Speech Recognition of a Moving Talker 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, Nara, 630-01 JAPAN

Abstract

This paper describes a new speech recognition algorithm to recognize speech from sound mixtures observed by a microphone array. The proposed algorithm extracts spatial information about sound sources using a microphone array, then performs Viterbi search on 3-D trellis space composed of talker directions, input frames, and HMM states. To evaluate the performance of the proposed algorithm, speech recognition experiments are carried out. These results show that the proposed algorithm works well by using a weight function based on pitch harmonics.

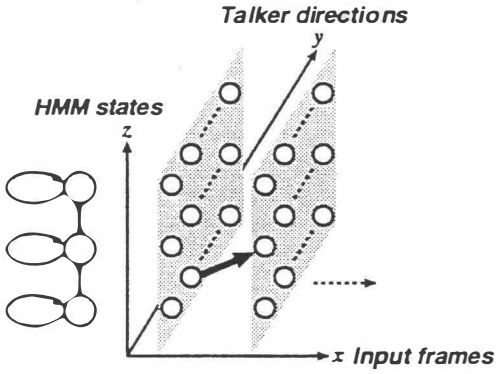

Figure 1: Viterbi search on 3-D trellis space

1 Introduction

It is very important to recognize speech from arbitrary sound mixtures observed in noisy and reverberant environments. Sound mixtures are physical phenomena in acoustical 3-dimensional space. Information about sound sources, such as directions, distances, and characteristics, plays an important role in perception of acoustical environments. Using such information, we can often recognize specific speech from sound mixtures, which typically include speech, non-speech sounds, and music. This paper describes a new speech recognition algorithm to recognize speech from sound mixtures observed by a microphone array.

A microphone array is one of the most promising methods for extracting information about sound sources. A microphone array is composed of multiple microphones spatially arranged. The output of each microphone has phase difference according to positions of sound sources. Utilizing this spatial information, a directional gain pattern sensitive to a talker direction can be formed to separate speech from sound mixtures. Several speech recognition systems using a microphone array have been proposed [Giuliani *et al.* , 1994; Lin *et al.* , 1994; Yamada *et al.* , 1996]. These systems localize a talker direction every frame, then form a directional gain pattern sensitive to the direction. However it is very difficult to localize a direction of a moving talker accurately in low SNR conditions and highly reverberant environments. Therefore the performance of these systems is degraded due to errors of talker localization.

In the conventional systems, speech recognition is carried out after localizing a talker direction. However these two procedures should be performed simultaneously like procedures in human perception. This paper introduces a new method to deal with talker localization and speech recognition simultaneously in a unified framework. This method finds an optimal combination of a transition of talker directions and a phoneme sequence of speech. In general, an HMM-based speech recognition algorithm performs Viterbi search on trellis plane composed of input frames and HMM states. As an extension of this algorithm, this paper proposes a speech recognition algorithm based on Viterbi search on 3-D trellis space composed of talker directions, input frames, and HMM states. To evaluate the performance of the proposed algorithm, speech recognition experiments are carried out.

2 Viterbi Search on 3-D Trellis Space

Figure 1 shows 3-D trellis space, where x-, y-, and z-axis represent input frames, talker directions, and HMM states, respectively. In general, an HMM-based speech recognition algorithm performs Viterbi search on the x-z trellis plane composed of input frames and HMM states.

The proposed algorithm is an extension of this algorithm and performs Viterbi search on the x-y-z trellis space composed of talker directions, input frames, and HMM states. As a result, a transition of talker directions and a phoneme sequence of speech are obtained by finding an optimal path with the highest likelihood. The likelihood is calculated as follows:

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

where d is the direction, n is the frame, and q is the state index. Also $a(q',q)$ is the transition probability from state q' to q, $a(d',d)$ is the transition probability from direction d' to d, and b is the output probability. Finally $\mathbf{x}(d,n)$ is the direction-frame sequence of parameter vectors obtained by steering a beamformer for each direction every frame. The transition probability $a(d',d)$ represents a change of talker directions. Because duration of a frame in speech recognition is about 10 msec, a talker can move to neighboring directions at most. Therefore it is reasonable to restrict range of movements as follows:

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

where $\theta(d)$ is the direction for index d and Θ is the range of movements.

