MEETING BROWSER: TRACKING AND SUMMARIZING MEETINGS

A. Waibel, M. Bett , M. Finke, R. Stiefelhagen
(waibel,mbett,finkem,stiefl)@cs.cmu.edu

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

ABSTRACT

To provide rapid access to meetings between human beings, transcription, tracking, retrieval and summarization of on-going human-to-human conversation has to be achieved. In DARPA and DoD sponsored work (projects GENOA and CLARITY) we aim to develop strategies to transcribe human discourse and provide rapid access to the structure and content of this human exchange.

The system consists of four major components: 1.) the speech transcription engine, based on the JANUS recognition toolkit, 2.) the summarizer, a statistical tool that attempts to find salient and novel turns in the exchange, 3.) the discourse component that attempts to identify the speech acts, and 4.) the non-verbal structure, including speaker types and non-verbal visual cues.

The meeting browser also attempts to identify the speech acts found in the turns of the meeting, and track topics. The browser is implemented in Java and also includes video capture of the individuals in the meeting. It attempts to identify the speakers, and their focus of attention from acoustic and visual cues.

1. THE MEETING RECOGNITION ENGINE

The speech recognition component of the meeting browser is based on the JANUS Switchboard recognizer trained for the 1997 NIST Hub-5E evaluation [3]. The gender independent, vocal tract length normalized, large vocabulary recognizer features dynamic, speaking mode adaptive acoustic and pronunciation models [2] which allow for robust recognition of conversational speech as observed in human to human dialogs.

1.1 Speaking Mode Dependent Pronunciation Modeling

In spontaneous conversational human-to-human speech as observed in meetings there is a large amount of variability due to accents, speaking styles and speaking rates (also known as the speaking mode [6]. Because current recognition systems usually use only a relatively small number of pronunciation variants for the words in their dictionaries, the amount of variability that can be modeled is limited. Increasing the number of variants per dictionary entry may seem to be the obvious solution, but doing so actually results in a *increase* in error rate. This is explained by the greater confusion between the dictionary entries, particularly, for short reduced words.

We developed a probabilistic model based on context dependent phonetic rewrite rules to derive a list of possible pronunciations for all words or sequences of words [2][4]. In order to reduce the confusion of this expanded dictionary, each variant of a word is annotated with an observation probability. To this aim we automatically retranscribe the corpus based on all allowable variants using flexible utterance transcription graphs (Flexible Transcription Alignment (FTA) [5]) and speaker adapted models. The alignments are then used to train a model of how likely which form of variation (i.e. rule) is and how likely a variant is, to be observed in a certain context (acoustic, word, speaking mode or dialogue) is.

For decoding, the probability of encountering pronunciation variants is then defined to be a function of the speaking style (phonetic context, linguistic context, speaking rate and duration). The probability function is learned through decision trees from rule based generated pronunciation variants as observed on the Switchboard corpus [2].

1.2 Experimental Setup

To date, we have experimented with three different meeting environments and tasks to assess the performance in terms of word accuracy and summarization quality: i.) Switchboard human to human telephone conversations, ii.) Research group meetings recorded in the Interactive Systems labs and iii.) Simulated crisis management meetings (3 participants) which also include video capture of the individuals. We report results from speech recognition experiments in the first two conditions.

1) Human to Human Telephone

The test set to evaluate the use of the flexible transcription alignment approach consisted of the Switchboard and CallHome partitions of the 1996 NIST Hub-5e evaluation set. All test runs were carried out using a Switchboard recognizer trained with the JANUS Recognition Toolkit (JRTk) [4].

The preprocessing of the system begins by extracting MFCC based feature vectors every 10 ms. A truncated LDA transformation is performed over a concatenation of MFCCs and their first and second order derivatives are determined. Vocal tract length normalization and cepstral mean subtraction are

computed to reduce speaker and channel differences.

The rule-based expanded dictionary that was used in these tests included 1.78 pronunciation variants/word, compared to 1.13 found in the baseform dictionary (PronLex). The first list of results in Table 1 is based on a recognizer whose polyphonic decision trees were still trained on Viterbi alignments based on the unexpanded dictionary. We compare a baseline system trained on the base dictionary with an expanded dictionary FTA trained system tested in two different ways: with the base dictionary and with the expanded one. It turns out, that FTA training reduces the word error rate significantly, which means, that we improved the quality of the transcriptions through FTA and pronunciation modeling. Due to the added confusion of the expanded dictionary the test with the large dictionary without any weighting of the variants yields slightly worse results than testing with the baseline dictionary.

