2009 11th IEEE International Symposium on Multimedia

Speaker Detection and Applications to Cross-Modal Analysis of Planning Meetings

Bing Fang
Virginia Polytechnic Institute and State Univ.
Blacksburg,VA 24061, USA
fangb@vt.edu

Yingen Xiong
Nokia Research Center
Palo Alto, CA 94304, USA
yingen.xiong@nokia.com

Francis Quek
Virginia Polytechnic Institute and State Univ.
Blacksburg,VA 24061, USA
quek@vt.edu

***Abstract*—Detection of meeting events is one of the most important tasks in multimodal analysis of planning meetings. Speaker detection is a key step for extraction of most meaningful meeting events. In this paper, we present an approach of speaker localization using combination of visual and audio information in multimodal meeting analysis. When talking, people make a speech accompanying mouth movements and hand gestures. By computing correlation of audio signals, mouth movements, and hand motion, we detect a talking person both spatially and temporally. Three kinds of features are extracted for speaker localization. Hand movements are expressed by hand motion efforts; audio features are expressed by computing 12 mel-frequency cepstral coefficients from audio signals, and mouth movements are expressed by normalized cross-correlation coefficients of mouth area between two successive frames. A time delay neural network is trained to learn the correlation relationships, which is then applied to perform speaker localization. Experiments and applications in planning meeting environments are provided.**

***Keywords*-meeting event detection; meeting analysis; speaker localization; mutimodal meeting analysis; mouth movement; hand motion; audio signal analysis; planning meeting;**

I. Introduction

A. Multimodal Analysis of Planning Meetings

Meetings are gatherings of humans for the purpose of communication. Meeting events are captured by cameras [1] in video-based multi-modal analysis of planning meetings. Considering the scenario, where there is access to long video/audio data streams of venues for meetings and joint planning, an analyst may have to browse the video and find when a particular meeting participant is speaking and the junctures where speaking-turn-exchange occurs. The analyst may request a segmentation of the underlying discussion by coherence of topical foci or index it with likely topic break points. These topical clusters may be weighted by the detection of chained referential, clustering of exchanges by subsets of participants, or by clustered references to artifacts of discussion (e.g. a particular locus along a map on the wall). In addition, the system should be able to classify the activity in the venues as either ordered meeting, sub-group caucusing, or social bantering. The analyst may be interested in locating video segments involving the highest ranking discussant by the communicative behavior of the discussants, or the formation of sub-group coalitions in the meetings. When speech information is not available (e.g. because of poor audio quality or for languages where automatic speech recognition is inadequate), the indexing may be accomplished solely on vision-based multi-modal behavior and prosody analysis. When automatic speech recognition is useful, the analyst may frame questions in terms of topical discussion content and chained references to the content using information from the multi-modal communication (speech, gesture and gaze) of the participants in meeting. Understanding of particular linguistic constructs like motion descriptions may be enriched with information from other modalities (such as path and ground information provided by the accompanying gestural behavior). Such multi-modal language analysis technology may be applied either to meetings of friendly cooperative discussants (e.g. to enrich video-based minutes of military planning meetings to provide commanders better understanding of evolving sub-plans formulated by distributed teams, or to evaluate the communicative effectiveness of trainees in war-gaming), or in surveillance video of subjects who are unwitting of the presence of the recording device. This paper addresses the aspect of speaker localization which is a key step for the detection of most meaningful meeting events in multi-modal analysis of planning meetings.

B. Speaker Localization

In speaker localization, there are three typical approaches: audio-based approach, vision-based approach and audio-vision-based approach.

Audio-based approaches, such as [2], [3], [4], locate speakers using arrays of microphones. Vision-based approach, such as [5], [6], [7], employ dynamic Bayesian networks or Bayesian network models for speaker detection and recognition. There models are used to combine four simple vision sensors: face detection, skin color, skin texture and mouth motion. [11] employs a duration dependent input output Markov model to localize speakers. [12] uses simple image processing techniques to detect face and face features of the speaker, and then employs visual measures of speech activity as well as audio energy to determine if the previously detected user is actually speaking. Audio-vision-based approaches, such as [8], [9], [10], detect speakers both spatially and temporally by employing vision and audio

978-0-7695-3890-7/09 $26.00 © 2009 IEEE
DOI 10.1109/ISM.2009.66

signal together.

