HANDS-FREE SPEECH RECOGNITION USING SPATIAL SUBTRACTION ARRAY

Yasuaki Ohashi, Koichi Sakamoto, Tsuyoki Nishikawa, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192, Japan
Email: {yasua-o, tsuyo-ni, sawatari, shikano}@is.naist.jp

1. Introduction

We propose a new microphone array system to realize a robust hands-free speech recognition under noisy environments. Many types of microphone arrays, e.g., Delay-and-Sum (DS) [1] and Griffith-Jim adaptive array (GJ) [2], have been proposed in the past research. Though GJ can achieve a superior performance relatively to others, GJ requires a huge amount of calculations for learning adaptive FIR-filters of thousands or millions of taps. In order to resolve this problem, we propose a spatial subtraction array (SSA). In the proposed method, the noise reduction is achieved by subtracting the estimated noise spectrum from the target speech spectrum to be enhanced in the mel-scale filter bank domain. Moreover, since the proposed method is performed in the mel-scale filter bank domain, the transform into mel-frequency cepstrum coefficient (MFCC) [3] becomes easier, which reduces the amount of calculation in SSA compared to that of GJ. The experimental results obtained under a real environment reveal that word accuracy of the proposed method is greater than those of DS and GJ even when the target user moves between ±20° around the microphone array.

2. Proposed SSA

Figure 1 shows the target speech enhancement procedure in the proposed SSA. In the main pass, the target speech signal is partly enhanced in advance by DS in mel-scale filter bank domain. On the other hand, in the reference pass, the noise signal is estimated by null beamformer (NBF) [4] in which the directional null steers in the direction of arrival (DOA) of the user. In the proposed method, the noise reduction is performed by subtracting the estimated noise spectrum from the enhanced target speech spectrum in the mel-scale filter bank domain as follows:

$$m(l) = \begin{cases} \sum_{k=k_{\mathrm{lo}}(l)}^{k_{\mathrm{hi}}(l)} W(k;l)\{|Y_{\mathrm{DS}}(k)|^2 - \alpha(l)\cdot\beta\cdot|Y_{\mathrm{NBF}}(k)|^2\}^{\frac{1}{2}} & (\text{if } |Y_{\mathrm{DS}}(k)|^2 - \alpha(l)\cdot\beta\cdot|Y_{\mathrm{NBF}}(k)|^2 > 0) \\ \sum_{k=k_{\mathrm{lo}}(l)}^{k_{\mathrm{hi}}(l)} W(k;l)\{\gamma\cdot|Y_{\mathrm{DS}}(k)|\} & (\text{otherwise}) \end{cases}, \quad (1)$$

where l denotes the order of mel-scale filter bank, k denotes the frequency bin, $k_{\mathrm{lo}}(l)$ and $k_{\mathrm{hi}}(l)$ are the lower and higher frequency bins of each triangle window, respectively. $m(l)$ is the output from mel-scale filter bank and $W(k;l)$ is a triangular window. Also, $Y_{\mathrm{DS}}(k)$ is the output signal from DS, i.e., the partly enhanced speech signal and $Y_{\mathrm{NBF}}(k)$ is the output signal from NBF in which the directional null steers in DOA of the user, i.e., the estimated noise signal. The system switches in two models depending on the conditions in Eq. (1). $m(l)$ is a function of the subtraction coefficient β and parameter $\alpha(l)$ which is determined during speech break. On the other hand, if the power spectrum takes negative value, $m(l)$ is obtained by using flooring processing where γ is the flooring coefficient. Since a common speech recognition is not sensitive against phase information, the proposed SSA which is performing subtraction processing in power-domain is more applicable for the speech recognition. GJ requires the adaptive learning of FIR-filters of thousands or millions of taps. On the other hand, in general, the order of the filter bank l is set to 24 and consequently the proposed method optimizes only 24 parameters. Moreover, since the proposed method is performed in the mel-scale filter bank domain, the transform into MFCC becomes easier without

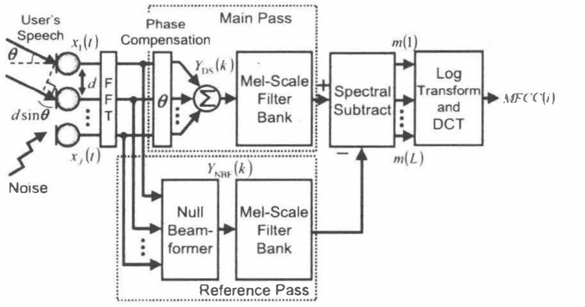

Figure 1: Speech enhancement procedure in the proposed SSA.

