Dynamically Localizing Multiple Speakers based on the time-frequency domain

Anonymous authors
Paper under double-blind review

Abstract

In this study we present a deep neural network-based online multi-speaker localisation algorithm based on a multi-microphone array. A fully convolutional network is trained with instantaneous spatial features to estimate the direction of arrival (DOA) for each time-frequency (TF) bin. The high resolution classification enables the network to accurately and simultaneously localize and track multiple speakers, both static and dynamic. Elaborated experimental study using simulated and real-life recordings in static and dynamic scenarios, demonstrates that the proposed algorithm significantly outperforms both classic and recent deep-learning-based algorithms.

1 Introduction

Locating multiple sound sources recorded with a microphone array in an acoustic environment is an essential component in various cases such as source separation and scene analysis. The relative location of a sound source with respect to a microphone array is specified in the term of the DOA of the sound wave originating from that location. DOA estimation and tracking are essential building blocks in all modern far-field speech enhancement and recognition for smart homes devices as well robot audition applications. In real-life environments, sound sources are captured by the microphones together with acoustic reverberation. While propagating in an acoustic enclosure, the sound wave undergoes reflections from the room facets and from various objects. These reflections deteriorate speech quality and, in extreme cases, its intelligibility. Furthermore, reverberation increases the time dependency between speech frames, making source DOA estimation a very challenging task.

A plethora of classic signal processing-based approaches have been proposed throughout the years for the task of broadband DOA estimation. The multiple signal classification (MUSIC) algorithm (Schmidt, 1986) applies a subspace method that was later adapted to the challenges of speech processing in (Dmochowski et al., 2007). The steered response power with phase transform (SRP-PHAT) algorithm (Brandstein & Silverman, 1997) used a generalization of cross-correlation methods for DOA estimation. These methods are still widely in use for both single- and multi-speaker localization tasks. However, in highly reverberant enclosures, their performance rapidly deteriorates.

Supervised learning methods can be potentially advantageous for this task since they are data-driven. Deep neural networks can be trained to find the DOA in different acoustic conditions. Moreover, if a network is trained using rooms with different acoustic conditions and multiple noise types, it can be made robust against noise and reverberation even for rooms which were not in the training set. Deep learning methods have recently been proposed for sound source localization. In (Xiao et al., 2015; Vesperini et al., 2016) simple feed-forward deep neural networks (DNNs) were trained using generalized cross correlation (GCC)-based audio features, demonstrating improved performance as compared with classic approaches. Yet, this method is mainly designed to deal with a single sound source at a time. Takeda & Komatani (2016) trained a DNN for multi-speaker DOA estimation. In high reverberation conditions, however, their performance is not satisfactory. Pujol et al. (2019) and Vera-Diaz et al. (2018) used time-domain features and they have shown performance improvement in highly-reverberant enclosures. Chakrabarty & Habets (2017) applied a convolutional neural network (CNN) based classification method in the short-time Fourier transform (STFT) domain for broadband DOA estimation, assuming that only a single speaker is active per time-frame. The phase component of the STFT coefficients of the input signal were directly provided as input to the CNN.

This work was extended by Chakrabarty & Habets (2019) to estimate multiple speakers' DOAs, and has shown high DOA classification performance.

The main drawback of most DNN-based approaches, however, is that they only use low-resolution supervision, namely at the time-frame level or even utterance-based level, and the network outputs a single localization decision for the entire time-frame. In speech signals, however, each time-frequency bin can be dominated by a different speaker, a property referred to as W-disjoint orthogonality (WDO) (Rickard & Yilmaz, 2002). In case of multiple speakers, each TF bin can therefore be associated with a different DOA. This high-resolution information can yield an improved DOA estimation also for the entire time-frame localization resolution, especially in the case of multiple speakers.

In this study we present a multi-speaker DOA estimation algorithm that is based on the U-net architecture that infers the DOA of each TF bin. The DOA decisions of all the frequency bands of a single time-frame are then aggregated to extract the active speakers at that time-frame. The TF-based classification also facilitates the tracking ability for multiple moving speakers. U-Net has been introduced in the medical imaging domain (Ronneberger et al., 2015a) and was recently successfully applied to various audio processing tasks, e.g. for speech dereverberation (Ernst et al., 2018), speaker separation (Chazan et al., 2019) and noise reduction (Grechkov et al., 2020), all in the STFT domain, and for speech enhancement in the time-domain (Zhang et al., 2019; Giri et al., 2019) also employing self-attention mechanism. In this study we show that U-net is the preferred network also for speaker localization. We tested the proposed method on simulated data, using publicly available room impulse responses (RIRs) recorded in a real room (Hadad et al., 2014), as well as real-life experiments. We show that the proposed algorithm significantly outperforms state-of-the-art competing methods.