Condition	SWB WER	CH WER
Baseline	32.2%	43.7%
FTA traing test w.basedict	30.7%	41.9%
FTA traing test w.expanded dict	31.1%	42.5%

Table 1 Recognition results using flexible transcription alignment training and label boosting. The test using the expanded dictionary was done without weighting the variants

Adding vowel stress related questions to the phonetic clustering procedure and regrowing the polyphonic decision tree based on FTA labels improved the performance by 2.6% absolute on SWB and 2.2% absolute on CallHome. Table 2 shows results for mode dependent pronunciation weighting. We gain an additional ~2% absolute by weighting the pronunciation based on mode related features.

Condition	SWB WER	CH WER
unweighted	28.7%	38.6%
Weighted p(r|w)	27.1%	36.7%
Weighted p(r|w,m)	26.7%	36.1%

Table 2 **Results using different pronunciation variant weighting schemes.**

2) Research Group Meetings

In a second experiment we used recorded during internal group meetings at our lab. We placed lapel microphones on three out of ten participants, and recorded the signals on those three channels. Each meeting was approximately one hour in length, for a total of three hours of speech on which to adapt and test.

Since we have no additional training data collected in this particular environment, the following unsupervised adaptation techniques was used to adapt a read speech, clean environment Wall Street Journal dictation recognizer to the meeting conditions:

1. MLLR based adaptation: In our system, we employed a regression tree, constructed using an acoustic similarity criterion for the defnition of regression classes. The tree is pruned as necessary to ensure sufficient adaptation data on each leaf. For each leaf node we calculate a linear transformation that maximizes the likelihood of the adaptation data. The number of transformations is determined automatically.

2. Iterative batch-mode unsupervised adaptation: The quality of adaptation depends directly on the quality of the hypotheses on which the alignments are based. We iterate the adaptation procedure, improving both the acoustic models and the hypotheses they produce. Significant gains were observed during the two iterations, after which performance converges.

3. Adaptation wth confidence measures: Confidence measures were used to automatically select the best candidates for adaptation. We used the stability of a hypothesis in a lattice as indicator of confidence. If, in rescoring the lattice with a variety of language model weights and insertion penalties, a word appears in every possible top-1 hypothesis, acoustic stability is indicated. Such acoustic stability often identifies a good candidate for adaptation. Using only these words in the adaptation procedure produces 1-2% gains in word accuracy over blind adaptation [9].

The baseline performance of the JRTk based WSJ Recognizer over the Hub4-Nov94 test set is about 7% WER. These preliminary experiments suggest that due to the effects of spontaneous human-to-human speech, significant differences in recording conditions, significant crosstalk on the recorded channels, significantly different microphone characteristics, and inappropriate language models the error rate on meetings is in a range of 40-50\% WER.

Speaker	Adaptation Iterations			
	0	1	2	Adaptation Gain
maxl	51.7	45.3	45.2	12%
fdmg	48.4	43.8	44.9	9%
flsl	63.8	59.5	59.6	7%
Total	54.8	49.6	49.9	

Table 3 **Error rates for three different speakers in a research group meeting using JRTk trained over WSJ dictation data.**

2. SUMMARIZATION

Based on transcripts that are produced manually or by recognizer output, we now wish to produce condensed informative summaries of these meetings. Rather than attempting a detailed linguistic analysis of the semantics of a meeting, we adopt a statistical approach, whereby we flag and select salient, relevant and informative passages from a meeting and present only a meeting's soundbites in varying detail. In the following experiments, we attempt to quantify the quality and compression achieved by this approach.

As a first metric for selecting salient, informative passages from a human dialog, we have explored the Maximal Marginal Relevance (MMR) metric [1] first introduced for text summarization in the TIPSTER project (Carbonell [1]). The MMR iteratively maximizes the similarity between a query and each section of a document while it minimizes the similarity among previously ranked document sections. It thereby identifies the most relevant, yet most diverse, non-redundant sections of a document.

