Noise Robust Hands-Free Speech Recognition Using Microphone Array and Kalman Filter as Front-End System of Conversational TV

M.Fujimoto and Y.Ariki
Department of Electronics and Informatics
Ryukoku University, Seta, Otsu-shi, Shiga, 520-2194, JAPAN
Email: masa@arikilab.elec.ryukoku.ac.jp, ariki@rins.st.ryukoku.ac.jp

***Abstract*—In this paper, we investigate hands-free speech recognition as front-end system of conversational TV. The conversational TV is one of machine conversation systems to retrieve the interesting information by inquiring it to the TV. To realize the natural machine conversation without consciousness of microphone, hands-free speech recognition is required. In the hands-free speech recognition system, the directions of the arriving signal are estimated by using a microphone array and the desired signal is enhanced by beam forming. Then, the user utterance section is detected automatically from continuously observed signal. Furthermore, by applying the noise reduction and noise adaptation, the enhanced speech signal is recognized accurately.**

I. Introduction

Recently, many TV news programs are broadcast throughout the world. However, they are now broadcast in one-way from broadcasting stations to the viewer(user). In this situation, the user cannot obtain the detail about the interesting information when it appears in the TV news. Furthermore, even when retrieving the information through the internet, the user feels inconvenient because he must prepare and activate the internet clients quickly for the information retrieval. To retrieve the interesting information without inconveniences, a conversational TV is required which can retrieve the interesting information through man-machine interaction.

In the conversational TV, the required information is retrieved by query words extracted by speech recognition of a user question. Here, it is desired for the user to gibe a question without consciousness of a microphone because the conversational TV is one of machine conversation systems. To realize the natural machine conversation, in this paper we propose hands-free speech recognition using a microphone array as a front-end system of the conversational TV.

In the hands-free speech recognition system, the DOA(Direction Of Arrival) of a desired speech signal(user utterance) is estimated by using a microphone array and the desired signal is enhanced by beam forming. Then, the user utterance section is detected automatically from continuously observed signals based on the time stability of the DOA. Furthermore, by applying a Kalman filter based noise reduction[1] and MLLR(Maximum Likelihood Linear Regression)[2] based noise adaptation, the enhanced speech signal is recognized accurately.

II. System Overview

Fig.1 shows the system overview of our proposed conversational TV. In the system, the user can retrieve his interesting information by inquiring it to the TV when it appears in the TV news. Then, to present the information to the user, the system works according to the following processes.

(1) Recognize the user utterance in hands-free environment.
(2) Extract the query words from the recognition results.
(3) Retrieve the information through the internet by using the query words.
(4) Synthesize the response speech by using the texts of the retrieved results.
(5) Present the retrieved results to the user using synthesized response speech.

Here, the above processes are classified into the front-end system and the back-end system. The front-end system includes the process (1), and the back-end system includes the other processes. In this paper, we investigate a hands-free speech recognition as the front-end system(process (1)) of the conversational TV.

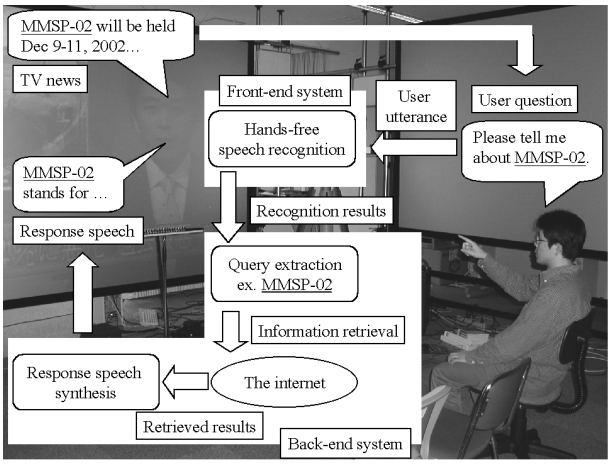

Fig. 1. System overview of the conversational TV

III. Hands-Free Speech Recognition

Fig.2 shows the overview of our proposed hands-free speech recognition as a front-end system of the conversational TV. In

the figure, the DOA of desired speech signal is estimated by using a microphone array and the desired speech signal is enhanced by beam forming. Then, the user utterance section is detected automatically from the continuously observed signal. Furthermore, by applying a Kalman filter based noise reduction and MLLR based noise adaptation, the enhanced speech signal is recognized accurately. In the below sections, the detailed process of the hands-free speech recognition is described.

