AN EFFECT OF ADAPTIVE BEAMFORMING ON HANDS-FREE SPEECH RECOGNITION BASED ON 3-D VITERBI SEARCH

Takeshi Yamada, Satoshi Nakamura, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan

ABSTRACT

To integrate the microphone array processing into speech recognition, we have proposed a speech recognition algorithm based on 3-D Viterbi search, which localizes a target talker considering the likelihood of HMMs (Hidden Markov Models) while performing speech recognition. The performance of the 3-D Viterbi search method depends on the improvement of the SNR (Signal to Noise Ratio) by the beamforming technique. This paper proposes a novel method based on an adaptive beamforming technique instead of the delay-and-sum beamformer used in our previous study. The speaker-dependent isolated-word recognition experiments were carried out on real environment data to evaluate the effect of the adaptive beamformer. These results showed that the adaptive beamformer drastically improves the recognition performance both for a fixed-position talker and for a moving-talker.

1. INTRODUCTION

In recent years, speech recognition systems with a microphone array have been proposed to realize hands-free speech interface [1,2]. Most of these systems localize a target talker by using short- and long-term power, then extract a frame sequence of parameter vectors for speech recognition by steering a beamform to the localized talker. However, localization of a moving talker is very difficult in low SNR (Signal to Noise Ratio) conditions and highly reverberant environments. The errors of talker localization seriously degrade the performance of speech recognition.

The conventional systems regard a microphone array as a pre-processor of speech recognition. However, talker localization and speech recognition should be performed simultaneously. To integrate the microphone array processing into speech recognition, we have proposed a speech recognition algorithm based on 3-D Viterbi search [3], which localizes a target talker considering the likelihood of HMMs (Hidden Markov Models) while performing speech recognition. The 3-D Viterbi search method extracts a direction-frame sequence of parameter vectors by steering a beamform to each direction in every frame. Then Viterbi search is performed in 3-dimensional trellis space composed of talker directions, input frames, and HMM states. A locus of the talker and a phone sequence of the speech are obtained by finding an optimal path with the highest likelihood. Real environment data experiments showed that the 3-D Viterbi search method works well even for the moving talker [3].

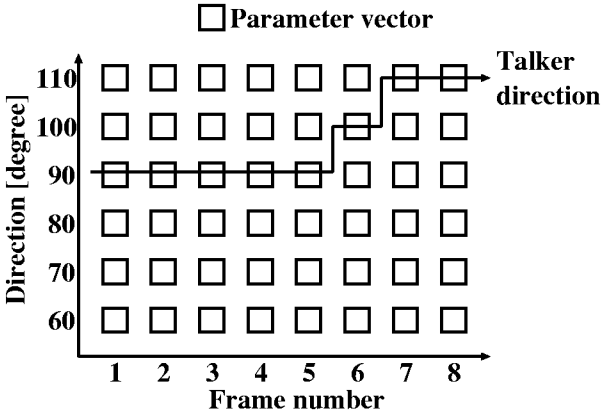

Figure 1: A direction-frame sequence of parameter vectors.

The performance of the 3-D Viterbi search method depends on the improvement of the SNR by the beamforming technique. The delay-and-sum beamformer [4] was used in our previous study. However, it is very difficult to make the beamform sharp, because many microphone elements are necessary. To improve the performance of the 3-D Viterbi search method in real environments, this paper proposes a novel method based on an adaptive beamforming technique. The speaker-dependent isolated-word recognition experiments are carried out on real environment data to evaluate the effect of the adaptive beamformer.

2. 3-D VITERBI SEARCH METHOD

The conventional systems regard a microphone array as a pre-processor of speech recognition. However, talker localization and speech recognition should be performed simultaneously. To integrate the microphone array processing into speech recognition, we have proposed a speech recognition algorithm based on 3-D Viterbi search [3], which localizes a target talker considering the likelihood of HMMs while performing speech recognition.

