JOINT USE OF DISTRIBUTED MICROPHONE ARRAY AND LASER RANGE FINDERS FOR SPEAKER IDENTIFICATION IN MEETING

Jani Even, Panikos Heracleous, Carlos Ishi and Norihiro Nogita

ATR Intelligent Robotics and Communication Laboratories
2-2-2 Hikaridai, Kyoto 619-0288, Japan
even@atr.jp

ABSTRACT

This paper presents a text-independent speaker identification system for meetings. During the meeting, all of the meeting participants carry a microphone while a human tracker monitors their movements. The human tracker is based on scanning laser range finder and gives the positions of all the participants at any time. The position information is used to track the geometry of the distributed microphone array formed by all of the microphones. Using the geometry of the distributed array it is possible to cancel interfering speeches and noises from the audio stream assigned to each of the participants. Then, using these processed audio streams, the participants are identified by means of Gaussian mixture models (GMM) that were trained before hand. The proposed system is able to perform identification of simultaneously speaking participants and is thus a good candidate system for meeting diarization task. In particular, the use of laser range finders is a novel approach that makes the position estimation immune to acoustic noise and reverberation. An experiment conducted with three subjects reproducing a meeting configuration demonstrates the performance of the system for identification.

1. INTRODUCTION

These last years, the speech recognition community has been intensively working on the transcription of meetings [1, 2, 3]. An important task in meeting transcription is speaker diarization (i.e. to find ”Who spoke when”).

In a meeting, it is desirable to impose the least constraints to the participants. For example participants should be allowed to seat freely. Thus a convenient speaker diarization system should be flexible relatively to the positioning of the participants. For hands-free diarization, single microphone [4] or multiple microphones [5] approaches were proposed. Using multiple microphones, it is possible to estimate the position of the speakers using the time differences of arrival [6]. However, in a real environment, the accuracy of the position estimation is reduced because of reverberation and noise. Moreover, prior to diarization, separating the audio streams captured by a distant microphone array requires heavy computation.

For a meeting rooms equipped with a microphone array or with distributed microphones, the observed audio streams are usually processed to obtain one stream for each active participant (for example with audio beamforming in [7]).

Nowadays, with the proliferation of portable devices (laptop computers, PDAs and smart phones), it is not rare that in a meeting situation, each of the participants may be carrying a device having a microphone. Thus the speaker localization and the acquisition of the data streams may be performed using these microphones [8, 9]. Such a set of microphones is referred to as a distributed array. These approaches usually require the different devices to communicate together in order to acquire all the audio streams, then distributed or centralized processing may be applied to perform localization, diarization or other tasks.

As the first step in developing a multi-modal front-end for speaker diarization exploiting a distributed array, this paper discusses the signal processing involved in the speaker identification task (at this first step networking problems are not treated yet). The proposed front-end exploits audio data from the tie microphones and position information given by a human tracker system based on laser range finders (LRF) [10]. During the meeting, the positions of the different participants are tracked using the LRF and one audio stream is obtained for each of the participants using a tie microphone (a microphone fixed in front of the torso). Then speaker identification is performed by using Gaussian mixture models (GMM) of the mel-frequency cepstral coefficients (MFCCs) extracted from theses audio streams [11]. For each participant, the tie microphone fixed on their torso is dominated by their voice when talking. But the speech signal from the tie microphones also contains environmental noises and interferences from the other participant voices if they are talking. If a participant is silent, the interfering voice of the person, or persons, talking at that moment is likely to be the dominant signal. Thus it is necessary to implement a accept/reject system to detect the active channels.

