HANDS-FREE SPEECH RECOGNITION BASED ON 3-D VITERBI SEARCH USING A MICROPHONE ARRAY

Takeshi Yamada, Satoshi Nakamura, and Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, 630-01 Japan

ABSTRACT

A microphone array is the promising solution for realizing hands-free speech recognition in real environments. Accurate talker localization is very important for speech recognition using the microphone array. However localization of a moving talker is difficult in noisy reverberant environments. The talker localization errors degrade the performance of speech recognition. To solve the problem, this paper proposes a new speech recognition algorithm which considers multiple talker direction hypotheses simultaneously. The proposed algorithm performs Viterbi search in 3-dimensional trellis space composed of talker directions, input frames, and HMM states. As a result, a locus of the talker and a phoneme sequence of the speech are obtained by finding an optimal path with the highest likelihood. To evaluate the performance of the proposed algorithm, speech recognition experiments are carried out on simulated data and real environment data. These results show that the proposed algorithm works well even if the talker moves.

1. INTRODUCTION

In recent years, the use of a microphone array for hands-free speech recognition has been investigated. The microphone array is composed of multiple microphones which are spatially arranged. It can take advantages of spatial information about sound sources to suppress noise signals and reverberations.

It is very important for hands-free speech recognition using the microphone array to localize a talker accurately. Recently, several talker localization techniques and the application of these techniques to hands-free speech recognition have been proposed [1, 2, 3, 4, 5]. Most of these systems localize a talker by using short- and long-term power, then extract parameter vectors for speech recognition by steering a beamform to the direction. However localization of a moving talker is very difficult in low SNR conditions and highly reverberant environments. The errors of talker localization degrade the performance of speech recognition.

In order to solve the problem, this paper proposes a new speech recognition algorithm which considers multiple talker direction hypotheses simultaneously. The proposed algorithm extracts the direction-frame sequence of the parameter vectors by steering the beamform to each direction every frame, then performs Viterbi search in 3-dimensional trellis space composed of talker directions, input frames, and HMM states. As a result, a locus of the talker and a phoneme sequence of the speech are obtained by finding an optimal path with the highest likelihood. To evaluate the performance of the proposed algorithm, speech recognition experiments are carried out on simulated data and real environment data.

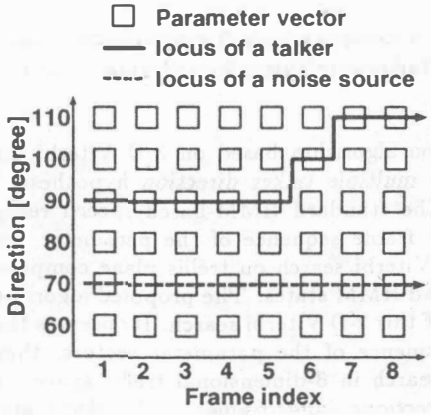

Figure 1: Direction-frame sequence of parameter vectors. The box depicts the parameter vector, the solid line is the locus of a talker, and the dotted line is the locus of a noise source.

2. SPEECH RECOGNITION ALGORITHM BASED ON 3-D VITERBI SEARCH

The direction-frame sequence of parameter vectors $\mathbf{x}(d, n)$ is obtained by steering a beamform to each direction every frame, where $\mathbf{x}(d, n)$ represents the parameter vector such as mel frequency cepstrum coefficients at the direction d and the frame index n. We now consider the direction-frame sequence of the parameter vectors as shown in Figure 1. In Figure 1, the box depicts the parameter vector, the solid line is the locus of a talker, and the dotted line is the locus of a noise source.

The conventional approach localizes a talker by using short- and long-term power. However it is very difficult to estimate the locus of the talker accurately in low SNR conditions and highly reverberant environments. The errors of talker localization degrade the performance of speech recognition.

To solve the problem, this paper proposes a speech

0-7803-4428-6/98 $10.00

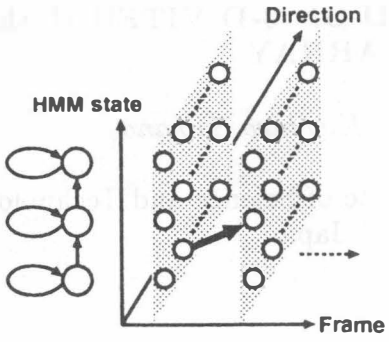

Figure 2: Viterbi search in 3-dimensional trellis space composed of talker directions, input frames, and HMM states.

