MULTIPLE SOUND SOURCES RECOGNITION BY A MICROPHONE ARRAY-BASED 3-D N-BEST SEARCH WITH LIKELIHOOD NORMALIZATION

Panikos Heracleous\textsuperscript{1,2}, Satoshi Nakamura\textsuperscript{1}, Kiyohiro Shikano\textsuperscript{2}

ATR Spoken Language Translation Research Labs\textsuperscript{1}
Graduate School of Information Science, Nara Institute of Science and Technology\textsuperscript{2}

ABSTRACT

This paper deals with the hands-free speech recognition and, particularly, with the simultaneous recognition of multiple sound sources. Our method is based on the 3-D Viterbi search, i.e., extended to 3-D N-best search method enabling the recognition of multiple sound sources. The baseline system integrates two existing technologies – 3-D Viterbi search and conventional N-best search – into a complete system. However, the first evaluation of the 3-D N-best search-based system showed, that new ideas are necessary in order to build a system for simultaneous recognition of multiple sound sources. Two factors found to have an important role in the performance of our system, namely the different likelihood ranges of the sound sources and the direction-based separation of the hypotheses. In order to solve these problems we implemented a likelihood normalization and a path distance-based clustering technique into the baseline 3-D N-best search-based system. The performance of our system was evaluated through experiments on simulated data for the case of two talkers. The experiments showed significant improvements by implementing the two techniques described above. The best results were obtained by implementing the two techniques and using a microphone array composed of 32 elements. More specifically, in that case the Word Accuracy for the two talkers was higher than 80\% and the Simultaneous Word Accuracy (both sources are correctly recognized simultaneously) higher than 70\%, which are very promising results.

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. This paper deals with the hands-free speech recognition, and particularly with the recognition of multiple sound sources (talkers or noise sources). The aim of our research is to build a system enabling the simultaneous recognition of multiple sound sources - located at fixed position or moving - in real environments.

Most of hands-free speech recognition 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 [Fig. 1]. 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 since it considers only the one best path it the 3-D Trellis space it can be applied only in the case of one sound source.

In this paper we propose 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 [3].

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 we cannot compare them accurately. 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. 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 steered 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 1 shows the general way the N hypotheses with the highest likelihoods are found.

\[
\alpha^N(q,d,t) = \text{sort}(\alpha^N(q',d',t-1) + \log a_1(q',q) + \log a_2(d',d) + \log b(q,x(d,t)))
\]  

(1)

Considering a node at time t, the overall \(\alpha^N(q',d',t - 1)\) 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 \(\alpha^N(q,d,t)\) 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.

2.1. Likelihood normalization of the hypotheses

The N-best hypotheses of a \((q, d)\) (state, direction) are found by sorting the overall arriving hypotheses and choose the top N. However, hypotheses arriving from different directions correspond to different sound sources with different likelihood dynamic ranges. Therefore, the comparison of the hypotheses according to their likelihoods can not be accurate. In order to avoid this problem we introduced a technique for likelihood normalization. The technique used for likelihood normalization is similar to the method proposed by Matsui T. et al.[4]. 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. More specifically, this model runs in parallel with the other models and its accumulated likelihood is used to normalize the likelihoods of the hypotheses involved. Two different techniques were implemented and compared. The two likelihood normalization techniques, L1 and L2 are the following:

- **L1 Likelihood Normalization Technique**
  In the first approach we normalize the likelihoods only at the last frame. The actual likelihoods \(P(q,d,T)\) of every state \(q\) and direction \(d\) are normalized at the last frame \(T\) by dividing with the likelihood \(P_G(d,T)\) of the one-state model. Considering logarithmic likelihoods, Eq. (2) gives the normalized likelihood \(\Lambda(q,d)\).

\[
\Lambda(q,d) = \sum_{t=0}^{T-1} \log P(q,d,t) - \sum_{t=0}^{T-1} \log P_G(d,t)
\]  

(2)

- **L2 Likelihood Normalization Technique**
  In this approach, the actual accumulated likelihoods \(P(q,d,t)\) 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 \(\Lambda(d,q,t_f)\) at time \(t_f\).

