THE ISL MEETING ROOM SYSTEM

Tanja Schultz[1,2] *Alex Waibel*[1,2] *Michael Bett*[1] *Florian Metze*[2] *Yue Pan*[1] *Klaus Ries*[1,2]
Thomas Schaaf[2] *Hagen Soltau*[2] *Martin Westphal*[2] *Hua Yu*[1] *Klaus Zechner*[1]
Interactive Systems Labs [1]Carnegie Mellon University and [2]Universität Karlsruhe

ABSTRACT

Oral communication is transient but many important decisions, social contracts and fact findings are first carried out in an oral setup, documented in written form and later retrieved. At Carnegie Mellons University's Interactive Systems Laboratories we have been experimenting with the documentation of meetings. This paper summarizes part of the progress that we have made in this test bed, specifically on the question of automatic transcription using LVCSR, information access using non-keyword based methods, summarization and user interfaces. The system is capable to automatically construct a searchable and browsable audiovisual database of meetings and provide access to these records.

1. INTRODUCTION

Humans spend a lot of time transforming oral communications into written documents. This process however is not only expensive and puts burden on the participants, it may also suffer from other fundamental flaws in the meeting scenario: A written record takes **time** to produce; it may loose **accuracy** since the minute preparer may not remember or interprete correctly or is biased; it looses the meetings **originality** and therefore many qualifications such as emotions, hedges, attention and the precise wordings; finally it may loose **completeness** for efficiency reasons and no selective probing for further details is possible. Even if meeting minutes are produced the meeting record can be used to ground the information presented in the meeting or the minutes can be produced by enhancing the record.

The focus of our work, starting with [1], is aiming at a realistic meeting scenario, the corresponding speech recognition problems, the analysis of retrieval performance and addition of non-keyword based features, the generation of readable summaries and a pratical user interface. Not covered in this paper is active work in our group on audio and vision based people identification [2] which is important to understand who attended a meeting, the determination of focus of attention [3] and the detection of emotion [18].

Other important projects on the problem of information access to spoken language are concentrated on the TREC-SDR task [4] which is focussed on the retrieval of broadcast news documents. The participants managed to show that keyword based retrieval can often be done successfully even if there is a significant word error rate by a speech recognizer. [5, 6] are spoken language access systems including a graphical browsing interface that have been evaluated in this domain and focus on named entity tagging, prosodic processing and salient keywords. [7] goes one step further and presents work on video summarization using key-shots and uses geographic and temporal information to add further indices for retrieval. [8] shows that additional features can be retrieved from whiteboards, online collaborative note-taking and slide presentations that help to browse and index lectures. [9] presents segments a dialogue and has a fast playback capability to skim segments.

Fig. 1 shows the components of our system that will be introduced in the next sections. A recording program with a speaker identification module sends the audio files to the speech recognizer (Janus) and meeting browser. The summarization, emotion and discourse module are called with the data they need from the meeting browser front end and send their results back for display. A meeting archive can be accessed.

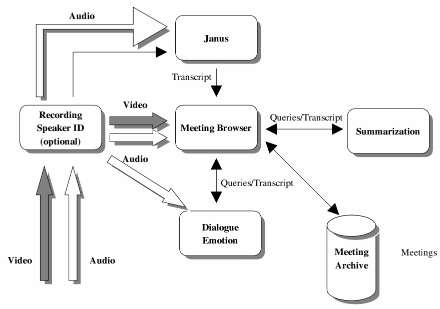

Fig. 1. **Components of the meeting room system**

2. SPEECH RECOGNITION

As already identified in previous works [10] meeting recognition is a very challenging LVCSR task parallel to Hub5 (Switchboard) and Hub4 (Broadcast News). The difficulty results basically from three reasons: First, the conversational style - meetings consists of uninterrupted continuous recordings with multiple speakers talking in a conversational style. Second, the lack of training data - meeting data is highly specialized depending on the topic and participants, therefore large databases can not be provided on demand. As a consequence our research has focused on the question on how to build LVCSR systems for new tasks and languages [11] using limit amounts of training data. Third, the degraded recording conditions: to minimize interference a clip-on lapel microphone was chosen instead of a close-talking headset. However, this comes at the cost of significant channel cross-talk. Quite often one can hear multiple speakers on a single channel.

2.1. Acoustic and Language Model Adaptation

For the purpose of building a speech recognition engine on the meeting task, we combined a limited set of meeting data with English speech and text data from various sources, namely Wall

Street Journal (WSJ), English Spontaneous Scheduling Task (ESST), Broadcast News (BN), Crossfire and Newshour TV news shows. The meeting data consists of a number of internal group meeting recordings (about one hour long each), of which 14 are used for experiments in this paper. A subset of three meetings are chosen as the testset.

