音声言語情報処理 29－15
(1999. 12. 20)

An Improvement to 3-D N-best Search Using Path-Distance Based Clustering for Recognizing Multiple Sound Sources

Panikos HERACLEOUS, Satoshi NAKAMURA, Kiyohiro SHIKANO

Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama, Ikoma, Nara, 630-01 JAPAN

E-Mail: (paniko-h,nakamura,shikano)@is.aist-nara.ac.jp

Abstract A problem which requires solution in the recognition of distant talking speech is the talker localization. In some approaches the talker is localized by using short- and long-term power. However, in the case of low SNR the talker localization appears to be difficult. The 3-D Viterbi search based method, proposed by Yamada T. et.al, integrates talker localization and speech recognition. Although, the method provides high recognition rates its applications are restricted only to the presence of one talker. In previous work we introduced a method able for simultaneous recognition of multiple sound sources. The method is an extension of the 3-D Viterbi search and performs N-best search in a 3-D trellis space composed of input frames, HMM states and direction. This paper describes the improvement of the method by using path-distance based clustering of the hypotheses.

keywords speech recognition, distant-talking speech,multiple sound sources, microphone array

キーワード

複数話者の音声認識における音源方向経路間距離を用いた3-D N-best 探索法の改善

Panikos HERACLEOUS, 中村 哲, 鹿野 清宏

Graduate School of Information Science, Nara Institute of Science and Technology

8916-5 Takayama, Ikoma, Nara, 630-01 JAPAN

E-Mail: (paniko-h,nakamura,shikano)@is.aist-nara.ac.jp

おらまし ハンズフリー音声認識において、話者の位置を推定することは非常に重要である。その方法として、短・長時間のパワーを用いて話者の位置を推定する方法がある。しかしこの方法では、SNRが低い環境下においては、話者の位置を推定することは難しいという問題がある。この問題を解決する方法として、これまでに話者位置推定と音声認識を統合した3-Dビタビ探索法を提案している。しかしこの方法は、話者が1人の場合には話者位置推定および音声認識において有効な方法であったが、複数の話者には対応できないとう問題があった。そこで著者らはこれまでに、複数の話者が同時に発話しても認識が可能である方法を提案している。その方法とは、3-Dビタビ探索法を拡張させて、入力フレーム、HMM状態、話者方向で構成される3-Dトレリス空間内で、N-best探索を行なうことである。本稿では、音源方向経路間の距離に基づいてクラスタリングを行なうことにより、3-D N-best探索法の改善を試みたので、その方法について報告する。

キーワード 音声認識、ハンズフリー音声、複数話者、マイクロホンアレー

key words

1 Introduction

The recognition of the distant talking speech is of an important role in any practical speech recognition system. Moreover, factors such as noisy and reverberant environment, presence of multiple sound sources, moving talkers, etc. should be, also, considered.Due to the fact that a microphone array can take advantage of the spatial and acoustical information of a sound source, most of those systems are microphone array-based.

A complex problem, which requires solution in a such system is the talker localization and the speech recognition. In some approaches [1, 2, 3, 4] the talker is localized by using short- or long-term power, then the 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, in low SNR the talker localization appears to be difficult. The 3-D Viterbi search method, proposed by Yamada et.al [5, 6] , integrates talker localization and speech recognition and performs Viterbi search in a 3-D trellis space, composed of input frames, HMM states and direction [Fig. 1]. By steering a beamformer to each direction a locus of the sound source and a feature vectors sequence can be obtained simultaneously. Although, the 3-D Viterbi search using adaptive beamforming based system provides high recognition rates, it can be applied only in the case of one sound source.

In previous works [7, 8] a novel method able for recognizing multiple sound sources simultaneously had been proposed. The method is based on the 3-D Viterbi search and extended to 3-D N-best search. The method performs full search in the direction and considers N-best word hypothesis and direction sequence. As a result, the algorithm provides an N-best list, which includes the direction sequences and the phoneme sequences of the multiple sound sources.

