2010年03月31日

Electronic & Information Engineering		ID	079302
Name	Alberto Yoshihiro Nakano		

Advisor	Seiichi Nakagawa

Title	Exploring spatial information for distant speech recognition under real environmental conditions

(800 words)

In the last decades, significant achievements in speech recognition for closed talking microphone were obtained, however, distant speech recognition is still a challenge problem most due to the mismatch between the training condition, performed with clean speech data recorded by a closed talking microphone, and the testing condition which consists of the noisy and reverberant speech data recorded by a distant microphone. In this thesis, we explored spatial cues/information estimated by a distributed microphone array network to improve the speech recognition performance in the distant talking case. To achieve our goal, we proposed a new method to estimate the spatial information of a directional acoustic source in a real environment and we proposed some approaches to use these estimated information.

In the proposed method which automatically estimates the spatial cues, defined as the position and orientation of a directional acoustic source in an enclosed environment, different combinations of the estimated parameters from the received signals and the microphone positions of each array were used as inputs to the artificial neural network (ANN). The estimated parameters were composed of time delay estimates (TDEs), source position estimates, distance estimates, and energy features. The outputs of the ANN were the source orientation (one out of four possible orientations shifted by 90 degree and either the best array (which is defined as the nearest to the source) or the source position in 2D/3D space. We studied the position and orientation estimation performances of the ANN for different input/output combinations (and different numbers of hidden units).

The estimated spatial information was then employed in speech recognition tasks. The estimated position of the speaker in an enclosed space was used to reestimate time delays for a delay-and-sum beamformer, thus enhancing the observed signal. On the other hand, the orientation angle was used to restrict the lexicon used in the recognition phase, assuming that the speaker faces a particular direction while speaking. To compensate the effect of the transmission channel inside a short frame analysis window, a new CMN (cepstral mean normalization) method based on a GMM (Gaussian Mixture Model) was investigated and showed better performance than the conventional CMN for short utterances. The performance of the method was evaluated through Japanese digit/command recognition experiments which showed improvements in the speech recognition performance.

Subjective experiments were also evaluated in the same test environment of the automatic estimation system because understanding the human visual and auditory system provides spatial cues to create/develop/improve virtual environments, human-computer interfaces among other applications. That was the reason to investigate the perception of the spatial cues of a directional acoustic source by blindfolded listeners and compare to our proposed automatic estimation method. In our case, we verified that the entire array network surpasses the human auditory system in terms of position localization and orientation estimation. However, when only one microphone array was available, the nearest array to the blindfolded listener, the automatic system had better performance only in terms of position estimation.