The main contribution of our work is casting the time-domain DOA estimation problem into a time-frequency segmentation problem. The proposed method improves the DOA estimation performance with respect to (w.r.t.) the state-of-the-art (SOTA) approaches, which are frame-based, and facilitates simultaneous tracking of multiple moving speakers. Unlike many other domains, the multi-speaker DOA task lacks a standard benchmark. Another contribution of this work is a benchmark data composed of training and test datasets that represent various real-life enclosures.[1]

2 Multiple-speakers localization algorithm

Multi-Microphone Time-frequency features. Consider an array with M microphones acquiring a mixture of N speech sources in a reverberant environment. The i-th speech signal $s^i(t)$ propagates through the acoustic channel before being acquired by the m-th microphone:

$$z_m(t) = \sum_{i=1}^{N} \{s^i * h_m^i\}(t), \qquad m = 1, \dots, M, \tag{1}$$

where h_m^i is the RIR relating the i-th speaker and the m-th microphone. In the STFT domain (1) can be written as (provided that the frame-length is sufficiently large w.r.t. the filter length):

$$z_m(l,k) = \sum_{i=1}^{N} s^i(l,k) h_m^i(l,k), \tag{2}$$

where l and k, are the time-frame and the frequency indices, respectively.

The STFT (2) is complex-valued and hence comprises both spectral and phase information. It is clear that the spectral information alone is insufficient for DOA estimation. It is therefore a common practice to use the phase of the TF representation of the received microphone signals, or their respective phase-difference, as they are directly related to the DOA in non-reverberant environments. We decided to use an alternative feature, which is generally independent of the speech signal and is mainly determined by the spatial information. For that, we have selected the relative transfer function (RTF) (Gannot et al., 2001) as our feature, since it is known to encapsulate the spatial fingerprint for each sound source. Specifically, we use the instantaneous relative transfer function (iRTF), which is the bin-wise ratio between the m-th microphone signal and the reference microphone signal $z_{\text{ref}}(l,k)$:

$$\text{iRTF}(m,l,k) = \frac{z_m(l,k)}{z_{\text{ref}}(l,k)}. \tag{3}$$

[1] The benchmark data will become publicly available upon paper acceptance.

Table 1: The TF-DOAnet multi-speaker localization algorithm.

- Compute the iRTF features from the multi-microphone recordings.
- Apply the U-net network to classify each TF bin to one of the possible DOAs.
- Based on the U-net results, decide the locations of the active speakers at each time frame.

Note, that the reference microphone is arbitrarily chosen. Reference microphone selection is beyond the scope of this paper (see Stenzel et al. (2014) for a reference microphone selection method). The input feature set extracted from the recorded signal is thus a 3D tensor $\mathcal{R}$:

$$\mathcal{R}(l,k,m) = [\mathfrak{Re}(\text{iRTF}(m,l,k)), \mathfrak{Im}(\text{iRTF}(m,l,k))]. \tag{4}$$

The matrix $\mathcal{R}$ is constructed from $L \times K$ bins, where L is the number of time-frames and K is the number of frequencies. Since the iRTFs are normalized by the reference microphone, the latter is excluded from the features. Then for each TF bin (l,k), there are $P = 2(M-1)$ channels, where the multiplication by 2 is due to the real and imaginary parts of the complex-valued feature. For each TF bin the spatial features were normalized to have a zero mean and a unit variance. Other feature extraction methods can be considered. In Section 3, we show that the features described above are a suitable choice for the localization task.

U-Net for DOA estimation. The WDO assumption (Rickard & Yilmaz, 2002) implies that each TF bin (l,k) is dominated by a single speaker. Consequently, as the speakers are spatially separated, i.e. located at different DOAs, each TF bin is dominated by a single DOA. We first accurately estimate the speaker direction at every TF bin from the given mixed recorded signal. Then, we extract the speakers' locations at each time-frame.

