RECOGNITION OF DISTANT-TALKING SPEECH BASED ON 3-D TRELLIS SEARCH USING A MICROPHONE ARRAY AND ADAPTIVE BEAMFORMING

Satoshi Nakamura, Takeshi Yamada, Panikos Heracleous, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan. Email:nakamura@is.aist-nara.ac.jp

ABSTRACT
Application of a microphone array is a promising solution for realizing distant talking speech recognition, since it can make use of spatial information of sound sources in the environment. Generally, accurate talker localization is very important for speech recognition using a microphone array. However localization of a moving talker is difficult in noisy reverberant environments. We have proposed a novel speech recognition algorithm which considers multiple talker direction hypotheses simultaneously. The algorithm performs Viterbi search in 3-dimensional trellis space composed of talker directions, input frames, and HMM states. As a result, the direction sequence of the talker and the phoneme sequence of the speech are obtained by finding an optimal combination with the highest likelihood. This paper describes the performance improvements by applying an adaptive beamforming technique, and the extension of the algorithm to recognition of simultaneous multiple talkers based on the N-best search.

1. INTRODUCTION
Speech recognition of distant moving talker is one of key technologies for natural human-machine interface. However, distant-talking speech contaminated by environment noise and room reverberation seriously degrades speech recognition performance. Application of a microphone array is a promising solution for realizing distant-talking speech recognition, since it can make use of spatial information of sound sources in the environment. In recent years, speech recognition systems with a microphone array have been proposed to realize distant-talking 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 beam to the localized talker. However, localization of a moving talker is very difficult in low SNR 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 while performing speech recognition. The 3-D trellis space is prepared to be composed of a direction-frame sequence of parameter vectors by steering a microphone array beam 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. The direction sequence of the talker and the phoneme sequence of the speech are obtained by finding an optimal path with the highest likelihood.

The performance of the 3-D Viterbi search method depends on the SNR enhancement by beamforming techniques. The delay-and-sum beamformer [4] was used in our previous study. However, it is very difficult to make the beamformer sharp, because many microphone elements are necessary. To improve the performance of the 3-D Viterbi search method in real environments, this paper describes a method based on an adaptive beamforming technique. Usually, there are many sound sources in the environments including target talkers’ speech. To deal with multiple sound sources, the 3-D Viterbi search is extended to the N-best search paradigm. 3-D N-best search can realize simultaneous recognition of multiple sound sources including the target talkers’ speech. The isolated-word recognition experiments are carried out to evaluate the effects of the adaptive beamformer and the N-best search of multiple sound sources.

2. 3-D VITERBI SEARCH
The 3-D Viterbi search algorithm has been proposed to deal with multiple direction hypotheses of the sound source. The direction sequence of the talker and the phoneme sequence of the speech are obtained as an optimal combination with the highest likelihood. A direction-frame sequence of parameter vectors is obtained by steering a microphone array beam to each direction in every frame. Given the direction-frame sequence of parameter vectors, talker localization and speech recognition can be performed simultaneously.
in the statistical framework as follows:

\[
(q_\text{d}, d) = \arg \max_{(q, d)} P(x \mid q, d, M),
\]

where \((q_\text{d}, d)\) is the optimal combination of the phoneme sequence \(q\) of the speech and the direction sequence \(d\) of the talker, \(M\) is the speech model, and \(x\) is the direction-frame sequence of parameter vectors. The optimal combination \((q_\text{d}, d)\) 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 1. The likelihood is calculated as follows:

\[
\alpha(q, d, n) = \max_{q', d'} \{\alpha(q', d', n - 1) + \log a_1(q', q) + \log a_2(d', d) + \log b(q, x(d, n))\},
\]

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:

\[
a_2(d', d) = \begin{cases} 
\frac{1}{2 \Delta d}, & |d - d'| \leq \Delta d \\
0, & |d - d'| > \Delta d
\end{cases},
\]

where \(\Delta d\) is the range of movements.

An example of the talker direction obtained by the 3-D Viterbi search for the moving talker in SNR 20 dB is shown in Figure 2, where the horizontal axis is the frame index and the vertical axis is the direction. The 3-D Viterbi search works well in the speech period.

### 3. ADAPTIVE BEAMFORMING

The performance of the 3-D Viterbi search method depends on the SNR improvement by using the microphone array beamforming. An adaptive beamforming is applied to improve 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 \(m\)th microphone element and \(H_m(\omega)\) is the frequency response of the \(m\)th filter. The frequency response \(F(\omega)\) of the adaptive beamformer to the desired signal is represented as follows:

\[
F(\omega) = \sum_{m=1}^{M} G_m(\omega)H_m(\omega),
\]

where \(M\) is the number of microphone elements. The concept of the adaptive beamformer is to minimize the output noise power under the constraint that \(F(\omega)\) is equal to the desired frequency response. In this paper, Equation (5) is used as the AMNOR constraint [4].

\[
D = \int |1 - F(\omega)|^2 d\omega,
\]
The AMNOR constraint attains maximum noise reduction while allowing a small distortion $D$ in the frequency response to the desired signal.