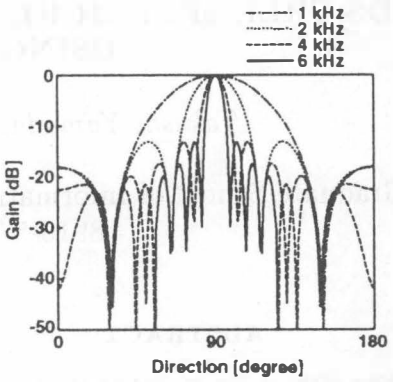

Figure 3: Directive gain pattern calculated for 6 kHz band-limited white Gaussian noise.

recognition algorithm based on 3-D Viterbi search which considers multiple talker direction hypotheses simultaneously. The standard HMM-based speech recognizers observe the frame sequence of the parameter vectors, then perform Viterbi search on trellis plane composed of input frames and HMM states. The proposed algorithm is an extension of this 2-D Viterbi search. It observes the direction-frame sequence of the parameter vectors, then performs Viterbi search in 3-dimensional trellis space composed of talker directions, input frames, and HMM states as shown in Figure 2. As a result, a locus of the talker and a phoneme sequence of the speech are obtained by finding an optimal path with the highest likelihood. The likelihood is calculated as follows:

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

where d is the direction, n is the frame index, and q is the HMM state index. $a(q',q)$ is the transition probability from the HMM state q' to q, $a(d',d)$ is the transition probability from the direction d' to d, and b is the output probability. The transition probability $a(d',d)$ represents how likely the talker moves. It can be assumed that the talker moves to neighboring directions at most, since a duration of the frame in speech recognition is about 10 msec. Therefore it is reasonable to restrict the range of movements as follows:

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

where Δd is the range of movements.

As mentioned above, the proposed algorithm finds an optimal path with the highest likelihood. Therefore, when the likelihood in the correct talker direction is lower than that in the other ones, the performance of the proposed algorithm is degraded. In such a case, it will be 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. In this paper, an weight function based on the pitch harmonics is introduced as follows:

$$w(d,n) = \log \frac{\sum_{n'=n-(\nu-1)}^{n} \{c(d;n')\}^{\mu}}{\sum_{d'=0}^{180} \sum_{n'=n-(\nu-1)}^{n} \{c(d';n')\}^{\mu}}, \quad (3)$$

where $c(d;n)$ is the maximum value of cepstrum coefficients in high quefrency region, which is extracted by cepstrum analysis for the direction d and the frame index n. 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. EXPERIMENTS AND RESULTS

3.1. Experiment Conditions

A speech recognizer is based on the tied-mixture HMM with 256 distributions. A speech corpus is the ATR Japanese speech database Set-A. 2620 words of the speaker MHT are used for training 54 context independent phoneme models and another 216 words are used for testing. Speech signals are sampled at 12 kHz and windowed by the 32 msec Hamming window every 8 msec. Then 16-order mel frequency cepstrum coefficients (MFCCs), 16-order ΔMFCCs, and a Δpower are calculated. A microphone array is a linear and equally spaced array composed of 14 microphones, where the distance between two adjacent microphones is 2.83 cm. As the microphone array signal processing, the delay-and-sum beamformer is used. The directional gain pattern calculated for 6 kHz band-limited white Gaussian noise is shown in Figure 3. The direction-frame sequence of the parameter vectors is computed every 10 degree.

3.2. Simulated data experiments

We consider two arrangements of the sound sources and the microphone array as shown in Figure 4.

(1) The positions of the talker and the noise source are fixed.

Table 1: Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] for the fixed talker on the simulated data.

	Clean		SNR 20 dB		SNR 10 dB	
	WA	TLA	WA	TLA	WA	TLA
Single microphone	96.2	—	80.0	—	25.9	—
Delay-and-sum beamformer	96.2	100.0	94.9	100.0	90.7	100.0
3-D Viterbi search 1 ($\Delta d = 10$)	96.2	99.0	72.6	40.9	28.2	17.9
3-D Viterbi search 2 ($\Delta d = 10, \mu = 40, \nu = 20$)	96.2	99.1	94.9	77.6	88.4	71.2

Table 2: Word recognition accuracy (WA) [%] and talker localization accuracy (TLA) [%] for the moving talker on the simulated data.