To achieve robust performance over a range of different tasks, we trained our baseline system on Broadcast News (BN) using JRTk [12]. The system deploys a pentphone model with 6000 distributions sharing 2000 codebooks. There are about 105k Gaussians in the system. Vocal Tract Length Normalization and cluster-based Cepstral Mean Normalization are used to compensate for speaker and channel variations. Linear Discriminant Analysis is applied to reduce feature dimensionality to 42, followed by a diagonalization transform (Maximum Likelihood Linear Transform). A 40k vocabulary and trigram language model are used. The baseline language model is trained on the Broadcast News (BN) corpus. The error rates on the meeting data are quite high as can be seen Tab. 1 but using acoustic and language model adaptation the error rate can be reduced by about 10.2% relative over the BN baseline system.

Baseline System WER on Different Tasks [%]	
Broadcast News (h4e98_1) F0-condition	9.6
Broadcast News (h4e98_1) all F-conditions	18.5
Newshour	20.8
Crossfire	25.6
Adaptation to Meeting Data	
ESST system	54.1
Baseline Broadcast News system	43.1
+ acoustic MAP Adaptation (10h meeting data)	40.4
+ language model interpolation (16 meetings)	38.7

Table 1. **Recognition Results:** The upper part evaluates the baseline BN system across different tasks. MAP (Maximum A Posterior) adaptation was used for domain adaptation. The language model was adapted by interpolating the BN model with a small meeting model. The ESST system [13] has been trained on clean speech in travel domain and is significantly smaller than the BN system.

2.2. Model Combination based Acoustic Mapping (MAM)

For the experiments on meeting data reported above we have used comparable recording conditions as each speaker in the meeting has been wearing his or her own lapel microphone. Frequently however this assumption does not apply. We have also carried out experiments aimed at producing robust recognition when microphones are positioned at varying distances from the speaker. In this case data, specific for the microphone distance and SNR found in the test condition is unavailable. We therefore apply a new method, Model Combination based Acoustic Mapping (MAM) originally proposed for recognition in different car noise environments [14] to the recognition of speech at different distances.

MAM estimates an acoustic mapping on the log-spectral domain in order to compensate for noise condition mismatches between training and test. During training, the generic acoustic models λ_k $(k = 1, 2, ..., n)$ and a variable noise model N are estimated. Then, model combination is applied to get new generic models $\hat{\lambda}_k = \lambda_k + N$, which correspond to noisy speech. During decoding of a given input x, the mapping process requires a classification as a first step. The score for each $class(model)$ is computed as $g_k(x) = P(k|x, \hat{\lambda}_k)$. In the second step x is reconstructed according to the calculated score, where μ refers to the mean vector: $\hat{x} = x + \sum_{k=1}^{n} g_k(x)(\mu_k - \hat{\mu}_k)$.

Word Error Rate [%]		Close	Distant
Baseline	Speaker 1	35.2	46.1
MAM	Speaker 1	-	38.2
Baseline	Speaker 2	13.8	31.2
MAM	Speaker 2	14.4	26.1

Table 2. Recognition results on Model Combination based Acoustic Mapping (MAM)

We applied MAM to data that was recorded simultaneously by an array of microphones positions at different distances from the speaker. Each speaker read several paragraphs of text from the Broadcast News corpus. The results of experiments with two speakers are summarized in Table 2. Experiments suggest that MAM effectively models the signal condition found in the test resulting in substantial performance improvements.

3. DIALOGUE ANALYSIS

The idea of the dialogue analysis module in the meeting room context is to use features other than keywords for information access to spoken communication. Traditional information retrieval methods focus only on a very narrow notion of topic as a bag of keywords where as spoken language is also happening in a certain situation and in a certain style [15]. In this paper we can only give one simplified example where the speaker identities and their dominance are important, namely in the selection of a meeting from the database. Other problems not covered here include the selection of a database out of a collection of databases, the segmentation of a meeting and the selection of a segment in a meeting. Also not covered is work on the detection of dialogue acts, games and activities [16, 17].

Five meetings in the meeting database have been annotated with topic segmentations. Selecting a meeting by a query that contains the precise time, all of the keywords or the precise information who was there and how much they talked would be trivial. On the other hand the location of the meeting is uninformative since they were all recorded around the conference table in our lab.

For dialogue selection it is assumed that the queries correspond to features of a dialogue segment and that each segment in the database is equally likely to be chosen as a query. A neural network that detects a dialogue identity for a segment has been build (Tab. 3). The network has been designed to create a probability distribution of meeting identities as its output which is tested using round robin over the whole database. To assess information access performance the reduction of empirical entropy for the meeting identity was measured in bit. This retrieval model is quite natural since we could assume that a user remembers just some part of the meeting and that most features are similar (yet not identical) in other segments of the meeting. The results show that keyword based methods are powerful but that alternatives such as speaker identity and activity exist that seem to be (a) more natural, (b) likely part of queries, (c) easy to visualize in a browsing task and (d) explain most of the word level information implicitly.