We formulated the DOA estimation as a classification task by discretizing the DOA range. The resolution was set to $5°$, such that the DOA candidates are in the set $\Theta = \{0°, 5°, 10°, \ldots, 180°\}$. Let $D_{l,k}$ be a random variable (r.v.) representing the active dominant direction, recorded at bin (l,k). Our task boils down to deducing the conditional distribution of the discrete set of DOAs in Θ for each TF bin, given the recorded mixed signal:

$$\mathcal{P}_{l,k}(\theta) = p(D_{l,k} = \theta | \mathcal{R}), \quad \theta \in \Theta. \tag{5}$$

For this task, we use a DNN. The network output is an $L \times K \times |\Theta|$ tensor, where $|\Theta|$ is the cardinality of the set Θ. Under this construction of the feature tensor and output probability tensor, a pixel-to-pixel approach Badrinarayanan et al. (2017) for mapping a 3D input 'image', $\mathcal{R}$ and a 3D output 'image', $\mathcal{P}$, can be utilized. A U-net is used to compute (5) for each TF bin. The pixel-to-pixel method is beneficial in two ways. First, for each TF bin in our input image the network estimates the DOA distribution separately. Second, the TF supervision is carried out with the spectrum of the different speakers. The U-Net hence takes advantage of the spectral structure and the continuity of the sound sources in both the time and frequency axes. These structures contribute to the pixel-wise classification task, and prevent discontinuity in the DOA decisions over time. In our implementation, we used a U-net architecture, similar to the one described in (Ronneberger et al., 2015b).

The input to the network is the feature tensor $\mathcal{R}$ (see (4)). In our U-net architecture, the input shape is (L, K, P) where $K = 256$ is the number of frequency bins, $L = 256$ is the number of frames, and $P = 2M - 2$ where M is the number of microphones.

TF bins in which there is no active speech are non-informative. Therefore, the estimation is carried out only on speech-active TF bins. As we assume that the acquired signals are noiseless, we define a TF-based voice activity detector (VAD) as follows:

$$\text{VAD}(l,k) = \left\{ \begin{array}{ll} 1 & |z_{\text{ref}}(l,k)| \geq \epsilon \\ 0 & \text{o.w.} \end{array} \right. , \tag{6}$$

where ϵ is a threshold value. In noisy scenarios, we can use a robust speech presence probability (SPP) estimator instead (Wang & Chen, 2018).

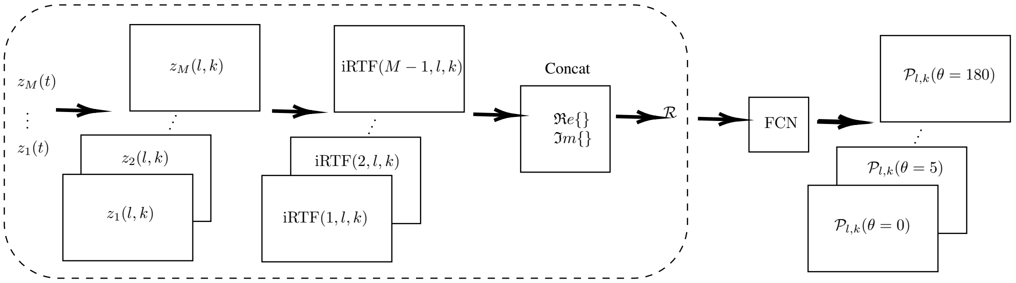


Figure 1: Block diagram of the TF-DOAnet algorithm. The dashed envelope describes the feature extraction step.

Table 2: Configuration of training data generation. All rooms are 2.7 m in height.

	Simulated training data				
	Room 1	Room 2	Room 3	Room 4	Room 5
Room size	(6×6) m	(5×4) m	(10×6) m	(8×3) m	(8×5) m
RT$_{60}$	0.3 s	0.2 s	0.8 s	0.4 s	0.6 s
Signal	Noiseless signals from WSJ1 **training** database				
Array position in room	6 arbitrary positions in each room				
Source-array distance	1.5 m with added noise with 0.1 variance				

The DOAs should only be estimated on a time-frame basis. Hence, we aggregate over all active frequencies at time-frame l to obtain a frame-wise probability:

$$p_l(\theta) = \frac{1}{K'} \sum_{k=1}^{K} \mathcal{P}_{l,k}(\theta) \text{VAD}(l,k), \qquad \theta \in \Theta \tag{7}$$

where K' is the number of frequency bands for which (6) exceed the threshold at the l-th time frame. We thus obtain for each time-frame a posterior distribution over all possible DOAs. If the number of speakers is known in advance, we can choose the directions corresponding to the highest posterior probabilities. If an estimate of the number of speakers is also required, it can be determined by applying a proper threshold. We dub our algorithm time-frequency direction-of-arrival net (TF-DOAnet). Figure 1 summarizes the TF-DOAnet network architecture. The algorithm is summarized in Table 1.

