報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　A Study on Utilization of Prior Knowledge in Underdetermined Source Separation and Its Application
（劣決定音源分離における事前情報の活用とその応用に関する研究）

氏　　名　　関　翔悟

論 文 内 容 の 要 旨

In the field of environmental sound recognition, source separation is one of core technologies, used to extract individual sound sources from mixed signals. Source separation is closely related to other acoustic technologies and is used to develop various applications such as automatic transcription systems for meetings, active music listening systems, and music arranging systems for composers. When a mixed signal is composed of more sources than the number of microphones, i.e., in an underdetermined source separation scenario, separation performance is still limited and there remains much room for improvement. Moreover, depending on the method used to extract the source signals, subsequent systems using the acoustic features calculated from the estimated source information can suffer from performance degradation. Supervised learning is a promising method which can be used to alleviate these problems. Training data composed of source signals, as well as mixed signals, is used to obtain as much prior information about the sound sources as possible into account. Supervised learning is essential for improving the performance of underdetermined source separation, however there are problems which remain to be addressed.

In this dissertation, I address two problems with the supervised learning approach for underdetermined source separation and its application. The first is how to improve the use of prior information, and the second is how to improve the representation ability of source models. To deal with the first problem, I focus on, 1) the characteristics of individual source signals in the spectral and feature domains, and 2) the temporal characteristics implicitly considered in time-frequency analysis. Furthermore, I also explore the use of deep generative models for prior information, to deal with the second problem.

Since synthesized music signals are often stereophonic signals and are generated as linear combinations of many individual source signals and their respective mixing gains, information about phase, or its differential, between

each channel, which represents the spatial characteristics of recording environments, cannot be utilized as acoustic clues for source separation. In order to address this problem, this dissertation proposes a super- vised source separation method for stereophonic music signals based on an extension of non-negative matrix factorization (NMF). NMF-based decomposition is applied to approximate the amplitude spectrogram of a music signal as linear combinations of mixing gains and the spectrograms of individual sources, in which source spectrograms are further decomposed into a set of spectral templates and respective activations. In addition to the conventional supervised approach, cepstral distance regularization (CDR) is further introduced to regularize the timbre information of each source. Experimental evaluations demonstrate that CDR yields significant performance improvements and provides better estimation for mixing gains.

While time-frequency masking is a powerful approach for source separation and speech enhancement in terms of signal recovery accuracy, e.g., signal-to-noise ratio, it can over-suppress and damage speech components, leading to limited performance in succeeding speech processing systems. To overcome this problem, this dissertation proposes a method of restoring missing components of time-frequency masked speech spectrograms using direct estimation of a time domain signal based on time-domain spectrogram factorization (TSF). This TSF-based method allows us to take the local interdependencies of the components of a complex spectrogram, derived from the redundancy of a time- frequency representation, into account, as well as the global structure of the magnitude spectrogram. Experimental results show that the proposed TSF-based method significantly out-performs conventional methods, and has the potential to estimate both phase and magnitude spectra simultaneously and precisely.

Multichannel non-negative matrix factorization (MNMF) is a well-known method used for underdetermined audio source separation, which adopts the NMF concept to model and estimate the power spectrograms of the sound sources in a mixed signal. While MNMF works reasonably well for particular types of sound sources, one limitation is that it can fail to work for sources with spectrograms that do not comply with NMF. In contrast to underdetermined cases, an improved variant of determined source separation methods, called the multichannel variational autoencoder (MVAE) method, was recently proposed, in which a conditional VAE (CVAE) is used instead of the NMF model for expressing source power spectrograms. While the original MVAE method was formulated for use in determined mixing scenarios, we propose a generalized version, combining the features of MNMF and MVAE so that it can also be used for underdetermined source separation. We call this method the generalized MVAE (GM- VAE) method. Experimental evaluations reveal that GMVAE outperformed baseline methods, including MNMF.