In multi-modal analysis of planning meetings, it is required that speakers locate with spatial and temporal information, so that the speakers' other communication behavior can be analyzed in both spaces. In this paper, we present our visual audio-based techniques to perform speaker localization in our meeting room. Three kinds of signals including audio, hand gesture, and mouth movement are employed in our approach. We apply a *time delay neural network* (TDNN) to fuse these signals and detect the talking person both spatially and temporally. The spatial information of the talking person is determined by his mouth position.

C. Organization of the Paper

In Section II, we introduce the work flow of our approach. The extraction of visual and audio features is described in Section III. An architecture of TDNN used in this paper is given in Section IV. An implementation of the approach in multimodal meeting analysis environments is discussed in Section V, followed by a summary of the paper in Section VI.

II. Summary of Our Approach

Figure 1 shows the summary of our approach. First, we extract features from video and audio signals. For hand gestures, we apply motion efforts to express hand movements. Hand motion efforts are computed from hand motion trajectories extracted from video. We detect mouth movements by normalized cross-correlation in mouth area. We use correlation coefficients to characterize the mouth movements. We compute mel-frequency cepstral coefficients as features of audio signals. Second, we analyze correlation relationships of hand motion efforts of hand movements, correlation coefficients of mouth movements, and features of audio signals. Next, we create a TDNN and use sample data to train the network to learn the correlation relationships between video and audio signals while people are speaking. Finally, we apply the TDNN to perform speaker localization. Results include talking persons and positions of mouths.

III. Extraction of Visual and Audio Features

A. Features of Hand Movements

In order to obtain features of hand movements, we extract hand motion trajectories using a parallel algorithm called Vector Coherence Mapping (VCM)[13]. The VCM algorithm is able to exploit spatial and momentum coherence and color constraints using a fuzzy image integration approach. The parallel nature of the algorithm and its robustness to motion blur and noise contribute to its effectiveness in gestural motion tracking. The algorithm has been applied to extract hand motion out of very long video sequences, some in excess of 74,000 frames of video. VCM tracks a large number of vectors (typically 600 to 2000 per frame) and integrates the fields. This averaging effect gives a smooth

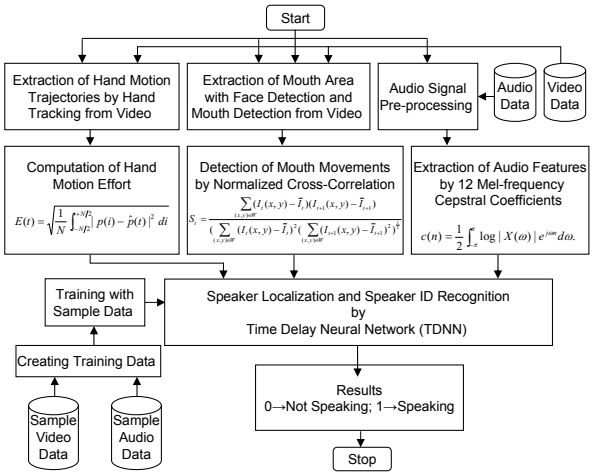


Figure 1. Approach of speaker localization

motion field that is temporally accurate (i.e. no oversmoothing across frames to degrade temporal resolution). These vectors are clustered to identify the moving hands. Clustering errors are fixed manually using an editing program that permits the user to initialize the vector clustering system when clustering errors occur. For the dataset captured with a monocular camera, we can obtain two dimensional (x and y) hand motion trajectories. If the dataset is captured in synchronized stereo with calibrated cameras, we can obtain three dimensional (x, y, and z) motion trajectories.