With this system, the participants are able to seat freely,

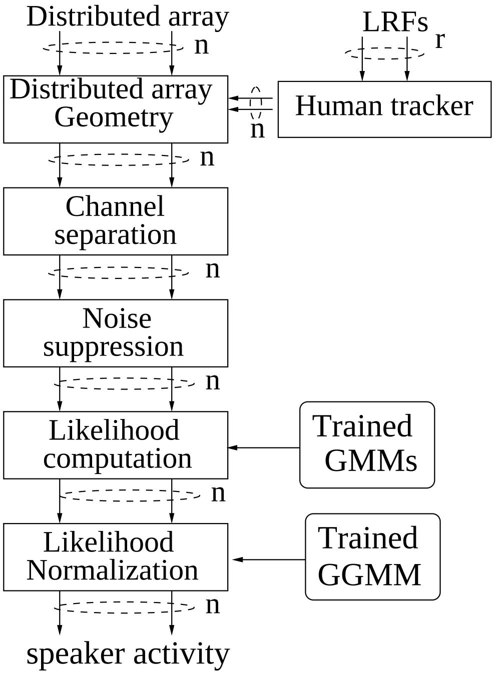

Fig. 1. Outline of the speaker identification system.

to stand and even move during the meeting because the LRFs tracks their positions. Moreover using LRF based tracking is preferable to using audio data for tracking as it insensitive to the acoustic noise and to the reverberation. It is also an interesting alternative to camera based tracking as it is very precise. Note that the number of participants is estimated by the human tracker independently of the fact that they are talking or not.

Experiments were conducted in a realistic meeting situation to demonstrate the capacity of the proposed front-end to identify active participants.

2. METHOD

Fig. 1 gives an outline of the proposed front-end for speaker identification. A presentation of the different modules follows.

2.1. Laser range finder

The motion of the participants in the meeting area is monitored using r LRFs mounted on pole around the meeting area's perimeter (represented by the circles in Fig. 5). The scanning laser range finders are mounted above the obstacles, like the table and chairs, to a height where the torso of the participants (sitting or standing) could be easily observed. To reduce the errors due to noise and occlusion, each person is tracked with a particle filter using a linear motion model with random perturbations. The likelihood is evaluated based on the potential occupancy of each particle's position. By computing a weighted average across all particles, the $\{x, y\}$ position is calculated at a frequency of approximately 37 Hz. Details of the algorithm are presented in [10].

At a given time t, the estimated number of participants in the meeting is $n(t)$ and their estimated positions are $\{x_i(t), y_i(t)\}_{i\in[1,n(t)]}$.

2.2. Noise cancellation

Each of the participants is wearing a tie microphone attached in the front of their torso. The position of these microphones are given by the LRF based tracking system that tracks the position of the torso of all the participants. In this paper, we assume for simplicity that the correspondence between a microphone and a given position is known. Thus the set of tie microphones defines a distributed microphone array whose geometry is known.

The goal of the noise cancellation module is to provide an audio stream for each of the $n(t)$ detected participants that contains less interference from the other participants and fewer environmental noise than the unprocessed streams from the tie microphones (the observed signals). These streams are obtained by filtering the observed signals in the frequency domain. After performing a F bins short time Fourier transform (STFT), the vector of observation in the fth frequency bin is

$$\mathbf{X}(f,k) = \begin{bmatrix} X_1(f,k) \\ X_2(f,k) \\ \vdots \\ X_n(f,k) \end{bmatrix}$$

where k denotes the frame index.

Let us define

$$\mathbf{S}(f,k) = \begin{bmatrix} S_1(f,k) \\ S_2(f,k) \\ \vdots \\ S_n(f,k) \end{bmatrix}$$

the vector containing the speech of all participants at frame index k and frequency bin f. Considering only direct path propagation we can write the mixing process as

$$\mathbf{X}(f,k) = \widehat{\mathbf{A}}(f,k)\mathbf{S}(f,k)$$

where $\widehat{\mathbf{A}}(f,k)$ is the matrix of general term

$$A_{ij}(f,k) = \frac{1}{4\pi r_{ij}(k)} e^{-j2\pi f r_{ij}(k)/c}$$

with c is the celerity of sound and $r_{ij}(k)$ the distance between the jth speech source (the mouth of the jth participant) and the ith microphone (fixed to the ith participant).