Model training. The supervision in the training phase is based on the WDO assumption in which each TF bin is dominated by (at most) a single speaker. The training is based on simulated data generated by a publicly available RIR generator software,[2] efficiently implementing the image method (Allen & Berkley, 1979). A four microphone linear array was simulated with $(8, 8, 8)$ cm inter-microphones distances. Similar microphone inter-distances were used in the test phase. For each training sample, the acoustic conditions were randomly drawn from one of the simulated rooms of different sizes and different reverberation levels RT$_{60}$ as described in Table 2. The microphone array was randomly placed in the room in one out of six arbitrary positions.

For each scenario, two clean signals were randomly drawn from the Wall Street Journal 1 (WSJ1) database (Paul & Baker, 1992) and then convolved with RIRs corresponding to two different DOAs in the range $\Theta = \{0, 5, \ldots, 180\}$. The sampling rate of all signals and RIRs was set to 16KHz. The speakers were positioned on a radius of $r = 1.5$ m from the center of the microphone array. To enrich the training diversity, the radius of the speakers was perturbed by a Gaussian noise with a variance of 0.1 m. The DOA of each speaker was calculated w.r.t. the center of the microphone array.

The contributions of the two sources were then summed with a random signal to interference ratio (SIR) selected in the range of SIR $\in [-2, 2]$ to obtain the received microphone signals. Next,

[2] Available online at `github.com/ehabets/RIR-Generator`.

Table 3: Configuration of test data generation. All rooms are 3 m in height.

	Simulated test data	
	Room 1	Room 2
Room size	(5×7) m	(9×4) m
RT$_{60}$	0.38 s	0.7 s
Source-array distance	1.3 m	1.7 m
Signal	Noiseless signals from WSJ1 **test** database	
Array position in room	4 arbitrary positions in each room	

we calculated the STFT of both the mixture and the STFT of the separate signals with a frame-length $K = 512$ and an overlap of 75% between two successive frames.

We then constructed the audio feature tensor $\mathcal{R}$ as described above. In the training phase, both the location and a clean recording of each speaker were known, hence they could be used to generate the labels. For each TF bin (l, k), the dominant speaker was determined by:

$$\text{dominant speaker} \leftarrow \underset{i}{\operatorname{argmax}}\, |s^i(l,k) h^i_{\text{ref}}(l,k)|. \tag{8}$$

The ground-truth label $D_{l,k}$ is the DOA of the dominant speaker. The training set comprised four hours of recordings with 30000 different scenarios of mixtures of two speakers. It is worth noting that as the length of each speaker recording was different, the utterances may also include non-speech or single-speaker frames. The network was trained to minimize the cross-entropy between the correct and the estimated DOA. The cross-entropy cost function was summed over all the images in the training set. The network was implemented in Tensorflow with the ADAM optimizer (Kingma & Ba, 2014). The number of epochs was set to be 100, and the training stopped after the validation loss increased for 3 successive epochs. The mini-batch size was set to be 64 images.

3 Experimental Study

Datasets We evaluated the TF-DOAnet and compared its performance to both classic and DNN-based algorithms. To objectively evaluate the performance of the TF-DOAnet, we first simulated two rooms that were different from the rooms in the training set. Then, we tested our TF-DOAnet with real RIR recordings in different rooms. Finally, a real-life scenario with fast moving speakers was recorded and tested. For each test scenario, we selected two speakers from the test set of the WSJ1 database (Paul & Baker, 1992), placed them at two different angles between $0°$ and $180°$ relative to the microphone array, at a distance of either 1m or 2m. The signals were generated by convolving the signals with RIRs corresponding to the source positions and with either simulated or recorded acoustic scenarios. The SIR was tested in accordance with the DOA literature.

