2005 IEEE/RSJ International Conference on Intelligent Robots and Systems

Noise-Robust Hands-free Speech Recognition Based on Spatial Subtraction Array and Known Noise Superimposition*

Yasuaki Ohashi, Tsuyoki Nishikawa, Hiroshi Saruwatari, Akinobu Lee, Kiyohiro Shikano
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma, Nara, 630-0192, Japan
sawatari@is.naist.jp

***Abstract*— We propose a spatial subtraction array (SSA) and known noise superimposition to achieve a noise-robust hands-free speech recognition wchich can be used in human-robot interaction. In the proposed SSA, noise reduction is achieved by subtracting the estimated noise power spectrum from the target speech power spectrum to be enhanced in the mel-scale filter bank domain. This offers a realization of error-robust spatial spectral subtraction with few computational complexities. In addition, we introduce known noise superimposition technique in the mel-scale filter bank domain, and utilize the matched acoustic model for the known noise. This can compensate the acoustic model mismatch and mask the residual noise component in SSA. The experimental results obtained under a real environment reveal that word accuracy of the proposed method is greater than that of the conventional method even when the target user moves between -10 and $+10$ degrees around the microphone array.**

***Index Terms*— Hands-free speech recognition, human-robot interaction, microphone array.**

I. Introduction

A hands-free speech recognition system [1] is indispensable for realizing an intuitive, unconstrained, and stress-free human-machine interface, especially in human-robot speech interaction [2], [3]. In this system, however, speech quality is always inferior to that of using close-talking-microphone such as a headset microphone. Therefore, the speech recognition performance is often degraded significantly. One approach for establishing a noise-robust speech recognition system is to enhance the speech signals using microphone array signal processing [4].

Delay-and-Sum (DS) array [5] is one of the simplest speech enhancement method which utilizes microphone array signal processing. To obtain the user's speech at the array output in DS, we compensate the time delay for each element and add the signals together to reinforce the target signal arriving from the look direction. On the other hand, null beamformer (NBF) [6] is one of the simplest noise reduction method. To reduce noise at the array output in NBF, we capture the user's speech with unit gain and steer the directional null to the direction of the noise signal. Moreover, Griffith-Jim adaptive array (GJ) [7] can achieve a superior performance relative to others. GJ comprises the main pass which enhances the target speech signal, the reference pass which estimates noise signal and the subtraction processing which is from the main pass to the reference pass on time domain. However, GJ requires a huge amount of calculations for learning adaptive FIR-filters of thousands or millions of taps.

In order to construct more feasible speech enhancement systems, we newly propose a spatial subtraction array (SSA) which is specifically designed for speech recognition application. In the proposed SSA, noise reduction is achieved by subtracting the estimated noise power spectrum from the target speech power spectrum to be enhanced in the mel-scale filter bank domain. Since a common speech recognition is not so sensitive against phase information, the proposed SSA which is performing subtraction processing only in the power-spectrum domain is more applicable to the speech recognition. Moreover, since the proposed method is performed in the mel-scale filter bank domain, the transform into mel-frequency cepstrum coefficient (MFCC) becomes easier, which requires less calculation in SSA.

*This work is partially supported by MEXT e-Society leading project in Japan.

II. Proposed SSA

Figure 1 shows the target speech enhancement procedure in the proposed SSA. $\boldsymbol{x}(t) = [x_1(t), \cdots, x_J(t)]$ is the observed signal vector and J is the number of array elements. In the main pass, the target speech signal is partly enhanced in advance by DS in mel-scale filter bank domain [8]. Subsequently, in the reference pass, the only noise signal is estimated by NBF in which the directional null steers in the direction of arrival (DOA) of the user.