The distance $r_{ij}(k)$ is decomposed in two terms d_i, the distance between the mouth of the ith participant and the

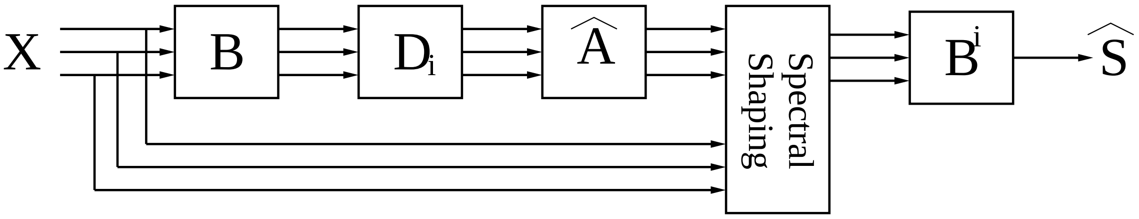

Fig. 2. Noise suppression.

microphone fixed to his/her torso (assumed constant), and $d_{ij}(k)$, the distance between the microphones i and j. We have

$$r_{ij}(k) = \sqrt{d_i^2 + d_{ij}^2(k)}.$$

The distances $d_{ij}(k)$ are obtained using the positions given by the human tracker whereas the distances d_i are assumed to be known.

A separation matrix is obtained by taking the inverse of the mixing matrix

$$\mathbf{B}(f,k) = \widehat{\mathbf{A}}^{-1}(f,k).$$

Then the separated audio streams for the n participants are

$$\mathbf{Y}(f,k) = \mathbf{B}(f,k)\mathbf{X}(f,k).$$

Rather than using these separated streams, better results were obtained by applying a post-filter approach as the ones in [12, 13] (see Fig. 2 where n = 3).

Let us define the noise estimate

$$\mathbf{N}_i(f,k) = \widehat{\mathbf{A}}(f,k)\mathbf{D}_i\mathbf{Y}(f,k)$$

where $\mathbf{D}_i$ is a diagonal matrix with all entries set to one except the ith entry which is null. $\mathbf{N}_i(f,k)$ is the estimate of the contribution in the observed signals of all the signals except the ith participant speech. Then an estimate of the contribution of the ith participant speech is obtained by using spectral shaping (We use a post-filter similar to the one used in [12, 13]) The gain for the ith signal is

$$G_i^{(j)}(f,k) \quad = \quad \frac{|X^{(j)}(f,k)|^2}{|X^{(j)}(f,k)|^2 + \alpha|N_i^{(j)}(f,k)|^2}$$

where the superscript (j) denotes the jth component and α is a parameter controlling the noise reduction. The ith component of the filtered target speech is

$$\widehat{Z}_i^{(j)}(f,k) \quad = \quad \sqrt{G_i^{(j)}(f,k)|X^{(j)}(f,k)|^2}\frac{X^{(j)}(f,k)}{|X^{(j)}(f,k)|}.$$

Finally the speech estimate $\widehat{S}_i(f,k)$ is obtained by taking

$$\widehat{S}_i(f,k) = \mathbf{B}^i(f,k)\widehat{\mathbf{Z}}_i(f,k)$$

where $\mathbf{B}^i(f,k)$ is the ith row of the matrix $\mathbf{B}(f,k)$ (the row corresponding to the ith participant).

2.3. Corpus and GMM

Text-independent speaker identification is performed by scoring the MFCCs (12 MFCCs and the energy, their derivatives and their accelerations) extracted from the audio streams of each participant by means of GMMs corresponding to the target speakers [11]. In this experiment, nine speakers were considered (5 females and 4 males). In the remainder of the paper, the speakers are designated by the letters $\{a, b, c, \cdots, i\}$. For each speaker a common training set of 100 Japanese sentences from the JNAS database [14] was recorded using a tie microphone while sitting at the table in the experiment room. Then a GMM was trained for each of the speakers using the 100 utterances. The GMM for all the speakers are designated by $\{\lambda_a, \lambda_b, \cdots, \lambda_i\}$. A general GMM was also trained using the 900 utterances (referred to as GGMM in Fig. 1). The general GMM is designated by λ_G.

