TOWARDS A MULTIMODAL MEETING RECORD

Ralph Gross, Michael Bett, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, Alex Waibel

Interactive Systems Laboratories
Carnegie Mellon University
Pittsburgh, PA 15213, USA

ABSTRACT

Face-to-face meetings usually encompass several modalities including speech, gesture, handwriting, and person identification. Recognition and integration of each of these modalities is important to create an accurate record of a meeting. However, each of these modalities presents recognition difficulties. Speech recognition must be speaker and domain independent, have low word error rates, and be close to real time to be useful. Gesture and handwriting recognition must be writer independent and support a wide variety of writing styles. Person identification has difficulty with segmentation in a crowded room. Furthermore, in order to produce the record automatically, we have to solve the assignment problem (who is saying what), which involves people identification and speech recognition. This paper will examine a multimodal meeting room system under development at Carnegie Mellon University that enables us to track, capture and integrate the important aspects of a meeting from people identification to meeting transcription. Once a multimedia meeting record is created, it can be archived for later retrieval.

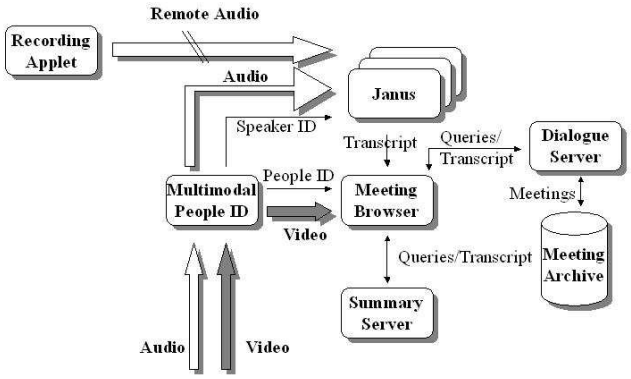

Figure 1: System overview of the multimodal meeting room

1. INTRODUCTION

The meeting room which we are developing at Carnegie Mellon University's Interactive Systems Laboratories is comprised of three separate components: a multimodal people identifier, a speech recognizer, and a meeting browser. It requires only a minimal amount of manual input. The system is able to automatically identify up to six distinct speakers in a meeting and automatically creates a transcript for read or conversational speech. Eventually the system will identify when a meeting begins and automatically start creating a complete meeting record. Once a meeting record is complete, the meeting browser allows us to archive and review previously recorded meetings.
In designing the system, we are striving to make:

- The interface as natural as possible

  Use of multiple modalities (speech, handwriting, and vision) when appropriate.

  All tasks automated as much as possible in order to minimize the burden on the user.

2. SYSTEM OVERVIEW

In the existing system architecture shown in Figure 1, audio and video streams are fed into the multimodal people id system. The identification is sent to both our Janus speech recognizer and to the meeting browser. Hypotheses flow from Janus to the meeting browser appearing as a transcript in this window. Transcripts are then summarized by the summary server or archived via the dialogue server.

The speech recognizer used in the multimodal meeting room is based on the Janus Switchboard recognizer trained for the 1997 NIST Hub-5E evaluation . The recognizer is described in detail in [14]. We have extended the system to support multiple recognizers simultaneously in order to improve on the speed of the speech recognition.

The people identification module continuously tracks and identifies meeting participants. We currently combine the outputs of three subsystems, namely speaker identification, sound source position estimation, and color appearance identification in a multimodal fusion framework. Using the results we are able to answer the question of "Who said what?" during a meeting. Our experiments show that the performance of the combined multimodal system is superior to the performance of the individual recognizers.

The meeting browser interface records meetings and displays meeting transcriptions, time-aligned to the corresponding audio and video files. Included in the meeting transcriptions are discourse features and emotions. The user can select all or a portion of these files for playback; text highlighting occurs in sync with the sound and video playback. These meetings are fully editable and searchable, allowing humans to annotate and correct recognition output as well as adding new informative streams manually. Once a meeting is complete, the meeting room automatically archives

the meeting for future use. Users are able to query this archive to create audio, video, and text dialogue summaries of the meetings which can then be mailed to other individuals for playback and review. The next three sections describe each of these modules in more detail.

3. MEETING ROOM SPEECH RECOGNITION

Meeting recognition is a challenging large vocabulary conversational speech recognition task parallel to Hub5 (Switchboard) [2] and Hub4 (Broadcast News) [1]. The difficulty mostly comes from the highly conversational style of meetings, and lack of training data. Since we are dealing with uninterrupted continuous recordings with multiple speakers (possibly using multiple microphones), our task requires three steps. First, we carefully partition the data into homogeneous segments and assign a "speaker" label to each segment. Second, we perform a first pass recognition which generates both a hypothesis and a confidence score. Finally, we do some unsupervised adaptation, and re-decode the utterances with the adapted model [5, 6].

