOA estimation in each frequency band before such information 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.

Based on these considerations, we propose the two-stage neural network (Fig. 5). The first stage extracts latent DOA features in each filter bank, by repeatedly applying *Subnet 1* on individual frequency regions that span all delays and all
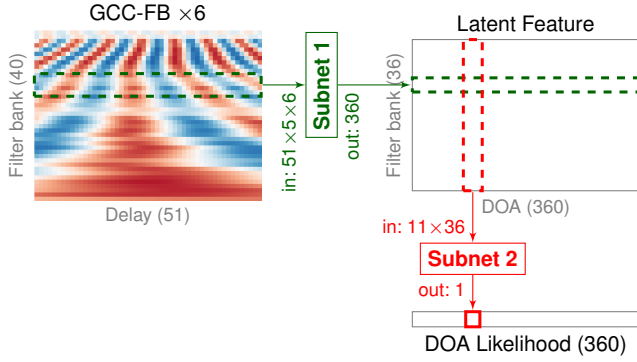
Fig. 5: NN architecture of two-stage neural network with GCCFB as input. The first and second stages are marked as green and red, respectively.

microphone pairs. The second stage aggregates information across all frequencies in a neighbor DOA area and outputs the likelihood of a sound being in each DOA. Similarly, the *Subnet 2* is repeatedly used for all DOAs in the second stage. To train such network, we adopt a two-step training scheme: First, we train the Subnet 1 in the first stage using the DOA likelihood as the desired latent feature. In such way, we obtain DOA and frequency-related features that help the NN to converge to a better result in the next step. During the second step, both stages are trained in an end-to-end manner. In our experiments, Subnet 1 is a 2-hidden-layer MLP, and Subnet 2 is a 1-hidden-layer MLP. All the hidden layers are of size 500.

III. EXPERIMENT

We implemented the proposed methods and compared them to the traditional SSL approaches with the data collected from a robot.

A. Datasets

For the development and evaluation of learning-based SSL methods, we collected two sets of real data: one with loudspeaker and the other with human subjects (see Table II).

We use Pepper for the recording of both sets. There are four microphones on the top of its head, forming a rectangle of 5.8×6.9 cm. The microphones are directional with a forward look direction. The audio signals received by the microphones are strongly affected by the robot's fan noise from inside the head. The sample rate is 48 kHz.

Recording with loudspeakers: We collected data by recording clean speech played from loudspeakers (Fig. 6a). The clean speech data were selected from the AMI corpus [14], which contains spontaneous speech of people interacting in meetings. The loudspeakers were attached with markers so that they can be automatically located by the camera on the robot. The data were recorded in rooms of different sizes, with the robot and loudspeakers put at random places. We programmed the robot to move its head automatically to acquire a large diversity of loudspeaker-to-robot positions.
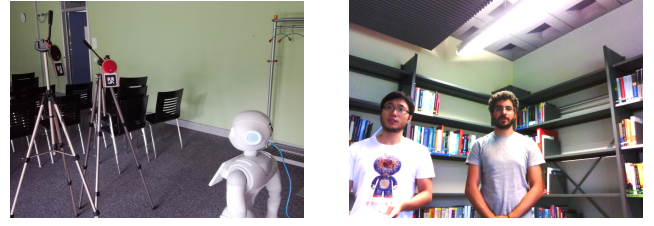
(a) Loudspeakers.

(b) Human subjects.

Fig. 6: Data collection with Pepper.

TABLE II: Specifications of the recorded data

	Loudspeaker		Human
	Training	Test	Test
# of files	4208	2393	21
- single source	2808	1597	–
- two sources	1400	796	21
# of male speakers	105	8	12
# of female speakers	43	8	2
Total duration	16 hours	8 hours	220 seconds
Azimuth (°)	$[-180, 180]$	$[-180, 180]$	$[-24, 23]$
Elevation (°)	$[-39, 56]$	$[-29, 45]$	$[-14, 13]$
Distance (m)	$[0.5, 1.8]$	$[0.5, 1.9]$	$[0.8, 2.1]$

Recording with human subjects: To evaluate SSL methods in real HRI, we collected the second dataset that involves human subjects (Fig. 6b). During the recording, the subjects spoke to the robot with phrases for interactions. This dataset includes recordings with single utterances as well as overlapping ones. We manually annotated the voice activity detection (VAD) labels and automatically acquired the mouth position by running a multiple person tracker [15] with detection from the convolutional pose machine (CPM) [16].

B. Evaluation Protocol

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

Known number of sources: 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). In addition, we consider the *accuracy* (ACC) as the percentage of correct predictions. By saying a prediction is correct, we mean the error of the prediction is less than a given admissible error E_a.