A. Mel-Scale Filter Bank Analysis

SSA includes mel-scale filter bank analysis, and outputs mel-frequency cepstrum coefficient (MFCC) [8]. The triangular window $W(k;l)$ $(l = 1, \cdots, L)$ to perform mel-scale filter bank analysis is designated as follows:

$$W(k,l) = \begin{cases} \dfrac{k - k_{\mathrm{lo}}(l)}{k_{\mathrm{c}}(l) - k_{\mathrm{lo}}(l)} & \left(k_{\mathrm{lo}}(l) \leq k \leq k_{\mathrm{c}}(l)\right) \\ \dfrac{k_{\mathrm{hi}}(l) - k}{k_{\mathrm{hi}}(l) - k_{\mathrm{c}}(l)} & \left(k_{\mathrm{c}}(l) \leq k \leq k_{\mathrm{hi}}(l)\right), \end{cases} \tag{1}$$

where $k_{\mathrm{lo}}(l)$, $k_{\mathrm{c}}(l)$, and $k_{\mathrm{hi}}(l)$ are the lower, center, and higher frequency bins of each triangle window respectively.

0-7803-8912-3/05/$20.00 ©2005 IEEE

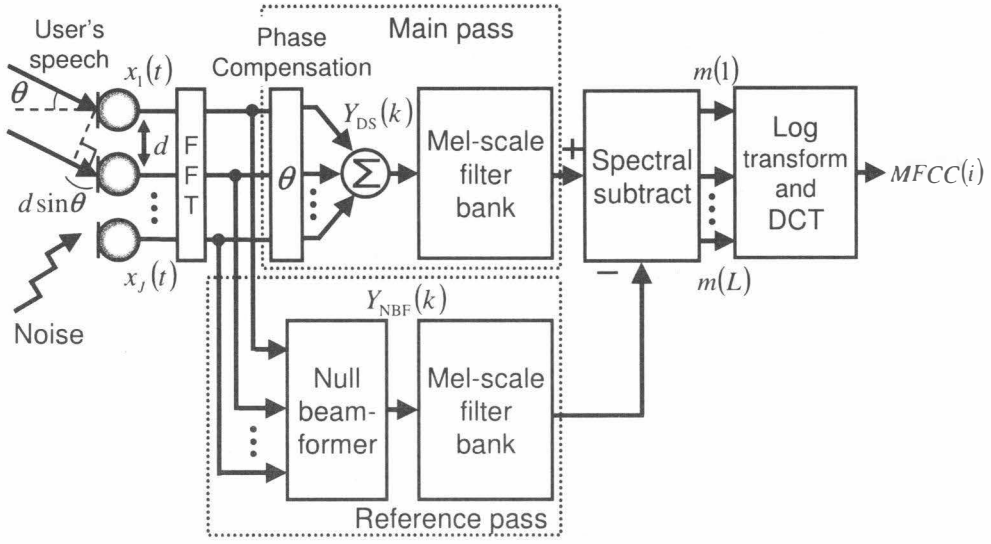

Fig. 1. Speech enhancement procedure in the proposed SSA.

They satisfy the relation among adjacent windows as

$$k_c(l) = k_{hi}(l-1) = k_{lo}(l+1). \quad (2)$$

Moreover, $k_c(l)$ is arranged in regular intervals on mel-frequency domain. Mel-scale frequency $Mel_{k_c(l)}$ for $k_c(l)$ is calculated as

$$Mel_{k_c(l)} = 2595 \log_{10}\left\{1 + \frac{k_c(l) f_s}{700 \cdot M}\right\}, \quad (3)$$

where f_s is the sampling frequency and M is the DFT size.

B. Noise Reduction Processing

In the proposed method, noise reduction is carried out by subtracting the estimated noise power spectrum from the enhanced target speech power spectrum in the mel-scale filter bank domain as

$$m(l) = \sum_{k=k_{lo}(l)}^{k_{hi}(l)} W(k;l)\left\{|Y_{DS}(k)|^2 - \alpha(l)\cdot\beta\cdot|Y_{NBF}(k)|^2\right\}^{\frac{1}{2}} \quad (4)$$

$$(\text{ if } |Y_{DS}(k)|^2 - \alpha(l)\cdot\beta\cdot|Y_{NBF}(k)|^2 > 0\ ) \quad (5)$$

$$m(l) = \sum_{k=k_{lo}(l)}^{k_{hi}(l)} W(k;l)\left\{\gamma\cdot|Y_{DS}(k)|\right\} \quad (\text{otherwise}), \quad (6)$$

where $m(l)$ is the output from the mel-scale filter bank, $Y_{DS}(k)$ is the output signal from DS, i.e., the partly enhanced speech signal, and $Y_{NBF}(k)$ 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 equations depending on the conditions in (4) and (6). $m(l)$ is a function of the subtraction coefficient β and the parameter $\alpha(l)$ which is determined during a speech break. On the other hand, if the power spectrum takes a negative value, $m(l)$ is obtained by using flooring processing where γ is the flooring coefficient.

