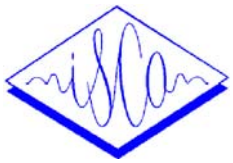

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

SIMULTANEOUS RECOGNITION OF MULTIPLE SOUND SOURCES BASED ON 3-D N-BEST SEARCH USING MICROPHONE ARRAY

Panikos HERACLEOUS[1], Takeshi YAMADA[2], Satoshi NAKAMURA[3], Kiyohiro SHIKANO[4]
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5, Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN

10.21437/Eurospeech.1999-21

ABSTRACT

The recognition of distant talking speech in a noisy and reverberant environments is key issue in any speech recognition system. A so-called hands-free speech recognition system plays an important role in the natural and friendly human-machine interface. Considering the practical use of a speech recognition system, we realize that such a system has to deal, also, with the case of the presence of multiple sound sources, including multiple talkers, as well as other noise sources. This paper proposes a novel method which recognizes multiple talkers simultaneously in real environments by extending the 3-D Viterbi search to a 3-D N-best search algorithm. While the 3-D Viterbi method finds the most likely path in the 3-D trellis space, the proposed method considers multiple hypotheses for each direction in every frame. Combinations of the direction sequence and the phoneme sequence of multiple sources are included in the N-best list. The paper investigates the performance of the proposed method through experiments using real utterances of multiple talkers.

1. INTRODUCTION

In recent years several works which focus on the hands-free speech recognition have been introduced. Most of those recognition systems are microphone array-based systems. The use of the microphone array is based on the fact that the microphone array can take advantage of the use of the spatial information about the sound sources to suppress noise signals and reverberations.

In order to achieve high quality sound acquisition by beamforming techniques it is essential to localize a talker accurately. Most of the speech recognizers using microphone array localize a talker by using short- and long-term power, and then extract a frame sequence of parameter vectors for speech recognition by steering a beamformer to the direction [1][2][3] . However localization of a moving talker is very difficult in low SNR conditions and highly reverberant environments.

One way to solve this problem is to integrate microphone array processing and speech recognition. In the last year, a speech recognition algorithm based on a 3-D Viterbi search has been proposed [4][5]. A direction-frame sequence of parameter vectors (e.g. mel-frequency cepstrum coefficients) can be obtained by steering a beamformer to each direction in every frame. The parameter vectors in the talker direction are extracted from high quality speech. Therefore, the talker direction may be estimated by matching between the direction-frame sequence of parameter vectors and HMMs. The 3-D Viterbi method performs talker localization and speech recognition simultaneously by finding the most likely path in a 3-dimensional trellis space composed of talker directions, input frames and HMM states. Speaker-dependent isolated-word recognition experiments have shown that word recognition rates of the 3-D Viterbi method with adaptive beamforming in a real room for a moving talker case are drastically improved compared with those of a remote single microphone [5]. Although the 3-D Viterbi search method is a promising way to realize hands-free speech recognition in a real environment, its applicable situations are restricted to those of only one talker. However, for practical use it is necessary to deal with multiple talker situations, too. This paper proposes a new algorithm for simultaneous recognition of multiple talkers by introducing the N-best paradigm in the 3-D Viterbi search.

2. 3-D VITERBI SEARCH

Due to the fact that the localization errors have an impact effect to the performance, several methods have been proposed in order to solve the localization problem. Most of the proposed methods are based on the extraction of the direction with the maximum power. The speech recognition algorithm based on 3-D Viterbi search approaches the problem in a different way. By steering a beamformer to every direction in each frame a direction-frame sequence of parameter vectors is obtained. Based on the fact that

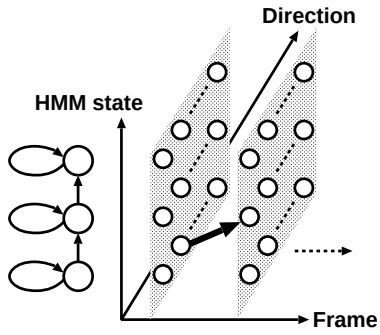

Figure 1. 3-D trellis space

the parameter vectors in the talker direction are extracted with high quality, the talker can be localized by matching between the direction-frame sequence of parameter vectors and the HMMs.