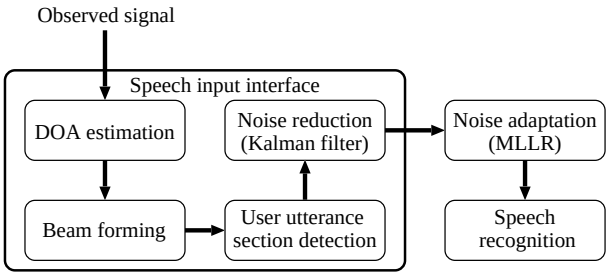

Fig. 2. Processing diagram of hands-Free speech recognition

A. Delay and Sum Beam Former

To capture the desired speech signal with high quality, a microphone array requires to form the directivity of the desired speech signal. In this paper, a delay and sum beam former[3] as shown in Fig.3 is employed to capture the desired speech signal with high quality. In the figure, $y_i(t)$ denotes the captured signal, M denotes the number of microphones, θ denotes the DOA of desired signal $x(t)$, τ denotes the delay of arrival of $x(t)$ and d denotes the microphone interval.

In the delay and sum beam former, the desired signal $x(t)$ from the direction θ is emphasized M times because the signals captured with multiple microphones are added after synchronizing them by using the estimated delay τ. On the other hand, the undesired signals are not emphasized M times because the directions of the undesired signal are different from that of the desired signal. Therefore, the directivity of the delay and sum beam former can be formed in only a direction θ.

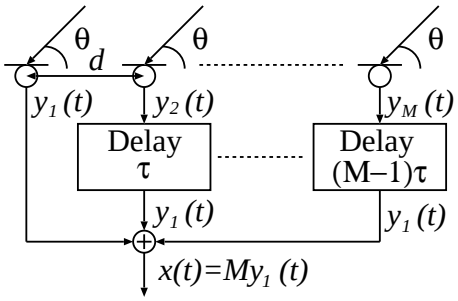

Fig. 3. Delay and Sum Beam Former

B. Delays of Arriving Signal Estimation

To estimate the delay τ, a CSP(Cross power Spectrum Phase analysis) method[4], [5] is employed and the CSP coefficients $CSP_{i,j}(k)$ are derived by the following equation.

$$CSP_{i,j}(k) = IDFT\left[\frac{DFT[y_i(t)]DFT[y_j(t)]^*}{|DFT[y_i(t)]|\,|DFT[y_j(t)]|}\right] \quad (1)$$

where $y_i(t)$ and $y_j(t)$ denote the signals captured by the microphones i and j. When a desired signal is observed, the delay τ is estimated by detecting the maximum value of $CSP_{i,j}(k)$ as the following equation.

$$\tau = arg\max_k (CSP_{i,j}(k)) \quad (2)$$

Then the DOA θ is computed by the following equation.

$$\theta = cos^{-1}\left(\frac{c\cdot\tau/f}{d}\right) \quad (3)$$

where c denotes the sound propagation speed and f denotes the sampling frequency.

C. User Utterance Section Detection

In the conversational TV, it is supposed that the user inquires about the interesting or unknown information to the TV when TV news programs are broadcast. Therefore, to recognize the user utterance in a hands-free mode in the TV news sound environments, a speech input interface is required which can barge into TV news sounds. Generally, this speech input interface is realized by detecting the user utterance section from continuously observed signal. We propose here the user utterance detector based on time stability of the DOA.

In this paper, it is supposed that the TV news sounds arrive from the back of the microphone array, as shown in Fig.4. In this case, the TV news sounds are captured with various reflections at the microphones. Therefore, the DOA of the TV news sounds is not stable in the time sequence. On the other hand, the DOA of the user utterance is stable because it arrives from the front of the microphone array. Under these assumptions, time section with more than 1 second DOA stability is detected as the user utterance region as shown in Fig.5.