This paper describes, also, the improvement to the proposed method by using path distance-based clustering. The speech recognition system based on this method operates in two pass. In a first forward-pass the required top N hypotheses are found. During this pass the direction sequences are, also, found and stored. An additional second pass is integrated, which traces back the direction sequences of the provided top N hypotheses and calculates for each hypothesis-pair a distance. Based on this distance the hypotheses are clustered into a pre-defined

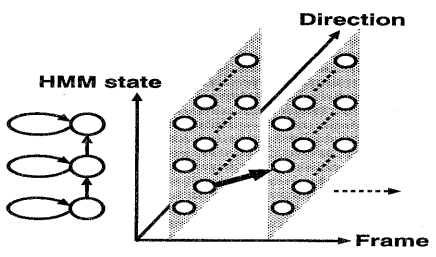

Figure 1: 3-D Trellis space

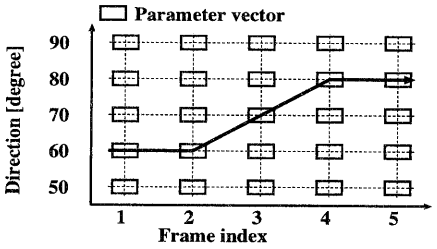

Figure 2: Locus of sound Source

number of clusters. As a result, the sound sources located in different direction sequence are included in different cluster. By finding the top N for each cluster the sound sources and their direction sequences can be obtained.

2 3-D Viterbi search

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

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

In the hypothesized path a direction sequence and a feature vector sequence can be obtained as Figure 2 shows. The direction sequence corresponds to the locus of the sound source and the feature vectors sequence to the uttered speech or to other sound source.

The speech recognition system based on 3-D Viterbi search and using adaptive beamforming provides high recognition rates and operates efficiently, even in the case of 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 to the 3-D N-best search capable for considering multiple sound sources.

3 3-D N-best search

The 3-D N-best search is an extension of the 3-D Viterbi search and it's based on the idea, that recognition of multiple sound sources can be performed by introducing the N-best paradigm. While the 3-D Viterbi search considers only the most likely path in the 3-D trellis space, the 3-D N-best search considers multiple hypotheses for each direction and in this way the N path with the highest likelihood can be obtained. In a similar way with the conventional 3-D Viterbi search the direction-feature vector sequences are extracted by steering the beamformer to each direction every time frame. After a direction-feature vector sequence is extracted, matching is followed between the extracted vector sequence and the HMM models.

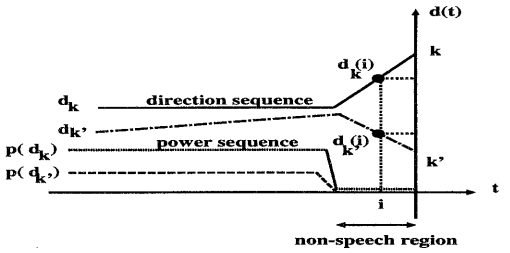

Figure 3: Path distance of a hypothesis-pair

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

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

Considering a node at time n , the overall $\underline{\alpha}^N(q',d',n-1)$ predecessor hypotheses are sorted and by adding to those the α_1 state and α_2 direction transition, and 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 the 3-D N-best search the overall provided word-hypotheses are sorted according to the likelihood and the top N with the highest likelihood are selected. The correct sound sources are included in the top N hypotheses and the direction sequences can be, also, obtained.

4 Clustering Hypotheses Using the Information of the Path Distance

The original 3-D N-best search was extended by implementing the path distance-based clustering. Using the information of the provided direction sequences the top N hypotheses are classified into several clusters. Figure 3 shows the direction and power sequences of two hy

Table 1: Results for top 4. Sorting according to the likelihood. Only one sound source is included in the N-best list.

