Statistical analysis of speech using moments

No.2181223 Keita Kawano
Supervised by Prof. Shogo Kiryu

ABSTRACT

We are surrounded by a variety of sounds, and we judge various information and situations from sounds. For example, we communicate with others through the sound of their voices, and we can detect danger from the sound of cars coming from outside our visual field when we are walking down the street. Humans have a remarkable ability called the cocktail party effect, which enables us to distinguish between two or more voices and certain sounds in noise. There are two techniques for mechanically distinguishing between these two types of sounds: sound source enhancement and sound source separation. I focused on the sound source separation because of its application to products such as improved speech recognition. In recent years, various methods of signal analysis have been used, such as frequency analysis and cepstrum analysis using Fourier transforms, to understand the characteristics of signals in the frequency domain. However, speech signals have partially unpredictable fluctuations such as noise, and it is difficult to separate them when the position and magnitude of the sound source to be observed is unknown. Therefore, it is necessary to use a method called blind source separation, which takes advantage of the independence of the signals. This is an analysis method that uses the method of moments to statistically estimate the population based on the assumption that multiple sound sources are independent of each other, and currently, fourth-order moments are used as evaluation functions that can be used for source separation. On the other hand, fifth order and higher-order moments are not widely used. In this study, we evaluated whether higher-order moments can be used as a new evaluation function, based on the idea that voices that are indistinguishable by fourth-order moments can be distinguished by using fifth order and higher-order moments. In this study, we used audio recordings of four men and women in their 20s reciting different literary works and conducted two types of analysis: one was conducted by varying the length of the audio recordings, and the other was conducted by varying the ratio of the audio recordings. The results of the analysis and evaluation showed that 10,000 samples were the most useful number of samples when the length of the voice was varied, and differences were observed in moments of higher order than the fourth order (sampling frequency: 44.1 kHz). When the ratio of voices was varied, if the sound pressure difference between voices was too large, the characteristics of the voice with the higher sound pressure were too large and approached the value of that voice. Therefore, if the sound pressure difference between voices is too large, source separation by moments is difficult. However, detailed evaluation is necessary to determine at what level of sound pressure recognition becomes difficult and to what extent the value of the mixed voice changes.