Because a common speech recognition is not so 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, the proposed method is performed in the mel-scale filter bank domain and the transform into MFCC as follows:

$$MFCC(i) = \sqrt{\frac{2}{L}} \sum_{l=1}^{L} \log\{m(l)\}\cos\left\{\left(l - \frac{1}{2}\right)\frac{i\pi}{L}\right\}, \quad (7)$$

where i denotes the dimension of MFCC. The proposed SSA doesn't require the transformation into the time-domain waveform. Therefore the amount of calculation of SSA is significantly reduced compared with that of GJ.

III. Known Noise Superimposition [12]

In the previous section, we describe noise reduction procedures which utilized in SSA. There still, however, exists a residual component of the original noise spectrum and spectral distortion in the noise reduced signal. In order to achieve an optimum recognition performance appropriately, we generally need to create matched acoustic models. However, since there are many different types of noise, then it would be impractical to create matched models for each of these noise. To solve this problem, we propose a superimposition of a known noise in mel-scale filter bank domain and introduce only one acoustic model matched with the known noise. Figure 2 shows a configuration of this procedure.

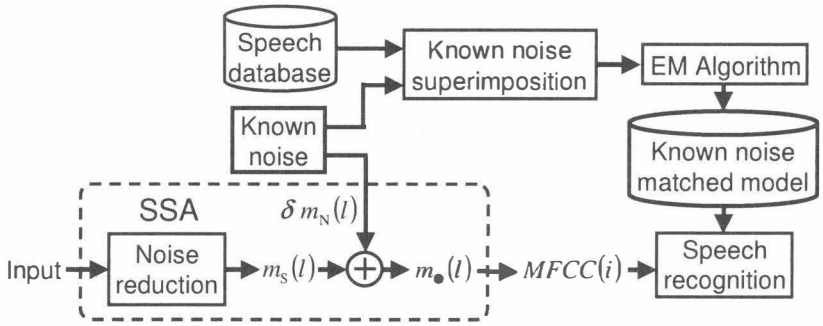

Fig. 2. Procedure of known noise superimposition and using known noise matched acoustic model with SSA.

First, we superimpose known noise to speech database and make the corresponding matched model trained by EM algorithm in advance. Secondly, we superimpose known noise to the noise reduced output from mel-scale filter bank in SSA. The superimposition coefficient δ is calculated as

$$\delta = \sqrt{\frac{1}{10^{SNR/10}} \cdot \frac{\langle \sum_{l=1}^{L} |m_S(l)|^2 \rangle}{\langle \sum_{l=1}^{L} |m_N(l)|^2 \rangle}}, \tag{8}$$

where SNR denotes the amount of noise superimposition, $m_S(l)$ is the output of speech from mel-scale filter bank, and $m_N(l)$ is the known noise processed with mel-scale filter bank analysis. The superimposed output $m_O(l)$ is given as

$$m_\bullet(l) = m_S(l) + \delta m_N(l). \tag{9}$$

Finally, we perform speech recognition using known noise matched model for the output of SSA.

IV. Experiments and Results

A. Experimental Setup

Figure 3 shows a layout of the reverberant room used in the experiment, and Table I shows the experimental conditions. In this experiment, we use the following signals as testing data: the original speech convoluted with the impulse responses which are recorded in the real environment, and added with PC noise which is included in the real environment with an average of 5 dB at the array input.