In the case of the 3-D Viterbi search based algorithm, the extraction of the direction-frame parameter vector is followed by the Viterbi search, which is performed in 3-D trellis space [fig.1] composed of talker directions, input frames and HMM states. Based on the maximum likelihood an optimal path can be found and, in this way a combination of talker direction sequence and phoneme sequence of the speech can be obtained.

The optimal combination of the direction and state sequence $(\hat{d}, \hat{q})$ can be found by using the formula

$$(\hat{q}, \hat{d}) = \underset{q,d}{argmax} \Pr(\mathbf{x}|d, q, M) \quad (1)$$

This likelihood can be calculated using the Viterbi formula

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

where M is the model, q,d,n are the state, direction and frame index respectively, b is the output probability, $a_1(q',q)$ is the transition probability from state q' to state q and $a_2(d',d)$ is the transition probability from direction d' to direction d. The $a_2(d',d)$ probability represents how likely the talker moves.

3. 3-D N-BEST SEARCH

The method is based on the idea that the recognition of multiple sound sources can be achieved by introducing the N-best paradigm. The recognition of multiple sound sources can be performed by considering multiple hypotheses for each direction in every frame. The proposed method is one-pass algorithm, which performs baseline N-best search in the 3-D trellis space.

In a similar way with the conventional 3-D Viterbi search based method, the direction-frame sequence of parameter vector is extracted by steering a beamformer to each direction in every frame.

The 3-D N-best search considers multiple hypotheses for each state and direction (q,d). The N-best $\underline{\alpha}^N$ hypotheses are found by considering all the predecessor hypotheses which end in (q,d) at frame n. The arrival hypotheses are merged and the unique ones with different direction sequence are sorted in order to find the N-best hypotheses. The formula 3. shows the general way to calculate the likelihood of the N-best hypotheses.

$$\begin{aligned}\underline{\alpha}^N(q,d,n) &= \underset{d',q'}{\mathbf{sort}}\{\underline{\alpha}^N(q',d',n-1) + \log a_1(q',q) \\ &\quad + \log a_2(d',d)\} + \log b(q, \mathbf{x}(d,n)) \quad (3)\end{aligned}$$

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.

The beamformer is steered to each direction and multiple hypotheses are taken into account. The system should deal with a huge number of hypotheses, which results high memory requirements and low recognition speed. Convetional beam pruning could be applied in order to reduce the number of the considered hypotheses.

An additional problem which the described 3-D N-best search faces is the case when the likelihood in the correct direction is lower than that in other directions. In this case the performance of the method is degraded. The effect of this problem can be reduced by introducing a weight function based on the power, which raises the likelihood in directions with speech-like characteristics. The introduced weight function, which results higher recognition rates is given by the following formula :

$$w(d,n) = \log \frac{\sum_{n'=n-(\nu-1)}^{n} \{p(d;n')\}^{\mu}}{\sum_{d'=1}^{D} \sum_{n'=n-(\nu-1)}^{n} \{p(d';n')\}^{\mu}}, \quad (4)$$

where $p(d;n)$ is the power. This value is extracted for the (d,n) direction, frame index respectively. The μ is the parameter to control the weight effect, ν is the parameter for adjusting the continuation and D is the number of directions.

4. EXPERIMENTS AND RESULTS

In order to evaluate the performance of the proposed method preliminary experiments were carried out on

Table 1: Results for 4-best. Sorting according to the likelihood. Both sound sources are included in the N-best list.

