A Unified Statistical Approach to Fast and Robust Multichannel Speech Separation and Dereverberation

Kouhei Sekiguchi

Abstract

This thesis describes a unified statistical approach to joint multichannel source separation and dereverberation. This technique is useful as a front end of various audio applications including smart speakers, conversational robots, and hearing aid systems because recorded signals usually consist of utterances of target and non-target speakers that often overlap, environmental noise, and reverberation. The objective of source (speech) separation is to recover multiple source (speech) signals from observed multichannel mixture signals. Speech enhancement is one important type of speech separation that aims to extract only utterances of a particular speaker from noisy mixture signals. In addition, it is necessary to recover anechoic (dry) speech signals for improving the speech intelligibility and the performance of automatic speech recognition.

A typical approach to multichannel source separation is to formulate and optimize a unified probabilistic model consisting of a *source model* representing the power spectral densities (PSDs) of sources and a *spatial model* representing their spatial covariance matrices (SCMs). Assuming that the PSDs of all sources have low-rank structures in the time-frequency domain, nonnegative matrix factorization (NMF) has often been used for formulating a source model. One of the most successful examples of this approach is multichannel nonnegative matrix factorization (MNMF) consisting of a *low-rank source model* based on NMF and a *full-rank spatial model* assuming the full-rankness of the SCMs for dealing with reverberation longer than a window size. Although MNMF is a versatile blind source separation method that has a tuning-free convergence-guaranteed iterative optimization algorithm, it has four major problems. 1) The low-rankness of the source PSDs does not always hold in reality, especially for speech sources having complicated dynamics. 2) MNMF tends to easily get stuck in a local optimum because of a high degree of freedom of the spatial model. 3) The optimization algorithm is too computationally expensive. 4) The performance of MNMF is severely degraded under a realistic echoic condition.

To solve the problems 1) and 2), in Chapter 3, we propose a semi-blind speech enhancement method called MNMF-DSP that uses a conventional low-rank model and a deep generative model as noise and speech models (source models), respectively. While the noise model is learned *on the fly* from observed noisy speech signals in an unsupervised manner, the speech model is learned *in advance*

from clean speech signals in an unsupervised manner and used as a prior of clean speech spectrogram. We experimentally show that MNMF-DSP outperformed MNMF and alleviates the initialization sensitivity.

To solve the problems 2) and 3), in Chapter 4, we propose a computationally-efficient variant of MNMF called FastMNMF based on a jointly-diagonalizable full-rank spatial model. Assuming the SCMs of all sources to be jointly diagonalizable, computationally-expensive MNMF dealing with the inter-channel covariance can be converted to light-weight nonnegative tensor factorization (NTF) based on the inter-channel independence. To explicitly consider the directivity or diffuseness of each source, we also propose rank-constrained FastMNMF that enables us to individually specify the ranks of SCMs. We experimentally show the superiority of FastMNMF over MNMF and the effectiveness of the rank constraint.

To solve the problem 4), in Chapter 5, we propose an extension of FastMNMF based on an autoregressive-moving average (ARMA) model called ARMA-FastMNMF for joint blind source separation and dereverberation. The early part of the reverberation is represented by the MA model, and the late part is mainly represented by the AR model, which is suitable for representing long reverberations. To derive efficient update rules, we introduce the joint-diagonalization constraint on the MA model. We experimentally show that ARMA-FastMNMF outperforms conventional methods in many situations.

In Chapter 6, we conclude this thesis with a brief look at future work for real-time joint separation and dereverberation of a varying number of sources.