We apply"effort" to measure hand movements [14]. The effort is roughly analogous to the kinetic energy of the hand movement. The kinetic energy can be computed by calculating the instantaneous velocity of the hand. Since it involves taking the first derivative of the hand trajectory, this method amplifies noise. Here, we compute the energy of motion using a sliding window.

If $p(i)$ is the position of hand at time i, the RMS energy $E(t)$ of window of width N at time t can be computed as:

$$E(t) = \sqrt{\frac{1}{N}\int_{t-N/2}^{t+N/2} |p(i) - \hat{p}(t)|^2 di} \qquad (1)$$

where $\hat{p}(t)$ is the average hand position in the window:

$$\hat{p}(t) = \frac{1}{N}\int_{t-N/2}^{t+N/2} p(i) di \qquad (2)$$

In practice the energy (effort) is computed for discretized time (video frames), so the integrals in the above formulas are substituted with summations and N specifies the width of the window in frames. Figure 2 shows an example of hand motion trajectories and efforts. In the figure, the top shows right hand trajectories in x and y directions and the bottom shows the corresponding motion efforts.

B. Features of Mouth Movements

In order to obtain features of mouth movements, we need to extract mouth positions for all participants in meetings.

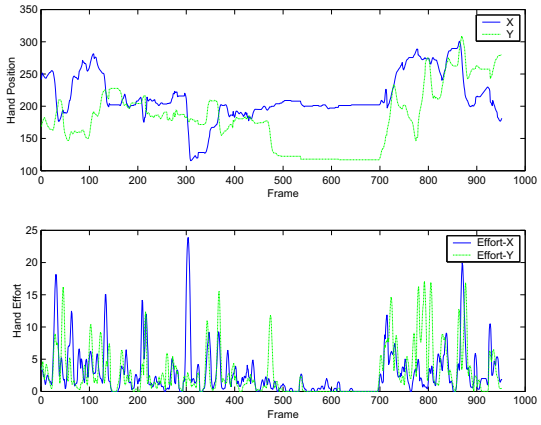


Figure 2. Hand motion trajectories and efforts

We build a skin color model and a skin color filter for each participant. Face regions can be segmented by the filters.

The skin color model theory is established by Yang and other researchers [15]. Xiong and Quek [16] apply skin color techniques to detect faces and build 3D texture map model for head tracking. A survey on skin color detection techniques is available in [17]. The skin color model theory tells us that under certain lighting conditions, a skin color distribution can be characterized by a multivariate Gaussian distribution in the normalized color space.

We build the face skin color model in RGB space. In order to reduce lighting effects, we convert original color images to chromatic color images. Suppose $\mathrm{x}(R,G,B)$ and $\mathrm{x}'(R_n,G_n,B_n)$ are pixels in the original color image and chromatic color image respectively.

$$R_n = \frac{R}{R+G+B}, B_n = \frac{B}{R+G+B}, G_n = \frac{G}{R+G+B} \tag{3}$$

In above, as $R_n + B_n + G_n = 1$, there are only two independent components, so we omit the third component. For each pixel, we have a color vector $\mathrm{x} = (R_n \ B_n)^T$. The two dimensional Gaussian distribution model is expressed as $N(\mu, \Sigma)$ i.e.

$$p(\mathrm{x}) = \frac{1}{2\pi|\Sigma|^{1/2}\exp[-\frac{1}{2}(\mathrm{x}-\mu)^T\Sigma^{-1}(\mathrm{x}-\mu)]} \tag{4}$$

with

$$\begin{cases} \mu = E\{\mathrm{x}\} \\ \Sigma = E\{(\mathrm{x}-\mu)(\mathrm{x}-\mu)^T\} \end{cases} \tag{5}$$

where,

μ is the mean vector;
Σ is the covariance matrix.

With skin color samples, we apply Maximum Likelihood Estimation approach to estimate these parameters $(\hat{\mu}, \hat{\Sigma})$.

