Abstract

The development in multimedia technologies has promoted dramatically the rapid growth of music data in recent years. There are various different applications for people's demands in music such as information retrieval, identification and handing. However, singing voice and background music are related to each other in the mixed music, the mutual interference has brought huge obstacles to music information processing. The problem of how to extract the audio information from music has become an important research topic. As the part of music information retrieval, the technologies of singing voice separation are facing unprecedented challenge.

The objective of this research is to deal with the problem of singing voice separation from monaural recordings. It is even more difficult than multichannel since the spatial information cannot be applied in the separation procedure. Singing voice separation is a technique for separating or extracting singing voice from a musical mixture, which has found many applications in the wide areas such as singer identification, singing evaluation and query by humming. This is a relatively easy sep-

aration task of the human auditory system, but it becomes more difficult when we attempt to simulate this problem in a computational method. To achieve the task of singing voice separation, this study mainly focuses on extended robust principal component analysis (RPCA) and deep learning.

RPCA has been recently proposed of popularization and effectiveness way of separation approach that separates singing voice and accompaniment from a mixture music. It decomposes a given amplitude spectrogram (matrix) of a mixture signal into the sum of a low-rank matrix (accompaniment) and a sparse matrix (singing voice). Since musical instruments reproduce nearly the same sounds every time, a given note is played in a given song, the magnitude spectrogram of these sounds can be considered as a low-rank structure. Singing voice, in contrast, varies significantly, but has a sparse distribution in the spectrogram domain to its harmonic structure. Although RPCA is an effective approach to separate singing voice from the mixed audio signal, it fails when there are significant differences in dynamic range among the different background instruments. Some instruments, such as drums, correspond to singular values with tremendous dynamic range; because it uses nuclear norm to estimate the rank of the low-rank matrix, RPCA

algorithm over-estimates the rank of a matrix that includes drum sounds. The accuracy of such separation results thus decreases, as drums may be placed in the sparse subspace instead of being low-rank. Thus, it motivates us to describe exactly the separated low-rank matrix.

To overcome the disadvantage of RPCA for singing voice separation, two extensions of RPCA algorithm are proposed in this book. One is called weighted robust principal component analysis (WRPCA). It uses different weighted values to describe the low-rank matrix for singing voice separation. Additionally, incorporating the proposed WRPCA with gammatone auditory filterbank for singing voice separation. The significance of WRPCA can describe different low-rank matrix under the conditions of human's auditory perceptual properties. Because the cochleagram is derived from non-uniform time-frequency transform whereas time-frequency units in low-frequency regions have higher resolutions than in the high-frequency regions, which closely resembles the functions of the human ear. Therefore, it is promising to separate singing voice via sparse and low-rank decomposition on cochleagram instead of the spectrogram.

Another extension of RPCA with rank-1 constraint called constraint RPCA (CRPCA). It utilizes the rank-1 constraint minimization of sin-

gular values in RPCA instead of minimizing the nuclear norm for separating singing voice from the mixture music. Thus, it not only provides a robust solution to large dynamic range differences among instruments but also reduces the computation complexity. Then, incorporating the proposed CRPCA with gammatone auditory filterbank on cochleagram for singing voice separation. In addition, constructing coalescent masking and vocal activity detection on CRPCA method to constrain the temporal segments that allowed to constrain singing voice from the mixed music datum. Finally, combining F0 and non-negative rank-1 constraint RPCA, which incorporates F0 and non-negative rank-1 constraint minimization of singular values in RPCA instead of minimizing the nuclear norm.

Deep learning-based approaches are very effective and popular for singing voice separation, especially for the U-Net architecture. However, there is still existing lots of room to improve the separation results by developing a deeper neural network. Therefore, a deep fully residual convolutional neural network (Fusion-Net) for singing voice separation is proposed. It combines U-Net and residual framework to improve the separation performance. The phase spectra feature with magnitude spectra feature are integrated to syntheses the singing voice and music.

As the process of post-processing, time-frequency masking is used ito improve the separation performance. Evaluation results show that the proposed Fusion-Net architecture can achieve better separation quality than U-Net architecture on the evaluation database.

In conclusion, this book proposes two extensions of the effective optimization algorithms concentrating on RPCA and Fusion-Net for singing voice separation. One is using different weighted value for describing the separated low-rank matrix. The other is exploring rank-1 constraint minimization of singular value in RPCA. In terms of source-to-artifact ratio, the previous is better than the later. However, CRPCA obtains better separation quality than WRPCA in singing voice separation. The outcomes of this research contribute to further improving the technologies related to music information retrieval. Additionally, the potential contribution of this research is to deal with the problems of noise reduction and speech enhancement by using the separated low-rank and sparse model. Since the background noise is assumed as the part of low-rank component and the human speech is regarded as the part of sparse component.