Multi-Speaker Localization Using Convolutional Neural Network Trained with Noise

Soumitro Chakrabarty
International Audio Laboratories
Erlangen,* Germany

Emanuël A. P. Habets
International Audio Laboratories
Erlangen, Germany

Abstract

The problem of multi-speaker localization is formulated as a multi-class multi-label classification problem, which is solved using a convolutional neural network (CNN) based source localization method. Utilizing the common assumption of disjoint speaker activities, we propose a novel method to train the CNN using synthesized noise signals. The proposed localization method is evaluated for two speakers and compared to a well-known steered response power method.

1   Introduction

In microphone array processing, the source location is an important parameter, which is generally unavailable and needs to be estimated. The location of the source with respect to the array is often given in terms of the direction-of-arrival (DOA) of the sound wave originating from the source position. Over the years, many array processing based methods have been proposed for the task of DOA estimation [1, 2, 3, 4]. Most of these methods however suffer from degradation in performance in reverberant and noisy conditions [5].

Supervised learning methods, being data-driven, provide a distinct advantage for this task, namely they can be adapted to different acoustic conditions via training. If training data from varying acoustic conditions are available, then these methods can also be made robust against adverse acoustic conditions. Recently, several supervised learning methods have been proposed for the task of sound source localization [6, 7, 8]. In [9], the current authors presented a convolutional neural network (CNN) [10, 11] based supervised learning method for the task of single speaker localization. The CNN was trained with synthesized noise signals, which enabled the creation of large amount of training data in a much more convenient manner than using real-world signals. However, for the case of multi-speaker localization, since the STFT phase components of individual signals are not additive for multiple simultaneously active speakers, the extension of the idea of training the CNN with synthesized noise signals is not straightforward.

To be able to train a CNN for multi-speaker localization using synthesized noise signals, we propose to use the assumption that speakers are not simultaneously active per time-frequency. This assumption is know as W-disjoint orthogonality, has been shown to hold approximately for speech signals [12], and is commonly used in speech separation.

Following a brief introduction to the complete system, we describe the procedure for creating the training data with synthesized noise signals for multi-speaker localization. In addition, we also provide preliminary results from simulated experiments.

---

*A joint institution of the Friedrich-Alexander-University Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS). E-mail: `firstname.lastname@audiolabs-erlangen.de`
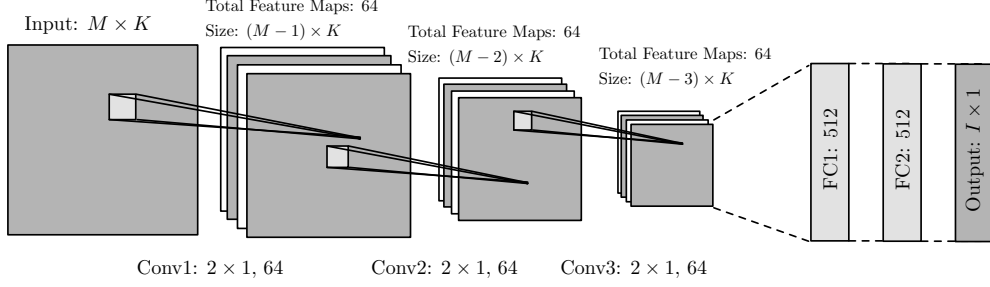
Figure 1: CNN architecture.

2 Speaker localization with CNNs

Considering an independent source DOA model, we formulate the problem of multi-speaker DOA estimation as a multi-class multi-label classification. The number of classes, I, and the class vector is formed based on a discretized set of possible DOA values, similar to [9].

The input representation chosen in this work is the same as [9], where the phase component of the STFT coefficients of the signal are given in the form of a matrix of size $M \times K$, where M and K are the number of microphones and frequency sub-bands, respectively.

The convolutional neural network architecture used in this work is shown in Fig. 1. In contrast to the architecture presented in [9], we use local filters of size 2×1, which leads to each filter learning from the phase correlations between the neighboring microphones for each frequency sub-band separately. This is done since we want to utilize the disjoint activity of speech signals for localization. Also, since we formulate our problem as a multi-label classification, the last layer of the CNN now consists of I sigmoid activations.

3 Generating the training data

For speech signals, it is commonly assumed that the time-frequency (TF) representation of two simultaneously active sources do not overlap. We utilize this assumption to generate training data from synthesized noise signals. In the following, we explain the procedure for generating the training data for two speaker localization per STFT time frame.

