

A SEMI-BLIND SOURCE SEPARATION METHOD FOR HANDS-FREE SPEECH RECOGNITION OF MULTIPLE TALKERS

Panikos Heracleous^{1,2}, Satoshi Nakamura², Kiyohiro Shikano¹

¹Graduate School of Information Science, Nara Institute of Science and Technology, Japan

²ATR Spoken Language Translation Research Labs, Japan

ABSTRACT

In this paper, we present a beamforming based semi-blind source separation technique, which can be applied efficiently for hands-free speech recognition of multiple talkers (including moving talkers, too). The main difference from the conventional blind source separation techniques lies in the fact that the proposed method does not attempt to separate explicitly the unknown signals in a pre-processing pass before speech recognition. In fact, localization of multiple talkers, separation of the signals, and speech recognition are integrated in a single pass. Each time frame, beams formed by a delay-and-sum beamformer are steered to every direction, and speech information is extracted. A modified Viterbi formula provides n -best hypotheses for each direction and word hypotheses. At the final frame, all hypotheses are clustered based on their direction information. The clusters, which correspond to the talkers include information about the recognized speech of the multiple talkers and about their direction. Experiments for recognition of two and three talkers showed very promising results. In the case of two talkers, and using simulated clean data we achieved for 'top 5' hypotheses a recognition rate of **95.02%** on average, which is very promising result.

1. INTRODUCTION

The hands-free speech recognition is a challenging research area. By solving this problem, the human-machine communication becomes user friendly and practical. To date, most of hand-free speech recognition systems are microphone array based [1, 2, 4]. The use of a single microphone provides limited speech enhancement, and especially in the case of multiple talkers cannot be used without physical steering.

A main problem in hands-free speech recognition is the localization of the talker. Omologo et al., proposed a talker localization method based on the Cross-power spectrum phase (CSP) [1]. Nishiura et al., proposed also a method based on CSP for localizing multiple talkers [5]. Those methods perform deterministic talker localization, and their main disadvantage is the difficulty of localizing moving talkers. Yamada et al., approached the problem by introducing the 3-D Viterbi search, which performs talker localization and speech recognition simultaneously, without deterministic talker localization [3]. Although, the 3-D Viterbi search achieved high recognition rates for moving talker, too, its applications are restricted to the recognition of speech of a single talker.

In this paper, we propose a novel semi-blind source separation technique for recognizing multiple talkers (including moving talker, too). The main difference from the conventional blind source separation techniques [6] lies in the way the signal mixtures

are de-mixed. The conventional blind source separation techniques attempt to estimate the signals using only the information of the mixed signals observed in each input channel. The de-mixed signals can be then processed in a conventional way, and speech of multiple talkers can be recognized.

Our approach is an extension of 3-D Viterbi search and does not attempt to separate explicitly the input mixtures, but integrates talker localization, signal separation, and speech recognition in a single pass. The method is based on beamforming and does not require deterministic talker localization. The proposed method performs 3-D N-best search, which keeps n -best hypotheses for each word and direction hypothesis. At the last stage of the recognition process, the multiple talkers can be separated and recognized based on their acoustic scores and the direction information.

2. PROPOSED METHOD

The proposed 3-D N-best search method is based on the idea that the 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 a 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 1 shows the general way in which the N hypotheses $\underline{\alpha}^N(q, d, t)$ with the highest likelihoods are found.

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

The initial conditions for the above recursion are

$$\underline{\alpha}^n(1, 1, 1) = 1 \quad (2)$$

and

$$\underline{\alpha}^n(q, d, 1) = a_1(1, q)a_2(1, d)b(q, \mathbf{x}(d, 1)) \quad (3)$$

for $1 < q < Q$, $1 < d < D$, and $n = 0, 1..N$. Q is the number of states, D is the number of directions, and N indicates the N-best hypotheses.