Here, we apply a modified version of the MMR to conversational dialogue to find the most relevant, non-redundant turns in a meeting transcript. The top N turns are then presented to the user in the original order of the meeting transcript as a summary of the meeting. For our first experiments, we have created summaries for dialogues from the Switchboard collection, a set of two person conversations. We demonstrate that the resulting segments presented provide a brief summary or gist of the meeting.

Since the MMR requires a query, it is necessary to generate a query around which we can center the summary. In previous work modifications have been made to the MMR to create generic summaries by submitting the most common words in a document in combination with the document title as the query. In our case we have chosen to use the most common word as the query. In the future we plan to explore using the top N common words or phrases as the query to generate an improved generic summary.

2.1 The Summarization Algorithm

The summarization algorithm takes as input a textual transcript that was generated manually or from an actual speech recognition run. It produces a summary consisting of n turns or utterances. The algorithm can be divided into the following steps:

1. Eliminate all stop words from consideration
2. Identify the most common stems from the set of remaining words
3. Weight each turn or utterance
4. Include the highest weighted turn in the summary
5. Eliminate the most common stem words and the included turn from consideration
6. If a preset summary size has not been reached, go to step 2.

In the first step, a set of more than 550 stop words consisting of the most common words in spoken English language is used to eliminate uninformative words from the dialogue and to focus on topical substance. We mark each of the stopwords as uninteresting essentially them from consideration.

Our goal in the second step is to identify from the set of remaining words the most common word stem. We are using a technique that is similar to those used in noun phrase summaries. Algorithms for noun phrase summaries generate lists of words or phrases that appear in a document ordered by the number of times a phrase appears. The purpose is to provide the reader with the gist of the document by presenting the most common words.

We also provide an ordered list of words by counting the occurrence of each word with a non zero weight. To save time, these words were not limited to noun phrases. In a more complete implementation, we would count both phrases and individual words. It is our contention that the word with the highest weight is the most unique and therefore most important. We believe that finding the common stem provides an indication of the most important remaining topic in the document.

As a further simplification, we are only considering the first four letters of a word to be significant. This technique has proven successful [8] as compromise taking the place of better stemming algorithms [7].

Once the common stem has been identified, in the third step each turn that has not previously been included in the summary is weighted. We count the number of occurrences of the common word stem in the turn. Future expansions of this weighting scheme could account for term expansion words and phrases such as "therefore" and "in conclusion" that indicate summary information in text and dialogues as well.

In the fourth step each of the turns weighted in step three is ranked. The highest-ranking turn is included in the summary and is marked as being included in the summary in order to exclude it during the next iteration.

In order to minimize redundance in the summary and to identify the most unique parts, the weight of all words that contain the most common stem is set to zero. This helps to avoid the repetition of similar topics in the summary.

While the summary size is less than the maximum allowable size, turns continue to be added to the summary in this manner until a preset size limit is reached.

2.2 Using Categorization to Test Summarization Results

In order to test the quality of the summarization, we conducted three separate experiments.

Subject	Total Correct/ Incorrect
A	29/1
B	28/2
C	28/2
D	29/1
E	24/6
F	29/1
Total	167/180

Table 4 Categorization results for the reference dialogues

The first experiment was use to determine if summaries automatically generated from the original reference text dialogues could be properly categorized into one of five narrow

categories

To test this we chose five related Switchboard topics and pick five documents from each topic an additional five documents randomly chosen from another five topics. Each subject was given thirty ten-turn summaries generated from the original reference text and was asked to identify to which of the six topics (the original five topics and a "none of the above") they belonged. The results of the experiment are shown in Table 4

The results of the experiment demonstrate that the subjects are able to successfully categorize the summarized dialogues 92.8% of the time within the context of the switchboard transcripts.

Based on this information we performed categorization using summaries automatically generated from transcripts obtained from speech recognizer runs. This second experiment presented the reader with fifteen ten-utterance summaries each from a different topic. The object was to match the summary with the correct topic. The results of this experiment are shown in Table 5

Subject	Number Correct/Incorrect
G	15/0
H	15/0
I	14/1
J	15/0
K	15/0
L	15/0

Table 5 **Categorization results for summarized dialogues**

From the results of both of these experiments we conclude that within the switchboard context a ten-utterance summary appears to be sufficient for topic identification. Moreover, in the second experiment we found that a summary automatically generated from speech recognizer also appears to be sufficient for identifying the topic.