Performance measures Two different measures to objectively evaluate the results were used: the mean absolute error (MAE) and the localization accuracy (Acc.). The MAE, computed between the true and estimated DOAs for each evaluated acoustic condition, is given by

$$\text{MAE}(^\circ) = \frac{1}{N \cdot C} \sum_{c=1}^{C} \min_{\pi \in S_N} \sum_{n=1}^{N} |\theta^c_n - \hat{\theta}^c_{\pi(n)}|, \tag{9}$$

where N is the number of simultaneously active speakers and C is the total number of speech mixture segments considered for evaluation for a specific acoustic condition. The term π is the permutation and S_N represents the permutation possibilities. The true and estimated DOAs for the n-th speaker in the c-th mixture are denoted by θ^c_n and $\hat{\theta}^c_n$, respectively.

The localization accuracy is given by

$$\text{Acc.}(\%) = \frac{\hat{C}_{\text{acc.}}}{C} \times 100 \tag{10}$$

where $\hat{C}_{\text{acc.}}$ denotes the number of speech mixtures for which the localization of the speakers is accurate. We considered the localization of speakers for a speech frame to be accurate if the angular distance between the true and the estimated DOA for all the speakers was less than or equal to 5°.

Compared algorithms We compared the performance of the TF-DOAnet with two frequently used baseline methods, namely the MUSIC and SRP-PHAT algorithms. In addition, we compared its performance with the CNN multi-speaker DOA (CMS-DOA) estimator (Chakrabarty & Habets, 2019).[3] To facilitate the comparison, the MUSIC pseudo-spectrum was computed for each frequency subband and for each STFT time-frame, with an angular resolution of 5° over the entire DOA domain. Then, it was averaged over all frequency subbands to obtain a broadband pseudo-spectrum followed by averaging over all the time frames L. Next, the two DOAs with the highest values were selected as the final DOA estimates. Similar post-processing was applied to the computed SRP-PHAT pseudo-likelihood for each time-frame.

Static simulated scenario We first generated a test dataset with simulated RIRs. Two different rooms were used, as described in Table 3. For each scenario, two speakers (male or female) were randomly drawn from the WSJ1 test database, and placed at two different DOAs within the range $\{0, 5, \ldots, 180\}$ relative to the microphone array. Since the length of each speaker recording is different, the test dataset also includes non-speech or single-speaker frames. We assume the minimum angle between 2 speakers to be 20°, which, for the radius of ≈ 1.5 meters from the microphone array implies that the speakers are practically standing shoulder to shoulder. Each speaker has a different signal length in the mixture. The microphone array was similar to the one used in the training phase. The assumption that we are familiar with the microphone array is fairly common and realistic. For instance, the microphone array in a conference room, in smart devices, or even in phones, is known in advance. Using the RIR generator, we generated the RIR for the given scenario and convolved it with the speakers' signals.

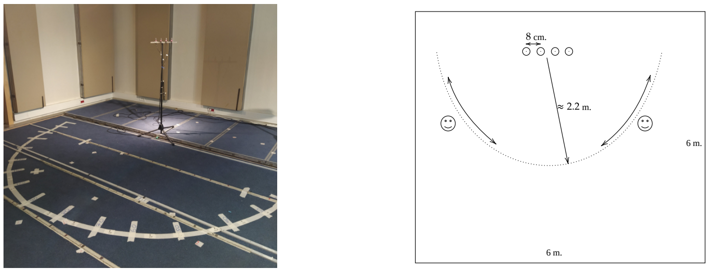
The results for the TF-DOAnet compared with the competing methods are depicted in Table 4. The tables demonstrate that the deep-learning approaches outperform the classic approaches. The TF-DOAnet achieved very high scores and outperforms the DNN-based CMS-DOA algorithm in terms of both MAE and accuracy. Note, that the results in Table 4 are reported at a frame-based resolution, where each frame may consist one or two speakers.

Static real recordings scenario The best way to evaluate the capabilities of the TF-DOAnet is testing it with real-life scenarios. For this purpose, we first carried out experiments with real measured RIRs from a multi-channel impulse response database (Hadad et al., 2014). The database comprises RIRs measured in an acoustics lab for three different reverberation times of $\text{RT}_{60} = 0.160, 0.360$, and 0.610 s. The lab dimensions are $6 \times 6 \times 2.4$ m.