The test set was recorded in the same room while monitoring the speaker movement with the LRF based human tracker system. Only three $\{a, b, c\}$ of the nine speakers were sitting around the table and were not constrained of any manner (see Fig. 5). Three different sets of 50 sentences from the JNAS database were prepared and each speaker was assigned one of these sets. Using these sets, 350 test utterances were recorded. First each of the speaker was reading alone its test set (the two other persons are sitting around the table but are remaining silent). These are the test sets T_a, T_b and T_c. Then the three combinations of two speakers simultaneously reading were recorded (test sets T_{ab}, T_{ac} and T_{bc}). Finally, the three speakers were reading simultaneously (test set T_{abc}).

Training and scoring were performed with Htk 3.41 [15] using the whole test utterances.

2.4. Activity detection

The GMMs are used to determine for each utterance which of the participants are active. For decision based on likelihood, it is usual to apply some sort of normalization [16, 17]. In this paper, for a given stream $\widehat{S}_k$ of a given test utterance the likelihood given by the GMMs are normalized using the following likelihood ration

$$\bar{p}(\widehat{S}_k|\lambda_i) = \log p(\widehat{S}_k|\lambda_i) - \log p(\widehat{S}_k|\lambda_G).$$

Then the accept/reject procedure is conducted by comparing the largest normalized likelihood

$$\bar{p}(\widehat{S}_k|\lambda_j) \quad = \quad \mathrm{max}_i \bar{p}(\widehat{S}_k|\lambda_i)$$

to a threshold ϵ

- if $\bar{p}(\widehat{S}_k|\lambda_j) \geq \epsilon$ then speaker j is active in stream $\widehat{S}_k$.
- if $\bar{p}(\widehat{S}_k|\lambda_j) < \epsilon$ then no speaker is active in stream $\widehat{S}_k$.

For each utterance, this test is conducted for all the audio streams.

3. EXPERIMENTS

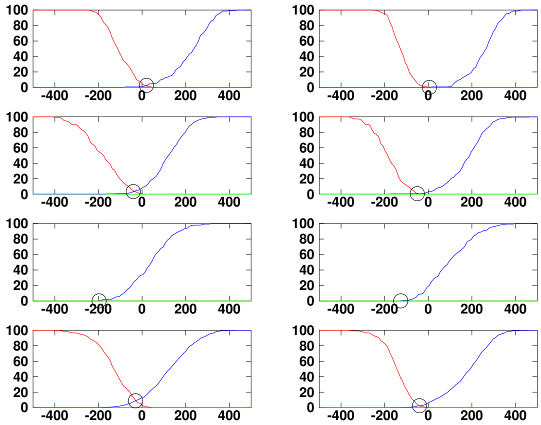

Fig. 3. Deletion (blue), insertion (red) and substitution (green) versus threshold ϵ for single speaker (top row), two speakers (second row), three speakers (third row) and all cases (bottom) for unprocessed (left) and processed audio streams (right).

The experiment setup is described in Fig. 5. The four circles in the corners represent the pole mounted LRFs used by the human tracker, the cross gives the position origin and the probability densities of the positions of the three speakers during the experiment also appear. In this first step, the tie microphones are still wired microphones connected to the same computer.