As mentioned above, the proposed algorithm performs Viterbi search for all directions. However, when the likelihood for the correct talker direction is lower than that for the other ones, the performance of the proposed algorithm is degraded. In such a case, it will be effective to raise the likelihood for directions with speech-like characteristics. Pitch harmonics of speech can be used as a measure of speech-like characteristics. In this paper, a weight function based on pitch harmonics is introduced as follows:

$$w(d,n)=\log\frac{\sum_{n'=n-(j-1)}^{n}\{c(d;n')\}^i}{\sum_{d'=1}^{D}\sum_{n'=n-(j-1)}^{n}\{c(d';n')\}^i}, \tag{3}$$

where $c(d,n)$ is the maximum value of cepstrum coefficients in high quefrency region, which is extracted by cepstrum analysis for direction d and frame n. This value becomes larger when pitch harmonics are remarkably included. Also i is the parameter to emphasize weight and j is the parameter to adjust continuation.

3 Experiments and Results

3.1 Experimental Conditions

The speech recognizer is based on tied-mixture HMM with 256 distributions. As a speech corpus, ATR Japanese speech database Set-A is used. 2620 words of the MHT-speaker are used for training 54 context independent phone models, and another 216 words for testing. Speech signals are sampled at 12 kHz and windowed by the 32 msec Hamming window every 8 msec, and then 16-order mel frequency cepstrum coefficients (MFCCs), 16-order ΔMFCCs, and a Δpower are calculated.

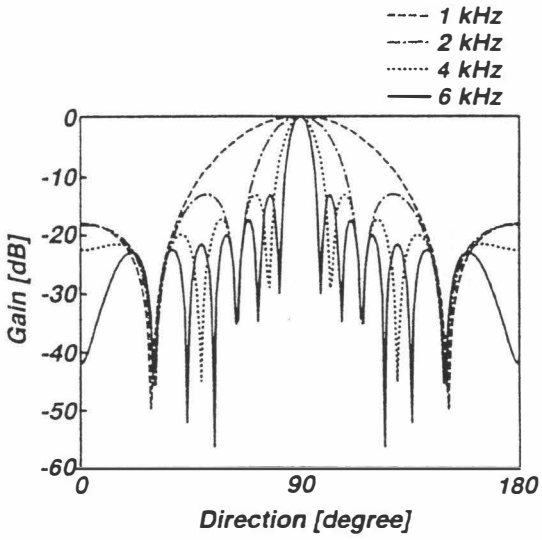

Figure 2: Directional gain pattern

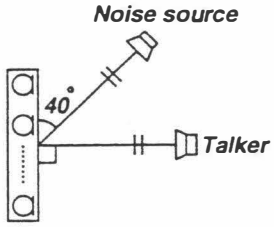

Figure 3: Sound sources and a microphone array

The microphone array is an equally spaced array composed of 14 microphones, where distance between two adjacent microphones is 2.83 cm. The output of each microphone is simulated considering only time difference. As microphone array signal processing, delay-and-sum beamformer is used [Pillai, 1989]. The directional gain pattern, calculated for 6 kHz band-limited Gaussian noise, is shown in Figure 2. Finally the direction-frame sequence of parameter vectors is computed every 10 degree.

3.2 Experiment 1

Sound sources and a microphone array are located as shown in Figure 3. The directions of the talker and the Gaussian noise source are at 90 degree and 40 degree.