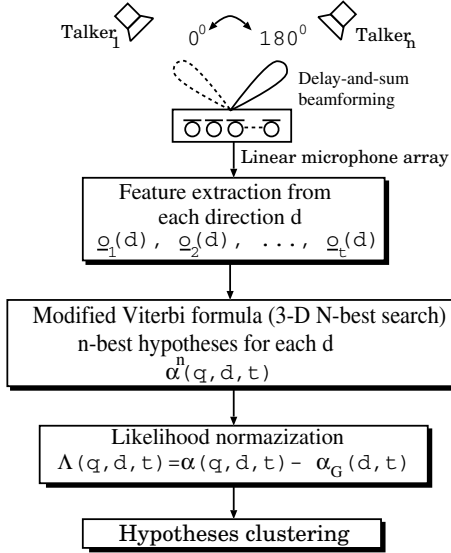


Fig. 1. Algorithm of the proposed method

By considering a node at time t , the overall $\underline{\alpha}^N(q', d', t-1)$ predecessor hypotheses are sorted after adding to them the a_1 state and a_2 direction transition as well as the b output probabilities. In this way, the $\underline{\alpha}^N(q, d, t)$ N-best hypotheses can be found. The $a_1(q', q)$ state transition probability can be trained from the training data. However, the $a_2(d', d)$ direction transition probability, which indicates how likely the talker moves, is very difficult to be trained automatically. Therefore, we use a heuristic approach for the estimation of a_2 and it is set as follows:

$$a_2(d', d) = \begin{cases} \frac{1}{2\Delta d} & , \quad |d - d'| \leq \Delta d \\ 0 & , \quad |d - d'| > \Delta d \end{cases} \quad (4)$$

where Δd is the range of the talker movements.

The initial value $a_2(1, d)$ is a uniform distribution of directions, because we assume the same equal probability for each direction.

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

To date, although, our method has been evaluated only for isolated word recognition, it can be applied for continuous speech recognition, too. Our method keeps hypotheses for every direction, and not only the one best among all the hypotheses. At the word boundaries, decisions about successor words are made for every direction, while considering the direction transition probabilities as well. Figure 1 shows the block diagram of the system.

The N-best hypotheses of a (q, d) (*state, direction*) are found by sorting the overall arriving hypotheses and choosing the ‘top N’. Hypotheses, however, arriving from different directions correspond to different talkers with different likelihood dynamic ranges. Therefore, comparisons of the hypotheses according to their likelihoods are inaccurate. In order to avoid this problem, we

introduce a technique for likelihood normalization. The likelihood normalization is based on a Gaussian mixture model (GMM). This model runs in parallel with all other models and its accumulated likelihood is used to normalize the likelihoods of the hypotheses involved. In this approach, the actual accumulated likelihoods $\alpha(q, d, t)$ of every state q and direction d are normalized at each time frame t by dividing them with the accumulated likelihood $\alpha_G(d, t)$ of the one-state model. Considering logarithmic likelihoods, Eq. (5) gives the normalized likelihood $\Lambda(d, q, t_f)$ at time t_f .

$$\Lambda(d, q, t_f) = \alpha(q, d, t_f) - \alpha_G(d, t_f) \quad (5)$$

In some cases, however, our algorithm faces an additional problem. Namely, if the likelihoods of the hypotheses of one direction happen 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 of 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. By using information on the provided direction sequences, the hypotheses are clustered into several clusters. Using Eq. (6), the path distance $D_k^{k'}$ can be calculated. The $D_k^{k'}$ is the Euclidean distance between the two direction sequences weighted by the power sequences and can be calculated as follows:

$$D_k^{k'} = \sum_{t=0}^{T-1} (d_k(t) - d_{k'}(t))^2 (p(d_k(t), t) + p(d_{k'}(t), t)) \quad (6)$$

In Eq. (6), 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 . The power is used in order to minimize the importance of the silence region. However, our algorithm cannot guarantee the correct direction in this region.

3. EXPERIMENTS FOR TWO TALKERS

3.1. Experimental results using simulated data

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.

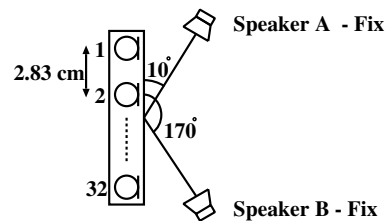


Fig. 2. Source positions

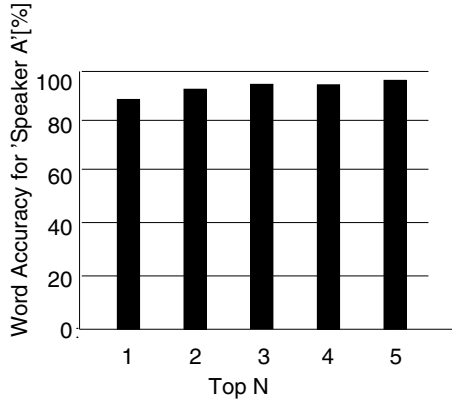


Fig. 3. Word accuracy for 'Speaker A'