As a first step, we generate the training signals for the single speaker case for different acoustic conditions, as explained in [9]. Then, for a specific source array setup, two multi-channel training signals, corresponding to different DOAs, are concatenated along the time axis. Following this, for each frequency sub-band separately, the time-frequency bins for all microphones are randomized to get a single training signal. This procedure is repeated for all combinations of DOAs for all different acoustic conditions considered for training. Finally, the phase map corresponding to each time frame for all training signals is extracted to form the complete training dataset.

While generating the training data, there are two important things to note regarding the randomization process. First, it is essential that the randomization of the TF bins is done separately for each frequency sub-band, such that the order of the frequency sub-bands remains the same for different time frames. Secondly, it is essential that for each frequency sub-band, the TF bins for all the microphones are randomized together, such that phase relations between the microphones for the individual TF bins are preserved.

An illustration of this procedure is shown in Fig. 3. From the rightmost figure, we can see that following the randomization procedure, at each time frame there are approximately equal number of TF bins with activity corresponding to the two DOAs. Therefore, at each frequency sub-band of the phase map input to the CNN, the phase of the STFT coefficients for all microphones correspond to a single source. This makes the assumption of disjoint activity of signals implicit within our framework.
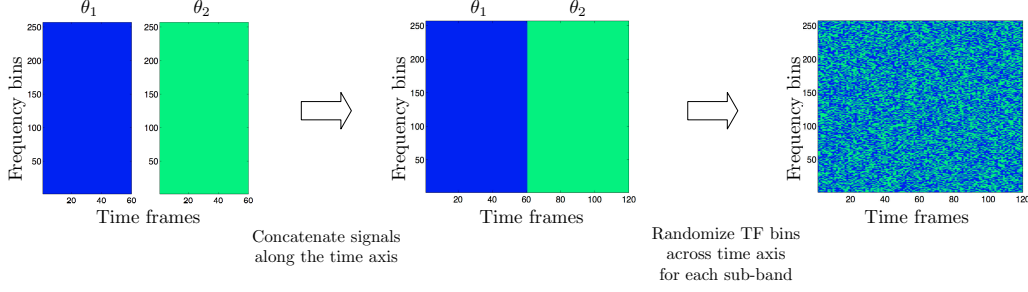
2

Figure 2: Illustration of the procedure for generating the training data. The STFT representation of two spectrally white noise signals corresponding to DOAs θ_1 and θ_2, of size $N \times K \times M$, where N, K, and M denote the number of time frames, frequency sub-bands and microphones, respectively are concatenated along the time axis and randomized to get the training signal for the specific DOA combination. From this, for each time frame the phase map is extracted to get the multiple training data samples for this combination. **Note**: Though the microphone dimension is not shown, the TF bins for all microphones are randomized simultaneously.

With this training input, the CNN can learn the relevant features for localizing multiple speakers at each time frame from the individual TF bins that contain the phase relations across the microphones for each source DOA separately.

4 Experimental results

We evaluated the performance of the noise trained CNN for the task of DOA estimation of two sources over a complete speech mixture. The posterior probabilities for each DOA class obtained from the CNN output at each time frame are averaged over all the frames, and then normalized to 1. Then the final DOA estimates are obtained by choosing the DOAs corresponding to the classes with the two highest averaged posterior probabilities. The performance was compared to the well-known Steered Response Power with the PHAse Transform (SRP-PHAT) method [3] with similar post-processing applied to the obtained frame-level probabilities. As an objective measure, we used the mean absolute error (MAE) between the true and estimated DOAs over all the speech mixtures in the test dataset.

We consider a ULA with $M = 4$ microphones with inter-microphone distance of 8 cm, and the input signals are transformed to the STFT domain using a DFT length of 512, with 50% overlap. To form the classes, we discretize the whole DOA range of a ULA with a $5°$ resolution to get $I = 37$ DOA classes. The room impulse responses (RIRs) required to simulate different acoustic conditions are generated using the RIR generator [13].

The details regarding the training acoustic conditions is given in Table 1. The training data is generated as explained in Section 3. Spatially uncorrelated Gaussian noise was added to the training signals with randomly chosen noise levels between 0 and 20 dB before extracting the phase maps. In total, the training data consisted of around 12.4 million time frames for all the different DOA combinations. We used cross-entropy as the loss function and the CNN was trained using the Adam gradient-based optimizer [14], with mini-batches of 512 time frames. During training, at the end of

Simulated training data	
Signal	Synthesized noise signals
Room size	R1: (6×6) m , R2: (5×4) m, R3: (10×6) m, R4: (8×3) m, R5: (8×5) m
Array positions in room	7 different positions in each room
Source-array distance	1 m and 2 m for each position
RT_{60}	R1: 0.3 s, R2: 0.2 s, R3: 0.8 s, R4: 0.4 s, R5: 0.6 s
SNR	Uniformly sampled from 0 to 20 dB

Table 1: Configuration for training data generation. All rooms are 2.7 m high.