In the third experiment we choose five Switchboard dialogues and created a list of questions from the reference dialogues. The subjects were given summaries automatically generated from speech recognizer transcripts. Subjects were given dialogue summaries that were 2,5,10 or unlimited number of turns long. The reader was asked to answer questions based on each summary.

The objective was to identify if the summary was conveying key information from the dialogue and to determine how long of a summary is required to maintain key content. This experiment is undoubtedly more subjective and harder to quantify. It was nonetheless carried out to explore the correspondence between summary size and adequacy of content

From Figure 1 we can see that there is an upward trend in the percentage of questions answered correctly as the number of turns in the summary is increased. We believe that this demonstrates the potential for speech recognition output to be used for summarization and to provide the user with the gist of a dialogue if not important elements of its content.

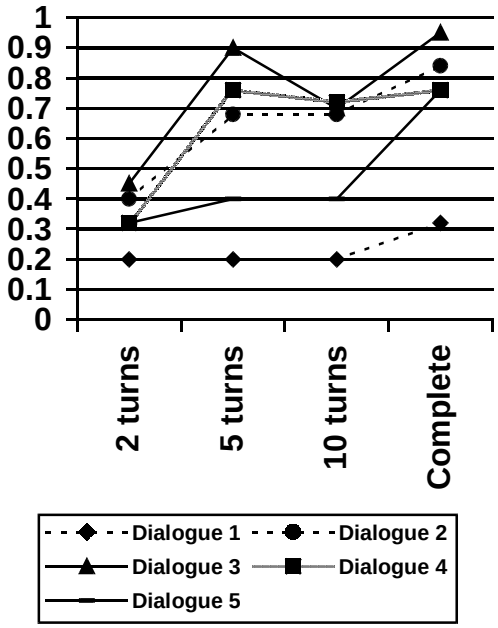

Figure 1 **Percent Correct on Average**

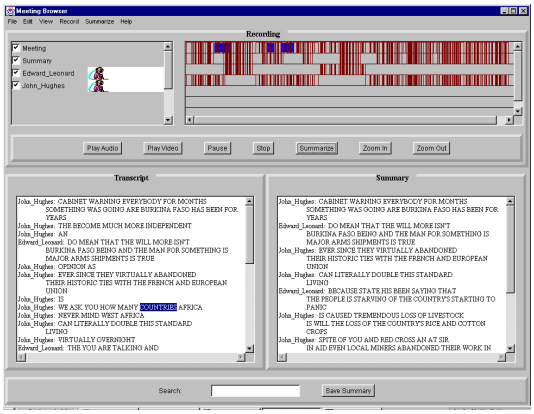

Figure 2 **Meeting Browser with Summarization**

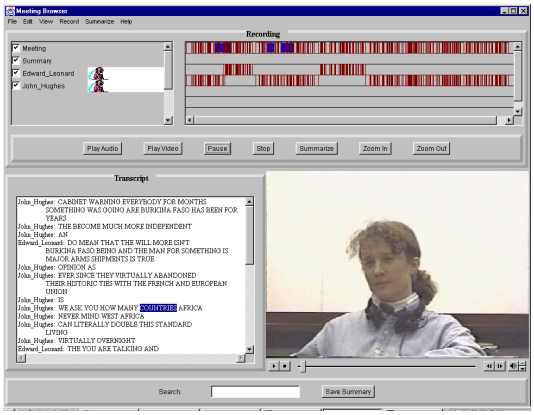

Figure 3 **Meeting Browser with video capture**

2.3 Meeting Browser Provides Maximum Flexibility

An important aspect of generating meeting summaries or minutes is the successful and efficient delivery of the result. We have developed a meeting browser that allows the user to review and browse transcribed and summarized meetings efficiently. The browser (pictured in Figure 2.) is implemented in Java. It also includes video capture of the individuals in the meeting for use in meeting rooms or video conferencing (Figure 3.)
In addition to being a tool for browsing and viewing meeting records, the meeting browser also attempts to provide tools for more rapid and informative access of key events in the meeting. Among the possibilities for informative access, we are experimenting with automatic detection of the speech acts found in the turns of the meeting, and topics tracking. We also attempt to identify the speakers, and their focus of attention from acoustic and visual cues.