The recordings were carried out with different DOA positions in the range of $[0^\circ, 180^\circ]$, in steps of 15°. The sources were positioned at distances of 1 m and 2 m from the center of the microphone array. The recordings were carried out with a linear microphone array consisting of 8 microphones with three different microphone spacings. For our experiment, we chose the [8, 8, 8, 8, 8, 8, 8] cm setup. In order to construct an array setup identical to the one in the training phase, we selected a sub-array of the four center microphones out of the total 8 microphones in the original setup. Consequently, we used a uniform linear array (ULA) with $M = 4$ elements with an inter-microphone distance of 8 cm.

The results for the TF-DOAnet compared with the competing methods are depicted in Table 5. Again, the TF-DOAnet outperforms all competing methods, including the CMS-DOA algorithm. Interestingly, for the 1 m case, the best results for the TF-DOAnet were obtained for the highest reverberation level, namely $\text{RT}_{60} = 610$ ms, and for the 2 m case, for $\text{RT}_{60} = 360$ ms. While surprising at the first glance, this can be explained using the following arguments. There is an accumulated evidence that reverberation, if properly addressed, can be beneficial in speech processing, specifically for multi-microphone speech enhancement and source extraction (Gannot et al., 2001; Markovich-Golan et al., 2009; Dokmanić et al., 2015) and for speaker localization (Deleforge et al., 2015; Laufer-Goldshtein et al., 2016). In reverberant environments, the intricate acoustic propagation pattern constitutes a specific "fingerprint" characterizing the location of the speaker(s). When reverberation level increases, this fingerprint becomes more pronounced and is actually more

[3] The trained model is available here `https://github.com/Soumitro-Chakrabarty/Single-speaker-localization`



(a) Room view.

(b) Speakers' trajectory.

Figure 2: Real-life experiment setup.

Table 4: Results for two different test rooms with simulated RIRs.

Test Room	Room 1		Room 2	
Measure	MAE	Acc.	MAE	Acc.
MUSIC (Dmochowski et al., 2007)	27.95	28.34	31.62	18.38
SRP-PHAT (Brandstein & Silverman, 1997)	28.25	27.23	36.61	36.28
CMS-DOA (Chakrabarty & Habets, 2019)	12.87	73.09	24.0	39.25
TF-DOAnet (our algorithm)	**1.58**	**97.45**	**2.76**	**93.0**

informative than its an-echoic counterpart. An inference methodology that is capable of extracting the essential driving parameters of the RIR will therefore improve when the reverberation is higher. If the acoustic propagation becomes even more complex, as is the case of high reverberation and a remote speaker, a slight performance degradation may occur, but as evident from the localization results, for sources located 2 m from the array, the performance for $\text{RT}_{60} = 610$ ms is still better than the performance for $\text{RT}_{60} = 160$ ms.

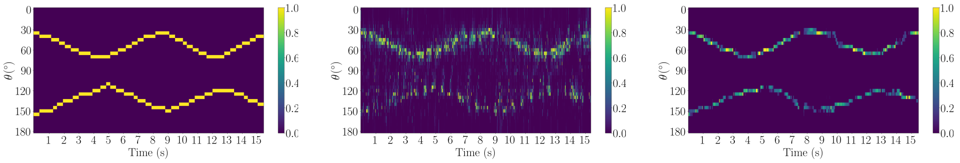
It is worth noting that the test samples were not part of the training phase. The network was not fine-tuned for these test conditions. Yet, since we trained the network with the same RIR generator (with different conditions) it is likely that the results on the simulated test set will be high. The RIR generator cannot capture the accurate sound propagation in real acoustic environments. Therefore, with real recordings, the network performance is likely to be inferior.

Real-life dynamic scenario To further assess the capabilities of the TF-DOAnet, we also carried out real dynamic scenarios experiments. Two rooms with different reverberation level, 390 ms and 720 ms, where used. The microphone array consisted of 4 microphones with an inter-microphone spacing of 8 cm. The speakers walked naturally on an arc at a distance of about 2.2 m from the center of the microphone array. For each RT_{60} two experiments were recorded. The two speakers started at the angles 20° and 160° and walked until they reached 70° and 100°, respectively, turned back and walked to their starting point. This was done several times throughout the recording. Figure 2a depicts the real-life experiment setup and Fig. 2b depicts a schematic diagram of the setup of this experiment. The ground truth labels of this experiment were measured with the Marvelmind indoor 3D tracking set.[4]

Figures 3 and 4 depict the results of the two experiments. It is clear that the TF-DOAnet outperformed the CMS-DOA algorithm, especially for the high RT_{60} conditions. Whereas the CMS-DOA fluctuated rapidly, the TF-DOAnet output trajectory was smooth and noiseless.