Feature	bit
speaking style	1.34
speaker identity detected by speaking style	1.13
most frequent 50 keywords	1.21
most frequent 1000 keywords	1.64
speaker identity, dominance weighted per segment	2.06
oracle	2.29

Table 3. **Empirical entropy reduction for meeting identity:** Speaking style in this table defines the distribution of the most frequent 50 words and parts of speech and explains a lot of the meeting identity. Speaking style however also serves as a very good speaker identity detector: The detected speakers can be used to detect the meeting identity and the result is almost the same as for the speaking style feature itself. $1.64 - 1.21 = 0.43$ bits of information are added by less frequent keywords and the real speaker identity is still the strongest feature.

4. SUMMARIZATION

The summarization system provides the meeting browser with a relevance ranked list of sentences. The GUI can thus display the most relevant passages of a meeting, the size of the summary being dependent on the user's choice. In the following we describe the five major components of the summarization system, the first four of which addressing important issues intrinsic to spoken as opposed to written language summarization. The system architecture is similar to the one described in detail in [19]. Since then we were able to use the Penn Treebank SWITCHBOARD corpus, annotated for disfluencies for automatic training of the components which before were mainly based on hand-crafted heuristic rules (Tab. 4).

Disfluency detection and removal Spoken language contains a significant amount of false starts, repetitions, filled pauses, discourse markers and speech repairs. Our goal is to detect and remove those to make the summary more readable for the user. We trained a version of Brill's part of speech (POS) tagger [20] which marks filled pauses, editing terms, discourse markers, and non-informational conjunctions. Further, we use a decision tree [21] to determine false starts, and a script based repetition filter to eliminate the majority of speech repairs.

Sentence boundary detection Unlike written language, LVCSR output does not contain punctuation markers. Turns often contain multiple sentences, and sometimes sentences span successive turns of one speaker. To determine both inter-turn and intra-turn sentence boundaries, we use a decision tree with POS, trigger word, and time features.

Detection of question-answer-pairs In dialogues, information units are sometimes shared across several speakers. A typical example is a question-answer-pair, where question or answer alone are much less meaningful than both of them together. The goal of this component is to render the summary more coherent. To decide whether a sentence is a question or not, we use a decision tree with POS, question-specific trigger words and length information features.

Relevance ranking with word error rate minimization For determining the relevance ranking of sentences, we use an adapted version of the maximal marginal relevance (MMR) algorithm [22], where the query vector is a vector of words within a topical segment. User-defined keywords can be emphasized to turn a generic summary into a query-specific summary. Since automatic meeting transcription is less than perfect, the summary will reflect many errors from the speech recognizer. As we have reported in [23], we are able to (a) significantly reduce the summary word error rate, and (b) substantially improve the summary accuracy by combining the LVCSR confidence scores with the relevance weighting scheme of the MMR algorithm.

Topic segmentation Given the nature of meetings (and other spoken dialogues) being multi-topical, we automaticaly segment the transcript into topically coherent passages, using a variant of Hearst's TextTiling algorithm [24].

Task	handcrafted	trained
Disfluency detection and removal (exluding false starts)	0.74	0.80
Sentence boundary detection	0.60	0.78
Question classification	0.34	0.56
Question answer pair classification	0.24	0.51

Table 4. **Training on SwitchBoard:** The performance comparisons below are made with respect to the system described in [19] after training on Switchboard data. Results are reported as $F_1 = \frac{2PR}{P+R}$-scores, combining precision and recall.

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 remotely 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.

When a meeting is being created, each participant may join either remotely or locally. Once the meeting has begun, speech is transmitted to Janus, our speech recognition engine. As the speech is recognized, the hypothesis is sent to the dialogue system 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 system 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 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 [25]. In the future we plan to add speech recognition as an additional error repair modality.

6. CONCLUSIONS AND FUTURE WORK

The meeting room scenario is surprisingly challenging when it comes to speech recognition but significant progress has been made using adaptation which is also the focus of ongoing work along with noise reduction. This and ongoing work on dialogue analysis and summarization are encouraging since the output of the speech recognizer may not be crucial for all applications and suboptimal speech recognition results can be used effectively. The meeting browser user interface has developed considerably over time and is presently subjected to usability studies. The overall system architecture is a significant departure from previous systems and accounts for the dialogue style of meetings and the desire for interactive access and drilldown capabilities.

7. REFERENCES

[1] Alex Waibel, Michael Bett, and Michael Finke, “Meeting browser: Tracking and summarising meetings,” in *Proceedings of the DARPA Broadcast News Workshop*, 1998.