The Meeting Browser interface displays meeting transcriptions, time-aligned to the corresponding sound and video files. The user can select all or a portion of these files for playback; text highlighting occurs in sync with the sound and video playback.
As software design, the Meeting Browser is built around information streams. Transcribed meeting text is just one such stream; the interface can accept streams from virtually any source that produces text output. These streams are fully editable and searchable, allowing humans to annotate and correct recognition output as well as adding new informative streams manually. Since the usefulness of a meeting transcription system is bounded by the usability of the user interface, the flexibility present in the Meeting Browser is extremely important for user acceptance of the meeting recording and transcription process.

2.4 Visual Cues Aid Meeting Browsing

Visual cues present another valuable source of information in structuring and human communicative events. We have introduced face tracking and gaze tracking to obtain a visual view of the speakers and their interaction with each other. Ones such non-verbal cue, that is used in assessing the dynamics of human interaction is to determine the target of a human speech event. When more than two agents are present, the object and adressee of a speech act is not necessarily known based on the speech stream alone. In Figure 4 we show an example in which we are using automatic face and gaze tracking algorithms to determine who a speaker is addressing or focussing on by the direction of gaze of the meeting participant.

Figure 4 **Examples in which gaze tracking tracks left, down, and right**

3. CONCLUSIONS AND FUTURE WORK

We have presented a meeting browser that attempts to provide tools for a human user attempting to rapidly review and search records of human interaction. We have reviewed the software of such a browser and discussed the challenges of automatc speech transcription, dialogue summarization and the extraction of verbal and non-verbal cues that describe the dynamics of human interaction. We believe that a great number of additional cues and structural information can be extracted automatically, that will make review and access to meeting records more efficient and easier to use.

4. ACKNOWLEDGEMENTS

This research is sponsored in part by the Defense Advanced Research Projects Agency under the Genoa project, subcontracted through the ISX Corporation under Contract No. P097047 and by the Department of Defense (project Clarity). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, ISX, DoD or any other party.

REFERENCES

1. Carbonell, J. G., Geng, Y., and Goldstein, J., Automated Query-Relevant Summarization and Diversity-Based Reranking, IJCAI-97 Workshop on AI and Digital Libraries, 1997.

2. Michael Finke and Alex Waibel, Speaking Mode Dependent Pronunciation Modeling in Large Vocabulary Conversational Speech Recognition,; Eurospeech 97, Rhodes, Greece.

3. M. Finke and J. Fritsch and P. Geutner and K. Ries and T. Zeppenfeld and A. Waibel, The JanusRTk Switchboard/Callhome 1997 Evaluation System, Proceedings of LVCSR Hub 5-e Workshop, May 1997,

4. M. Finke, The JanusRTk Switchboard/Callhome 1997 Evaluation System: Pronunciation Modeling, Proceedings of LVCSR Hub 5-e Workshop, May 1997

5. M. Finke and A. Waibel, Flexible Transcription Alignment, 1997 IEEE Workshop on Speech Recognition and Understanding, Santa Barbara, California, December 1997

6. M. Ostendorf and B. Byrne and M. Bacchiani and M. Finke and A. Gunawardana and K. Ross and S. Roweis and E. Shriberg and D. Talkin and A.Waibel and B. Wheatley and T. Zeppenfeld, Systematic Variations in Pronunciation via a Language-Dependent Hidden Speaking Mode, ICSLP, Philadelphia, USA, 1996

7. Porter, An Algorithm for Suffix Stripping ,Computer Lab, July 1980, vol 14, no. 3, p 130-137.

8. Klaus Zechner, Fast Generation of Abstracts from General Domain Text Corpora by Extracting Relevant Sentences,

Proceedings of COLING-96, pp. 986-989, Kopenhagen, 1996

9. R. Stiefelhagen, Jie Yang and Alex Waibel, Towards Tracking Interaction Between People, Intelligent Environments AAAI Spring Symposium, March 1998.

10. T. Zeppenfeld and M. Finke and K. Ries and M. Westphal and A. Waibel, Recognition of Conversational Telephone Speech using the JANUS Speech Engine, IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, IEEE, 1997