AUDIO TRACKING FOR SMALL MEETINGS USING LASER RANGE FINDERS AND LOCAL AUDIO SCANS

Jani Even, Panikos Heracleous, Carlos Ishi, Takahiro Miyashita and Norihiro Nogita

ATR Intelligent Robotics and Communication Laboratories, Kyoto, Japan
even@atr.jp

ABSTRACT

This papers presents a system designed for separating and tracking the voices of a few persons talking around a table. During the meeting, the locations of the participants are monitored by a human tracker system based on laser range finders (LRFs). Then using a uniform circular array (UCA) of microphones, audio localization is performed to estimate the most powerful sound source, usually the mouth, in the neighborhood of each of the detected participants. Finally, beamforming is applied to obtain an audio stream for each of the detected participants. The use of LRF based human tracker enables the system to assign a continuous audio track to each of the participants. Experimental results using real meeting data show the efficiency of the proposed approach.

1. INTRODUCTION

An important task in meeting transcription is speaker diarization (i.e. to find "Who talked when") [1]. It is quite common to use a microphone array or distributed microphones to obtain one stream for each active participant (for example with audio beamforming in [2] or using the directions of arrival [3]). Then in order to create the diary of the meeting, speaker identification and activity detection is performed using these streams.

In this paper, we propose an extension of the multi-modal approach to this problem we presented in [4]. Contrary to other approaches, the speaker localization is not performed using only the audio signals. In addition to the audio signals recorded by a uniform circular microphone array (UCA), laser range finders (LRFs) are used to obtain distances. First the locations of the participants are estimated by a human tracker system based on laser range finders (LRF) [5] then these positions are refined using the audio signals to scan with a beamformer the neighborhood of each of the positions given by the human tracker (this was not done in [4]). This scan, referred to as local scan, is based on the broadband MUSIC algorithm (see [6] for details on the different broadband MUSIC approaches). In particular, the power of the MUSIC pseudo spectrum is used to determine the activity of the participants. After localization and activity detection, the audio data are processed in order to obtain an enhanced audio stream for each of the active participants at all time. A specificity of the proposed method is that silent participants are also assigned an audio stream because they are detected by the LRF based human tracker. Experiments were conducted in a realistic meeting situation to demonstrate the efficiency of the proposed method. In order to underline the gain of using the human tracker, a conventional broadband MUSIC algorithm was also used. For this algorithm, the tracking is performed by using spatial information but also Gaussian mixture models (GMMs) [7] that were trained before hand.

2. METHOD

2.1. Localization

The motion of the participants in the meeting area is monitored using 4 LRFs mounted on poles around the meeting area's perimeter. 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 (see [5]). The human tracker gives the position $\{x, y\}$ of the torso of each of the participants in the room. However, the positions that matter are not the positions of the participants but the positions of their mouths. Consequently, the positions given by the human tracker have to be refined. In particular, the z coordinates have to be estimated.

For this purpose, a local audio scan is applied around each of the positions given by the human tracker to estimate the position of the mouth (see Fig. 1). This local audio scan is based on the MUSIC algorithm.

The raw audio signals, referred to as the *observed signals* in the remainder, are acquired by a uniform circular array (UCA) of $m = 16$ microphones positioned on a table in the middle of the meeting area. The position of the microphone array is assumed to be known.

For the localization purpose, the frequency domain observation is obtained by using a short time Fourier transform with a hanning window of 51 points, a shift of 25 points and an fft size of 64 points. The localization is performed every 200 ms corresponding to 128 frequency frames. The vector

THIS WORK WAS SUPPORTED BY KAKENHI (21118003).

of observed signals in the fth frequency bin is

$$\mathbf{X}_L(f,k) = \left[\begin{array}{ccc} X_1(f,k), & \cdots & , X_m(f,k) \end{array} \right]^T$$

where k denotes the frame index.

For a selected number of frequency bins, the narrow band MUSIC pseudo power spectrum $\mathbf{P}_{nb}(f,x,y,z)$ is obtained by

- performing a singular value decomposition of the observation covariance $\mathbf{\Gamma}(f) =< \mathbf{X}_L(f,k)\mathbf{X}_L^H(f,k) >_k$,
- creating the projector $\mathbf{P}_K(f)$ on the space spanned by the K least powerful singular values,
- scanning the space around the LRF position by using a beamformer $\mathbf{W}(f,x,y,z)$
- estimating the pseudo power by

$$\mathbf{P}_{nb}(f,x,y,z) = \frac{1}{\mathbf{W}(f,x,y,z)\mathbf{P}_K(f)\mathbf{W}^H(f,x,y,z)}.$$

Then the broadband MUSIC pseudo spectrum is obtained by averaging the narrow band pseudo spectra

$$\mathbf{P}_{bb}(x,y,z) =< \mathbf{P}_{bb}(f,x,y,z) >_f .$$

For each of the positions $\{x_0, y_0\}$ given by the human tracker, the updated position $\{x,y,z\}$ gives the maximum of the broadband MUSIC pseudo spectrum estimated in the space around $\{x_0, y_0\}$. The participant is considered active if that maximum pseudo spectrum is above a threshold ϵ_p.