Unlike many typical speech recognition tasks, there is not enough data available to train a domain-specific recognizer for the meeting recognition task. We experimented with several systems that we developed at the Interactive Systems Laboratories for different tasks. The word error rate (WER) on group meeting data (internal research meeting recorded with lapel microphones) is in the 40% range.

Each of the systems was built upon the Janus Recognition Toolkit (JRTk), which is summarized in [7]. Incorporated into our continuous HMM (Hidden Markov Model) system are techniques like linear discriminant analysis (LDA) for feature space dimension reduction, vocal tract length normalization (VTLN) for speaker normalization, cepstral mean normalization (CMN) for channel normalization, and wide-context phone modeling (Polyphone modeling) See [10] for a technical description of each of these.

Recently, we are leveraging the large amount of data in the Broadcast News (BN) domain to build a robust BN recognizer. The BN data includes a wide range of background conditions (clean / noise / music), planned / spontaneous speech, field speech /telephone interview, etc. We succesfully used the BN system to recognize discussion-style TV news shows (Newshour and Crossfire) which are similar to meetings. The results on this data can be found in Table 1.

Show Type	WER (1st pass)	WER (after adaptation)
Newshour	26.9	26.3
Crossfire	36.0	34.6

Table 1: Word error rates (WER) in percent on Newshour and Crossfire shows

4. MULTIMODAL PEOPLE IDENTIFICATION

The purpose of the people identification module is to continuously track and identify meeting participants within a room. In order to increase the robustness and efficiency of the identification process we have taken a multimodal approach and integrated a number of recognizers that use audio and video information. As shown in Figure 2, the system is comprised of five components: people segmentation, color appearance ID, speaker ID, face ID and multimodal information fusion. The face identification module is not currently incorporated in the system. The following paragraphs describe the modules in more detail (see [15] for a more technical description).

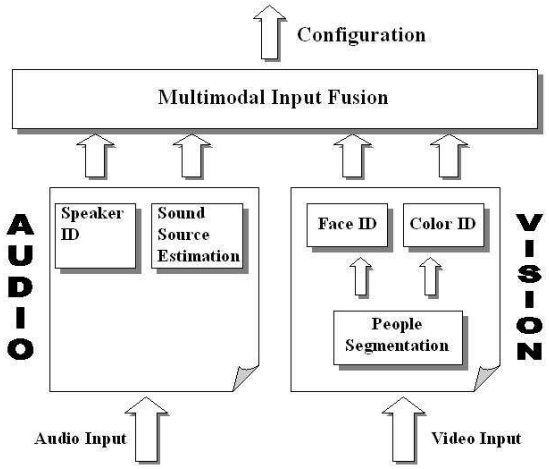

Figure 2: Overview over the multimodal people identification

4.1. People Segmentation

The ability to identify an object in a given image or image sequence requires the availability of an internal representation of said object. Assuming that such a model is given, it could be utilized to locate and identify objects in one unified step. Unfortunately the search space that the recognizer would have to tackle in each run is too large to meet the real time requirements of an interactive system. We therefore use a motion-based preprocessing step to segment people from the background before we try to identify them. Our approach uses four different stages, namely background subtraction, noise removal, region growing, and background update.

4.2. Color Appearance Identification

Based on the segmentation derived by the people segmentation module, we create models for the different meeting participants using color histograms. As noted by other researchers before [11] color histograms provide a stable object representation, which is largely unaffected by occlusions or changes in view. A major obstacle in the use of color for object identification is the fact that colors change with illumination. In order to reduce this sensitivity of the color models we use a perceptually motivated color encoding scheme, the so-called tint-saturation color space [12].

4.3. Speaker Identification and Sound Source Position

The speaker ID module has to solve the problem of finding out which meeting participant is speaking at any given time, independent of what they are saying. This can be seen as a text-independent close-set speaker identification task. We consider both convolution and additive noise as consistent, except for occasional events - phone ringing, door clapping etc. The limited training and

test sets are collected in the same noise environment [4]. Our experiments show that if training and testing are done on the same noise conditions, the performance is comparable with the performance achieved on clean speech. The major challenge in this task is how to achieve high performance in real-time with a relatively small amount of training data. The results of our system are shown in Table 2.

Recording	Test length 3 sec	6 sec
Clear	97.8%	100.0%
Noisy	96.6%	100.0%

Table 2: Identification performance on 30 speakers

In order to combine audio and visual information we need an estimation of the sound source position. In our initial system setup this estimation is based on a model of the speech energy pair obtained from two microphones. As for the speaker identification module we use Gaussian Mixture Models for this task.