Input	MHT /ogosoka/	FTK /yotsukado/

Best	Word	Likelihood
1	/ogosoka/	-77.0445
2	/yotsukado/	-77.3181
3	/monosugoi/	-77.3501
4	/naosara/	-77.4224

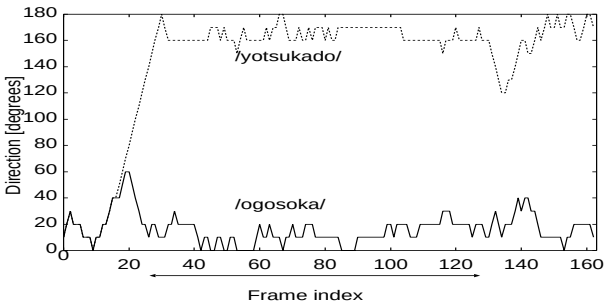

Figure 2. Transition of the two sound sources /ogosoka/ and /yotsukado/

simulated data (only time delay).

4.1. Experimental conditions

The speech recognizer is based on tied-mixture HMM with 256 distributions. The 54 context dependent phoneme models are trained with the 64 speakers ASJ speaker-independent database. The testing data are 216 phoneme balanced words of the MHT- and FTK-speaker of the ATR database SetA. The feature vectors are of length 33 (16 MFCC, 16 ΔMFCC and Δpower). A linear array composed of 64 microphones is used and the distance between them is 2.83 cm. The sound sources are located in fixed position at 10 and 170 degrees, respectively.

4.2. Experiment I

The two talkers, MHT- and FTK-speaker pronounce

Table 2: Results for 4-best. Sorting according to the likelihood. Only one sound source is included in the N-best list.

Input	MHT /ikioi/	FTK /kakurepyuuritaN/

Best	Word	Likelihood
1	/kakurepuuritaN/	-79.2763
2	/hiQkurikaesu/	-80.3220
3	/oiharu/	-80.3851
4	/akegata/	-80.4967

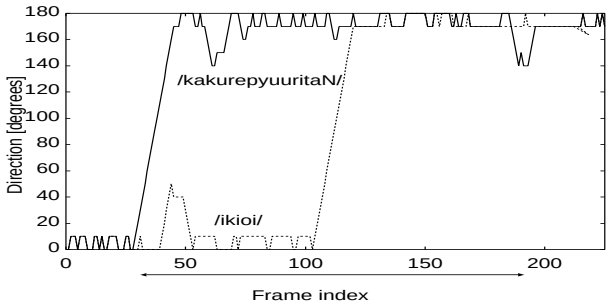

Figure 3. Transition of the two sound sources /ikioi/ and /kakurepyuuritaN/

Table 3: Results for 4-best. Sorting according to the likelihood. Both sound sources are included in the N-best list.

Input	MHT /ikioi/	FTK /kakurepyuuritaN/

Best	Word	Likelihood
1	/ikioi/	-82.3510
2	/oNnamyoori/	-82.8542
3	/nakanaori/	-83.1201
4	/kakurepyuuritaN/	-83.6749

a different word, respectively. Table 1 shows the achieved N-best list,which includes both pronounced words and figure 2 shows the transition of the two words. The N-best list is sorted according to the likelihood. Table 2 shows an example in which the two sound sources aren't included and figure 3 shows the transition of the two sound sources in this case. A possible reason which results, that the sound sources aren't included in N-best list is the different duration of the sound sources. Since the duration of the word */kakurepyuuritaN/* is much longer than the duration of the word */ikioi/* the hypotheses about the second word can't survive. However, if a number of the last frames is discarded, then we see that the search finds both words. Table 3 shows the achieved N-best list and fig. 4 the transition of the two sound sources.

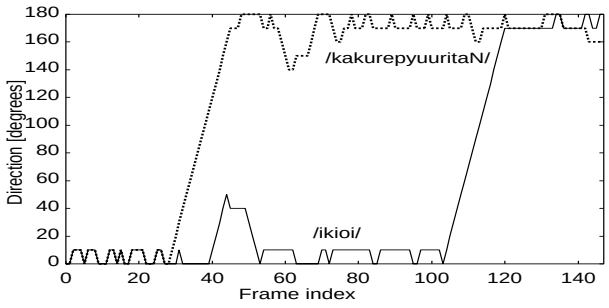

Figure 4. Transition of the two sound sources /ikioi/ and /kakurepyuuritaN/. Last frames are discarded.