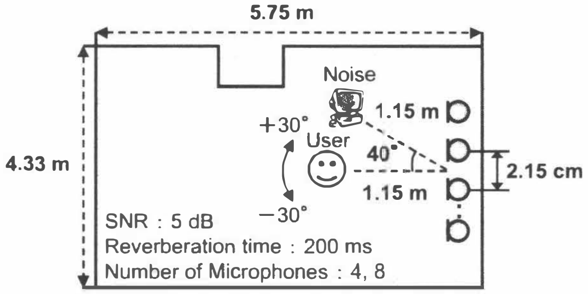

Figure 2: Layout of reverberant room used in experiments.

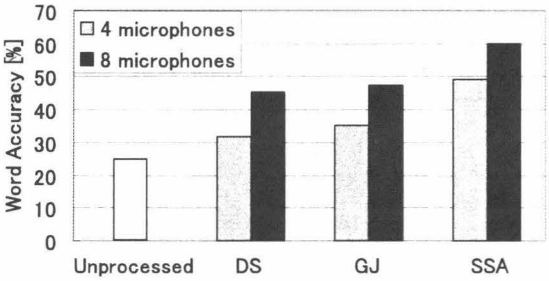

Figure 3: Results of word accuracy in each method.

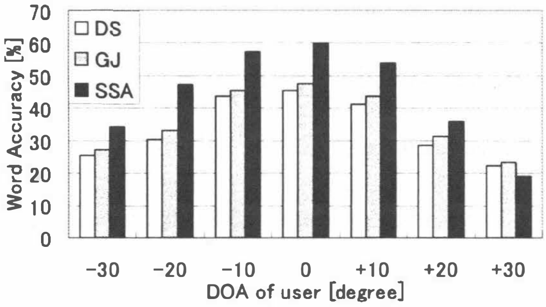

Figure 4: Robustness against user's movement (8 mics.).

the transformation into time-domain waveform. Therefore the amount of calculation of SSA can be significantly reduced compared to that of GJ.

3. Experiments and results

3.1 Experimental setup

Figure 2 shows the layout of the reverberant room used in the experiments. In this paper, we compare DS, GJ, and the proposed SSA on the basis of a large vocabulary continuous speech recognition task (20-k newspaper dictation). Regarding the decoder, JULIUS [5] is used. We use a Phonetic Tied Mixture (PTM) model[6] trained via 260 speakers selected from JNAS [7] database. The test sets include 200 sentences. The sampling frequency of the input data is 16 kHz.

3.2 Results of word accuracy

First we compare DS, GJ, and the proposed SSA in word accuracy scores. Figure 3 shows the experimental results, where the user' position is fixed in front of the microphone array. "Unprocessed" refers to the result without noise reduction processing using one microphone. From these results, the word accuracy of the proposed SSA remarkably overtakes those of the conventional methods in both 4-microphone and 8-microphone conditions. This is mainly due to the fact that there are differences in subtracting the noise, i.e., GJ performs the noise subtraction in terms of both amplitude and phase spectra. On the other hand, since SSA works in only power-spectrum domain, it becomes robust for estimation of the parameters.

3.3 Robustness against user's movement

Figure 4 shows the results of the word accuracy for different DOAs of user. In this experiment, we use 8 microphones and the same parameters of GJ and SSA which were estimated in the experiments of the previous section. From these results, the word accuracy of SSA is superior to those of the conventional methods in the case that DOAs of user are within ±20°. Therefore the proposed SSA is more applicable compared to the conventional approaches.

4. Conclusion

In this paper, we propose an SSA to realize a robust hands-free speech recognition under noisy environments. Since the noise reduction of the proposed method is performed in the mel-scale filter bank domain, the amounts of calculation of SSA can be remarkably reduced. The experimental results obtained under the real environment reveal that word accuracy of the proposed method is greater than those of DS and GJ even when target user moves between ±20° around the microphone array.

Acknowledgement Part of this work is supported by MEXT e-Society leading project.

References

[1] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. America*, vol.78, no.5, pp.1508–1518, 1985.

[2] L. J. Griffith, and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagation*, vol30, no.1, pp.27–34, 1982.

[3] S. B. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, Signal Proc.*, vol.ASSP-28, no.4, pp.357–366, 1982.

[4] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Applied Signal Proc.*, vol.2003, no.11, pp.1135–1146, 2003.

[5] A. Lee, T. Kawahara, K. Shikano, "Julius – An open source real-time large vocabulary recognition engine," *EUROSPEECH*, pp.1691–1694, 2001.

[6] A. Lee, T. Kawahara, K. Takeda, K. Shikano, "A new phonetic tied-mixture model for efficient decoding," *ICASSP*, vol.III, pp.1269–1272, 2000.

[7] K. Itou, M. Yamamoto, K. Takeda, T .Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS : Japanese speech corpus for large vocabulary continuous speech recognition research," *J. Acoust. Soc. Jpn (E)*, vol.20, no.3, pp.199–206, 1999.