A direction-frame sequence of parameter vectors is obtained by steering a beamform to each direction in every frame as shown in Figure 1. In Figure 1, the box depicts the parameter vector and the solid line is the locus of a talker. Given the direction-frame sequence of parameter vectors, talker localization and speech recognition can be performed simultaneously in the statistical framework as

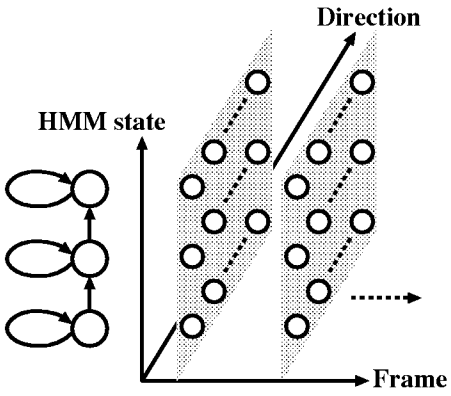

Figure 2: 3-dimensional trellis space composed of talker directions, input frames, and HMM states.

follows:

$$(\widehat{\mathbf{q}},\widehat{\mathbf{d}}) = \underset{(\mathbf{q},\mathbf{d})}{\operatorname{argmax}} P(\mathbf{x},\mathbf{q},\mathbf{d} \mid \mathbf{M}), \tag{1}$$

where $(\widehat{\mathbf{q}},\widehat{\mathbf{d}})$ is the optimal combination of the phone sequence $\mathbf{q}$ of the speech and the locus $\mathbf{d}$ of the talker, $\mathbf{M}$ is the speech model, and $\mathbf{x}$ is the direction-frame sequence of parameter vectors. The optimal combination $(\widehat{\mathbf{q}},\widehat{\mathbf{d}})$ is obtained by the Viterbi search algorithm which finds the most likely path in 3-dimensional trellis space composed of talker directions, input frames, and HMM states as shown in Figure 2. The likelihood is calculated as follows:

$$\begin{aligned}\alpha(q,d,n) &= \max_{q',d'}\{\alpha(q',d',n-1) \\ &\qquad + \log a_1(q',q) + \log a_2(d',d)\} \\ &\quad + \log b(q,\mathbf{x}(d,n)),\end{aligned} \tag{2}$$

where q is the HMM state index, d is the direction, and n is the frame index. $a_1(q',q)$ is the transition probability from the HMM state q' to q, $a_2(d',d)$ is the transition probability from the direction d' to d, and b is the output probability. The transition probability $a_2(d',d)$ represents how likely the talker moves. Since the talker moves to neighboring directions at most for a duration of the frame (about 10 msec), it is reasonable to restrict the range of movements as follows:

$$a_2(d',d) = \begin{cases} \dfrac{1}{2\Delta d} & , \quad |d-d'| \le \Delta d \\ 0 & , \quad |d-d'| > \Delta d \end{cases}, \tag{3}$$

where Δd is the range of movements.

When the likelihood in the correct talker direction is lower than that in the other directions, the performance of the 3-D Viterbi search method is degraded. In such a case, it is effective to raise the likelihood in directions with speech-like characteristics. The pitch harmonics of speech can be used as a measure of speech-like characteristics. An weight function based on the pitch harmonics is introduced as

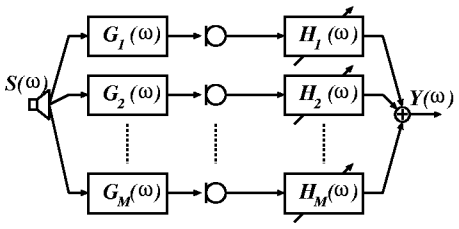

Figure 3: A block diagram of the adaptive beamformer.

follows:

$$w(d,n) = \log \frac{\sum_{n'=n-(\nu-1)}^{n} \{c(d,n')\}^{\mu}}{\sum_{d'=0}^{180} \sum_{n'=n-(\nu-1)}^{n} \{c(d',n')\}^{\mu}}, \tag{4}$$

where $c(d,n)$ is the maximum value of cepstrum coefficients in pitch quefrency region. This value becomes larger when the pitch harmonics exist. μ is the parameter to control the weight effect and ν is the parameter to adjust the continuation.

3. ADAPTIVE BEAMFORMING

The performance of the 3-D Viterbi search method depends on the improvement of the SNR by the beamforming technique. The delay-and-sum beamformer [4], which forms a super-directional gain pattern to the target direction, was used in our previous study [3]. However, it is very difficult to make the beamform sharp, because many microphone elements are necessary. In this paper, an adaptive beamforming technique is used to achieve the significant improvement of the SNR without increasing the number of microphone elements.