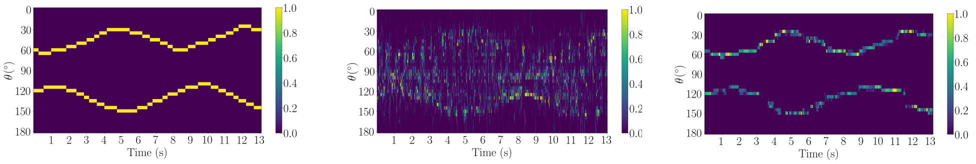
A major component of our system is the pre-processing of the multi-microphone recordings. We used the real and imaginary part of the iRTF (4). Wang et al. (2018) added the power spectral density (PSD) of the speech signal to the spatial features. We next check whether the PSD improves the performance. First, we used the proposed features as described in (4). The compared approach was a variant of our approach with the spectrum added ('TF-DOAnet with Spec.'). All features were crafted from the same training data described in Sec. 2. We tested the different approaches in the test

[4] `https://marvelmind.com/product/starter-set-ia-02-3d/`



(a) Ground truth. (b) CMS-DOA. (c) The TF-DOAnet.

Figure 3: Real-life recording of two moving speakers in a $6 \times 6 \times 2.4$ room with $\text{RT}_{60} = 390$ ms.



(a) Ground truth. (b) CMS-DOA. (c) The TF-DOAnet.

Figure 4: Real-life recording of two moving speakers in a $6 \times 6 \times 2.4$ room with $\text{RT}_{60} = 720$ ms.

Table 5: Results for three different rooms at distances of 1 m and 2 m with measured RIRs.

Distance	1 m						2 m					
RT_{60}	0.160 s		0.360 s		0.610 s		0.160 s		0.360 s		0.610 s	
Measure	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.	MAE	Acc.
MUSIC	18.7	57.6	19.2	53.2	21.9	42.9	18.4	54.1	26.1	35.8	25.4	32.2
SRP-PHAT	9.0	39.0	13.9	39.4	18.6	29.9	9.7	36.0	16.5	24.7	27.7	21.3
CMS-DOA	1.6	76.3	7.3	75.2	8.4	71.9	5.1	79.5	9.7	60.1	17.5	40.0
TF-DOAnet	**1.3**	**97.5**	**3.5**	**83.5**	**0.9**	**98.3**	**5.0**	**89.5**	**1.7**	**95.7**	**4.8**	**84.2**

conditions described in Table 3. In this experiment we used frames with two speakers active, namely $N = 2$.

First, all the features with our high resolution TF model outperformed the frame-based CMS-DOA algorithm, as reported in Table 4. This confirms that the TF supervision is beneficial for the task at hand. Finally, it is interesting to note that adding the PSD features slightly deteriorates the localization results.

Table 6: Ablation study results with different features.

Test Room	Room 1		Room 2	
Measure	MAE	Acc.	MAE	Acc.
TF-DOAnet with Spec.	0.6	98.4	3.3	86.7
TF-DOAnet	**0.3**	**99.5**	**1.7**	**94.3**

4 Conclusions

A joint Time-frequency approach was presented in this paper for the DOA estimation task. Instantaneous RTF features were used to train the model. The high TF resolution facilitated the simultaneous tracking of multiple moving speakers. A comprehensive experimental study was carried out with both simulated and real-life recordings. The proposed approach outperformed both the classic and CNN-based SOTA algorithms in all experiments. Training and test datasets with different real-life scenarios were constructed as a DOA benchmark and will become available upon publication.

REFERENCES

Jont B. Allen and David A. Berkley. Image method for efficiently simulating small-room acoustics. *The Journal of the Acoustical Society of America*, 65(4):943–950, 1979.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

Michael S. Brandstein and Harvey F. Silverman. A robust method for speech signal time-delay estimation in reverberant rooms. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1997.

