論文の内容の要旨

Music Signal Processing Exploiting Spectral Fluctuation of Singing Voice Using Harmonic/Percussive Sound Separation

（調波音打楽器音分離による

歌声のスペクトルゆらぎに基づく音楽信号処理の研究）

氏名　橘　秀幸

（本文）Singing voice is one of the most impressive component in music signals. Extracting singing voice from mixed music signals has a significance because of the potential needs for content-based music search and potential applicability to a technical component of interactive music players. The thesis contains the descriptions of novel music signal processing techniques on singing voice, especially singing voice enhancement, pitch estimation of singing melody, automatic real-time audio-to-audio karaoke generation system (singing voice suppression), discrimination of speech and singing voice, and some subsidiary technical discussions on fundamental techniques.

The thesis specifically focuses on the spectral fluctuation of singing voice, which is one of the principal characteristics of singing voice along with harmonicity and singing formants (characteristic spectral envelope), etc. Although there have been many studies on singing voice extraction exploiting harmonicity and spectral envelope, as well as many studies on singing voice detection and some related discrimination tasks exploiting singing fluctuation, the fluctuation of singing voice has not necessarily been explicitly exploited in the literature of source separation despite its significance. This thesis describes a promising approach to the fluctuation of singing voice for source separation. Since these properties of singing voice are supposed to be “orthogonal” each other, it is supposed that joint use of these properties may enrich the toolbox of singing signal processing techniques in the future.

In order to capture the fluctuation of singing voice, the thesis first characterizes a music signal component into three typical classes; harmonic (quasi-stationary, narrowband), fluctuating (intermediate), and percussive (non-stationary, wideband). The thesis first shows a separation technique of harmonic and percussive sound separation (HPSS) ignoring the fluctuating component for simplicity, then considers the extension the approach of HPSS to handle the intermediate component, namely the fluctuating component, under the same frame work. The idea here is the use of two differently-resolved spectrograms, one of which has rich temporal resolution and poor frequency resolution, while the other has the opposite resolution. That is, the idea is that the behavior of intermediate component is dependent on the time-scale of spectrogram on which HPSS is executed. Indeed, the spectral shapes of singing voice are quite different on these two spectrograms, because of the fluctuation. On the former spectrogram the shapes of singing voices are similar to those of sustained harmonic instruments, while on the latter spectrogram it is more similar to those of percussions. On the basis of the idea above, this thesis describes a novel singing voice enhancement technique, which is called two-stage HPSS. The experimental evaluations show that SDR improvement, a commonly-used criterion on the singing voice enhancement, indicated around 4 dB, which is a considerably higher level than some existing methods. The result shows the effectiveness of this approach. The idea beneath the technique is the most important contributions of the dissertation.

In addition to singing voice enhancement, two-stage HPSS is applied to following two problems, both of which are of importance in music information retrieval and music applications, respectively; estimation of the fundamental frequency of singing-melody in mixed music signals; singing voice suppression based on two-stage HPSS, and its application to audio-to-audio karaoke system. On the former application, it is verified that two-stage HPSS basically improves the accuracy of pitch estimation of a simple pitch estimation technique. The technique, tandem connection of two-stage HPSS and a simple pitch tracking algorithm, is evaluated in MIREX, an exchange on music information retrieval. The experimental results in MIREX show that the proposal pitch tracking technique is effective especially in low SNR (voice to accompaniment) conditions, comparing to other participants in MIREX. This is possibly because of the preprocessing by two-stage HPSS. This result also proves the effectiveness of two-stage HPSS. On the latter application, it is qualitatively verified that the singing voice suppression based on two-stage HPSS fairly suppresses the singing voice. The quantative

performance in terms of SDR is basically identical to that of singing voice enhancement. In this dissertation the system is actually implemented in C++, and it is verified the system works in real-time, which is advantageous considering the streaming-based client-side applications. Moreover, due to the efficiency of two-stage HPSS, it is verified that the system works even on a net book in real time.

The thesis further discusses the potential applicability of the idea above to characterizing fluctuation on differently-resolved spectrograms, which decompose a signal into many components according to the characteristic time-scale of fluctuation, and discusses a novel audio signal feature, Characteristic Fluctuation Time-scale (CFTS). The feature is applied to speech/singing discrimination, and its effectiveness comparing to MFCC, a standard audio feature, is described.

The thesis finally considers the improvements of HPSS which forms the basis of all the proposal techniques above, in order to make their performance better. Considering long-term relation on spectrogram, as well as the strict reconstructivity constraint, three variants of HPSS are derived. Especially the third HPSS, which is mentioned as modified HPSS 2, has advantages that the global optimality and the uniqueness of the solution are guaranteed. All the reformation above accelerated the computation of HPSS without significant loss of separation performance. The experiments on singing voice enhancement using the modified HPSS is also carried out, and it is shown that the performance is not much different in terms of SDR. This result shows that the modified HPSS can accelerate the singing voice enhancement and its applications, without significant loss of separation performance.

In summary, the contribution of this dissertation is as follows. The principal contribution is on “two-stage HPSS” and the underlying idea of exploiting the spectral fluctuation of singing voice in source separation problem using differently-resolved spectrograms, as well as the applications of the technique to audio melody extraction and a karaoke application. A subsidiary contribution includes the extended idea of “two-stage” to “multiple concurrent,” which is also promising approach to detection/recognition tasks. Another subsidiary contribution is the fundamental studies on HPSS to improve, namely accelerate, the techniques above all.