For each of the 200 ms block, the proposed method detect if a participant is active and at the same time give a refined estimate of the mouth position of this active participant. Note that the activity of the participants along the 200 ms blocks is tracked by the human tracker even if the participants are silents.

2.2. Audio stream

At any time, an audio stream is assigned to each of the Q active participants. The desired streams are obtained by processing the observed signals in the frequency domain. For the beamforming purpose, the frequency domain observation is obtained by using a short time Fourier transform with a hanning window of 401 points, a shift of 200 points and an fft size of 512 points. The beamforming is performed every 200 ms corresponding to 16 frequency frames. The vector of observed signals in the fth frequency bin is

$$\mathbf{X}(f,k) = \left[\begin{array}{ccc} X_1(f,k), & \cdots & , X_m(f,k) \end{array} \right]^T$$

where k denotes the frame index.

First, the refined positions (in the microphone array referential) are used to estimate a set of delay and sum (DS) beamformers. Only considering the delays for a direct path propagation we can write the set of DS beamformers as

$$\mathbf{Y}_{\mathrm{DS}}(f,k) = \left[\begin{array}{c} \mathbf{w}_1(f,k) \\ \vdots \\ \mathbf{w}_Q(f,k) \end{array} \right] \mathbf{X}(f,k)$$

where $\mathbf{Y}_{\mathrm{DS}}(f,k)$ are the beamformed audio streams and the $Q \times m$ matrix has general term

$$w_{ij}(f,k) = e^{-j2\pi f \frac{r_{ij}(k) - r_{i1}(k)}{c}}$$

with c the celerity of the sound and $r_{ij}(k)$ the distance between the mouth of the ith participant and the jth microphone (the first microphone is used as reference).

Then an audio stream for each of the participants is obtained by applying a linearly constrained minimum variance (LCMV) beamformer.

The LCMV beamformer weights for the ith participant are given by

$$\mathbf{w}_{\mathrm{LCMV},i}(f,k) = \frac{\mathbf{w}_i(f,k)\mathbf{K}^-1(f)}{\mathbf{w}_i(f,k)\mathbf{K}^-1(f)\mathbf{w}_i^H(f,k)}$$

where $\mathbf{K}(f)$ is the estimate of the noise and interference covariance and $\mathbf{w}_i(f,k)$ is the steering vector pointing to the ith participant.

The estimate of the noise and interference covariance is composed of two parts

$$\mathbf{K}(f) = \mathbf{\Gamma}(f) + \sum_{j=1, j\neq i}^{Q} \mathbf{w}_j^H(f,k)\mathbf{w}_j(f,k)\sigma_j^2(f).$$

The first term $\mathbf{\Gamma}(f)$ is the estimate of the noise covariance obtained when only the noise is present. The second term represents the contribution of the other participants (the interferences). It is a sum of the contributions made by each interfering participants. The interfering participants are represented by point sources located at the positions given by the human tracker. For each of these point sources, the DS beamformer is used to obtain the power which is estimated by

$$\sigma_j(f) \quad = \quad \mathrm{var}\left\{ \mathbf{Y}_{\mathrm{DS}}^{(j)}(f,k) \right\}.$$

Note that for a silent participant, this power is likely to be small.

Finally, the audio stream of the ith participant is

$$Y_{\mathrm{LCMV},i}(f,k) = \mathbf{w}_{\mathrm{LCMV},i}(f,k)\mathbf{X}(f,k)$$

The LCMV beamformers provide an audio stream for each of the detected participants that contains less interference from the other participants and fewer environmental noise than the DS beamformer streams. In the remainder, we refer to $Y_{\mathrm{LCMV},i}(f,k)$ by $Y_i(f,k)$ for convenience.