### 4. 3-D N-BEST SEARCH

The 3-D Viterbi method is able to extend to an 3-D N-best search for simultaneous recognition of multiple sound sources. The 3-D N-best search considers multiple hypotheses for each state and direction $(q,d)$. The N-best hypotheses are found by considering all the predecessor hypotheses which end in $(q,d)$ at frame $n$. The Equation (6) shows the general way to calculate the likelihoods of the N-best hypotheses.

$$
\alpha_N(q,d) = \text{sort}\{\alpha_N(q',d',n-1) + \log \alpha_1(q',q) \\
+ \log \alpha_2(d',d) + \log b(q,d,n)\} 
$$

As a result of the 3-D N-best search multiple hypotheses can be obtained and in this way multiple sound sources can be localized and recognized simultaneously.

### 5. EXPERIMENT RESULTS

#### 5.1 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. The talker and the noise source are located at 70 degree and 140 degree, respectively.
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 4.

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.

#### 5.2 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 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 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 noise exists.
(B) The white Gaussian noise source and the ambient noise exists.

#### 5.3 Results of Single Talker

The word recognition accuracies are 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$\&$sum to the correct talker direction and AMNOR to the correct talker direction indicate the case that the correct talker direction is known. However, no experiments of Delay$\&$sum beamforming to the correct talker direction and AMNOR to the correct talker direction is carried out, since the correct the talker direction 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$\&$sum.

These results show that the adaptive beamformer drastically improves the recognition performance both for a fixed-position talker and for a moving-talker.
Table 1: Word recognition accuracy [%]

<table>
<thead>
<tr>
<th>SNR[dB]</th>
<th>Fixed Position Talker</th>
<th>Moving Talker</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td><strong>Single microphone</strong></td>
<td>89.8</td>
<td>76.8</td>
</tr>
<tr>
<td><strong>Delay&amp;Sum to the correct talker direction</strong></td>
<td>92.1</td>
<td>86.5</td>
</tr>
<tr>
<td><strong>AMNOR to the correct talker direction</strong></td>
<td>94.4</td>
<td>91.2</td>
</tr>
<tr>
<td><strong>3-D Viterbi search with delay&amp;sum</strong></td>
<td>92.5</td>
<td>79.1</td>
</tr>
<tr>
<td><strong>3-D Viterbi search with AMNOR</strong></td>
<td>93.9</td>
<td>89.8</td>
</tr>
</tbody>
</table>

Table 2: Results for 5-best. The sound sources are included in the list.

<table>
<thead>
<tr>
<th>Input</th>
<th>MHT</th>
<th>FTK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>/yuumo/a</td>
<td>/omowazu/</td>
</tr>
<tr>
<td>Best</td>
<td>Word</td>
<td>Likelihood</td>
</tr>
<tr>
<td>1</td>
<td>omowazu</td>
<td>-75.5039</td>
</tr>
<tr>
<td>2</td>
<td>imagoro</td>
<td>-75.5795</td>
</tr>
<tr>
<td>3</td>
<td>yuumoa</td>
<td>-75.8902</td>
</tr>
<tr>
<td>4</td>
<td>uyamau</td>
<td>-75.9920</td>
</tr>
<tr>
<td>5</td>
<td>ikioi</td>
<td>-75.9998</td>
</tr>
</tbody>
</table>

Table 3: Results for 5-best. The sound sources are not included in the list.

<table>
<thead>
<tr>
<th>Input</th>
<th>MHT</th>
<th>FTK</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>/ikioi/</td>
<td>/kakurepyuuritaN/</td>
</tr>
<tr>
<td>Best</td>
<td>Word</td>
<td>Likelihood</td>
</tr>
<tr>
<td>1</td>
<td>kakurepyuuritaN</td>
<td>-79.2763</td>
</tr>
<tr>
<td>2</td>
<td>hiQkurikaesu</td>
<td>-80.3220</td>
</tr>
<tr>
<td>3</td>
<td>oiharau</td>
<td>-80.3852</td>
</tr>
<tr>
<td>4</td>
<td>akegata</td>
<td>-80.4967</td>
</tr>
<tr>
<td>5</td>
<td>atarimae</td>
<td>-80.5042</td>
</tr>
</tbody>
</table>

5.4 Results of Multiple Talker

In order to evaluate the performance of the simultaneous recognition experiments were carried out with simulated data. The testing data are 216 phoneme balanced words of the MHT-talker and FTK-talker. The sound sources are located in fixed position as is shown in the Figure 5. Each talker pronounces a different word respectively.

Table 2 shows results when the words are included in the N-best list. Table 3 shows results when the correct word isn’t included. A problem, which seems to have an impact effect to the performance is the different duration of the sound sources. The current N-best algorithm attempts to find the word which lasts longer than the others.

6. CONCLUSION

To improve the performance of the 3-D Viterbi search method in real environments, this paper described 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. The 3-D Viterbi search was extended to N-best search in 3-D trellis space so as to recognize speech of multiple talkers at the same time. The preliminary experiments for simultaneous speech recognition of two fixed talkers are carried out. Although the N-best search results the possibility of simultaneous recognition of multiple sound sources, the algorithm still needs improvement.

7. REFERENCES

2. T. Yamada, et al, “Robust speech recognition with speaker localization by a microphone array”, ICSLP96

Figure 5: Position of sound sources