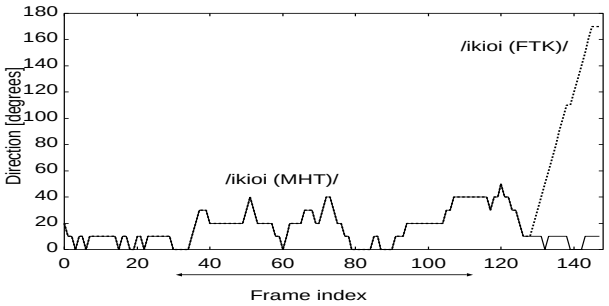

Figure 5. Transition of the two sound sources. Speaker-independent models

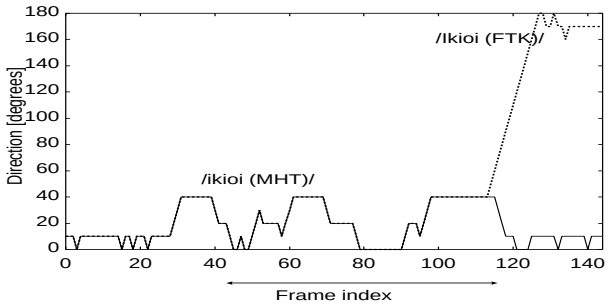

Figure 6. Transition of the two sound sources. Speaker-dependent models

Table 4: N-best hypotheses sorted according to the likelihood. The pronounced words are included in the N-best list.

	MHT	FTK
Input	/ikioi/	/ikioi/

Best	Word	Likelihood
1	/ikioi/	-73.1649
2	/omoshiroi/	-74.5365
3	/iyoiyo/	-74.6903
4	/nakanaori/	-74.6947

4.3. Experiment II

In this experiment the two talkers are the MHT- and FTK-talker respectively, who pronounces the same word. In this case the recognition is performed using the speaker-independent models. Table 4 shows an example of an achieved N-best list sorted according to likelihood and figure 5 shows the transition of the two pronounced words.

The experiment is modified in the sense that the recognition is performed using speaker-dependent MHT- and FTK-trained models. The 54 phoneme models are trained using 2620 words from the FTK- and MHT-speaker of the ATR SetA database, respectively. Figure 6 shows an example, when the recognition is performed using the FTK-trained models.

Figure 5 and figure 6 shows similarity. As the two figures show only the path corresponding to the MHT-talker survives. A possible reason is that the MHT-trained models match better the uttered words. The problem, which appear here could be solved by normalizing the likelihoods.

5. CONCLUSION AND FUTURE WORK

A method for simultaneous recognition of multiple sound sources has been proposed. Examples have been introduced illustrating that the proposed method can perform simultaneous recognition of multiple sound sources. However, problems such as the effect of the different duration of the sound sources and the likelihood normalization's problem are still remaining. As future work, we will try to improve the performance by offering solution to the described problems. Moreover, we will carry out experiments for recognition of non-speech sound sources.

REFERENCES

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

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

[3] T. Hughes, H. Kim, J. DiBiase, H. Silverman, "Using a real time, tracking microphone array as input to an HMM speech recognizer", ICASSP98, pp. 249-252, May 1998.

[4] T. Yamada, S. Nakamura, K. Shikano, "Hands-free Speech Recognition Based on 3-D Viterbi Search Using a Microphone Array", ICASSP98, pp. 245-248, May 1998.

[5] T. Yamada, S. Nakamura, K. Shikano, "An Effect of Adaptive Beamforming on Hands-free Speech Recognition Based on 3-D Viterbi Search", ICSLP98, pp. 381-384, Dec. 1998.