A MICROPHONE ARRAY-BASED 3-D N-BEST SEARCH ALGORITHM FOR THE SIMULTANEOUS RECOGNITION OF MULTIPLE SOUND SOURCES IN REAL ENVIRONMENTS

Panikos Heracleous[1,2], *Satoshi Nakamura*[1], *Kiyohiro Shikano*[2]

ATR Spoken Language Translation Research Labs[1]
Graduate School of Information Science, Nara Institute of Science and Technology[2]

ABSTRACT

This paper deals with the recognition of distant talking speech and, particularly, with the simultaneous recognition of multiple sound sources. A problem that must be solved in the recognition of distant talking speech is talker localization. In some approaches, the talker is localized by using short- and long-term power. The 3-D Viterbi search based method proposed by Yamada et al., integrates talker localization and speech recognition. This method provides high recognition rates but its application is restricted to the presence of one talker. In order to deal with multiple talkers, we extended the 3-D Viterbi search method to a 3-D N-best search method enabling the recognition of multiple sound sources. This paper describes our baseline 3-D N-best search-based system and two additional techniques, namely, a likelihood normalization technique and a path distance-based clustering technique. The paper also describes experiments carried out in order to evaluate the performance of the system.

1. INTRODUCTION

The recognition of distant talking speech plays an important role for any practical speech recognition system. Factors that should be considered include, noisy and reverberant environments, the presence of multiple sound sources, moving talkers, etc. Most of systems are microphone array-based, since a microphone array can take advantage of the spatial and acoustical information of a sound source. More specifically, a microphone array can form multiple beams and therefore can be electronically steered simultaneously to multiple directions each time. In contrast, the use of a single microphone provides limited directional sensitivity and can not be applied for the localization of multiple sound sources without physical steering.

A complex problem that must be solved for speech recognition system for distant talking speech involves talker localization and the speech recognition. In some approaches [1] the talker is first localized by using short- or long-term power. Then a beamformer is steered to the hypothesized direction and recognition is performed by extracting the feature vectors in this direction. However, these approaches face a serious problem, namely, the localization of the talker appears to be difficult under low SNR conditions. The 3-D Viterbi search method proposed by Yamada et al. [2], integrates talker localization and speech recognition and performs Viterbi search in a 3-D Trellis space composed of input frames, HMM states, and directions. A beamformer is steered to each direction at each time, and this enables a locus of the sound source and a feature vector sequence to be obtained simultaneously. A 3-D Viterbi search-based system using adaptive beamforming can provide high recognition rates, but it can be applied only in the case of one sound source.

In our previous works [3, 4], we proposed a novel method able to recognize multiple sound sources simultaneously. The method is based on the 3-D Viterbi search method, i.e., extended to a 3-D N-best search method. The method performs full search in all directions and considers N-best word hypotheses and direction sequences. As a result, the algorithm provides an N-best list, which includes the direction sequences and the phoneme sequences of multiple sound sources.

This paper describes the method along with two techniques implemented in a baseline 3-D N-best-based system. The two techniques are as follows :

- Likelihood normalization technique
  The N-best hypotheses are found by sorting hypotheses originated from different sound sources. However, the different sound sources have different likelihood dynamic ranges and therefore the accuracy in comparing them is poor. The proposed likelihood normalization technique enables the hypotheses to be compared.
- Path distance-based clustering technique
  In the case of the baseline system, there is only one N-best list which includes hypotheses originated from different sound sources. However, if the likelihoods are high in one direction the N-best list is occupied by the hypotheses of the sound source located in this direction. We try to solve this problem by implementing a path distance-based clustering technique, which separates the hypotheses according to their directions and provides one N-best list for each sound source. By finding the top N for each cluster the sound sources and their direction sequences can be obtained.

2. 3-D VITERBI SEARCH

3-D Viterbi search attempts to solve the problem of localization in the case of low SNR values, by integrating talker localization and speech recognition. The algorithm performs Viterbi search in a 3-D Trellis space and finds the optimal $(\hat{d}, \hat{q})$ path with the highest likelihood as the Eq. (1) shows. In this equation, q is the state, d is the direction, M is the HMM model, and $\underline{X}$ is the feature vector.

$$(\hat{q}, \hat{d}) = \underset{q,d}{argmax} \Pr(\underline{\mathbf{X}}|d, q, M) \tag{1}$$

In the hypothesized path, a direction sequence and a feature vector sequence can be obtained. The direction sequence corresponds to the locus of the sound source and the feature vector sequence to the uttered speech or to other sound sources. A speech recognition system based on 3-D Viterbi search and using adaptive beamforming can provide high recognition rates and operate efficiently, even in the case of a moving talker. However, the system focuses on the presence of one sound source only. In order to avoid this disadvantage, we extended the 3-D Viterbi search method to a 3-D N-best search method capable of considering multiple sound sources.