Simulated test data	
Signal	Speech signals from TIMIT
Room size	$(9 \times 4 \times 3)$ m
Array positions in room	1 arbitrary position
Source-array distance	1.8 m
RT_{60}	0.70 s

(a)

SNR	10 dB	20 dB	30 dB
Proposed	14.3	6.1	1.8
SRP-PHAT	27.1	21.6	18.2

(b)

Table 2: (a) Configuration for test data. (b) Mean absolute error ($°$) for different levels of spatially white noise.

the three convolution layers and after each fully connected layer, a dropout procedure [15] with a rate of 0.5 was used to avoid overfitting.

To evaluate the performance for all possible DOA combinations, our test dataset consisted of 666 speech signal mixtures, each of length 2 seconds, each corresponding to a specific DOA combination. The acoustic condition for the test case is presented in Table 2a.

The results of our preliminary experiment with three different levels of spatially white noise are presented in Table 2b. Since the mixture signal for all the DOA combinations was different, we also averaged over the MAE for the two DOAs for each speech mixture. From these results, we can see that the CNN trained with the synthesized noise signals clearly outperforms the SRP-PHAT for all cases. In addition, it is promising to see that even with a simple post processing of the frame-level probabilities, the CNN is able to localize both the sources with such low errors.

In addition, to the MAE, we also show an example result in Fig.3. From the qualitative result, the reason for the big difference in performance between the proposed method and SRP-PHAT becomes clear. It can be seen that for the proposed method there are clear peaks in the distribution whereas for the SRP-PHAT it is much flatter which makes it difficult to obtain accurate final DOA estimates.
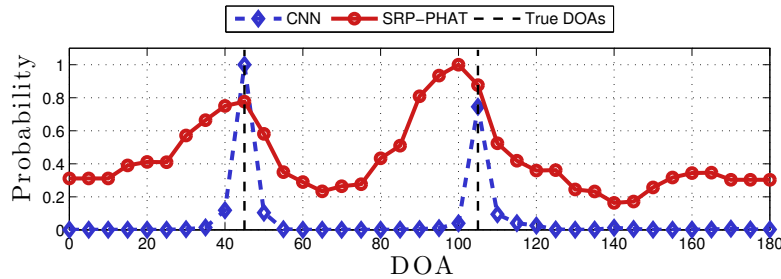

Figure 3: DOA probabilities over the speech mixture when the sources are placed at $45°$ and $105°$.

5 Conclusion and further work

We presented a CNN-based method trained with synthesized noise signals for the task of multi-source DOA estimation by utilizing the assumption of disjoint activity of speech sources in the STFT domain. Being able to train with synthesized signals enables us to create a large amount of training data conveniently.

Preliminary experimental results obtained in simulated unmatched acoustic conditions are very promising, especially in terms of the superiority in performance when compared to SRP-PHAT. Future work will focus on more detailed experiments to identify the source of the superior performance as well as limitations of training with synthesized noise signals for this specific task. Finally, we also aim to study the performance of the proposed method in real acoustic conditions.

References

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

[2] Y. Huang, J. Benesty, G. W. Elko, and R. M. Mersereau, "Real-Time Passive Source Localization: A Practical Linear-Correction Least-squares Approach," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 943–956, Nov. 2001.

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

[4] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, 1986.

[5] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

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

[7] N. Ma, G. Brown, and T. May, "Exploiting deep neural networks and head movements for binaural localisation of multiple speakers in reverberant conditions," in *INTERSPEECH 2015*, 2015, pp. 160–164.

[8] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "A neural network based algorithm for speaker localization in a multi-room environment," in *IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2016, pp. 1–6.

[9] S. Chakrabarty and E. A. P. Habets, "Broadband DOA estimation using convolutional neural networks trained with noise signals," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems*, 2012, pp. 1106–1114.

[11] Y. LeCun and Y. Bengio, "The handbook of brain theory and neural networks," M. A. Arbib, Ed. Cambridge, MA, USA: MIT Press, 1998, ch. Convolutional Networks for Images, Speech, and Time Series, pp. 255–258. [Online]. Available: http://dl.acm.org/citation.cfm?id=303568.303704

[12] S. Rickard and O. Yilmaz, "On the approximate W-disjoint orthogonality of speech," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2002, pp. 529–532.

[13] E. A. P. Habets. (2016) Room Impulse Response (RIR) generator. [Online]. Available: https://github.com/ehabets/RIR-Generator

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, 2014.

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, Jan. 2014.