様式VI

博士学位論文調査報告書

論 文 題 目

A Unified Statistical Approach to Fast and Robust Multichannel Speech Separation and Dereverberation（高速かつ頑健な多チャネル音声分離・残響除去のための統合的・統計的アプローチ）

申請者氏名
關口 航平

最 終 学 歴　　平成29年3月
京都大学 大学院情報学研究科 知能情報学専攻 修士課程 修了
令和3年3月
京都大学 大学院情報学研究科 知能情報学専攻 博士後期課程
研究指導認定見込

学 識 確 認　　令和　　年　　月　　日（論文博士のみ）

論文調査委員　京都大学 大学院情報学研究科
（調査委員長）　吉井 和佳 准教授

論文調査委員　京都大学 大学院情報学研究科
河原 達也 教授

論文調査委員　京都大学 大学院情報学研究科
西野 恒 教授

論文調査委員　京都大学 大学院情報学研究科
田中 利幸 教授

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	關口　航平
論文題目	A Unified Statistical Approach to Fast and Robust Multichannel Speech Separation and Dereverberation（高速かつ頑健な多チャネル音声分離・残響除去のための統合的・統計的アプローチ）		

（論文内容の要旨）

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 include overlapping utterances of target and non-target speakers, 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 target 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 modern standard approach to multichannel audio source separation is to perform maximum-likelihood estimation for a unified probabilistic model that consists of *source and spatial models* representing the power spectral densities (PSDs) and spatial covariance matrices (SCMs) of sources, respectively. Assuming the low-rankness of source PSDs in the time-frequency domain, nonnegative matrix factorization (NMF) has often been used as a source model. One of the most successful examples of this approach is multichannel nonnegative matrix factorization (MNMF) that consists of a low-rank source model based on NMF and a full-rank spatial model that can deal with reverberation longer than the window size. Although MNMF is a versatile blind source separation (BSS) method with a convergence-guaranteed iterative optimization algorithm, it has four major problems. 1) The low-rankness of source PSDs does not always hold in reality, especially for speech sources with complicated frequency and temporal dynamics. 2) MNMF tends to easily get stuck in a local optimum because the log-likelihood function to be maximized is highly non-convex and the spatial model has a high degree of freedom. 3) The optimization algorithm is computationally expensive because of a large number of SCM inversions. 4) The performance of MNMF is severely degraded under a realistic echoic condition.

As the basis of the thesis, Chapter 2 provides comprehensive reviews of existing techniques of blind and non-blind sound source separation, speech enhancement, and dereverberation based on deep learning and/or probabilistic modeling. The multichannel source separation and speech enhancement methods are categorized in terms of the complexity of the spatial model.

To solve the problems 1) and 2), Chapter 3 proposes 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. It has experimentally been shown that MNMF-DSP outperformed MNMF

in terms of the signal-to-distortion ratio (SDR) and alleviated the initialization sensitivity of MNMF in speech enhancement.

To solve the problems 2) and 3), Chapter 4 proposes 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 covariance of channels in the spatial domain is converted to light-weight nonnegative tensor factorization (NTF) assuming the independence of channels in the transformed domain, where the transform matrix (diagonalizer) is estimated efficiently via iterative projection (IP). To explicitly consider the directivity of each source, rank-constrained FastMNMF called RC-FastMNMF that can individually specify the ranks of SCMs is proposed. It has experimentally been shown that FastMNMF outperformed MNMF in terms of the SDR in speech separation and that RC-FastMNMF with rank-1 speech SCMs and full-rank noise SCMs worked better than FastMNMF in speech enhancement.

To solve the problem 4), Chapter 5 proposes 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, the joint-diagonalization constraint is introduced on the MA model. It has experimentally been shown that ARMA-FastMNMF outperformed its ablated version based on the AR or MA model only (AR- or MA-FastMNMF) and the cascading method that sequentially uses AR-based dereverberation and MA-FastMNMF in terms of the SDR.

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

（続紙 ２ ）

（論文審査の結果の要旨）

ロボットやスマートスピーカーが、実環境下で人と音声でインタラクションを行うには、周囲の雑音・残響を抑圧し、複数の音声を分離・強調する技術が求められる。本論文は、マイクロフォンアレイで観測される多チャネル音響信号に対して、音源モデル・空間モデルからなる確率モデルを定式化し、尤度最大化という統一的な基準のもとで、高速かつ頑健な統計的音源分離・残響除去に取り組んだ研究をまとめたものである。主な成果は以下の通りである。

1. 多チャネル非負値行列因子分解 (MNMF) は、低ランクパワースペクトル密度に基づく音源モデルと、フルランク空間共分散行列に基づく空間モデルから構成される汎用ブラインド音源分離手法である。しかし、NMFに基づく音源モデルで仮定される低ランク性は音声のパワースペクトル密度に対しては成立せず、また、尤度関数の非凸性と空間モデルの自由度の高さから、性能が悪い局所解に陥りやすい問題があった。この問題に対して、雑音の音源モデルとしてNMFを、音声の音源モデルとして大量のクリーンな音声データから予め学習した深層生成モデルを用いるMNMF-DSPを提案した。音声強調実験により、MNMFの初期値依存性が解消され、音声強調の性能が向上することを確認した。

2. フルランク空間モデルは、残響をある程度表現できる一方、空間相関行列の逆行列計算に多大な時間を要した。この問題に対して、各周波数における空間相関行列を同時対角化可能なものに限定するFastMNMF1を提案した。さらに、対角化のための変換行列が、ランク1空間モデルに基づく独立成分分析 (ICA) や独立低ランク行列分析 (ILRMA) で用いられる分離行列と同様の役割を担うことから、全ての周波数で音源方向分布を共有するように同時対角化を行うFastMNMF2を提案した。音声分離実験により、MNMFと比較して最大10倍程度の高速化を達成し、音声強調の性能がFastMNMF1,FastMNMF2と順に向上することを確認した。

3. 同時対角化制約付きフルランク空間モデルは、直接音と残響を明示的に区別せずに音の伝播過程を表現しているため、分離音には残響が含まれており、長い残響が存在する環境では性能が大幅に劣化する問題があった。この問題に対して、初期反射に対しては音源ごとに移動平均 (MA) モデルを、後部残響に対しては環境固有の自己回帰 (AR) モデルを仮定することで、音源分離と残響除去を同時に行うARMA-FastMNMFを提案した。さらに、MNMF-DSPと同様に、音声の深層生成モデルを音源モデルに導入することにより、ARMA-FastMNMF-DSPを考案した。音声分離・強調実験により、提案法の有効性を確認した。

以上のように本論文は、物理現象である音の生成過程に対する深い洞察に基づき、多チャネル観測信号の確率モデルを定式化し、最尤推定のための収束保証付きの最適化アルゴリズムを導出するという統一的な音源分離・残響除去技術を提示したもので、学術上・実用上寄与するところが少なくない。よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和3年2月16日に論文とそれに関連した内容に関する口頭試問を行った結果、合格と認めた。なお、本論文は、京都大学学位規程第14条第2項に該当するものと判断し、公表に際しては、当面の間、当該論文の全文に代えてその内容を要約したものとすることを認める。

要旨公開可能日：　　年　　月　　日以降