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`

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.