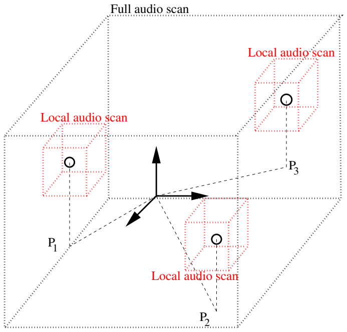

Fig. 1. Full and local audio scans for three positions P_1, P_2 and P_3

3. EXPERIMENTS

Two different cases were compared where the audio stream of each participant is obtained by: using localization based only on audio signals (MUSIC; the full scan in Fig. 1) and based on human tracker and audio signals (LRF + MUSIC; the local scans in Fig. 1).

3.1. Experimental setup

The experiment setup is described in Fig. 3. The four circles in the corners represent the pole mounted LRFs used by the human tracker, the cross gives the position of the microphone array and the probability densities of the positions of the three speakers during the experiment also appear (note that the densities are sharp even if the speakers were not told to limit their movements). The experiment setup consists of four pole mounted LRFs (Fig.2 right) in the corner of the monitored area and of a table top UCA (Fig.2 left).

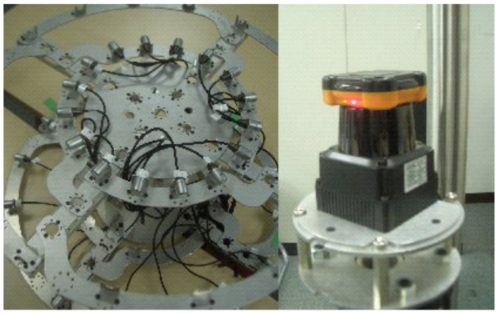

Fig. 2. Table top microphone array (left) and pole mounted LRF (right).

3.2. Data set

In this experiment, three participants were considered (2 females and 1 males). In the remainder of the paper, the speakers are designated by the letters $\{a, b, c\}$. Two test sets were recorded in a room while monitoring the speaker movement with the LRF based human tracker system. The three participants were sitting around a table (the participants were not given any instruction concerning their movements). A first test set, referred to as *reading set* is obtained by letting the participants read some sentences from the JNAS database. First b and c are reading at the same time then after a short pause a and b are reading at the same time. The second test set, referred to as *conversation test*, is extracted form a real conversation between the three participants and includes speech and interjections. The activity of the participants was hand labeled for both of the test sets. The observed signal from microphone 1 is given for each test set in Fig.4.

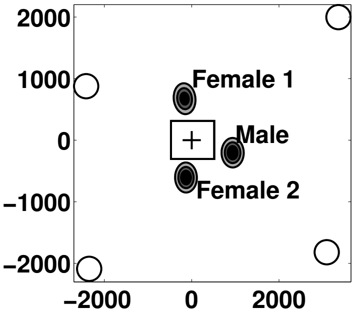

Fig. 3. Microphone array (cross), pole mounted LRFs (circles), table (rectangle) and probability densities of the three speakers position (distances are in mm)

3.3. Conventional broadband MUSIC

To show the advantage of using human tracker system for the diarization, a conventional broadband MUSIC approach was also used.

For the conventional broadband MUSIC algorithm, only one broadband pseudo spectrum is obtained by scanning the whole space then in the selected frequency bins. Then the number of audio sources is determined by finding the local maxima of the pseudo power spectrum that are above the threshold ϵ_p. The localization is also performed every 200 ms using 128 frames. Then audio streams are obtained for each of the detected audio sources using the same beamforming technique as for the LRF + MUSIC case.

However, a big difference is that the detected audio sources from each of the 200 ms blocks have to be combined together to create the audio tracks. For audio sources active in consecutive blocks, the distance between the sources is used to combine them: sources that did not move much are considered the same. For combining sources that are inactive for several blocks, it was necessary to use speaker identification based on GMMs [7]. The features extracted from the audio streams are the MFCCs (12 MFCCs and the