D. Kalman Filter Based Noise Reduction

The delay and sum beam former described in Sec.III-A enhances the desired speech signal. However, additive noise components(residual noise) still exists in the enhanced speech signal and they degrade the speech recognition rate. To solve this problem, estimation of the clean speech signal from the speech signal enhanced by the delay and sum beam former is employed by using a Kalman filter based noise reduction method. The Kalman filtering algorithm is defined based on a speech state transition model. A speech state transition model represents the state transition of speech component included in the noisy speech and is modeled by using the first order Taylor expansion[1].

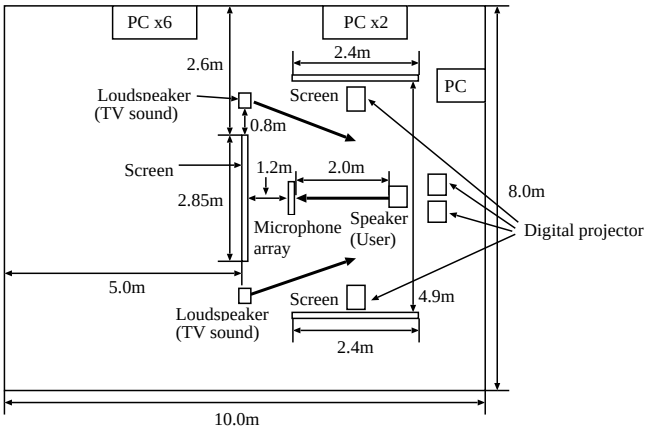

Fig. 4. Experimental room environment

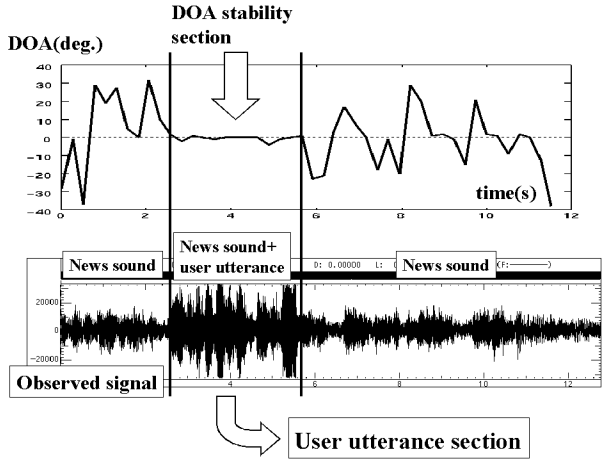

Fig. 5. An example of user utterance section detection

At the kth frame, let $\mathbf{X}(k)$, $\mathbf{S}(k)$ and $\mathbf{N}(k)$ denote the vectors of power spectra of noisy speech, clean speech and (residual) noise respectively, and superscript $\mathbf{l}$ denote the log-spectral domain, then the Kalman filtering algorithm is obtained as follows:

$$\hat{\mathbf{S}}(k) = \mathbf{F}_{k-1}\hat{\mathbf{S}}(k-1) + \mathbf{K}_k(\mathbf{X}(k) - \mathbf{F}_{k-1}\hat{\mathbf{S}}(k-1)) \quad (4)$$

$$\mathbf{K}_k = \mathbf{Q}_k \left[\mathbf{Q}_k + \Sigma_{\mathbf{N}(k)}\right]^{-1} \quad (5)$$

$$\begin{aligned}\mathbf{Q}_k &= \mathbf{F}_{k-1}(\mathbf{I} - \mathbf{K}_{k-1})\mathbf{Q}_{k-1}\mathbf{F}_{k-1}^T \\ &\quad + \mathbf{G}_{k-1}\Sigma_{\mathbf{W}(k-1)}\mathbf{G}_{k-1}^T \end{aligned} \quad (6)$$

$$\mathbf{F}_k = \mathbf{I} + \Delta\mathbf{X}^{\mathbf{l}}(k) \quad (7)$$

$$\Delta\mathbf{X}^{\mathbf{l}}(k) = \mathbf{X}^{\mathbf{l}}(k+1) - \mathbf{X}^{\mathbf{l}}(k) \quad (8)$$

$$\mathbf{G}_k = \mathbf{N}(k) \quad (9)$$

where $\hat{\mathbf{S}}(k)$ denotes the estimation of $\mathbf{S}(k)$ and $\mathbf{Q}_k$ denotes the diagonal co-variance matrix of the estimating error respectively.