3. THE PROPOSED 3-D N-BEST SEARCH METHOD

The proposed 3-D N-best search method is an extension of 3-D Viterbi search and it is based on the idea that recognition of multiple sound sources can be performed by introducing the N-best paradigm. While 3-D Viterbi search considers only the most likely path in a 3-D Trellis space, 3-D N-best search considers multiple hypotheses for each direction and in this way the N paths with the highest likelihoods can be obtained. In a similar way to the conventional 3-D Viterbi search approach the direction-feature vector sequences are extracted by steering the beamformer to each direction at every time frame.

The baseline 3-D N-best search is a one-pass search algorithm, which performs full search in all directions. At each time frame, the arriving hypotheses to a node are considered and the N-best are found by sorting the unique ones with different directions. Equation 2 shows the general way the N hypotheses with the highest likelihoods are found.

$$\begin{aligned}\underline{\alpha}^N(q, d, n) = \underset{d',q'}{\textbf{sort}}\{&\underline{\alpha}^N(q', d', n-1) + \log a_1(q', q) \\ &+ \log a_2(d', d)\} + \log b(q, \mathbf{x}(d, n))\end{aligned} \tag{2}$$

Considering a node at time n , the overall $\underline{\alpha}^N(q', d', n-1)$ predecessor hypotheses are sorted. Then, by adding to those the a_1 state and a_2 direction transition as well as the b output probabilities, the $\underline{\alpha}^N(q, d, n)$ N-best hypotheses can be found.

At the last stage of the recognition system based on 3-D N-best search, the overall provided word-hypotheses are sorted according to their likelihoods and the top N with the highest likelihoods are selected. The correct sound sources are included in the top N hypotheses and the direction sequences can also be obtained.

3.1. Likelihood Normalization

The technique used for likelihood normalization is similar to the method proposed by Matsui T. et al.[5]. That method

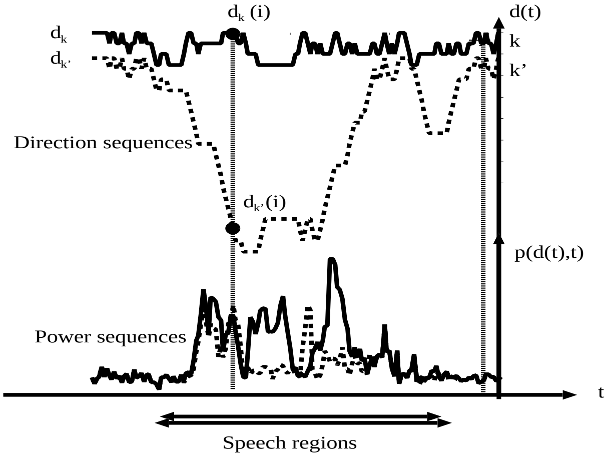

Figure 1: Path distance of a hypothesis-pair

was used for speaker verification, but it was found that it can also be efficiently applied for our task. Our one-state Gaussian mixture (GM) (1 state, 64 mixtures) model is close to that proposed by Matsui T. et al., but its objective is different. This model runs in parallel with the HMM models and its accumulated likelihood is used to normalize the likelihoods of the hypotheses involved. In our approach, the actual accumulated likelihoods $P_s(q, d, t)$ of every state q and direction d are normalized at each time frame t by dividing them with the accumulated likelihood $P_G(d, t)$ of the one-state model. Considering logarithmic likelihoods, Eq. (3) gives the normalized likelihood $\Lambda_s(d, q, t_f)$ at time t_f.

$$\Lambda_s(d, q, t_f) = \sum_{t=0}^{t_f} logP_s(q, d, t) - \sum_{t=0}^{t_f} logP_G(d, t) \tag{3}$$

3.2. Clustering hypotheses using information on path distances

The original 3-D N-best search was extended by implementing the proposed path distance-based clustering. By using information on the provided direction sequences, the top N hypotheses are classified into several clusters. Figure 1 shows the direction and power sequences of two hypotheses. In this experiment, two talkers were located at fixed positions at 10 and 170 degrees. The solid lines describe the hypotheses originated for the 170-degree direction and the dotted line the hypotheses originated from the 10-degrees direction. As can be seen in the speech region, the two hypotheses are well separated based on their directions. In the silence region, the directions of the two hypotheses appear to approach each other, but using the power information we can minimize the effect of this region. Using Eq. (4), the path distance $D(k, k')$ for the two hypotheses can be calculated as follows.