Soumitro Chakrabarty and Emanuël A. P. Habets. Broadband DOA estimation using convolutional neural networks trained with noise signals. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

Soumitro Chakrabarty and Emanuël A. P. Habets. Multi-speaker DOA estimation using deep convolutional networks trained with noise signals. *IEEE Journal of Selected Topics in Signal Processing*, 13(1):8–21, 2019.

Shlomo E. Chazan, Hodaya Hammer, Gershon Hazan, Jacob Goldberger, and Sharon Gannot. Multi-microphone speaker separation based on deep DOA estimation. In *European Signal Processing Conference (EUSIPCO)*, 2019.

Antoine Deleforge, Florence Forbes, and Radu Horaud. Acoustic space learning for sound-source separation and localization on binaural manifolds. *International journal of neural systems*, 25(01): 1440003, 2015.

Jacek P. Dmochowski, Jacob Benesty, and Sofiene Affes. Broadband music: Opportunities and challenges for multiple source localization. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2007.

Ivan Dokmanić, Robin Scheibler, and Martin Vetterli. Raking the cocktail party. *IEEE journal of selected topics in signal processing*, 9(5):825–836, 2015.

Ori Ernst, Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger. Speech dereverberation using fully convolutional networks. In *The 26th European Signal Processing Conference (EUSIPCO)*, pp. 390–394, 2018.

Sharon Gannot, David Burshtein, and Ehud Weinstein. Signal enhancement using beamforming and nonstationarity with applications to speech. *IEEE Transactions on Signal Processing*, 49(8): 1614–1626, 2001.

Ritwik Giri, Umut Isik, and Arvindh Krishnaswamy. Attention wave-U-net for speech enhancement. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 249–253, 2019.

S. D. Grechkov, V. P. Semenov, and A. A. Bezrukov. Comparative analysis of the usage of neural networks for sound processing. In *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pp. 1389–1391, 2020.

Elior Hadad, Florian Heese, Peter Vary, and Sharon Gannot. Multichannel audio database in various acoustic environments. In *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bracha Laufer-Goldshtein, Ronen Talmon, and Sharon Gannot. Semi-supervised sound source localization based on manifold regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(8):1393–1407, 2016.

Shmulik Markovich-Golan, Sharon Gannot, and Israel Cohen. Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals. *IEEE Tran. on Audio, Speech, and Language Processing*, 17(6):1071–1086, August 2009. ISSN 1558-7916. doi: 10.1109/TASL.2009.2016395.

Douglas B. Paul and Janet M. Baker. The design for the Wall Street Journal-based CSR corpus. In *Workshop on Speech and Natural Language*, 1992. URL `https://www.aclweb.org/anthology/H92-1073`.

Hadrien Pujol, Eric Bavu, and Alexandre Garcia. Source localization in reverberant rooms using deep learning and microphone arrays. In *International Congress on Acoustics (ICA)*, 2019.

Scott Rickard and Ozgiir Yilmaz. On the approximate w-disjoint orthogonality of speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015a.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 2015b.

Ralph Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

Sebastian Stenzel, Jürgen Freudenberger, and Gerhard Schmidt. A minimum variance beamformer for spatially distributed microphones using a soft reference selection. In *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2014.

Ryu Takeda and Kazunori Komatani. Discriminative multiple sound source localization based on deep neural networks using independent location model. *IEEE Spoken Language Technology Workshop (SLT)*, 2016.

Juan Manuel Vera-Diaz, Daniel Pizarro, and Javier Macias-Guarasa. Towards end-to-end acoustic localization using deep learning: From audio signals to source position coordinates. *Sensors*, 18 (10):3418, 2018.

Fabio Vesperini, Paolo Vecchiotti, Emanuele Principi, Stefano Squartini, and Francesco Piazza. A neural network based algorithm for speaker localization in a multi-room environment. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016.

DeLiang Wang and Jitong Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

Zhong-Qiu Wang, Jonathan Le Roux, and John R. Hershey. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

Xiong Xiao, Shengkui Zhao, Xionghu Zhong, Douglas L Jones, Eng Siong Chng, and Haizhou Li. A learning-based approach to direction of arrival estimation in noisy and reverberant environments. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

Y. Zhang, Q. Duan, Y. Liao, J. Liu, R. Wu, and B. Xie. Research on speech enhancement algorithm based on SA-Unet. In *The 4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, pp. 818–8183, 2019.