In the proposed SSA, phase compensation of the main pass is done corresponding to the user's direction which is in front of the microphone array (0°), and we use NBF of the reference pass in which the null steered toward 0° and unit gain steered toward ±90° with 2 elements. Noise reduction coefficients are decided by the results of speech recognition under the clean model conditions. The matched model is created by adding known noise to speech database with 25 dB signal-to-noise ratio (SNR). Known noise obtained in office room is different from PC noise in Fig. 3 and office room noise is comparatively stationary signal whose bias of spectrum is low, so this is suitable to mask various residuals of noise. The amount of superimposition is decided by the results of speech recognition with known noise 25 dB

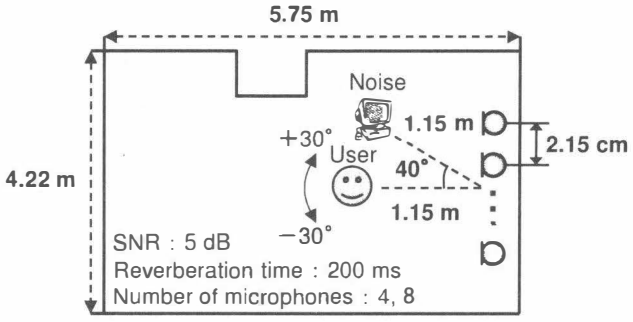

Fig. 3. Layout of reverberant room used in experiments.

matched model. In the conventional method, we superimposed known noise in time domain.

In the experiments on robustness against user's movement, we use the speech data where the target user moves between ±30° around the microphone array.

B. Results of Word Accuracy

First we compare DS, GJ, and the proposed SSA on the basis of word accuracy scores. Figure 4 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 only 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 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 power-spectrum domain, it becomes robust for estimation of the parameters.

C. Effect of Using Known Noise Superimposition

Figure 5 shows the experimental results using known noise superimposition and known noise matched model. "1-microphone" refers to the result performing known noise superimposition with 30 dB and using known noise 25 dB matched model for one microphone.

TABLE I
EXPERIMENTAL CONDITIONS

Database	JNAS [9],306 speakers (150 sentences / 1 speaker)
Task	20-k newspaper dictation
Acoustic model	phonetic tied mixture (PTM) [10] (clean model, known noise 25 dB matched model)
Number of training speakers	260 speakers (150 sentences / 1 speaker)
Number of testing speakers	46 speakers (200 sentences)
Decoder	JULIUS ver.3.4.2 [11]
Sampling frequency	16 kHz
Frame size	20 ms (400 sample)
Filter size	32 ms (512 tap)
Noise reduction coefficients	β:1.8 γ:0.2 (4-microphone), β:1.9 γ:0.3 (8-microphone)
Known noise	office room noise
Amount of superimposition	30 dB (4-microphone), 35 dB (8-microphone)

From these results, we can see that this method is effective in whole, and can mask the residual of noise. In the proposed SSA, word accuracy is about 5% higher than the results in the previous section. Thus the effectiveness of superimposing known noise in mel-scale filter bank can be asserted.

D. Robustness against User's Movement

Figures 6 and 7 show the results of the word accuracy for different DOAs of user. In this experiment, we use 4 or 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 $\pm 10°$. Therefore the proposed SSA is more applicable compared to the conventional approaches. However, the results of word accuracy in the case of that when a user moves over $\pm 20°$ is almost the same or lower than the conventional method. We speculate that since the leakage of the speech signal is included in the reference pass owing to the user's movement, SSA performs subtraction not only noise signals but also speech signals, which leads to the ravaged speech components. This result indicates that the DOA estimator is required if we confronted with the large movement of user. The combination of DOA estimation processing still remains as an open problem for future study.

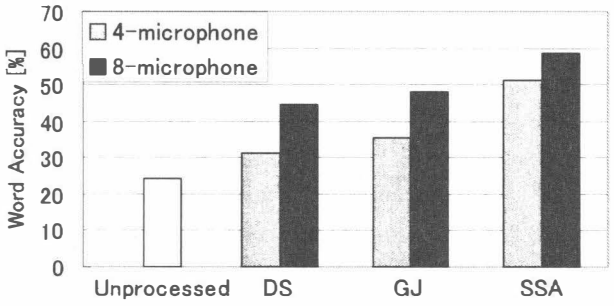

Fig. 4. Results of word accuracy in each method.