Based on the skin color model, we create a skin color filter for each participant with a threshold. With the skin color filter, face regions can be segmented. We create a mouth template to detect mouth positions in the face regions from each frame of video. The mouth positions are also applied to express the positions of speakers.

After obtaining mouth positions, we can define mouth areas with a window (a rectangle) and detect mouth (lip) movements. Since we only need to know whether the mouth has motion and we do not need to know how much motion. We use a normalized cross-correlation to measure changes of the mouth area between two frames. Suppose F_t and F_{t+1} are two frame images at time t and $t+1$. The normalized cross-correlation coefficient C_t can be computed as below.

$$C_t = \frac{\sum_{(x,y)\in W}(F_t(x,y)-\overline{F}_t)(F_{t+1}(x,y)-\overline{F}_{t+1})}{(\sum_{(x,y)\in W}(F_t(x,y)-\overline{F}_t)^2\sum_{(x,y)\in W}(F_{t+1}(x,y)-\overline{F}_{t+1})^2)^{\frac{1}{2}}} \tag{6}$$

where W is a windowing function in F_t.

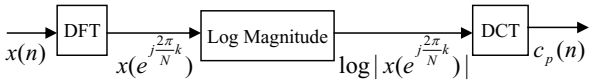
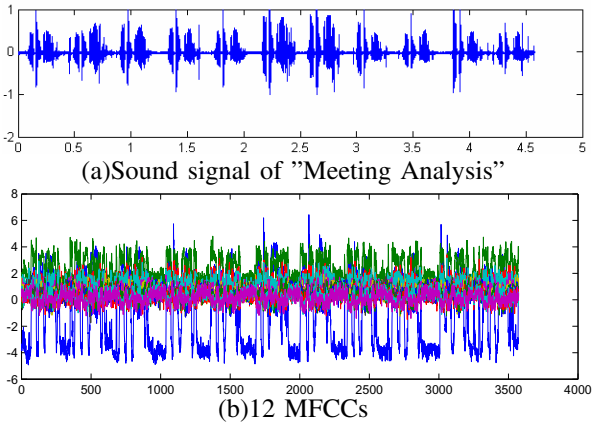


Figure 3. A procedure of computation of MFCCs



(a)Sound signal of "Meeting Analysis"

(b)12 MFCCs

Figure 4. An example of an audio signal and its 12 MFCCs

C. Audio Features

Mel Frequency Cepstrum Coefficients (MFCCs) of an audio signal are used as audio features, which are commonly used in speech recognition systems [18]. Figure 3 shows a procedure of computation of MFCCs. First, divide audio signals into frames. Second, obtain the amplitude spectrum for each frame using Discrete Fourier Transform (DFT); and then, take the logarithm and compute cepstral coefficients,

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} log|X(e^{j\omega})|e^{j\omega n}d\omega, \tag{7}$$

where, $|X(\cdot)|$ is the power spectrum of the audio signal $x(n)$. Finally, we perform mel-frequency warping and transform them into MFCCs by DCT. We compute 12 MFCCs in a window as features of audio signals.

Figure 4 (a) shows an audio signal for a person saying "meeting analysis" for 11 times and Figure 4 (b) shows its