Figure 3 shows a block diagram of the adaptive beamformer. In Figure 3, $S(\omega)$ is the Fourier transform of the desired signal and $Y(\omega)$ is the Fourier transform of the output signal. $G_m(\omega)$ is the acoustic transfer function from the desired sound source to the mth microphone element and $H_m(\omega)$ is the frequency response of the mth filter. The frequency response $F(\omega)$ of the adaptive beamformer to the desired signal is represented as follows:

$$F(\omega) = \sum_{m=1}^{M} G_m(\omega) H_m(\omega), \tag{5}$$

where M is the number of microphone elements. The concept of the adaptive beamformer is to minimize the output noise power while constraining $F(\omega)$ to the desired frequency response. In this paper, the AMNOR constraint [5] as Equation (6) is used.

$$D = \int |1 - F(\omega)|^2 d\omega. \tag{6}$$

The AMNOR constraint attains maximum noise reduction while allowing a small distortion D in the frequency response to the desired signal.

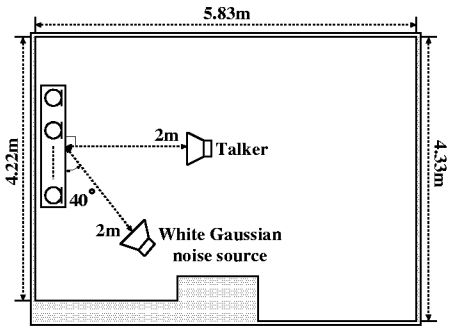

Figure 4: An arrangement of the sound sources and the microphone array. The position of the talker is fixed.

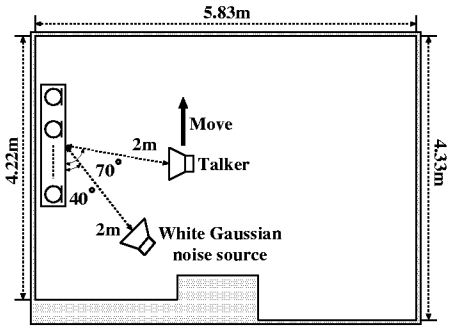

Figure 5: An arrangement of the sound sources and the microphone array. The talker moves from 70 degree to 140 degree while uttering each word.

4. REAL ENVIRONMENT EXPERIMENTS

4.1 Real Environment Data

Real environmental data were collected in an experiment room that the reverberant time is about 0.18 sec. These data include one talker, one white Gaussian noise source, and ambient noises such as computer-fans and air-conditioners. Two arrangements of the sound sources and the microphone array are considered.

(1) The positions of the talker and the white Gaussian noise source are fixed as shown in Figure 4.

(2) The talker moves from 70 degree to 140 degree while uttering each word, and the position of the white Gaussian noise source is fixed as shown in Figure 5.

Two loud speakers are used instead of the talker and the white Gaussian noise source. The loud speakers face the microphone array. The microphone array is a linear and equally spaced array composed of 14 microphones, where the distance between two adjacent microphones is 2.83 cm.

4.2 Experiment Conditions

A speech recognizer is based on the tied-mixture HMM with 256 distributions. A speech corpus is the ATR

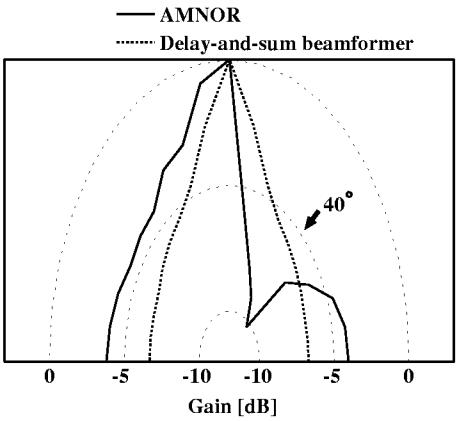

Figure 6: The directional gain patterns.

Japanese speech database Set-A. 2620 words of the speaker MHT are used for training 54 context independent phone models and another 216 words are used for testing. Speech signals are sampled at 12 kHz and windowed by the 32 msec Hamming window in every 8 msec. Then 16-order mel frequency cepstrum coefficients (MFCCs), 16-order ΔMFCCs, and a Δpower are calculated. The cepstrum mean normalization technique [6] is also applied to the speech recognizer. The direction-frame sequence of the parameter vectors is computed every 10 degree.

The filter coefficients of the AMNOR are calculated using pre-recorded noise signals in the two situations.

(A) The ambient noises exist.

(B) The white Gaussian noise source and the ambient noises exist.

The directional gain patterns are shown in Figure 6. In Figure 6, the directional gain pattern of the AMNOR is calculated in the situation (B). The delay-and-sum beamformer forms the super-directional gain pattern to the target direction (90 degree). The gain of the AMNOR to the white Gaussian noise source direction (40 degree) is lower than that of the delay-and-sum beamformer.