	Clean		SNR 20 dB		SNR 10 dB	
	WA	TLA	WA	TLA	WA	TLA
Single microphone	95.8	—	79.6	—	22.2	—
Delay-and-sum beamformer	95.8	100.0	91.6	100.0	86.5	100.0
3-D Viterbi search 1 ($\Delta d = 10$)	96.2	86.7	74.5	44.7	26.8	21.0
3-D Viterbi search 2 ($\Delta d = 10, \mu = 40, \nu = 10$)	96.7	84.0	93.9	63.1	84.7	51.9

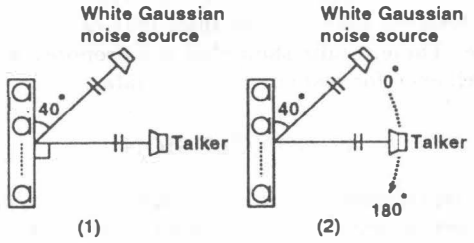

Figure 4: Arrangement of the sound sources and the microphone array. (1) The positions of the talker and the noise source are fixed. (2) The talker moves from 0 degree to 180 degree while uttering each word. The position of the noise source is fixed.

(2) The talker moves from 0 degree to 180 degree while uttering each word. The position of the noise source is fixed.

The outputs of each microphone are simulated considering only time difference of the wavefront arrival without reverberations.

Word recognition accuracy (WA) and talker localization accuracy (TLA) for the fixed talker are shown in Table 1. The TLA is defined as follows:

$$\text{TLA} = \frac{\text{number of correct frames}}{\text{number of total speech frames}} \times 100[\%], \qquad (4)$$

where the number of correct frames is the number of frames that the correct talker direction is selected. In Table 1, the delay-and-sum beamformer indicates that the correct talker direction is known. The frame sequence of the parameter vectors is obtained by steering the beamform only to the correct talker direction. The 3-D Viterbi search 1 indicates that Viterbi search in 3-dimensional trellis space is performed without the weight function. The 3-D Viterbi search 2 indicates that the pitch harmonics weight function is used.

The WA and TLA for the moving talker are shown in Table 2. In Table 2, the TLA is defined based on the number of frames that the correct talker direction is selected within 5 degree difference.

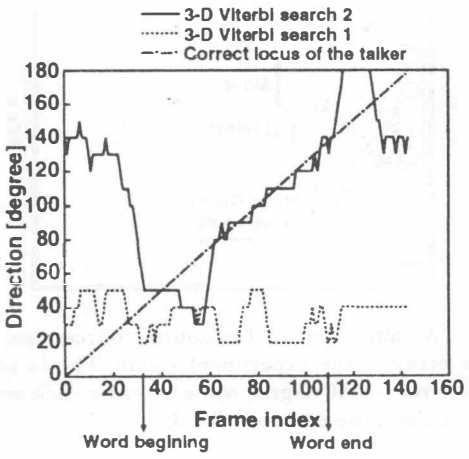

Figure 5: Example of the locus of the talker obtained by 3-D Viterbi search 1 and 2 for the moving talker in SNR 20 dB

These results are summarized as follows:

- The WA of the delay-and-sum beamformer is improved by steering the beamform only to the correct talker direction compared with that of the single microphone.
- The WA of the 3-D Viterbi search 2 is almost equal to that of the delay-and-sum beamformer, while the WA of the 3-D Viterbi search 1 is degraded due to low TLA.

An example of the locus of the talker obtained by the 3-D Viterbi search 1 and 2 for the moving talker in SNR 20 dB is shown in Figure 5, where the horizontal axis is the frame index and the vertical axis is the direction. The locus of the talker obtained by the 3-D Viterbi search 1 is degraded, because the likelihood in the correct talker direction is lower than that in the other ones. The 3-D Viterbi search 2 works

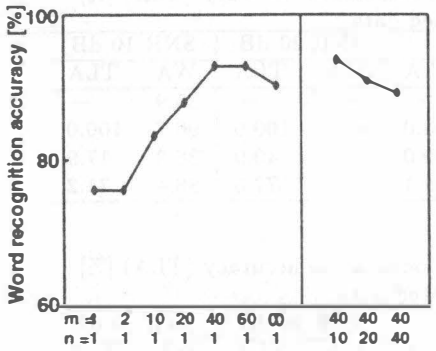

Figure 6: Word recognition accuracy obtained by 3-D Viterbi search 2 in SNR 20 dB for several combinations of the parameter μ and ν.