The initial values for Eq.(4)$\sim$(6) are given as follows:

$$\hat{\mathbf{S}}(0) = \mathbf{0} \quad (10)$$

$$\mathbf{Q}_0 = \mathbf{0} \quad (11)$$

In Eq.(6), $\Sigma_{\mathbf{W}(k)}$ denotes the diagonal co-variance matrix of $\mathbf{W}(k)$. $\Sigma_{\mathbf{W}(k)}$ is computed by the following equation under the assumption that $\mathbf{W}(k)$ follows zero mean Gaussian process.

$$\mathbf{W}(k) = \Delta\mathbf{X}^{\mathbf{l}}(k) - \Delta\mathbf{N}^{\mathbf{l}}(k) \quad (12)$$

$$\Delta\mathbf{N}^{\mathbf{l}}(k) = \mathbf{N}^{\mathbf{l}}(k+1) - \mathbf{N}^{\mathbf{l}}(k) \quad (13)$$

$$\Sigma_{\mathbf{W}(k)} = \mathbf{W}(k)\mathbf{W}(k)^T \quad (14)$$

On the other hand, in Eq.(5), $\Sigma_{\mathbf{N}(k)}$ denotes the diagonal co-variance matrix of $\mathbf{N}(k)$. $\Sigma_{\mathbf{N}(k)}$ is computed by the following equation under the assumption that $\mathbf{N}(k)$ follows zero mean Gaussian process as $\mathbf{W}(k)$ does.

$$\Sigma_{\mathbf{N}(k)} = \mathbf{N}(k)\mathbf{N}(k)^T \quad (15)$$

In Eq.(14) and Eq.(15), to compute the $\Sigma_{\mathbf{W}(k)}$ and $\Sigma_{\mathbf{N}(k)}$, the value of $\mathbf{N}(k)$ is estimated by using linear 14th order predictive estimation.

E. Unsupervised MLLR Adaptation

The noise reduction method described in Sec.III-D reduces the additive noise components. However, the method does not consider about the reverberant and the convolutional noise. To cope with them, adaptation of the acoustic models is employed by using an (on-line) unsupervised MLLR adaptation[2] to the reverberant and the convolutional noise.

In the unsupervised MLLR adaptation, speaker independent monophone HMMs are adapted to the speech signal estimated by the noise reduction method. To make this adaptation feasible, monophone labels are required for the estimated speech signal. To obtain these labels, the speech recognition is carried out using the monophone HMMs before adaptation, to the speech signal estimated by the noise reduction method. Then the MLLR is applied to the HMMs using the labels and the estimated speech signal. In this paper, an input speech signal itself is used as the adaptation material and the number of Gaussian distribution clusters included in the monophone HMMs was set to 1.

IV. Experiments

We evaluated the hands-free speech recognition in the conversational TV environment.

A. Experimental Setup

The experimental materials are 100 sentences spoken by 5 Japanese male subjects and include 20 keywords appeared in the TV news(Each subject speaks 20 sentences.). Then each utterance arrives from the subject in front of a microphone array. The distance from the subject to the microphone array is 2m. By using these materials, we evaluated the hands-free speech recognition by sub-word model based keyword spotting. The noise sources are the TV news sounds and the fan noise generated from 4 digital projectors and 9 PCs (Noise level is about 55dB.). The TV news is NHK TV news broadcast at 12:00 on November 30 in 2001.