Input	MHT /omoshiroi/	FSU /wagamama/

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

potheses. Using the Equation 3 the path distance $D(k,k')$ for the two hypotheses is calculated.

$$D(k,k') = \sum_{i=0}^{N-1} (d_k(i) - d_{k'}(i))^2 (p(d_k(i),i) + p(d_{k'}(i),i)) \quad (3)$$

In the Equation 3 N is the number of total frames, k and k' the directions at final frame of the two hypotheses, d_k is the direction sequence which ends at k and $p(d_k)$ is the power sequence corresponding to d_k. The power sequences are used in the calculation of the path distance in order to avoid the impact effect of the non-speech region. The path distance provides the measure the clustering is based on. By using the path distance the top N hypotheses are classified in different clusters, which corresponds to the sound sources. The number of the clusters corresponds to the number of the sound sources and the sound sources can be found by picking up the top N of each cluster. The direction of the sound sources can be obtained by examining the direction sequences of the hypotheses included in each cluster.

Figure 4 shows the transitions of four hypotheses in the case of the pronounced words **/omoshiroi/** and **/wagamama/**. The two words were pronounced by different talkers. The talkers are located in fixed position at 10 and 170 degrees respectively. The aim is to classify those hypotheses into two clusters based on their direction sequences. It's expected that the words **/omoshiroi/** and **/wagamama/** will be included in a different cluster and in high order.

Due to its implementation simplicity the bottom-up method has been chosen as clustering method. The algorithm operates as following :

- Find the minimum of the overall path distances.
- Merge the hypotheses with the minimum distance. 1st cluster is generated.
- Find the 2nd minimum of the remained path distances.
- Merge the hypotheses with the 2nd minimum distance. 2nd cluster is generated or the hypotheses are merged with previous clusters.
- Continue the previous steps until the required number of clusters is generated.

A very difficult problem, which must be solved is to find the number of the clusters necessary for our task. In this paper the number of the clusters is pre-defined and is the same as the known number of the sound sources.

5 Experiment and Results

To perform evaluation of the 3-D N-best search method, isolated-word recognition experiments were carried out on simulated data (only time delay). Results are provided for both of cases, namely with and without clustering.

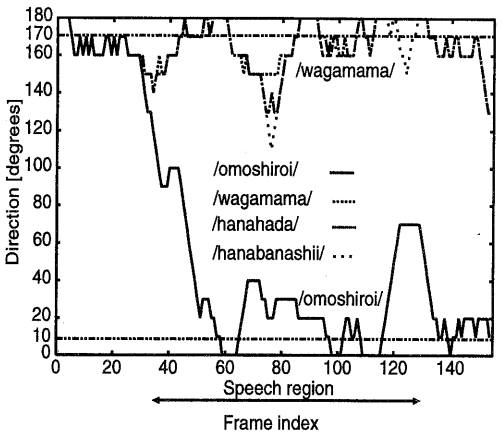

Figure 4: Results show the transitions of hypotheses

Table 2: Results for top 4. The hypotheses are classified using the path distance.

Input	MHT /omoshiroi/	FSU /wagamama/

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

5.1 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 FSU-speaker of the ATR database SetA. The feature vectors are of length 33 (16 MFCC, 16 ΔMFCC and Δpower). A linear array composed of 16 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 as Figure 5 shows.

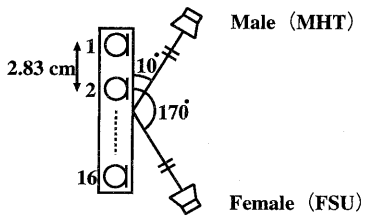

Figure 5: Source positions

Table 3: Results for top 4. Sorting according to the likelihood. Only one sound source is included in the N-best list.