Word recognition accuracy (WA) and talker localization accuracy (TLA) are shown in Table 1. The TLA is

Table 1: Word recognition accuracy (WA) and talker localization accuracy (TLA) [%]

	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 ($\Theta = 10$)	96.2	99.3	72.6	24.9	28.2	10.8
3-D Viterbi search 2 ($\Theta = 10, i = 40, j = 20$)	96.2	99.4	94.9	50.4	88.4	45.7

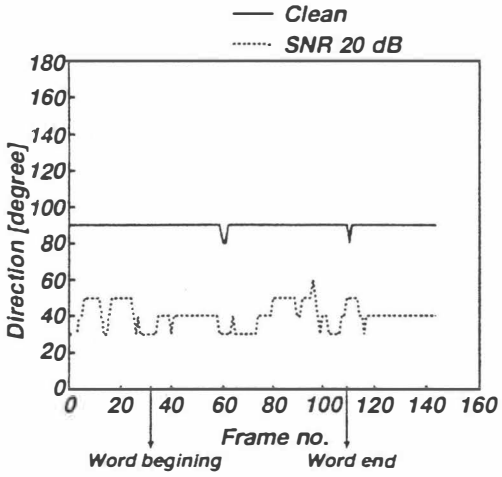

Figure 4: Example of a transition of talker directions obtained by 3-D Viterbi search 1 in clean and SNR 20 dB

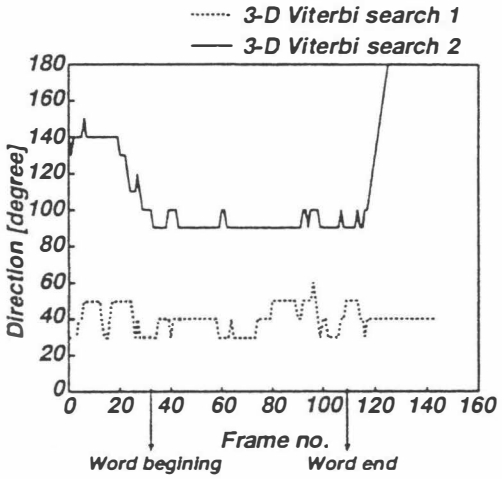

Figure 5: Example of a transition of talker directions obtained by 3-D Viterbi search 1 and 2 in SNR 20 dB

defined as follows:

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

where the number of correct frames is the number of frames detected the correct talker direction. In Table 1, delay-and-sum beamformer indicates that the correct talker direction is known. The frame sequence of parameter vectors is computed only for the correct talker direction, then the conventional Viterbi search is performed. 3-D Viterbi search 1 indicates that Viterbi search on 3-D trellis space is performed. 3-D Viterbi search 2 indicates that a weight function based on pitch harmonics is used.

These results show that performance of 3-D Viterbi search 2 is almost equal to that of delay-and-sum beamformer which is the case that the correct talker direction is known, while performance of 3-D Viterbi search 1 is degraded due to low TLA. An example of a transition of talker directions obtained by 3-D Viterbi search 1 in clean and SNR 20 dB is shown in Figure 4, where horizontal and vertical axis represent frame index and direction. Errors of talker localization in SNR 20 dB are increased compared with those in clean. For example, 60 degree is localized at frame index 97 in SNR 20 dB. This is because of the insufficient beamwidth. One way to solve this problem is to make the directional gain pattern more sharper. However the computational cost is very large for the broadband signals like speech. Another way is to raise the likelihood for directions with speech-like characteristics using a weight function such as Eq. (3). An example of a transition of talker directions obtained by 3-D Viterbi search 1 and 3-D Viterbi search 2 in SNR 20 dB is shown in Figure 5. The performance of 3-D Viterbi search 2 is improved by using a weight function compared with that of 3-D Viterbi search 1.

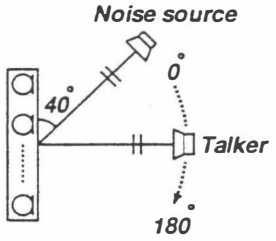

Figure 6: Sound sources and a microphone array

3.3 Experiment 2

Sound sources and a microphone array are located as shown in Figure 6. The direction of the Gaussian noise source is at 40 degree. The talker moves from 0 degree to 180 degree while uttering each word.