$$D(k, k') = \sum_{i=0}^{N-1} (d_k(i) - d_{k'}(i))^2 (p(d_k(i), i) + p(d_{k'}(i), i)) \tag{4}$$

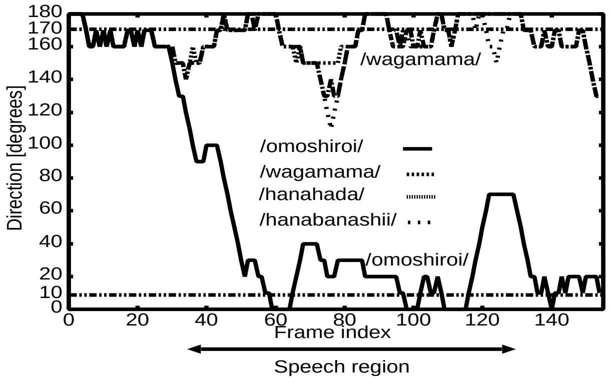

Figure 2: Results show the transitions of hypotheses

In the Eq. (4), N is the total number of frames, k and k' the directions at the final frames of the two hypotheses, d_k the direction sequence ending at k, and $p(d_k)$ the power sequence corresponding to d_k.

The path distance provides the measure that the clustering is based on. By using the path distance, the top N hypotheses are classified into different clusters, which correspond to the sound sources. The number of clusters corresponds to the number of sound sources, and the sound sources can be found by picking up the top N of each cluster. The directions of the sound sources can be obtained by examining the direction sequences of the hypotheses included in each cluster.

Figure 2 shows the transitions of four hypotheses in the case of the pronounced words **/omoshiroi/** and **/waga – mama/**. The two words were pronounced by different talkers. The talkers were located at fixed positions at 10 and 170 degrees. The aim was to classify the hypotheses into two clusters based on the direction sequences. The words **/omoshiroi/** and **/wagamama/** were expected to be included in different clusters and in high orders.

Table 1 shows the results obtained for the case described by Figure 2. The hypotheses are sorted according to the likelihood and no clustering is implemented. As Table 1 shows, only the one sound source is included in the top 4. Table 2 shows results when clustering is implemented. As can be seen, the two sound sources are included in different clusters and both are of the 1st order.

Table 1: Top 4 results. Sorting according to the likelihood. Only one sound source was included in the N-best list.

Input	Speaker A /omoshiroi/	Speaker B /wagamama/

Top	Word	Likelihood
1	**/wagamama/**	-78.5579
2	/hanahada/	-78.9105
3	/hanabanashii/	-78.9776
4	/wazawaza/	-79.2003
..	..	..
7	**/omoshiroi/**	-79.5485

Table 2: Top 4 results. The hypotheses are classified using the path distance.

Input	Speaker A /omoshiroi/	Speaker B /wagamama/

Top	1st Cluster	2nd Cluster
1	**/omoshiroi/**	**/wagamama/**
2	-	/hanahada/
3	-	/hanabanashii
4	-	/wazawaza/

Due to implementation simplicity, a bottom-up method was chosen as the clustering method. A very difficult problem that must be solved is to find the number of the clusters necessary. In this paper, the number of clusters is predefined and is the same as the known number of sound sources.

4. EXPERIMENT AND RESULTS

4.1. Experimental Conditions

The speech recognizer is based on tied-mixture HMM with 256 distributions. Fifty-four context dependent phoneme models are trained with the 64-speaker ASJ speaker independent database. The one-state GMM is also trained using the same database. The test data includes 216 phoneme balanced words of the ATR database SetA, which forms 215 word-pairs. Several speaker- and word-pairs are used. The feature vectors are of length 33 (16 MFCC, 16 ΔMFCC, and Δpower). A linear delay-and-sum array composed of 16 microphones is used and the distance between them is 2.83 cm.