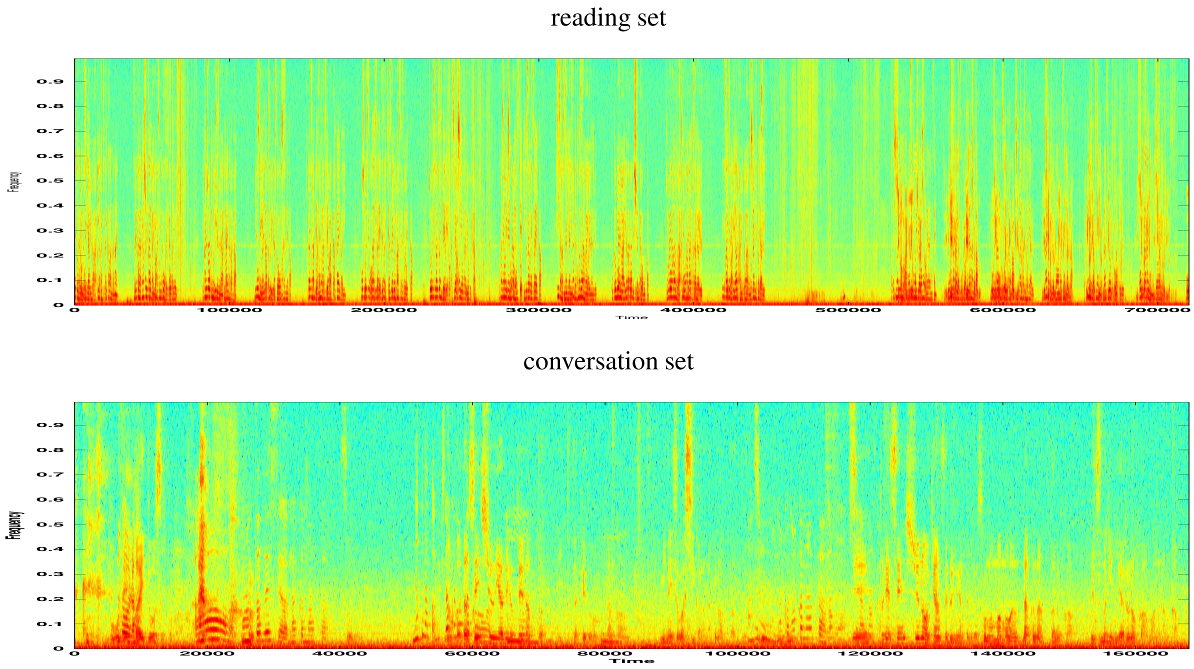

Fig. 4. Observations for reading set (top) and conversation set (bottom).

log spectral energy, their derivatives and their accelerations). For each speaker a common training set of 100 Japanese sentences from the JNAS database [8] was recorded using a close talking microphone while sitting at the table in the experiment room. A set of GMMs was trained for each of the speakers using these 100 utterances and a general GMM was also trained using the 300 utterances (referred to as GGMM). The GMMs for all the speakers are designated by $\{\lambda_a, \lambda_b, \lambda_c\}$ and the GGMM by $lambda_G$. The number of mixtures was set to 512 after testing several values. Training and testing were performed with HTK 3.41 using the whole utterances.

The GMMs are used to determine for each block which of the participants is active. In this paper, for a given block the likelihoods are normalized using the following likelihood ratio (For decision based on likelihood, it is usually necessary to apply a normalization [9, 10])

$$\mathcal{L}(Y_q|\lambda_i) = \log p(Y_q|\lambda_i) - \log p(Y_q|\lambda_G).$$

where λ_G is the general GMMs estimated on all training utterances.

The decision rule is to select for each of the block the speaker whose model has the largest likelihood

$$\mathcal{L}(Y_q|\lambda_j) \quad = \quad \max_i \mathcal{L}(Y_q|\lambda_i)$$

as the active speaker.

Table 1. Deletion, insertion and correct percentages for the MUSIC method.

	reading set			conversation set		
	del	ins	cor	del	ins	cor
a	0.0	6.5	93.5	0.5	23.7	75.8
b	1.1	23.7	75.2	10.2	14.3	75.5
c	0.0	18.0	81.9	10.8	8.2	81.1
avg.	0.4	16.1	83.5	7.2	15.4	77.4

Table 2. Deletion, insertion and correct percentages for the LRF + MUSIC method.

	reading set			conversation set		
	del	ins	cor	del	ins	cor
a	0.2	9.2	90.6	1.4	18.3	80.3
b	0.6	29.7	69.7	2.5	17.3	80.2
c	0.4	12.2	87.4	1.8	13.7	84.5
avg.	0.4	17.0	82.6	1.9	16.4	81.7

3.4. Diarization

The results of the meeting diarization are given in terms of deletion, insertion errors in Table 1 and 2:

- An insertion error occurs when a speaker is detected for the audio stream of a silent participant.
- A deletion error occurs when an active participant is not detected.