The acoustic models of the speech recognition are the speaker independent monophone HMMs. Their structure is composed of 5 states with 3 loops and 12 mixtures for each state. They were trained using 21,782 sentences spoken by 137 Japanese males. These speech data were taken from the database of Acoustical Society of Japan. The feature parameters are composed of 39 MFCCs with 12 MFCCs, log energy and their first and second order derivatives. Table I, II and III summarize the experimental conditions for the acoustic analysis and phoneme HMM.

TABLE I

ACOUSTIC ANALYSIS CONDITIONS FOR DOA ESTIMATION

Microphone array	Linear type, 16 microphones(2cm intervals)
Sampling frequency	16kHz, 16Bit
Feature parameter	CSP coefficients(4096th order)
Analysis frame length	256ms
Analysis frame shift	256ms
Analysis window	Hamming window

TABLE II

ACOUSTIC ANALYSIS CONDITIONS FOR SPEECH RECOGNITION

Sampling frequency	16kHz, 16Bit
Pre-emphasis	$1 - 0.97z^{-1}$
Feature parameter (Noise reduction)	FFT spectra(512th order)
Feature parameter (Recognition)	MFCC(12th order) + Log-Power + Δ + $\Delta\Delta$
Analysis frame length	20ms
Analysis frame shift	10ms
Analysis window	Hamming window

TABLE III

STRUCTURE OF PHONEME HMM

Number of states	5
Number of loops	3
Number of mixtures	12
Number of phonemes	41
Type	Left to right HMM

B. Experimental Results

Table IV shows the results of the user utterance section detection and the DOA estimation. The DOA estimation rate allows the estimation error of ± 5 degrees. In the table, 89 user utterance sections were correctly detected and the DOA estimation rate of correctly detected sections was about 96%.

TABLE IV

RESULTS OF USER UTTERANCE SECTION DETECTION AND DOA ESTIMATION

User utterance section detection rate (%)	The number of false alarm section	DOA estimation rate(%)
89.0(89/100)	5	95.5

Table V shows the speech recognition results of the correctly detected sections. In the table, by using the proposed method, the keyword extraction rate was about 70%. However, to realize the natural machine conversation, these results are not sufficient. Especially, the reverberant and the convolutional noises affect the speech recognition rate. From this fact, it is required for the speech recognition method to be robust for these noises.

TABLE V

SPEECH RECOGNITION RESULTS

	Keyword extraction rate(%)	The number of false alarm words
Without beam forming	22.4(20/89)	16
With beam forming	48.3(43/89)	7
Proposed method	**69.7(62/89)**	**6**

V. CONCLUSIONS

In this paper, we proposed a hands-free speech recognition as a front-end system of conversational TV. As the evaluation results, the proposed method showed that the user utterance section detection rate was 89%, the DOA Estimation was about 96% and the keyword extraction rate was about 70 %. In future, to improve the speech recognition rate in hands-free environments, we are planning to develop more robust hands-free speech recognition method to the reverberant and the convolutional noises. Furthermore, we will study the information retrieval method as a back-end system of conversational TV.

REFERENCES

[1] M. Fujimoto and Y. Ariki: "Continuous Speech Recognition under Non-stationary Musical Environments Based on Speech State Transition Model", Proc. ICASSP01, Vol.I, pp.297-300(2001).

[2] C.L. Leggetter and P.C. Woodland: "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models", Computer Speech and Language, Vol.9, pp.171-185(1995).

[3] J.L. Flanagan, J.D. Jhonston, R. Zhan and G.W. Elko: "Computer-Steered Microphone Arrays for Sound Transduction in Large Rooms", J.Acoust. Soc. Am., Vol.78, No.5, pp.1508-1518(1985).

[4] M. Omologo and P. Svaizer: "Acoustic Event Localization Using a Crosspower-Spectrum Phase Based Technique", Proc. ICASSP94,Vol.I.pp.273-276(1994).

[5] P. Svaizer, M. Matassoni and M. Omologo, "Acoustic Source Location Three-dimensional Space Using Crosspower Spectrum Phase", Proc. ICASSP97,Vol.I.pp.231-234(1997).