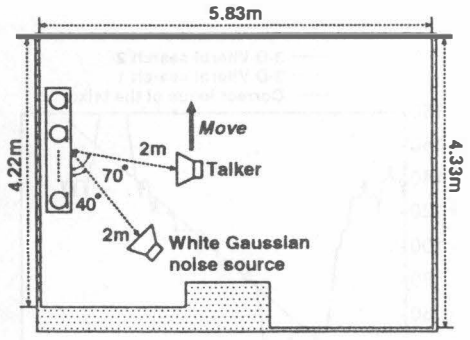

Figure 7: Arrangement of the sound sources and the microphone array in the experiment room. The talker moves from 70 degree to 140 degree while uttering each word. The position of the noise source is fixed.

well by using the weight function compared with the 3-D Viterbi search 1 in the speech period.

Finally we describe the effect of the parameter μ and ν in the weight function, where μ is the parameter to control the weight effect and ν is the parameter to adjust the continuation. Figure 6 shows the WA obtained by the 3-D Viterbi search 2 in SNR 20 dB for several combinations of the parameter μ and ν. When $\nu = 1$, the WA becomes higher by increasing the parameter μ. However, when $\mu = \infty$, which is the case that only one direction is selected as the talker direction according to the maximum value of cepstrum coefficients in high quefrency region, the WA is degraded. These results show that the approach of considering multiple talker direction hypotheses is better than the conventional approach. When $\mu = 40$, the WA becomes lower by increasing the parameter ν, because the talker moves.

3.3. Real environment data experiments

The arrangement of the sound sources and the microphone array in the experiment room is shown in Figure 7. The reverberant time in the room is about 0.18 sec. The talker moves from 70 degree to 140 degree while uttering each word. The position of the noise source is fixed. Two loud speakers are used instead of the talker and the noise source. The loud speakers face the microphone array. The cepstrum mean normalization technique is also applied to the speech recognizer described in section 3.1.

Table 3: Word recognition accuracy [%] for the moving talker on the real environment data.

	SNR [dB]		
	Clean	20	10
Single microphone	92.5	77.7	38.4
3-D Viterbi search 2 ($\Delta d = 10, \mu = 80, \nu = 10$)	89.3	81.9	52.3

The word recognition accuracy is shown in Table 3. In Table 3, the SNRs are calculated using the power values at the positions of the sound sources. Therefore the SNRs of the received signals are lower than these values. The experiments of the delay-and-sum beamformer weren't carried out, since the correct locus of the talker couldn't be measured. The WA of the 3-D Viterbi search 2 in SNR 20 dB and SNR 10 dB is higher than that of the single microphone. These results show that the proposed algorithm works well even for real environment data.

4. CONCLUSION

In order to consider multiple talker direction hypotheses, we proposed a speech recognition algorithm based on 3-D Viterbi search. To evaluate the performance of the proposed algorithm, speech recognition experiments were carried out on simulated data and real environment data. These results show that the proposed algorithm works well even if the talker moves by using the weight function based on the pitch harmonics. As a future work, we try to apply N-best algorithm for searching in 3-dimensional trellis space to recognize speech of multiple talkers at the same time.

5. 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. Am., 78, 5, pp. 1508–1518, Nov. 1985.

[2] H. F. Silverman and S. E. Kirtman, "A two-stage algorithm for determining talker location from linear microphone array data", Computer Speech and Language, 6, pp. 129–152, June 1992.

[3] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis", Proc. ICASSP96, pp. 921–924, May 1996.

[4] D. Giuliani, M. Omologo, and P. Svaizer, "Experiments of speech recognition in a noisy and reverberant environment using a microphone array and HMM adaptation", Proc. ICSLP96, pp. 1329–1332, Oct. 1996.

[5] T. Yamada, S. Nakamura, and K. Shikano, "Robust speech recognition with speaker localization by a microphone array", Proc. ICSLP96, pp. 1317–1320, Oct. 1996.