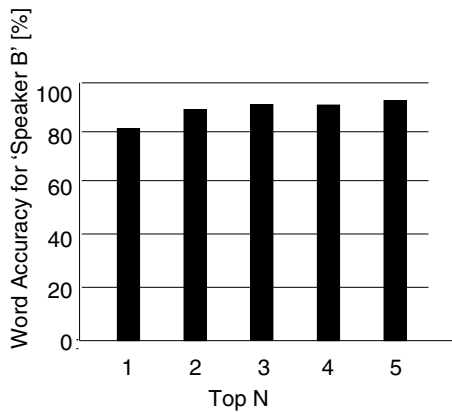


Fig. 4. Word Accuracy for 'Speaker B'

The test data includes 216 phoneme balanced words of the ATR database SetA. The speakers MHT, MAU, FTK, and FSU of ATR SetA are used. The four speakers form 12 combinations, and each speaker-combination forms 46440 (216*215) word-pairs (same words are avoided). We choose 10 speaker-combinations randomly, and for each one we choose 215 word-pairs. In total, we use 2150 test samples. The feature vectors are of length 33 (16 MFCC, 16 Δ MFCC, and Δ power). A linear microphone array composed of 32 microphones is used. We assume that the talkers are located far enough from the microphone array and that the speech signals are received by each microphone at the same angle. A total of 19 directions are considered. In these experiments, we use simulated data with only time delay and without reverberation to investigate the baseline performance of our method. The Δd range of the talker movement is set to 10 degrees. The two talkers are located at fixed positions at 10 and 170 degrees as shown in Figure 2. Figures 3 and 4 show the achieved results for the case of two fixed talkers. Results show the effectiveness of our method. For 'top 5' hypotheses we could achieve a recognition rate of **95.02%** on average, which is very promising result.

In order to make our system more general and not have any restrictions in terms of application we also consider the case of a moving talker. In this experiment, one of two talkers is located at

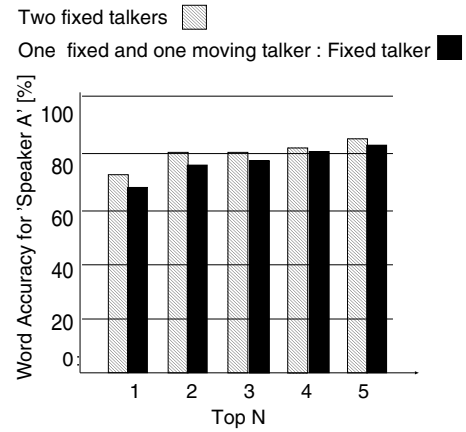


Fig. 5. Word Accuracy for 'Speaker A'

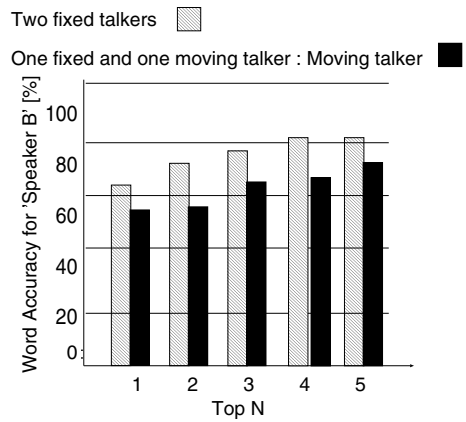


Fig. 6. Word Accuracy for 'Speaker B'

a fixed position at 10 degrees ('Speaker A'), while the other one moves from 0 to 180 degrees while uttering a word ('Speaker B'). A linear microphone array composed of 16 microphones is used. The distance between the microphones is 2.83 cm. This experiment uses 215 samples. The fixed talker is talker FTK and the moving talker is talker MHT. Figures 5 and 6 show the obtained results, in comparison with the case when two fixed talkers are used. By using the 3-D N-best search-based system, the word accuracy of the moving talker for 'top 5' hypotheses is 72.01%. Accordingly, 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 when one of two talkers is moving.