4.4. Face Identification

While people identification based on color appearance works reasonably well in most situations, it fails when meeting participants are dressed similarly. To overcome this problem we are developing face identification as part of the system. Among the numerous face recognition algorithms introduced in recent years, the eigenface approach proposed by Turk and Pentland [13] is one of the most influential ones. It uses Principal Components Analysis (PCA) to linearly project the high dimensional image space to a lower dimensional feature space. Recognition is performed by computing the Euclidean distances between the test image and the reference images in feature space. While the eigenface approach performs well in mug-shot settings, it has difficulties handling occlusions. We developed a new algorithm, called Dynamic Space Warping (DSW), which specifically addresses the problem of occlusion as it typically occurs during meetings [8]. The algorithm converts the input image into a sequence of subimages using a moving window. The subimages are then projected onto a sequence of points in the lower dimensional feature space. During the recognition process, the template set of points is compared to the unknown set of points in a procedure similar to dynamic time warping (DTW) used in speech recognition [10].

Tested on manually labeled data recorded in the meeting room, DSW outperforms the eigenface approach as shown in Table 3.

#Train images	3	5	7	10
PCA w/o occl	76.2%	79.9%	80.0%	82.7%
DSW w/o occl	82.0%	86.8%	86.5%	89.4%
PCA w occl	25.3%	31.6%	29.0%	30.8%
DSW w occl	45.5%	49.9%	47.5%	48.6%

Table 3: Comparison of the recognition rates for PCA and DSW on databases with and without occlusion

4.5. Multimodal Input Fusion

During input fusion we try to find the most probable configuration of people locations, identities in the room and assignment of a speaker. Assuming conditional independence of the input signals from the color appearance identification, speaker identification and sound source position estimation, we can directly combine the probabilities estimated by these modules (see [15] for details). To demonstrate the feasibility of the framework, we set up a simple meeting. In our experiment, we collected 2990 audio and video inputs. For both inputs, we found the optimal configuration with information fusion. We also computed the optimal configuration without fusion, i.e. using the models individually. In this experiment, the configuration error rate drops by 2% absolute after information fusion (from 12.51% to 10.67%).

5. MEETING BROWSER

An important part of meeting recognition is the ability to efficiently capture, manipulate and review all aspects of a meeting. To that end we have developed a meeting browser that lets users:

- Create meeting records and transcriptions of meetings with participants disjointly located.

  Create and customize dialogue, audio, and video summaries to the user's particular needs.

  Create a database of corporate knowledge.

  Quickly and accurately create and disseminate a list of conclusions and action items

  Provide rapid access to meeting records to allow browsing and reviewing existing meetings.

  Identify for each utterance the speaker properties (type, social relationships, and emotion) as well as the discourse structure and type.

Our meeting browser, shown in Figure 3, is written in Java. It is a powerful tool that allows us to create, review or summarize a meeting or search a set of existing meetings for a particular speaker, topic, or idea.

5.1. Meeting Creation

When a meeting is being created, each participant may join either remotely or locally. Participants joining locally start the people id system described in Section 4. Once the meeting has begun, speech along with the speaker id derived from people id flows to Janus, our speech recognition engine. As the speech is recognized, the hypothesis is sent to the dialogue server where it is assembled into a meeting format. The meeting browser displays the transcript for the current meeting. The meeting transcript can be sent to the summarization server which will create a summary of the current dialogue. Finally, a user may elect to save a meeting including any summaries in the meeting archive from within the meeting browser.

At the end of meetings, it is customary to reiterate a set of action items. Using speech recognition, we recognize the items and mail them out to each of the meeting participants. Likewise, we can mail complete meetings, meeting segments, or summaries including the audio portion directly from within the meeting browser to meeting participants or any other interested parties. Each of

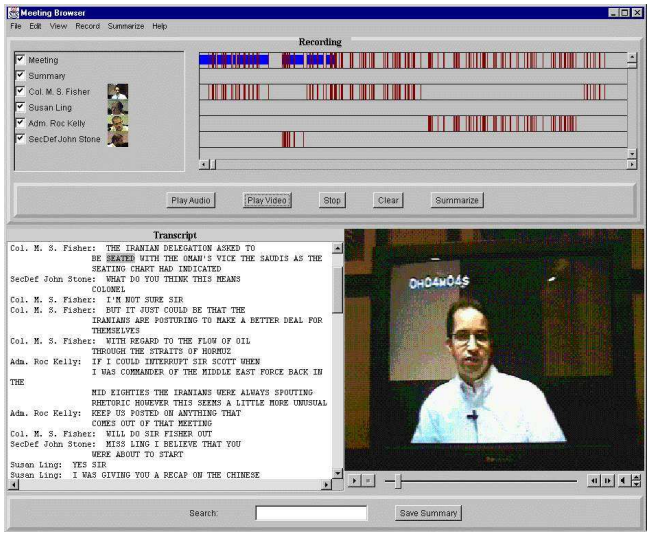

Figure 3: The meeting browser main window consists of three sections, an upper graphical display which shows the meeting over time, a lower left window that shows a meeting transcript, and a lower right window which displays either a video of the current meeting or a dialogue summary.

these may include annotations, comments or corrections. Corrections can be done by using a keyboard or handwriting recognition using a handwriting recognizer developed in our lab [9]. In the future we plan to add speech recognition as an additional error repair modality.