Input	MHT /omoshiroi/	FSU /daidokoro/

Top	Word	Likelihood
1	**/daidokoro/**	-76.6489
2	/daihyoo/	-76.6958
3	/zairyoo/	-76.8843
4	/imagoro/	-76.9619
..	..	..
10	**/omoshiroi/**	-77.5794

Table 4: Results for top 4. The hypotheses are classified using the path distance.

Input	MHT /omoshiroi/	FSU /daidokoro/

Top	1st Cluster	2nd Cluster
1	**/omoshiroi/**	**/daidokoro/**
2	-	/daihyoo/
3	-	/zairyoo
4	-	/imagoro/

5.2 Experimental Results

Table 1 shows the achieved results for the case described by Figure 4. The hypotheses are sorted according to the likelihood and clustering isn't implemented. As Table 1 shows only one sound source included in the top 4. Table 2 shows the results when clustering is implemented. As can be seen the two sound sources are included in different cluster and both are on the 1st order. Table 3 and Table 4 show another example. As can be seen by implementing clustering the correct words are recognised as top 1. Table 5 and 6 show an example when the clustering fails to classify correctly the words. Table 7 and Table 8 show the results of the experiments were carried out. As can be seen by implementing the clustering algorithm an improvement in accuracy was achieved in almost all cases, because two clusters have been generated and therefore, two top 1 and two top 5 exist.

5.3 Error's Investigation

As Table 7 and Table 8 show results aren't efficiently high. In order to improve the performance of the recognition

Table 5: Results for top 4. Sorting according to the likelihood. Only one sound source is included in the N-best list.

Input	MHT /omoshiroi/	FSU /hyoojuN/

Top	Word	Likelihood
1	**/hyoojuN/**	-77.8488
2	**/omoshiroi/**	-78.1610
3	/ryoogae/	-78.2200
4	/gikochinai/	-78.2943

Table 6: Results for top 4. The hypotheses are classified using the path distance.

Input	MHT /omoshiroi/	FSU /hyoojuN/

Top	1st Cluster	2nd Cluster
1	/oboe/	**/hyoojuN/**
2	-	**/omoshiroi/**
3	-	/ryoogae/
4	-	/gikochinai/

Table 7: Experimental results without clustering

	Accuracy [%]	
	Top 1	Top 5
One source	72.09	93.05
One source - Correct direction	64.65	91.21
Both sources	*	37.16
Both sources -Correct direction	*	30.55

system based on 3-D N-best search error's investigation is necessary. The following factors appear to have an impact effect to the performance of the system :

- Different duration of the sound sources. The search attempts to search the sound source which has longer duration. A possible solution to this problem is the implementation of word-spotting.
- Likelihood normalization appears to be necessary.
- The clustering accuracy appears to be lower as the expected. A possible solution is the implementation of different clustering method.
- Due to the fact that the delay and sum beamformer isn't sharp enough, the performance of the system decreases. The use of the adaptive beamformer could offer a solution to this problem.

An additional problem is the computation amount necessary. A possible solution to this problem is to consider only hypotheses from the directions where power is over a given threshold.

6 Conclusion

In this paper the 3-D N-best search method was described. The clustering technique based on the path distance was, also, implemented to the original 3-D N-best search, which offers recognition rate improvement. However, results aren't efficiently high. Several problems must be solved in order to improve the performance of the system . As future work error's investigation will be performed and it's expected that by offering solutions to the described problems the performance will be increased. Moreover, as future work we'll consider the case of moving talker, too.

Table 8: Experimental results using clustering

	Accuracy [%]	
	Top 1	Top 5
One source	78.14	96.76
One source - Correct direction	71.62	84.72
Both sources	25.46	52.56
Both sources - Correct direction	25.46	37.20
Clustering Accuracy	87.56	

References

[1] M. Omologo, P. Svaizer, "Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis', ICSLP94, pp. 1243–1246, Sep. 1994.

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

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

[4] 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.

[5] 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.

[6] 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.

[7] P. Heracleous, T. Yamada, S. Nakamura, 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.

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