4.2. Definitions

In order to describe the results, three definitions are necessary. Namely, we define the Word Accuracy (WA), Clustering Accuracy, (CA) and Simultaneous Word Accuracy (SWA) as :

$$WA = \frac{Correct\ Words}{Total\ Test\ Wordpairs} \qquad (5)$$

$$CA = \frac{[Word\ \textbf{AND}\ Cluster]\ Correct}{Total\ Correct\ Words} \qquad (6)$$

$$SWA = \frac{[Both\ Word\ \textbf{AND}\ Both\ Cluster]\ Correct}{Total\ Test\ Wordpairs} \qquad (7)$$

4.3. Experimental results for fixed position talkers

The two talkers were located at fixed positions at 10 and 170 degrees. Several speaker- and word-pairs were used. Figures 3 and 4 show the WA of the two speakers and Figure 5 shows the SWA. As can be seen for 'Top 5', the WAs are higher than 80 % and the SWAs close to 80 %. The figures also show the improvements achieved by implementing

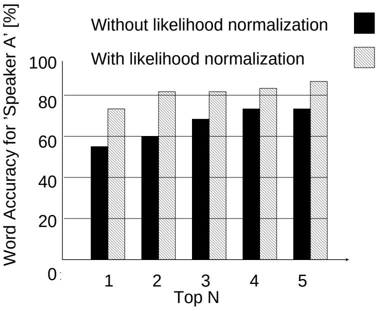

Figure 3: Word Accuracy for 'Speaker A'

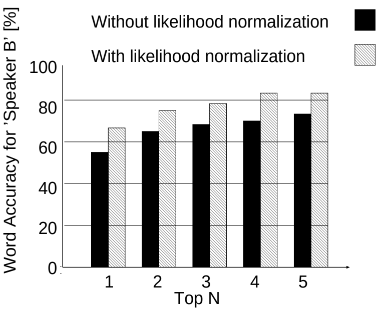

Figure 4: Word Accuracy for 'Speaker B'

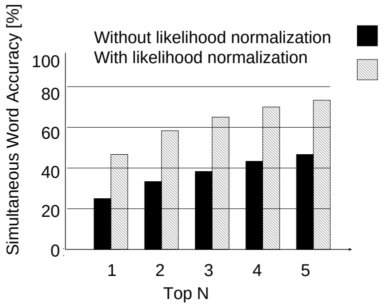

Figure 5: Simultaneous Word Accuracy

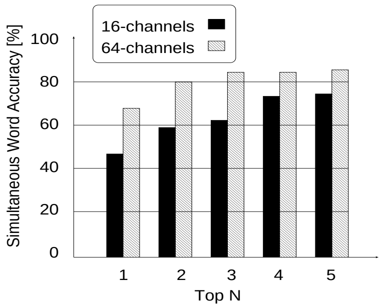

Figure 6: Simultaneous Word Accuracy

the likelihood normalization technique. Figure 6 shows the SWAs in the case of microphone arrays composed of 16 and 64 microphones. In this experiment the two talkers were the MHT- and FSU-talker of ATR database SetA. As can be seen by using more elements significant improvement was achieved. Although further improvements are possible, the obtained results are very promising and they justify the existence of our idea.

4.4. Experimental results for a moving talker

In this experiment one of the two talkers was located at fixed position at 10 degrees, the other one moved from 0 to 180 degrees uttering a word. The WA of the moving talker for 'Top 5' was **72.01%**. Compared with 2-D Viterbi search, 3-D N-best search has the additional advantage of being able to recognize a moving talker in an unknown direction. On the other hand, the performance of speech recognition systems using conventional localization methods, such as CSP, strongly depends on the accurate localization of the talker. With these systems, however, accurate localization appears to be very difficult with a moving talker. The described results show that our proposed 3-D N-best-based method performs efficiently, even in the case one of two talkers is moving.

5. CONCLUSION

In this paper, a 3-D N-best search method was described. A likelihood normalization technique and a clustering technique were also implemented to the baseline 3-D N-best search, to improve the recognition rate. As future work, we plan to carry out experiments using delay-and-sum beamformer composed of more elements and experiments using adaptive beamforming instead of delay-and-sum beamforming. Furthermore, we plan to use conventional localization methods and to compare their results with the results given by the 3-D N-best search method.

6. REFERENCES

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

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

[3] P. Heracleous, T. Yamada, S. Nakamura, and K. Shikano, "Simultaneous Recognition of Multiple Sound Sources based on 3-D N-best Search", Acoustical Society of Japan 1999, pp. 91-92, March 1999.

[4] P. Heracleous, T. Yamada, S. Nakamura, and K. Shikano, "Simultaneous Recognition of Multiple Sound Sources based on 3-D N-best Search using Microphone Array," Eurospeech99, pp. 69-72, Sept. 1999.

[5] T. Matsui and S. Furui,"Likelihood Normalization for Speaker Verification using a Phoneme- and Speaker-independent Model," Speech Communication 17 (1995), pp. 109-116.