SIMULTANEOUS RECOGNITION OF DISTANT-TALKING SPEECH OF MULTIPLE SOUND SOURCES BASED ON 3-D N-BEST SEARCH ALGORITHM

Panikos Heracleous, Satoshi Nakamura

ATR, Spoken Language Translation Research Labs, Japan

Kiyohiro Shikano

Nara Institute of Science and Technology, Japan

ABSTRACT

This paper deals with the simultaneous recognition of distant-talking speech of multiple talkers using the 3-D N-best search algorithm. We describe the basic idea of the 3-D N-best search and we address two additional techniques implemented into the baseline system. Namely, a path distance-based clustering and a likelihood normalization technique appeared to be necessary in order to build an efficient system for our purpose. In previous works we introduced the results of experiments carried out on simulated data. In this paper we introduce the results of the experiments carried out using reverberated data. The reverberated data are those simulated by the image method and recorded in a real room. The image method was used to know the accuracy-reverberation time relationship, and the real data was used to evaluate the real performance of our algorithm. The obtained Top 3 results of the Simultaneous Word Accuracy was 73.02% under 162ms reverberation time and using the image method.

1. INTRODUCTION

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, 2], 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. [3], 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.

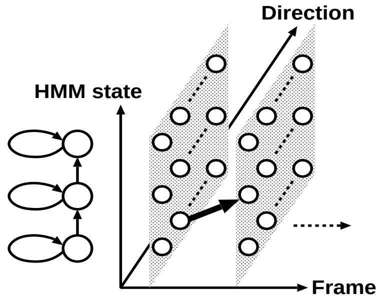

Fig. 1. 3-D Trellis space

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 [4]. The beauty of our proposed method is that does not require the deterministic talker localization, but it uses a probabilistic approach based on the input signal acoustic information. In the baseline system two additional techniques were implemented, which significantly improved the performance of our 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 can not 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.

In previous work [5] we introduced results obtained by experiments carried out on simulated data. In this paper we introduce the results of the experiments carried out on reverberated data. The reverberated data in one case were simulated using the **Image Method** [6], and in the other case were recorded in our experimental room.

2. 3-D VITERBI SEARCH EXTENDED TO 3-D N-BEST SEARCH METHOD

The 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 $\underline{\mathbf{X}}$ is the feature vector.

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

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 focus 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.

The proposed 3-D N-best search method it is based on the idea that recognition of multiple sound sources can be performed by introducing the N-best paradigm into the 3-D Viterbi search. The 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.

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. Equation 2 shows the general way the N hypotheses $\underline{\alpha}^N(q, d, t)$ with the highest likelihoods are found.

$$\underline{\alpha}^N(q, d, t) = \underset{d',q'}{\mathbf{sort}}\{\underline{\alpha}^N(q', d', t-1) + \log a_1(q', q) + \log a_2(d', d)\} + \log b(q, \mathbf{x}(d, t)) \tag{2}$$

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

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 direction 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 [7]. The technique used for likelihood normalization is similar to the method proposed by Matsui T. et al.[8]. 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.

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 that 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 a path distance-based clustering [5]. By using information on the provided direction sequences, the top N hypotheses are clustered into different clusters, which correspond to the sound sources. The path distance is calculated using the following equation:

$$D(k, k') = \sum_{t=0}^{T-1} (d_k(t) - d_{k'}(t))^2 (p(d_k(t), t) + p(d_{k'}(t), t)) \tag{3}$$

In the Eq. (3), 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 $p(d_k)$ the power sequence corresponding to d_k.

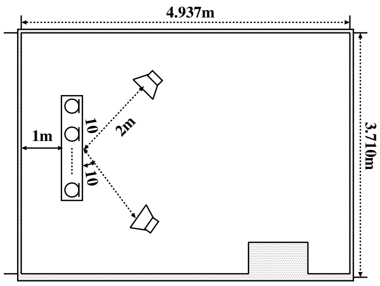

Fig. 2. Experimental arrangement for reverberant environment

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 form 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 microphone array composed of 32 microphones was used. The distance between the microphones was $2.83cm$. The two talkers were located at fixed positions at 10 and 170 degrees as Figure 2 shows.

3.2. Results

We carried out two kinds of experiments. In the first case the reverberated data were simulated using the image method. That method was developed by Allen et al. [6] and can be used to simulate small-room acoustics. In order to know the relationship between the reverberation time and the accuracy we carried out experiments under $T_{[60]}$ reverberation time 162, 200 and $240ms$. Figures 3, 4, and 5 show the obtained results. As results show in the case of reverberation time $T_{[60]} = 162ms$ the obtained results are promising. More specifically, the Simultaneous Word Accuracy (both talkers are simultaneously recognized) in Top 5 hypotheses was $\mathbf{80}\%$. The results show also that the performance of the system degrades as the reverberation time becomes longer. Namely, in Top 5 hypotheses the Simultaneous Word Accuracy under $T_{[60]} = 200ms$ becomes $\mathbf{78.5}\%$ and under $T_{[60]} = 240ms$ becomes $\mathbf{67}\%$.