Note that for all test sets except the test set T_{abc}, at least one of the speakers is silent. The results of the speaker identification experiment are given in terms of deletion, insertion and substitution errors:

- An insertion error occurs when the largest normalized likelihood associated to the audio stream of a silent speaker is larger than the threshold ϵ,

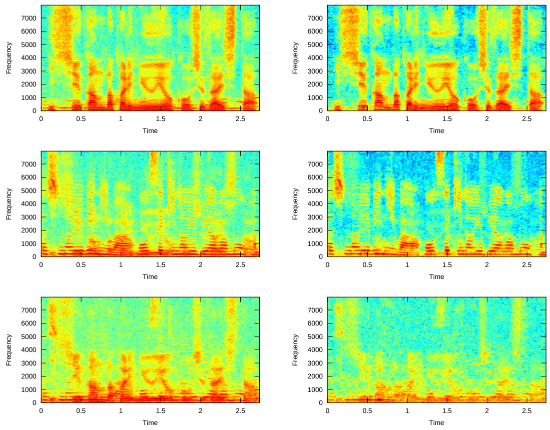

Fig. 4. Spectra of unprocessed audio streams (left) and processed audio streams (right) when speakers a and b are talking.

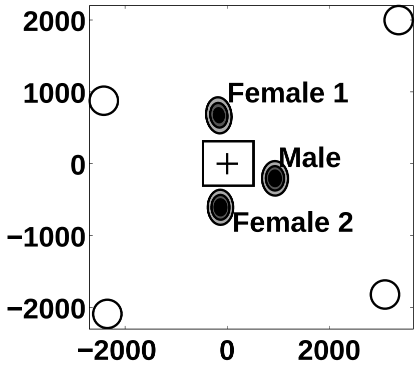

Fig. 5. Pole mounted LRFs (circles), table (rectangle), position origin (cross) and probability densities of the three speakers position (distances are in mm)

- A deletion error occurs when the largest normalized likelihood associated to the audio stream of an active speaker is smaller than the threshold ϵ,
- A substitution error occurs when the largest normalized likelihood associated to the audio stream of an active speaker is larger than the threshold ϵ but is not the correct one (for example $\bar{p}(\widehat{S}_k|\lambda_a)$ is the largest normalized likelihood but the audio $\widehat{S}_k$ is associated to the speaker b).

Two different cases were compared where the audio stream of each speaker is obtained by

- her or his own tie microphone (unprocessed),
- the processed stream $\widehat{S}_k$ she or he is assigned (processed with $\alpha = 17$).

Table 1. Deletion percentage for selected threshold.

	unprocessed	processed
one speaker	2.67	0
two speakers	3.33	0.67
three speakers	0	0
all	8.83	2.5

Table 2. Insertion percentage for selected threshold.

	unprocessed	processed
one speaker	2	0
two speakers	4	0.67
three speakers	0	0
all	10	2

The insertion, deletion and substitution percentages are plotted for different values of the threshold ϵ in Fig. 3. First, we can see that no substitution occurred for this experiment. In all figures, the black circle represent the threshold for which a trade off between insertion and deletion errors is obtained. The percentages for these thresholds are given in Tables 1 and 2. In particular, the processed audio streams give a better performance when considering one unique threshold for all the test sets (bottom of fig 3 and last rows of tables 1 and 2), which is the operating condition.

The spectra of the audio streams are given in Fig. 4 for the test set T_{ab}. The channel assigned to speaker a (top) and b (middle) contain less noise and fewer interference after processing. The processing also reduces the amount of speech that leaks in the bottom channel assigned to the silent speaker c.

4. DISCUSSION

The human tracker is a fast and accurate way of obtain the position of the speakers in the $\{x, y\}$ plane but we have no access to the z coordinate of the mouth or the tie microphone. In this paper, we assumed that all the tie microphones were at the same height and also used an approximation of the distance between a speaker mouth and tie microphone. But during the test recording, the three subjects fixed the tie microphone as they desired. Despite this mismatch, the proposed approach was able to improve significantly the performance for the considered task.

The experiment in this paper was conducted using tie microphones connected to the same computer in order to deal with the signal processing part only. In a real situation, the participants are likely to use microphones connected to devices that communicate using a wireless network. Then for usual approaches one of the most important problem is to synchronize the audio data in order to perform collaborative array processing (like beamforming for estimating the positions)[9]. But with the proposed approach, the localization is performed by the human tracker thus synchronization may be a less sensitive issue.