3.2. Experimental results using noisy and reverberant data

The two sound sources are located at fixed positions at 10 and 170 degrees respectively. The speech data are played back through loudspeakers. A linear microphone array composed of 32 micro-

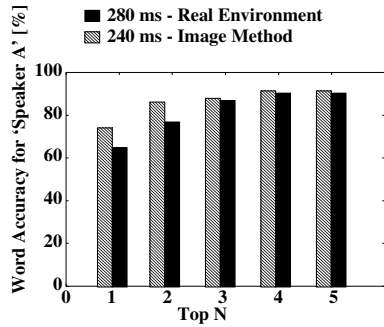


Fig. 7. Word Accuracy for 'Speaker A'

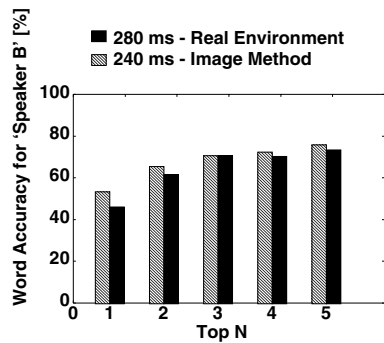


Fig. 8. Word Accuracy for 'Speaker B'

phones is used. The distance between the microphones is 2.80 cm . The distance between the loudspeakers and the microphone array is 2 m . The reverberation time ($T_{[60]}$) in the experimental room is 280 ms . Figures 7 and 8 show the obtained using real data results in comparison with the achieved results by using the image method [7]. Although the performance of our system in real environments is decreased, the achieved results are comparable with those when simulated reverberated data were used. However, in the case of using real data we should consider also the presence of ambient noise and the longer reverberation time. Therefore, we can conclude that the obtained results using real data are expected lower than the case when we use simulated data, and the comparison between the two cases is reasonable. In this experiments, we could achieve for 'top 5' hypotheses a 77.45% word accuracy on average.

4. EXPERIMENTS FOR THREE TALKERS

In this section we describe the experiments carried out for the simultaneous recognition of speech of three talkers. The three talkers are located at fixed positions at 10 , 90 , and 170 degrees, as Fig. 9 shows. In total we use 645 test utterances. The microphone array is linear and it is composed of 32 microphones. The distance between two microphones is 2.83 cm . The 'top 5' word accuracy for the three talkers was 82.22% , 81.79% , and 74.84% , respectively. Comparing with the two talkers case, the performance is degraded, since the used delay-and-sum beamformer cannot separate efficiently the speech signals of the three talkers. However,

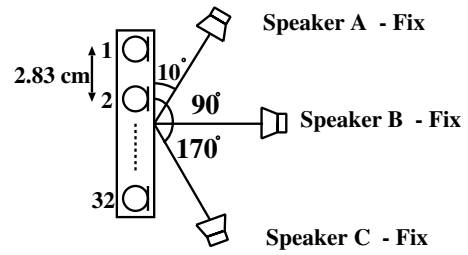


Fig. 9. Sources condition for three talkers

results show that our system performs well even in the case of three talkers, too.

5. CONCLUSION

In this paper, we presented a novel semi-blind source separation technique, which is applicable for hands-free speech recognition of multiple talkers. The proposed method is based on the 3-D N-best search, and experiments showed its effectiveness. Using simulated clean data we achieved for the speech recognition of two talkers a 95.02% recognition rate for 'top 5' hypotheses. We also carried out experiments using noisy and reverberant data, and experiments for the speech recognition of three talkers, with very promising results.

6. REFERENCES

- [1] M. Omologo, and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique", In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 273–276, 1994.
- [2] T. Yamada, S. Nakamura, and K. Shikano, "Robust speech recognition with speaker localization by a microphone array", In *Proceedings of International Conference on Spoken Language Processing*, pages 1317–1320, 1996.
- [3] T. Yamada, S. Nakamura, and K. Shikano, "Hands-free speech recognition based on 3-D Viterbi search using a microphone array", In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 245–248, 1998.
- [4] M. Omologo. "Hands-free speech recognition: Current activities and future trends", In *Proceedings of International Workshop on Hands-free Speech Communication*, pages 23–26, 2001.
- [5] T. Nishiura, T. Yamada, S. Nakamura, and K. Shikano, "Localization of multiple sound sources based on a CSP analysis with a microphone array", In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1053–246, 2001.
- [6] P. Common, "Independent component analysis, a new concept?", In *Signal Processing*, Vol. 36, pages 287–1994.
- [7] 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.