The real performance of our proposed algorithm was evaluated through experiments carried out on reverberated data recorded in a real room with reverberation time $T_{[60]} = 280ms$. The ambient noise level was 24.3 dBA. Figures 6, 7, and 8 show the obtained results. The results are shown

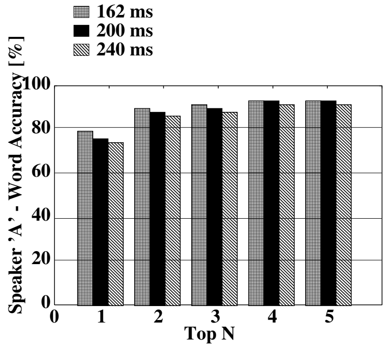

Fig. 3. Speaker 'A' Word Accuracy

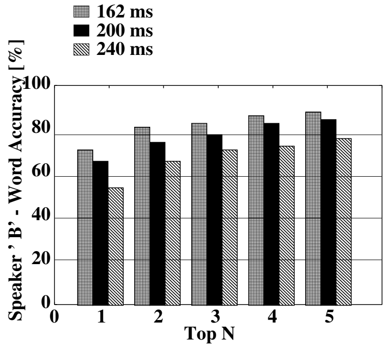

Fig. 4. Speaker 'B' Word Accuracy

in comparison with the case of simulated data. The Simultaneous Word Accuracy in Top 5 hypotheses was $\mathbf{63}\%$. Comparing with the results achieved by using the image method the real performance degrades. However, considering the presence of ambient noise and the longer reverberation time, the comparison between the two cases is reasonable.

4. CONCLUSION AND FUTURE WORK

In previous works the performance of our 3-D N-best search-based system was evaluated through experiments carried out on simulated data (only time delay). In this paper we introduced the results achieved by experiments carried out on reverberated data. In the first experiments the image method was used in order to simulate the reverberant environments. More specifically, we carried out experiments under $T_{[60]}$ reverberation time 162, 200, and $240ms$. Although, under reverberant conditions the performance of our system degraded, the obtained results are promising. In the second experiments real data were used recorded in our experimental room. In that case the $T_{[60]}$ reverberation time was $280ms$. As future work we plan to carry out experiments for the recognition of three sound talkers, including

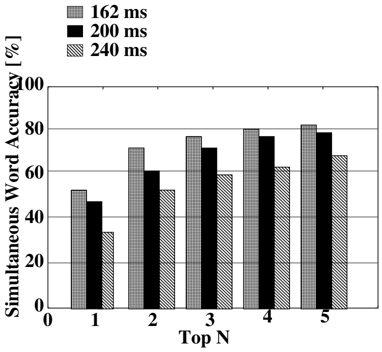

Fig. 5. Simultaneous Word Accuracy

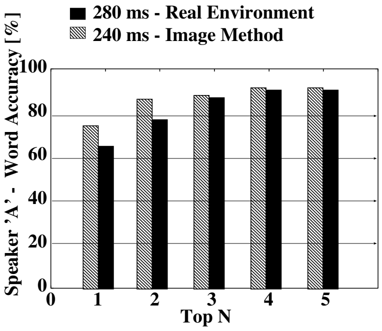

Fig. 6. Speaker 'A' Word Accuracy

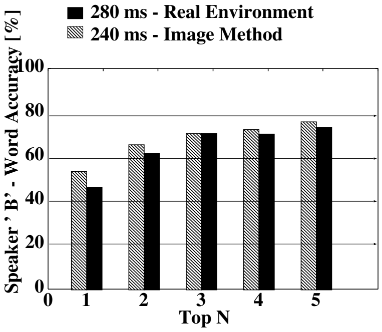

Fig. 7. Speaker 'B' Word Accuracy

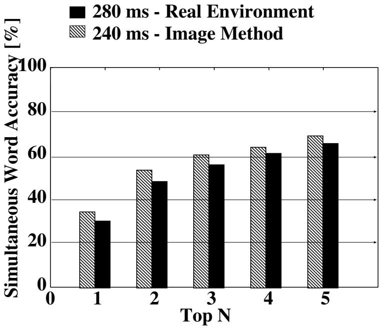

Fig. 8. Simultaneous Word Accuracy

two talkers and one noise source. Our system is already able to deal with this problem and the related experiments are in progress.

5. REFERENCES

[1] M. Omologo and P. Svaizer, "Acoustic source location in noisy and reverberant environment using CSP analysis," *Proc. ICASSP*, pages 921–924, 1996.

[2] T. Yamada, S. Nakamura, K. Shikano, "Robust speech recognition with speaker localization by a microphone array," *Proc. ICSLP*, pages 1317-1320, 1996.

[3] T. Yamada, S. Nakamura, K. Shikano, "Hands-free Speech Recognition Based on 3-D Viterbi Search Using a Microphone Array," *Proc. ICASSP*, pages 245–248, 1998.

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

[5] P. Heracleous, S. Nakamura, and K. Shikano, "Multiple Sound Sources Recognition by a Microphone Array-based 3-D N-best Search with Likelihood Normalization," *Proc. International Workshop on Hands-free Speech Communication*, pages 103–107, 2001.

[6] J. B.Allen and D. A.Berkley. Image method for efficiently simulating small-room acoustics. *Journal of Acoustical Society of America, vol. 65, No 4*,pages 943–950, 1979.

[7] P. Heracleous, S. Nakamura, and K. Shikano. A technique for likelihood normalization in the 3-D N-best search for simultaneous recognition of multiple sound sources. *Proc. of Acoustical Society of Japan*, pages 117–118, 2000.

[8] T. Matsui and S. Furui, "Likelihood Normalization for Speaker Verification using a Phoneme- and Speaker-independent Model," *Speech Communication 17* pages 109-116, 1995.