5.2. Summary Server

The meeting browser has the capability to create audio, video, and text summaries. In each case, a summary is created on the basis of the recognized text dialogue, then the appropriate portions of the audio or video are clipped in order to create a summary. A user specifies the summary size as well as the central topic (if any) of the summary. This information along with the meeting dialogue is sent to a summary server which runs remotely. The server analyzes the dialogue and returns a summary to the meeting browser. In this way we can create summaries that allow the user to drill down from a general summary to a very specific topic or area of interest. The algorithm for the summarization techniques is based on the MMR (Maximal Marginal Relevance) [3]. This is a uniqueness measure that ranks the turns in the dialogue by topic and includes only turns for which topics have not previously been included. The summary server identifies the set of topics and returns a marked dialogue to the meeting browser. The summary server eliminates redundant turns from the dialogue without loss of meaning. See [14] for a technical description of the algorithm.

5.3. Meeting Archive

An important part of meeting tracking is the creation of corporate knowledge that is archived and available for later reference. The meeting archive presents meetings in a tree format. It allows individuals to search for meetings based on any combination of participants, topics discussed, keywords, meeting length, and meeting date. In addition, if there is a summary for a meeting, the user can review it without loading the entire meeting in the meeting browser. The summary can be topic based, turn based or both.

6. FUTURE WORK

We need to continue to improve our meeting recognition results and to experiment with a variety of microphones in multiple settings. Furthermore we plan to employ more video cameras in order to improve the person tracking and identification.

7. ACKNOWLEDGEMENTS

We would like to thank Robert Malkin, Klaus Ries, Thomas Polzin, and Klaus Zechner for their support. We would also like to thank our sponsors at DARPA. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, or any other party.

8. REFERENCES

[1] *Proceedings of ARPA Speech and Natural Language Workshop*. Morgan Kaufman Publishers, 1995.

[2] *Proceedings of LVCSR Workshop*. Maritime Institute of Technology, October 1996.

[3] J. Carbonell, G. Geng, and J. Goldstein. Automated query-Relevant Summarization and Diversity-Based Reranking. In *IJCAI-97 Workshop on AI and Digital Libraries*, 1997.

[4] F. Bimbot et al. Speaker Verification in the Telephone Network: Research Activities in the Cave Project. Technical report, PTT Telecom, ENST, IDIAP, KTH, KUN, and Ubilab, 1997.

[5] Hua Yu et. al. Experiments in automatic meeting transcription using jrtk. In *Proc. ICASSP '98*, 1998.

[6] Hua Yu et. al. Progress in Automatic Meeting Transcription. In *Eurospeech '99*, 1999.

[7] M. Finke, J. Fritsch, P. Geutner, K. Ries, T. Zeppenfeld, and A. Waibel. The Janusrtk Switchboard/Callhome 1997 Evaluation System. In *Proceedings of LVCSR Hub 5-e Workshop*, 1997.

[8] R. Gross, J. Yang, and A. Waibel. Face recognition in a Meeting Room. In *Proceedings of the Fourth International Conference on Automatic Face and Gesture Recognition*, 2000.

[9] S. Manke, M. Finke, and A. Waibel. NPen++: A Writer Independent, Large Vocabulary On-line Cursive Handwriting Recognition System. In *ICDAR*, 1995.

[10] L.R. Rabiner and B.-H. Juang. *Fundamentals Of Speech Recognition*. Prentice Hall, 1993.

[11] M.J. Swain and D.H. Ballard. Color Indexing. In *International Journal of Computer Vision*, volume 7, 1991.

[12] J.-C. Terrillon, M. David, and S. Akamatsu. Automatic Detection of Human Faces in Natural Scene Images by use of a Skin Color Model and of Invariant Moments. In *Proceedings of the Third International Conference On Automatic Face And Gesture Recognition*, 1998.

[13] M. Turk and A. Pentland. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, 3(1), 1991.

[14] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen. Meeting Browser: Tracking and Summarizing Meetings. In *Broadcast News Transcription and Understanding Workshop*, Landsdowne, Virginia, 1998.

[15] J. Yang, X. Zhu, R. Gross, J. Kominek, Y. Pan, and A. Waibel. Multimodal People ID for a Multimedia Meeting Browser. In *Proceedings of ACM Multimedia*, 1999.