Unknown number of sources: We consider the ability of both detection and localization. To do this, we make predictions based on Eq. 4, and compute the *precision* vs. *recall* curve by varying the prediction threshold ξ. The precision is the percentage of correct predictions among all predictions. And, the recall is the percentage of correct detection out of all ground truth sources.

C. Network Training

We trained the NN with the loudspeaker training set, which includes a total of 506k frames of no source, one source, or two sources. We used the Adam optimizer [17] with mean squared error (MSE) loss and mini-batch size of 256. The

MLP-GCC and CNN-GCCFB were trained for ten epochs. We trained the TSNN-GCCFB for four epochs for the first stage and another ten epochs for the end-to-end training.

D. Baseline Methods

We include the following popular spatial spectrum-based methods for comparison:

- SRP-PHAT: steered response power with phase transform [3];
- SRP-NONLIN: SRP-PHAT with a non-linear modification of the score, it is a multi-channel extension of GCC-NONLIN from [18];
- MVDR-SNR: minimum variance distortionless response (MVDR) beamforming [19] with signal-to-noise ratio (SNR) as score [18];
- SEVD-MUSIC: multiple signal classification (MUSIC) [2], assuming spatially white noise and one signal in each bin;
- GEVD-MUSIC: MUSIC with generalized eigenvector decomposition [2, 20], assuming noise is pre-measured and one signal in each TF bin.

For all the above methods, the empirical spatial covariance matrices are computed using blocks containing 7 frames of 2048 samples with 50% overlap, so that each block is 170ms long.

E. Results

Table III shows the results of localization with a known number of sources. On the loudspeaker dataset, all three proposed NN models achieve on average less than 5° error and more than 90% accuracy, while the best baseline method, (SRP-PHAT) has 21.5° error and only 78% accuracy. For the human subject dataset, the baseline methods have slightly better MAE on frames with a single source. However, the proposed methods outperform the baseline methods in terms of accuracy, especially on frames with overlapping sources. Note that, the loudspeaker dataset is in general more challenging because it contains samples with lower SNR and wider range of azimuth directions. The sources from the rear are difficult to detect due to the directivity of the microphones.

In terms of simultaneous detection and localization with an unknown number of sources, our proposed methods outperform the baseline methods, achieving approximately 90% precision and recall on both datasets (Fig. 7 and 8). Among the three proposed models, the TSNN-GCCFB achieves the best results with its better performance on overlapping frames. This justifies that the usage of the sub-band feature and two-stage structure is beneficial for multiple SSL. We also notice that, unlike signal processing approaches, our NN-based methods are not affected by the condition of an unknown number of sources. This indicates that our output coding and data-driven approach are effective for detecting the number of sources. A demonstration video is accompanied with this paper.

IV. CONCLUSION

This paper has investigated neural network models for simultaneous detection and localization of speakers. We have proposed a likelihood-based output coding, making it possible to train the NN to detect an arbitrary number of overlapping sound sources. We have collected a large amount of real data, including recordings with loudspeakers and humans, for training and evaluation. The results of the comprehensive evaluation show that our proposed methods significantly outperform the traditional spatial spectrum-based methods.

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. Future work will explore network models that can generalize for multiple sound sources with limited training data. We will also explore the robustness of the NN to other more challenging noise, such as cocktail party noise. Furthermore, we will investigate the incorporation of temporal context, which was omitted in our experiments.

REFERENCES

[1] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, Aug. 1976.

[2] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, Mar. 1986.

[3] M. S. Brandstein and H. F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Apr. 1997, pp. 375–378 vol.1.

[4] B. P. Yuhas, "Automated sound localization through adaptation," in *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, vol. 2, Jun. 1992, pp. 907–912 vol.2.

[5] M. S. Datum, F. Palmieri, and A. Moiseff, "An artificial neural network for sound localization using binaural cues," *The Journal of the Acoustical Society of America*, vol. 100, no. 1, pp. 372–383, Jul. 1996.

[6] K. Youssef, S. Argentieri, and J. L. Zarader, "A learning-based approach to robust binaural sound localization," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Nov. 2013, pp. 2927–2932.

[7] N. Ma, G. J. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," *Proceedings of Interspeech 2015*, pp. 3302–3306, 2015.

[8] X. Xiao, S. Zhao, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "A learning-based approach to direction of arrival estimation in noisy and reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2015, pp. 2814–2818.

[9] R. Takeda and K. Komatani, "Sound source localization based on deep neural networks with directional activate function exploiting phase information," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 405–409.

[10] N. Yalta, K. Nakadai, and T. Ogata, "Sound Source Localization Using Deep Learning Models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, Feb. 2017.

[11] R. Takeda and K. Komatani, "Discriminative multiple sound source localization based on deep neural networks using independent location model," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2016, pp. 603–609.