\[
\Lambda(d,q,t_f) = \sum_{t=0}^{t_f} \log P(q,d,t) - \sum_{t=0}^{t_f} \log P_G(d,t)
\]  

(3)

2.2. Clustering hypotheses using information on path distances

By implementing the likelihood normalization technique the hypotheses can be compared accurately and the performance of the system is improved. However, in some cases our algorithm faces with an additional problem. Namely, if the likelihoods of the hypotheses of one direction happens to be much higher than those of the other directions, the N-best list is occupied by hypotheses of one direction only. In this case the algorithm fails and can not consider all the sound sources.

In order to solve this problem 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 clustered into several clusters. Figure 2 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-degree direction. As can be seen in the speech region, the two hypotheses are well separated based on their directions. However, in the silence region, the directions of the two hypotheses appear to be similar. The reason is that our algorithm can not guarantee the correct direction in the silence region. In order to minimize the importance of the silence region we use the power information. Using Eq. (4), the path distance $D_k^{k'}$ can be calculated. The $D_k^{k'}$ is the Euclidean distance between the two direction sequence weighted by the power sequences and can be calculated as follows:

$$D_k^{k'} = \sum_{t=0}^{T-1} \left( d_k(t) - d_{k'}(t) \right)^2 \left( p(d_k(t), t) + p(d_{k'}(t), t) \right)$$

(4)

In the Eq. (4), $T$ 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 $p(d_k(t), t)$ 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 3 shows the transitions of four hypotheses in the case of the pronounced words /omoshiroi/ and /wagamama/. The two words were pronounced by different talkers located at fixed positions at 10 and 170 degrees, respectively. 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 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.

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

### 3. EXPERIMENTS AND RESULTS

#### 3.1. Experimental Conditions

The speech recognizer is based on tied-mixture HMM with 256 distributions. Fifty-four context independent 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). Two linear delay-and-sum array composed of 16 and 32 microphones are used and the distance between them is 2.83 cm. The two talkers were located at fixed positions at 10 and 170 degrees as in Figure 4. Several speaker- and word-pairs were used.

#### 3.2. Experimental results for fixed position talkers

Figure 5 shows the comparison of the two likelihood

---

Table 1: Top 4 results. The hypotheses are classified using the path distance.

<table>
<thead>
<tr>
<th>Input</th>
<th>Speaker A</th>
<th>Speaker B</th>
</tr>
</thead>
<tbody>
<tr>
<td>/omoshiroi/</td>
<td>/wagamama/</td>
<td>/wazawaza/</td>
</tr>
<tr>
<td>/hanahada/</td>
<td>/hanabanashi/</td>
<td>/wazawaza/</td>
</tr>
<tr>
<td>/hanahada/</td>
<td>/hanabanashi/</td>
<td>/wazawaza/</td>
</tr>
<tr>
<td>/omoshiroi/</td>
<td>/wagamama/</td>
<td>/wazawaza/</td>
</tr>
</tbody>
</table>

---

105
normalization techniques described in Section 2.1. As can be seen the likelihood normalization in every time frame is more effective and therefore we choose this technique for our task. Figures 6 and 7 shows the WA (Word Accuracy) of the two speakers and Figure 8 shows the SWA (Simultaneous Word Accuracy) in the case of 32-elements microphone in comparison with the case of the 16-elements microphone. As can be seen significant improvements were achieved by increasing the number of elements.

4. CONCLUSION

In this paper we proposed the 3-D N-best search, a novel method for simultaneous recognition of multiple sound sources. We described two problems that the baseline 3-D N-

best search faced and we introduced their possible solutions. Namely, a likelihood normalization technique and a clustering technique were also implemented into the baseline 3-D N-best search-based system. The performance of our system was evaluated through experiments on simulated data for two talkers located both at fixed position. Although, further improvements are possible the obtained results are very promising. As future work we will carry out experiments on real data and we will try to implement new ideas in order to speed-up the system.

5. REFERENCES