[2] Michael Bett, Ralph Gross, Hua Yu, Xiaojin Zhu, Yue Pan, Jie Yang, and Alex Waibel, “Multimodal meeting tracker,” in *Proceedings of RIAO2000*, Paris, France, April 2000.

[3] Rainer Stiefelhagen, Jie Yang, and Alex Waibel, “Simultaneous tracking of head poses in a panoramic view,” in *International Conference on Pattern Recognition (ICPR)*, Barcelona, Spain, September 2000.

[4] J. Garofolo, C. Auzanne, and E. Voorhees, “The TREC spoken document retrieval track : A success story,” in *Text Retrieval Conference (TREC) 8*, E. Voorhees, Ed., Gaithersburg, Maryland, USA, 1999, November 16-19.

[5] Francis Kubala, Sean Colbath, Daben Liu, and John Makhoul, “Rough’n’Ready: a meeting recorder and browser,” *ACM Computing Surveys*, vol. 31, no. 7, September 1999, Article No. 7.

[6] John Choi, Donald Hindle, Fernando Pereira, Amit Singhal, and Steve Whittaker, “Spoken content-based audio navigation (SCAN),” in *Proceedings of the ICPhS-99*, 1999.

[7] Howard D. Wactlar, “Auto summarization and visualization across multiple video documents and libraries,” `http://www.informedia.cs.cmu.edu/dli2/`.

[8] Gregory D. Abowd, “Classroom 2000: An experiment with the instrumentation of a living educational environment,” *IBM Systems Journal, Special issue on Pervasive Computing*, vol. 38, no. 4, pp. 508–530, October 1999.

[9] Barry Arons, “Speechskimmer: A system for interactively skimming recorded speech,” *ACM Transactions on Computer Human Interaction*, vol. 4, no. 1, pp. 3–28, March 1997.

[10] Hua Yu, Michael Finke, and Alex Waibel, “Progress in automatic meeting transcription,” in *Proceedings of the EUROSPEECH*, September 1999.

[11] Tanja Schultz and Alex Waibel, “Polyphone decision tree specialization for language adaptation,” in *Proceedings of the ICASSP*, Instanbul, Turkey, June 2000.

[12] Torsten Zeppenfeld, Michael Finke, Klaus Ries, and Alex Waibel, “Recognition of conversational telephone speech using the Janus Speech Engine,” in *Proceedings of the ICASSP’97*, München, Germany, 1997.

[13] Alex Waibel, Hagen Soltau, Tanja Schultz, Thomas Schaaf, and Florian Metze, “Multilingual speech recognition,” in *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer-Verlag, 2000.

[14] Martin Westphal *Robust Continuous Speech Recognition in Changing Environments*, University of Karlsruhe, Ph.D. thesis, 2000.

[15] M. M. Bahktin, *Speech Genres and other late Essays*, chapter Speech Genres, University of Texas Press, Austin, 1986.

[16] Michael Finke, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, and Klaus Zechner, “Clarity: Automatic discourse and dialogue analysis for a speech and natural language processing system,” in *AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, March 1998.

[17] Klaus Ries, Lori Levin, Liza Valle, Alon Lavie, and Alex Waibel, “Shallow discourse genre annotation in callhome spanish,” in *Proceecings of the International Conference on Language Ressources and Evaluation (LREC-2000)*, Athens, Greece, May 2000.

[18] Thomas S. Polzin and Alex Waibel, “Detecting emotions in speech,” in *Proceedings of the CMC*, 1998.

[19] Klaus Zechner and Alex Waibel, “DIASUMM: Flexible summarization of spontaneous dialogues in unrestricted domains,” in *Proceedings of COLING*, Saarbrücken, Germany, 2000.

[20] Eric Brill, “Some advances in transformation-based part of speech tagging,” in *Proceeedings of AAAI-94*, 1994.

[21] J. Ross Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1992.

[22] Jaime Carbonell and Jade Goldstein, “The use of MMR, diversity-based reranking for reordering documents and producing summaries,” in *Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval, Melbourne, Australia*, 1998.

[23] Klaus Zechner and Alex Waibel, “Minimizing word error rate in textual summaries of spoken language,” in *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics, NAACL-2000, Seattle, WA, April/May*, 2000, pp. 186–193.

[24] Marti A. Hearst, “Texttiling: Segmenting text into multi-paragraph subtopic passages,” *Computational Linguistics*, vol. 23, no. 1, pp. 33–64, March 1997.

[25] Stefan Manke, Michael Finke, and Alex Waibel, “Npen++: A writer independent, large vocabulary on-line cursive handwriting recognition system,” in *ICDAR*, 1995.