4.3 Experiment Results

The word recognition accuracy for the arrangement (1) (the fixed-position talker) is shown in Table 1. In Table 1, when the SNR is 21 dB, the white Gaussian noise source does not exist. *Single microphone* is the center microphone element of the microphone array. *Delay-and-sum beamforming to the correct talker direction* and *AMNOR to the correct talker direction* indicate the case that the correct talker direction is known. The frame sequence of the parameter vectors is obtained by steering each beamform only to the correct talker direction. *3-D Viterbi search method with delay-and-sum beamforming* and *3-D Viterbi search method with AMNOR* indicate the case that the correct talker direction is unknown. The direction-frame sequence of the parameter vectors is obtained by each beamforming. These results are summarized as follows:

Table 1: Word recognition accuracy [%] for the arrangement (1) (the fixed-position talker).

	SNR [dB]		
	21	18	10
Single microphone	89.8	76.8	37.0
Delay-and-sum beamforming to the correct talker direction	92.1	86.5	75.0
AMNOR to the correct talker direction	94.4	91.2	89.3
3-D Viterbi search method with delay-and-sum beamforming	92.5	79.1	53.2
3-D Viterbi search method with AMNOR	93.9	89.8	83.3

Table 2: Word recognition accuracy [%] for the arrangement (2) (the moving talker).

	SNR [dB]		
	21	18	10
Single microphone	92.5	77.7	38.4
Delay-and-sum beamforming to the correct talker direction	—	—	—
AMNOR to the correct talker direction	—	—	—
3-D Viterbi search method with delay-and-sum beamforming	89.3	81.9	52.3
3-D Viterbi search method with AMNOR	92.5	88.8	81.0

- The word recognition accuracy of *AMNOR to the correct talker direction* is improved 2.3 % in SNR 21 dB, 4.7 % in SNR 18 dB, 14.3 % in SNR 10 dB compared with that of *Delay-and-sum beamforming to the correct talker direction.*
- The word recognition accuracy of *3-D Viterbi search method with AMNOR* is improved 1.4 % in SNR 21 dB, 10.7 % in SNR 18 dB, 30.1 % in SNR 10 dB compared with that of *3-D Viterbi search method with delay-and-sum beamforming.*

The word recognition accuracy for the arrangement (2) (the moving talker) is also shown in Table 2. No experiments of *Delay-and-sum beamforming to the correct talker direction* and *AMNOR to the correct talker direction* is carried out, since the correct locus of the talker could not be measured. These results are summarized as follows:

- The word recognition accuracy of the 3-D Viterbi search method for the moving talker is almost equal to that for the fixed-position talker.
- The word recognition accuracy of *3-D Viterbi search method with AMNOR* is improved 3.2 % in SNR 21 dB, 6.9 % in SNR 18 dB, 28.7 % in SNR 10 dB compared with that of *3-D Viterbi search method with delay-and-sum beamforming.*

These results show that the adaptive beamformer drastically improves the recognition performance both for a fixed-position talker and for a moving-talker.

5. CONCLUSION

To improve the performance of the 3-D Viterbi search method in real environments, this paper proposed the novel method based on the adaptive beamforming technique. The speaker-dependent isolated-word recognition experiments were carried out on real environment data to evaluate the effect of the adaptive beamformer. These results showed that the adaptive beamformer drastically improves the recognition performance both for a fixed-position talker and for a moving-talker, when no noise sources moves.

As a future work, to recognize speech of multiple talkers at the same time, we try to apply N-best algorithm to searching in 3-dimensional trellis space.

6. REFERENCES

1. Giuliani, M. Omologo, P. Svaizer, "Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation", Proc. Internat. Conf. on Spoken Language Process., pp. 1329–1332, Oct. 1996.
2. T. Yamada, S. Nakamura, K. Shikano, "Robust speech recognition with speaker localization by a microphone array", Proc. Internat. Conf. on Spoken Language Process., pp. 1317–1320, Oct. 1996.
3. T. Yamada, S. Nakamura, K. Shikano, "Hands-free Speech Recognition Based on 3-D Viterbi Search Using a Microphone Array", Proc. IEEE Internat. Conf. Acoust. Speech Signal Process., pp. 245–248, May 1998.
4. S. U. Pillai, "Array Signal Processing", Springer-Verlag, New York, 1989.
5. Y. Kaneda, J. Ohga, "Adaptive microphone array system for noise reduction", IEEE Trans. Acoustics, Speech, and Signal Processing, 34, 6, pp. 1391–1400, Dec. 1986.
6. B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J. Acoust. Soc. Am., 55, 6, pp, 1304–1312, June 1974.