[12] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

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
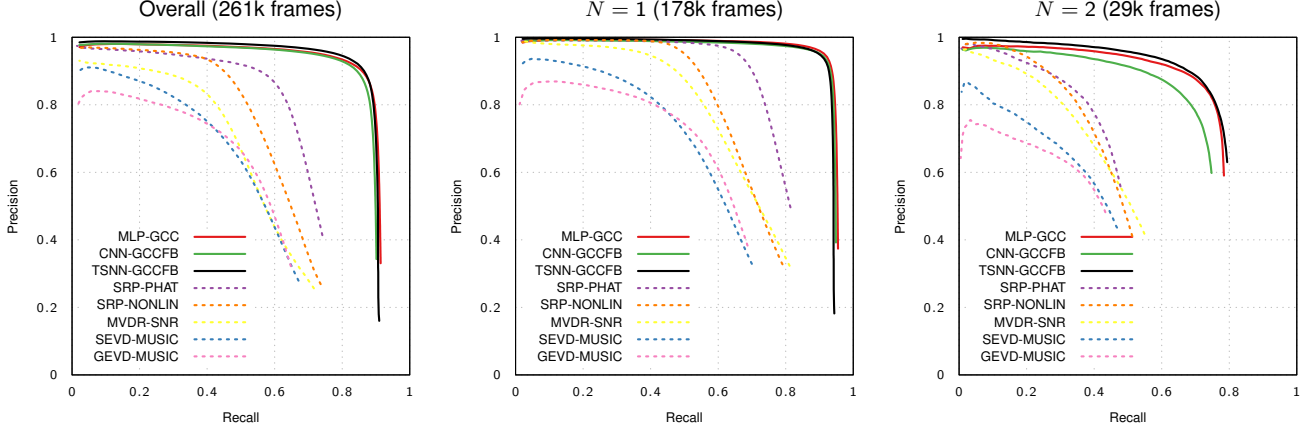
Fig. 7: Detection and localization performance on recordings with loudspeakers. $E_a = 5°$.
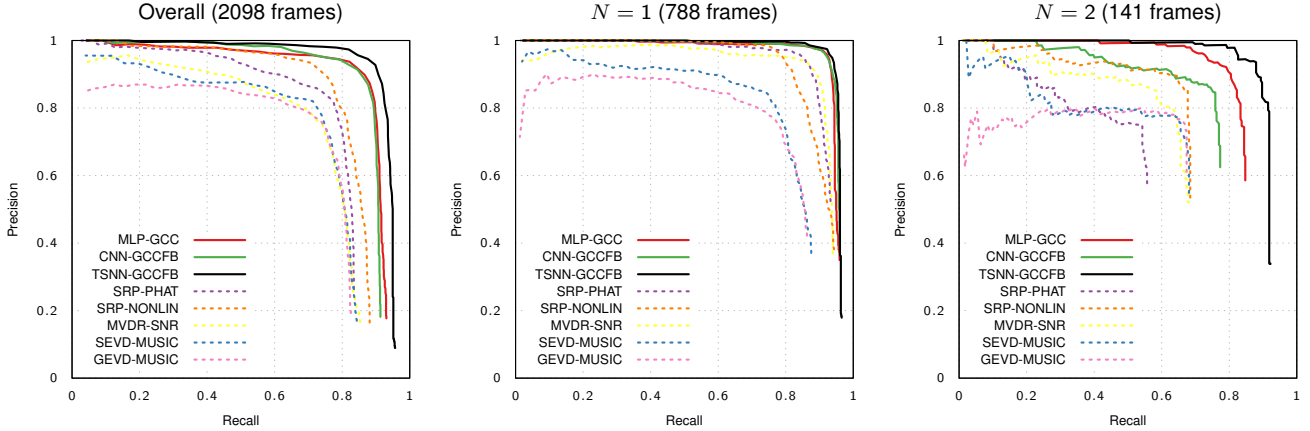
Fig. 8: Detection and localization performance on recordings with human subjects. $E_a = 5°$.

[13] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *PMLR*, Jun. 2015, pp. 448–456.

[14] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, and others, "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[15] V. Khalidov and J.-M. Odobez, "Real-time Multiple Head Tracking Using Texture and Colour Cues," Idiap, Tech. Rep. Idiap-RR-02-2017, 2017.

[16] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional Pose Machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[17] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Dec. 2014, arXiv: 1412.6980.

[18] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA Estimation in Reverberant Audio Using Angular Spectra and Clustering," *Signal Process.*, vol. 92, no. 8, pp. 1950–1960, Aug. 2012.

[19] H. Krim and M. Viberg, "Two decades of array signal processing research: the parametric approach," *IEEE Signal Processing Magazine*, vol. 13, no. 4, pp. 67–94, Jul. 1996.

[20] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2009, pp. 664–669.