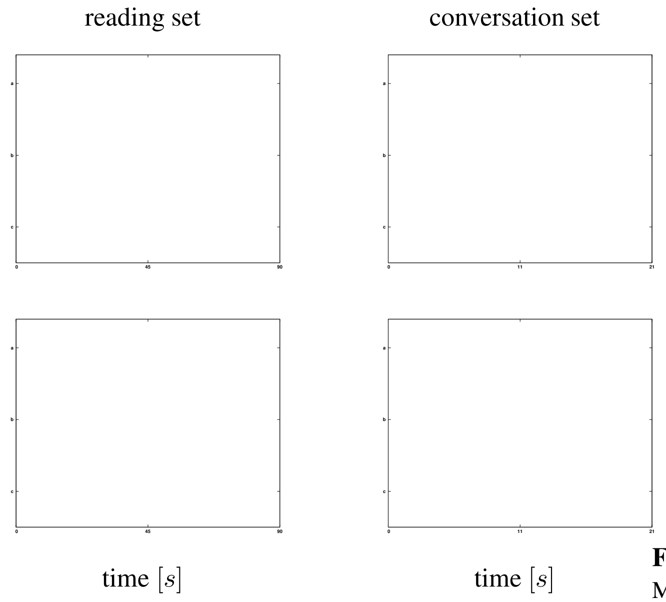

Fig. 5. Result of the diarization for the two test sets with MUSIC (top) and LRF + MUSIC (bottom).

These percentages are computed by comparing the hand labeled activity with the activity given by both of the methods. Figure 5 gives a graphical representation of the diarization results. For each of the sub-figure, one row correspond to one of the three participants. The color code shows the deleted samples (red), the inserted samples (blue) and the correctly detected samples (green).

We can see that using both the LRFs and the audio data for the localization gives the best performance for the conversation set but for the reading set there is not much difference (it is also faster than the full audio scan).

3.5. Localization

Figure 6 shows the repartition of the detected block power in the space for the three participants (a in blue, b in red and c in green) in the two data sets. We can especially see that for the conversation set, the MUSIC method has a bad estimate for the speakers b and c that are the two female speakers as the GMMs trained on reading conditions are not good for the interjections present in the conversation set.

4. CONCLUSION

This paper presents a multi-modal approach to the diarization problem that combines LRF base human tracker with microphone array. In particular using LRF is an efficient way to perform the tracking the participants and merge the detected audio blocks together.

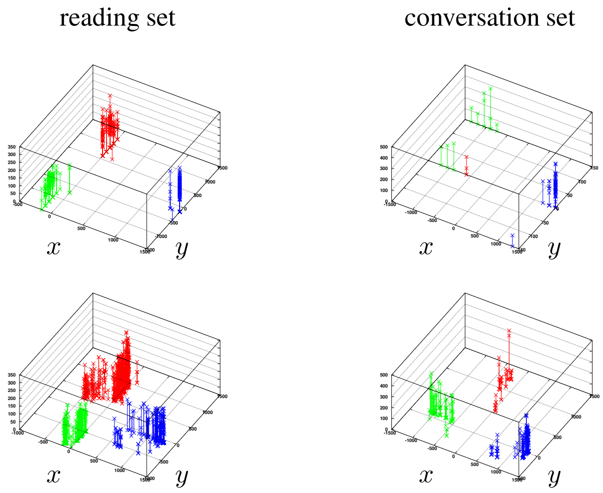

Fig. 6. MUSIC Pseudo power in space for the two test sets with MUSIC (top) and LRF + MUSIC (bottom).

5. REFERENCES

[1] 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.

[2] F. Asano et al., “Detection and separation of speech events in meeting recordings using a microphone array,” *EURASIP Journal on Audio, Speech, and Music Processing*, vol. ID 27616, 2007.

[3] S. Araki, M. Fujimoto, K. Ishizuka, H. Sawada, and S. Makino, “A doa based speaker diarization system for real meeting,” *HSCMA 2008, Trento, Italy*, pp. 29–32, 2008.

[4] J. Even, P. Heracleous, C. Ishi, and N. Hagita, “Multimodal front-end for speaker activity detection in small meetings,” *Proceedings of 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 536–541, 2011.

[5] D.F. Glas et al., “Laser tracking of human body motion using adaptive shape modeling,” *IROS 2007, San Diego, USA*, pp. 602–608, 2007.

[6] S. Argentieri and P. Danès, “Broadband variations of the music high-resolution method for sound source localization in robotics,” *IROS-2007, San Diego, USA*, pp. 2009–2014, 2007.

[7] 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.

[8] K. Ito et al., “Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research,” *The Journal of Acoust. Soc. of Japan*, vol. 20, pp. 196–206, 1999.

[9] 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.

[10] 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.