Distant-Talking Speech Acquisition Using Microphone Array

Yuki Denda

Hands-free speech interfaces, that allow users to easily utilize some informational machineries or equipments without specific skill, are expected as a novel technology for the infrastructure of informational society. However, the performance of the hands-free speech interfaces is seriously degraded in real noisy environment so that ambient noise and room reverberations distort the distant-talking speech.

Microphone array is widely used to overcome this problem in recent years. It is indispensable to robustly implement the following three technologies in real noisy environment for high-quality acquisition of distant-talking speech using microphone array: (1) talker direction estimation that estimates the direction of talker, (2) voice activity detection that detects speech segments from the captured signal that corrupted by noise, (3) beamforming that steers the high-sensitive directivity to the talker direction. Accordingly, this thesis discusses robust estimation of the talker direction and voice activity, and acquiring the high-quality speech by the beamforming using microphone array.

Firstly, this thesis proposes a noise-robust talker direction estimation method based on: (1) average speech spectrum-based WCSP (Weighted Cross-power Spectrum Phase) analysis, (2) CSP coefficient subtraction that eliminates noise distribution in acoustic space, (3) time-sequential ML (maximum likelihood) estimation of talker direction. As a result of evaluation experiments in real noisy environment, the author confirmed the effectiveness of the proposed method.

Secondly, this thesis proposes two noise-robust voice activity detection methods, adaptive zero crossing detection and adaptive short-time energy thresholding, that extract the temporal features based on spatial features extracted by the proposed talker direction estimation method. The evaluation experiments in real noisy environment revealed that the proposed method sufficiently detects voice activity than the ETSI (European Telecommunications Standards Institute) AFE (Advanced Front End).

Thirdly, this thesis proposes a noise reduction method based on: (1) multistage noise reduction based on delay and sum beamformer and Fourier/wavelet spectral subtraction, (2) instantaneous estimation of non-stationary noise based on subtractive beamformer. As a result of evaluation experiments in real noisy environment, the author confirmed that the proposed method improves noise reduction and speech recognition performance.

Finally, this thesis applies the above proposed methods for automatic multimedia transcription of the videoconferencing, and thus, proposes a noise-robust audio-visual talker direction estimation method to assign speaker ID for individual users as a first step of the transcription. The evaluation experiments in real noisy environment proved that the proposed method is superior to the conventional methods.