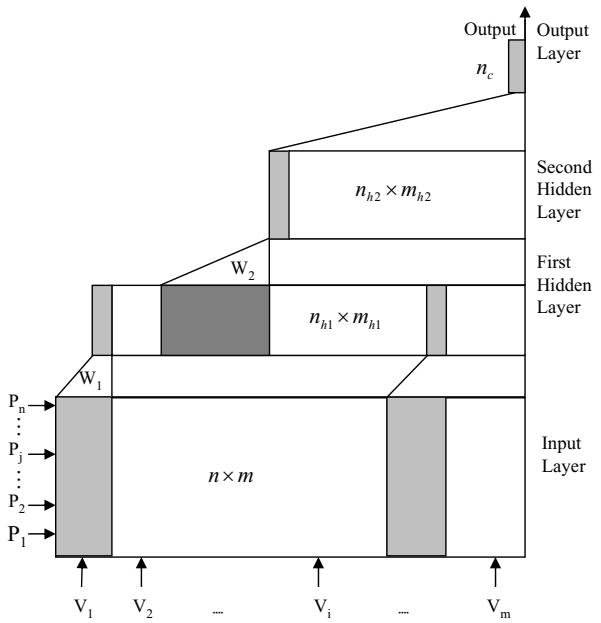


Figure 5. Architecture of TDNN in our approach

12 MFCCs. Of course, if there is no any speech signal, the MFCCs should be zero. Therefore, we can use MFCCs to determine whether we have speakers at this moment. Feather more, different people have different characteristics of MFCCs. We can apply this property to identify speakers.

IV. Architecture of TDNN in Our Approach

Figure 5 shows an architecture of TDNN used in our approach. This is a four layer neural network including an input layer, two hidden layers, and an output layer. The input layer has $n \times m$ neurons. The n is determined by the number of features and the m is determined by the segmentation of video (i.e. a segment of video has m frames). The time delay between the first hidden layer and the input layer is W_1. There are two hidden layers. There are $n_{h1} \times m_{h1}$ neurons in the first hidden layer and $n_{h2} \times m_{h2}$ neurons in the second hidden layer. The time delay between the two hidden layers is W_2. The output layer has one neuron for the state of the subject. The result 1 indicates that the subject is speaking, 0 not speaking. We adopt the Standard Back Propagation (SBP) algorithm as the learning method to train this TDNN.

V. Example Applications

A. Experimental Setup and Meeting Room Configuration

Figure 6 shows the configuration of our meeting room. There are five participants labeled C, D, E, F, G in the meeting. Ten movie cameras labeled C1, C2, C3, C4, C5, C6, C7, C8, C9, C10 are installed to record the meeting events. T1 and T2 are two table microphones to record speech. Each camera is installed in a fixed position on the ceiling of the meeting room, so that each camera can see certain

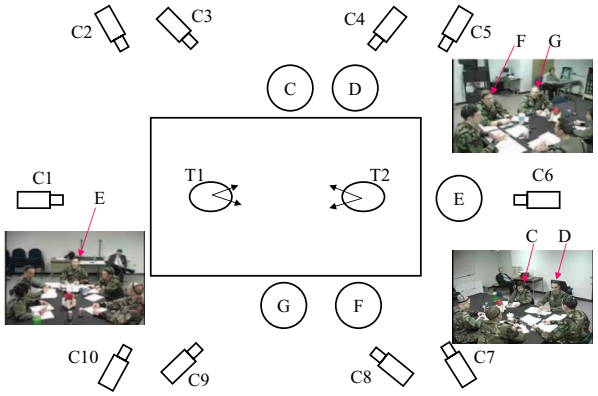


Figure 6. Meeting room configuration

participants at the same time. For speaker localization, we use three videos from cameras C1, C5, and C7 respectively. We let camera C1 see E, C5 for F, G, and C7 for C, D, so that we can see participants' front faces. We locate the participants' mouths as their positions. In the meeting room, we also installed a Vicon motion capture system (we did not show the system in Figure 6) to provide us ground truth data. Eight Vicon infra red cameras are installed in fixed positions on the celing of the meeting room. These infra red cameras can track Vicon markers mounted on targets to provide 3D motion data.

B. Experimental Work Flow

As we mentioned in Section V-A, there are five participants in the meeting. We need to detect the talking person both spatially and temporally during the meeting. First, we perform face detection to determine face positions for all participants in the meeting room by a head tracking approach [16]which is developed for meeting analysis. The participants' face positions can be used for the detection of their mouths. Second, we compute three kinds of features including hand efforts with hand movements, cross-correlation coefficients with mouth movements, and MFCCs with audio signals for each participant. All participants share the audio signals recorded by table microphones T1 and T2. Next, we create a TDNN for each participant. In our case, we create five TDNNs for the five participants in the meeting room. Before we apply these networks to detect the speaking person, we need to create sample data to train them and let them learn the complicated relationships of video and audio signals while people are talking. Finally, We input the three kinds of features into the TDNNs. The networks can decide which one is a talking person at a given time. While we compute mouth movements for each participant, we detect the mouth position which can also be used as the location of the talking person.