5. CONCLUSION

In this paper, we proposed an experiment to test the use of LRF based human tracker in a multi-modal front-end for speaker diarization in a meeting situation. Since the positions of all the participants are known at each instant, it is possible to use this information for monitoring a set of tie microphones worn by the participants. Then applying appropriate array processing techniques to this distributed microphone array, it was possible to improve the accuracy in a speaker detection and identification task.

6. ACKNOWLEDGEMENTS

This work was supported by the Ministry of Internal Affairs and Communication.

7. REFERENCES

[1] A. Waibel, H. Yu, M. Westphal, H. Soltau, T. Schultz, T. Schaaf, Y. Pan, F. Metze, and M. Bett, “Advances in meeting recognition,” *HLT ’01: Proceedings of the first international conference on Human language technology research*, pp. 1–3, 2001.

[2] J. Carletta, “Unleashing the killer corpus: experiences in creating a multi-everything ami meeting corpus,” *Language ressource and evaluation journal*, vol. 41, no. 2, pp. 181–190, 2007.

[3] J.G. Fiscus, J. Ajot, and J.S. Garofolo, “The rich transcription 2007 meeting recognition evaluation,” *Lecture note in computer science*, vol. 4625, pp. 373–389, 2008.

[4] H. Sun et al., “Speaker diarization system for rt07 and rt09 meeting room audio,” *ICASSP 2010*, pp. 4982–4985, 2010.

[5] H. Sun et al., “Speaker diarization for meeting room audio,” *INTERSPEECH 2009*, pp. 900–903, 2009.

[6] M.S. Brandstein and H.F. Silverman, “A robust method for speech signal time-delay estimation in reverberant rooms,” *ICASSP 1997*, pp. 375–378, 1997.

[7] A. Stolcke, G. Friedland, and D. Imseng, “Leveraging speaker diarization for meeting recognition from distant microphones,” *ICASSP 2010*, pp. 4390–4393, 2010.

[8] T.S. Wada, E. Robledo-arnuncio, G. Yue, and B.H. Juang, “Immersive acoustic signal processing for intelligent collaboration,” *Proc. 9th Western Pacific Acoustics Conference*, p. 653, 2006.

[9] Y. Jia, Y. Luo, Y Lin, and I. Kozintsev, "Distributed microphones arrays for digital home and office," *ICASSP 2006*, pp. 1065–1068, 2006.

[10] D.F. Glas et al., "Laser tracking of human body motion using adaptive shape modeling," *Proceedings of 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 602–608, 2007.

[11] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE transaction on speech and audio processing*, vol. 3, no. 1, pp. 72–82, 1995.

[12] Y. Takahashi, Y. Uemura, H. Saruwatari, K. Shikano, and K. Kondo, "Structure selection algorithm for less musical-noise generation in integration systems of beamforming and spectral subtraction," *2009 IEEE Workshop on Statistical Signal Processing SSP2009, Cardiff, Wales, UK*, pp. 701–704, 2009.

[13] J. Even, H. Saruwatari, K. Shikano, and T. Takatani, "Speech enhancement in presence of diffuse background noise: Why using blind signal extraction?," *International Conference on Acoustics, Speech, and Signal Processing ICASSP 2010, Dallas, USA*, pp. 4770–4773, 2010.

[14] K. Ito, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of Acoustical Society of Japan*, vol. 20, pp. 196–206, 1999.

[15] S.J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*, Cambridge University Press, 2006.

[16] A. Rosenberg, J. DeLong, C. Lee, B.H. Juang, and F. Soong, "The use of cohort normalized scores for speaker verification," *Proc. ICSLP*, pp. 599–602, 1992.

[17] T. Matsui and S. Furui, "Likelihood normalization for speaker verification using a phoneme- and speaker-independent model," *Speech communication*, vol. 17, no. 1-2, pp. 109–116, 1995.