Table 2: Word recognition accuracy (WA) and talker localization accuracy (TLA) [%]

	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 ($\Theta = 10$)	96.2	76.4	74.5	32.3	26.8	15.5
3-D Viterbi search 2 ($\Theta = 10, i = 40, j = 10$)	96.7	72.0	93.9	41.5	84.7	33.7

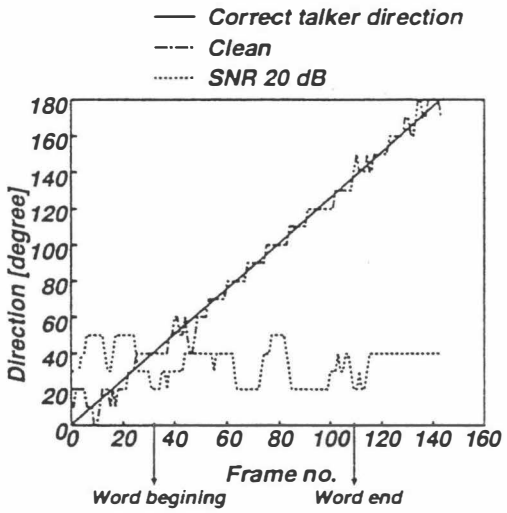

Figure 7: Example of a transition of talker directions obtained by 3-D Viterbi search 1 in clean and SNR 20 dB

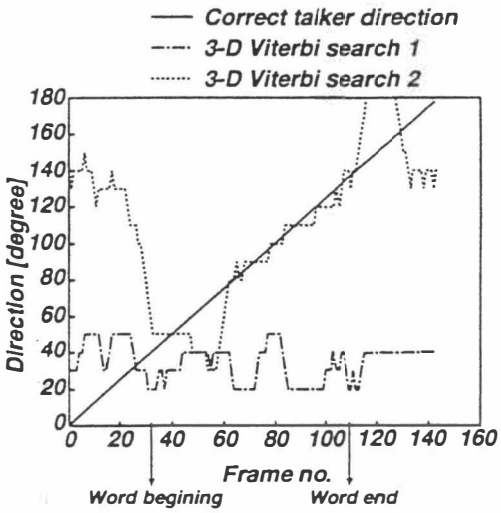

Figure 8: Example of a transition of talker directions obtained by 3-D Viterbi search 1 and 2 in SNR 20 dB

Word recognition accuracy (WA) and talker localization accuracy (TLA) are shown in Table 2. The TLA is defined based on the number of frames detected the correct talker direction within 5 degree difference.

These results show that the performance of 3-D Viterbi search 2 is almost equal to that of delay-and-sum beamformer which is the case that the correct talker direction is known, while performance of 3-D Viterbi search1 is degraded due to low TLA. An example of a transition of talker directions obtained by 3-D Viterbi search 1 in clean and SNR 20 dB is shown in Figure 7, where the solid line is the transition of correct talker directions. Errors of talker localization in SNR 20 dB are increased compared with those in clean. An example of a transition of talker directions obtained by 3-D Viterbi search 1 and 3-D Viterbi search 2 in SNR 20 dB is shown in Figure 8. The performance of 3-D Viterbi search 2 is improved by using a weight function compared with that of 3-D Viterbi search 1.

4 Conclusion and Future Work

This paper proposed a new speech recognition algorithm based on Viterbi search on 3-D trellis space. To evaluate the performance of the proposed algorithm, speech recognition experiments were carried out. These results show that the proposed algorithm works well by using a weight function based on pitch harmonics.

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

References

[Giuliani *et al.* , 1994] D. Giuliani, M. Omologo, and P. Svaizer. *Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis.* In *Proc. ICSLP94*, pp. 1243–1246, Sep. 1994.

[Lin *et al.* , 1994] Q. Lin, E. Jan, C. Che, B. Vries. *System of microphone arrays and neural networks for robust speech recognition in multimedia environment.* In *Proc. ICSLP94*, pp. 1247–1250, Sep. 1994.

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

[Pillai, 1989] S. U. Pillai. *Array signal processing.* Springer-Verlag, New York, 1989.