C. Experimental Results and Analysis

We applied our approach to the experimental dataset AFIT010705. This dataset comprised 74,000 frames video

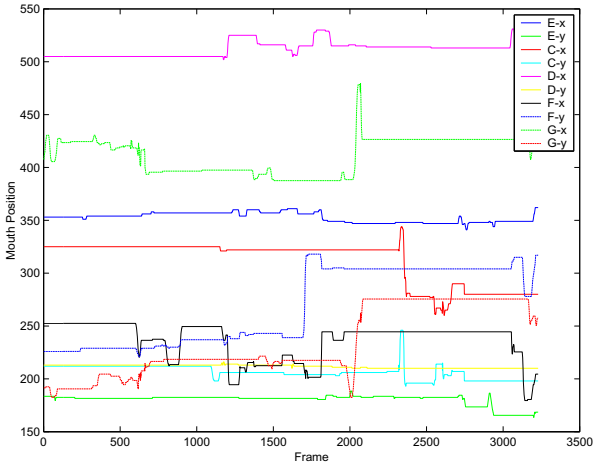


Figure 7. Mouth positions of dataset AFIT010705

(41.11 minutes) and audio. The topic of this meeting is Foreign Material Exploitation.

We have processed the whole dataset. Since it is too long, we choose a focus section which there are a lot of talking, gestures, and mouth movements to display the results. This section comprised 3229 frames.

First, we extract audio features by computing 12 MFCCs using a 10 ms window with the audio signals shown in Figure 8 (a). The results are shown in Figure 8 (b). Second, we segment face regions with skin color models and filters and detect mouth positions for all participants. Figures 8 (c), (d), (e), (f), and (g) show the results. Next, we extract hand motion trajectories for all participants with the VCM algorithm and compute hand motion efforts which are shown in Figures

After training, we input these three kinds of features into the TDNN to detect who is speaking. Figure 9 shows the results. In the figure, "1" expresses "speaking" and "0" for "not speaking".

From the results we can see that subject E talked a lot because he is a leader of this meeting. Subject G has a lot of talking at the beginning of this section and subject C has a lot of talking near the end of the section. Subject D did not say anything in this section. He just listens to other's talking and gives eye gazes to other subjects.

We compared the results with ground truth in the Final Cut Pro system. The results are satisfying. We obtain some errors which happened when the mouths are occluded by hands and the heads turned away from the cameras so that we can not see the mouths. Low resolution is another reason for the errors.

VI. CONCLUSIONS AND DISCUSSION

Speaker localization is very important for the detection of meeting events in multimodal analysis of planning meetings. It enables us to focus on the analysis of speaker's communication behavior during the meeting and detect meaningful