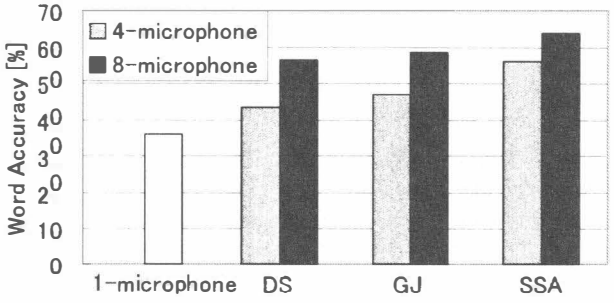

Fig. 5. Effect of using known noise superimposition.

V. CONCLUSIONS

In this paper, we proposed an SSA and known noise superimposition to realize a robust hands-free speech recognition under noisy environments. In the proposed SSA, since the noise reduction of the proposed method is performed in the mel-scale filter bank domain, the amounts of calculation of SSA is remarkably reduced. Moreover, in the proposed known noise superimposition, the residual noise and spectral distortion included output in SSA be homogenized in particular noise, and output can match against the acoustic model.

The experimental results obtained under the real environment reveal that the word accuracy of the proposed method is greater than those of DS and GJ even when target user moves between $\pm 10°$ around the microphone array.

REFERENCES

[1] B. H. Juang and F. K. Soong, "Hands-free telecommunications," *Proc. International Conference on Hands-Free Speech Communication*, pp.5–10, 2001.

[2] R. Nishimura, T. Uchida, A. Lee, H. Saruwatari, K. Shikano, and Y. Matsumoto, "ASKA: Receptionist robot with speech dialogue system," *Proc. IROS-2002*, pp.1314–1317, 2002.

[3] R. Prasad, H. Saruwatari, and K. Shikano, "Robots that can hear, understand and talk," *Advanced Robotics*, vol.18, pp.533–564, 2004.

[4] G. W. Elko, "Microphone array systems for hands-free telecommunication," *Speech Communication*, vol.20, pp.229–240, 1996.

[5] 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.

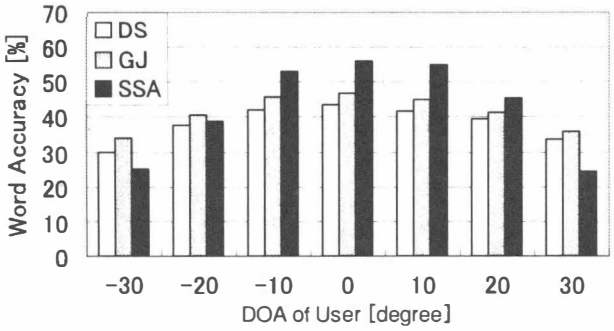

Fig. 6. Robustness against user's movement (4-microphone).

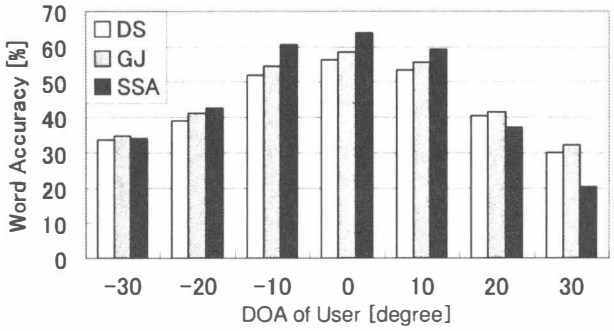

Fig. 7. Robustness against user's movement (8-microphone).

[6] 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.

[7] L. J. Griffith, and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagation*, vol.30, no.1, pp.27–34, 1982.

[8] 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.

[9] 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.

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

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

[12] S. Yamade, A. Lee, H. Saruwatari, and K. Shikano, "Unsupervised speaker adaptation based on HMM sufficient statistics in various noisy environments," *Proc. EUROSPEECH*, pp.II-1493–1496, 2003.