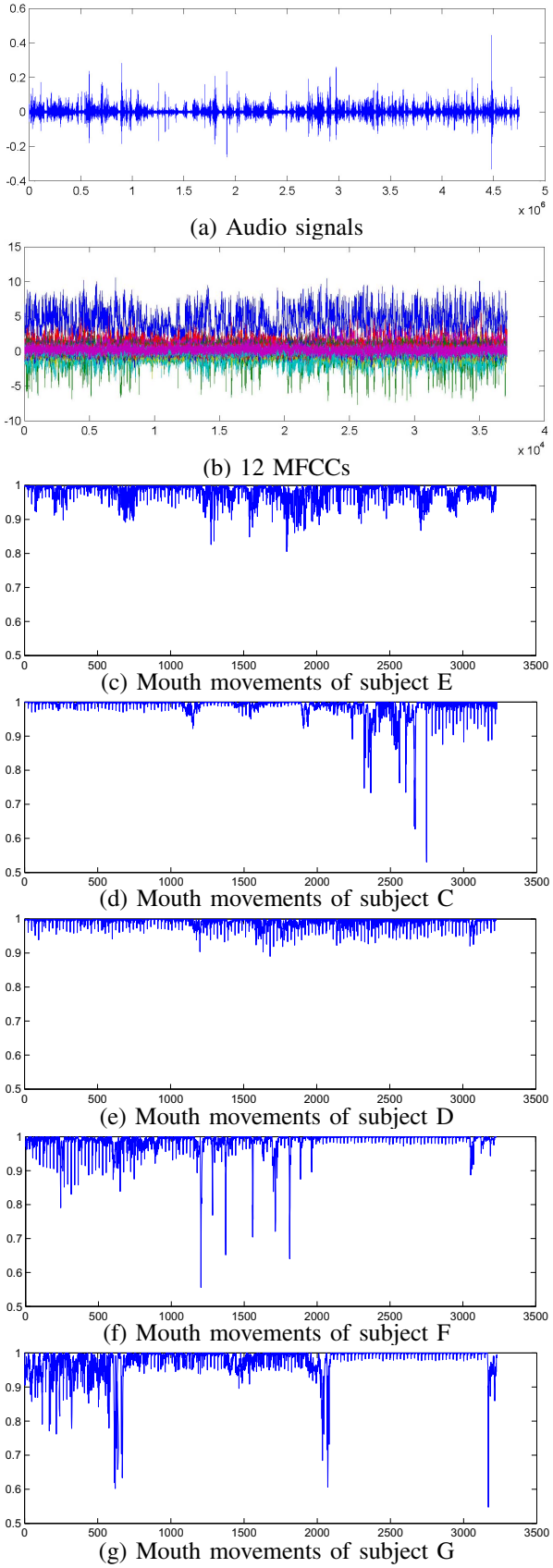


Figure 8. Results for dataset AFIT010705

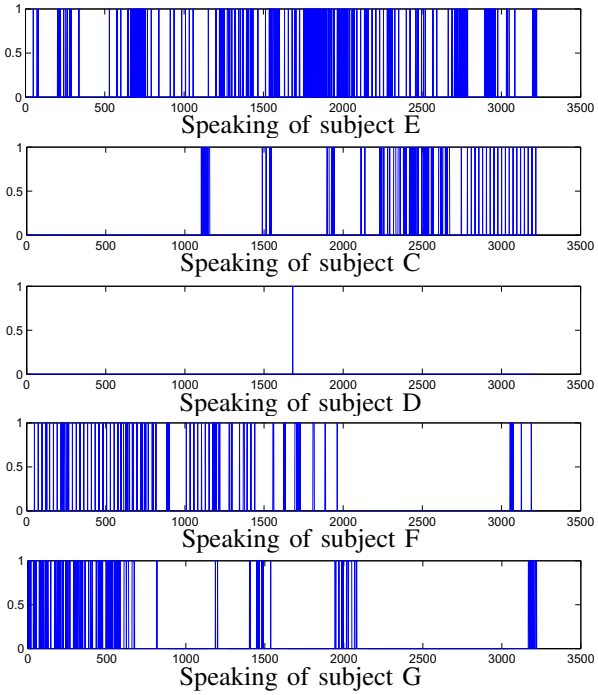


Figure 9. Speaking for all subjects of dataset AFIT010705

information units of the meeting. By following the speakers, we can easily understand the contents of the meeting. We presented our techniques of speaker localization in multimodal analysis of planning meetings. The basic idea of the approach is that speech accompanies with mouth movements and sometimes hand gestures. In our approach, we create a TDNN for each participant in the meeting to decide whether he/she is speaking or not. Three kinds of features are used for speaker localization.

REFERENCES

[1] Yingen Xiong and Francis Quek, "Meeting room configuration and multiple camera calibration in meeting analysis," in *ACM Seventh International Conference on Multimodal Interfaces, ICMI2005, October 04-06*, Trento, Italy, 2005.

[2] Andr G. Adami, Sachin S. Kajarekar, and Hynek Hermansky, "A new speaker change detection method for two-speaker segmentation," in *ICASSP02*, Orlando, Florida, May 13 - 17 2002, vol. 4, pp. 3908–3911.

[3] Frederick Weber, Linda Manganaro, Barbara Peskin, and Elizabeth Shriberg, "Using prosodic and lexical information for speaker identification," in *ICASSP2002*, Orlando, Florida, May 13 - 17 2002, vol. 1, pp. 141–144.

[4] Himanshu Vajaria, Tanmoy Islam, Sudeep Sarkar, Ravi Sankar, and Ranga Kasturi, "Audio segmentation and speaker localization in meeting videos," in *ICPR '06: Proceedings of the 18th International Conference on Pattern Recognition*, Washington, DC, USA, 2006, pp. 1150–1153, IEEE Computer Society.

[5] Jim Rehg Ashutosh Garg, Vladimir Pavlovic and Thomas S. Huang, "Speaker detection using input/output hidden markov model," in *3rd international conference on Multimodal Interfaces, (ICMI 2000)*, Berlin: Springer, 2000.

[6] James M. Rehg, Kevin P. Murphy, and Paul W. Fieguth, "Vision-based speaker detection using bayesian networks," in *CVPR99*, Ft. Collins, June 1999, pp. 110–116.

[7] Gerald Friedland, Hayley Hung, and Chuohao Yeo, "Multimodal speaker diarization of real-world meetings using compressed-domain video features," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 0, pp. 4069–4072, 2009.

[8] Ross Cutler and Larry Davis, "Look who's talking: Speaker detection using video and audio correlation," in *IEEE International Conference on Multimedia and Expo (ICME)*, Manhattan, New York, July 2000, vol. 3, pp. 1589–1592.

[9] Tanzeem Choudhury, James M. Rehg, Vladimir Pavlovi, , and Alex Pentland, "Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection," in *IEEE 16 th International Conference on Pattern Recognition (ICPR'02)*, Quebec City, QC, Canada, August 11 - 15 2002, vol. 3, pp. 30789–30794.

[10] Athanasios Noulas and Ben J. A. Krose, "On-line multimodal speaker diarization," in *ICMI '07: Proceedings of the 9th international conference on Multimodal interfaces*, New York, NY, USA, 2007, pp. 350–357, ACM.

[11] Ashutosh Garg Milind Napahade and Thomas S. Huang, "Duration dependent input output markov models for audio-visual event detection," in *international conference on Multimedia and Expo, (ICME 2001)*, Tokyo, Japan, August 22-25 2001.

[12] P.deCuetos, C.Neti, and A.Senior, "Audio-visual intent-to-speak detection for human-computer interaction," in *ICASSP2001*, Beijing, China, May 2001.

[13] F. Quek, X. Ma, and R. Bryll, "A parallel algorithm for dynamic gesture tracking," in *ICCV'99 Wksp on RATFG-RTS.*, Corfu, Greece, Sept.26–27 1999, pp. 119–126.

[14] R. Bryll, F. Quek, and A. Esposito, "Automatic hand hold detection in natural conversation," in *the IEEE Workshop on Cues in Communication*, Kauai, Hawaii, December 2001.

[15] J. Yang and A. Waibel, "A real-time face tracker," in *Proceedings of the Third Workshop on Applications of Workshop on Computer Vision (WACV'96)*, Sarasota, Florida, 1996.

[16] Yingen Xiong and Francis Quek, "Head tracking with 3d texture map model in planning meeting analysis," in *International Workshop on Multimodal Multiparty Meeting Processing (ICMI-MMMP'05)*, Trento, Italy, October, 07 2005.

[17] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *Proc. Graphicon-2003*, Moscow, Russia, September 2003, pp. 85–92.

[18] Ben Gold and Nelson Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*, John Wiley and Sons, Inc., 1999.