A Unified Structure-Based Framework for Indexing and Gisting of Meetings

T. Kristjansson*, T.S. Huang
Beckman Institute for
Advanced Science and Technology
University of Illinois at Urbana-Champaign
ttkristj@uwaterloo.ca, huang@ifp.uiuc.edu

P. Ramesh, B.H. Juang
Bell Laboratories, Lucent Technologies
Murray Hill, NJ
padma@research.bell-labs.com,
bjuang@lucent.com

Abstract

A variety of media involve the spoken interaction of people. For this media to be useful, indexing and browsing facilities must be provided to the user. In this paper we present a unified framework for indexing and gisting spoken interactions of people. We use speaker identification, prosody analysis and word spotting as preprocessing steps to find the structure of the meeting. The structure is modeled using a stochastic approach based on the Hidden Markov Model. The result of the analysis is an outline or table of content, as well as a rich set of visual queues for navigating the media. In addition to the automatic analysis, we provide the user with tools for browsing the meeting, as well as tools for directing the analysis and editing the results. We present early results using the proposed framework.

1. Introduction

With the advent of efficient compression techniques and the drastic reduction of cost of digital data storage space, it is possible to record the audio and video from meetings and other spoken interactions. Digital computer networks facilitate the distribution of the material from central databases. In order for this information to be useful there is a need for indexing techniques and tools for fast browsing of the material.

Meetings contain information that is valuable to an organization. The contents of a meeting may include arguments for a decision, deadlines, progress reports, deferred discussions etc., all of which may be of interest for review. A large part of corporate knowledge is contained in meetings. The ability to track a project's history or review performance and decision making, may allow corporations to optimize teamwork, as well as shorten the orientation period of new team members.

*This research was supported by the NCSA ISAAC project; Trausti Kristjansson is a Fulbright Scholar.

The traditional way of making a record of the content of a meeting is to designate a secretary that has the role of summarizing the discussion and producing the *meeting minutes*. Most meetings are not recorded in this way. When minutes are produced, they are often inaccurate and incomplete.

We present a novel framework for automatically finding the structure of a meeting, directly from the audio recording, and present early results of this method. The structure of the meeting is then displayed as a hierarchical structure similar to a Table of Contents, that allows the user of the system to easily navigate and find salient portions or portions of interest.

There are three aims in finding the structure of a discourse. The first is to find the structure itself. The second is to find units of discourse where the same topic is being discussed. The third is to find salient and synoptic segments. Finding the structure supports the second and third goals, since discourse boundaries correlates directly to topic boundaries, and salient segments can be found with reference to position within the discourse structure.

In addition to the meeting structure, we present the user with a rich set of visual cues that assist the user in locating portions of interest as well as tools for manipulating the results of the analysis and adding semantic information.

2. Related work

The content based approach to indexing meetings requires a fairly complete transcription of the meeting. When a transcription has been produced, traditional text summarization techniques can be applied. A problem with this approach is the poor accuracy of current speech recognition systems. Text based methods are also difficult to apply because of the often incomplete and un-grammatical utterances produced in fluent speech. However, some promising work has been done based on this approach [1].

Other research has focused on the use of prosody as a means of finding structure in monologs and dialogs. Arons [2] attempts to segment monologs into portions by the use

of a simple activity metric of the pitch within the segment. Chen et al. [3] find emphatic portions of the speech by use of hidden Markov models. They use the result to extract audio summaries of a meeting.

Speaker Identification [4] has been used to segment dialogs into utterances, for the purpose of allowing users to listen to utterances rather than arbitrary segments.

For video media, methods of finding a Video Table of Contents have been proposed in [5]. Their approach is to first find shot boundaries. The shots are then grouped according to some similarity criterion, and presented in a tree or graph structure.

From the above it is clear that indexing spoken dialogs can be approached by analyzing various aspects of speech. The challenge of the current work is to merge these modalities into one framework that is more powerful than the component systems.

In the current work we use speaker identification, word spotting and prosody analysis as subsystems to find the overall structure of a meeting.

Much work has been done in the areas of each subsystem, i.e. speaker identification, prosody analysis, and word spotting. We will highlight prior work in these areas in the respective sections.

2.1. Structure of spoken interaction

There is structure at all levels of language. Structure is an indispensable aid in understanding the content.

Within linguistics, Grosz et al. [6] describe discourse in terms of *linguistic*, *intentional* and *attentional* structure. They use *cue words* to identify types of discourse, e.g. *digression* is indicated by "By the way" or "incidentally" etc. Whereas their analysis attempts to describe the role of each sentence in the discourse, the object of this research is to find the large scale structure of the meeting automatically. However, we also use *structural keywords* as as an aid for identifying discourse states.

Meetings can have various degrees of structure. At one end of the spectrum are highly structured meetings, such as parliamentary gatherings. In this case an arbitrator or facilitator dictates who speaks at any instant, and the items on the agenda are predetermined. At the other end of the spectrum are relatively unstructured sessions such as "brain storming" sessions where the subject matter flows freely and the participants talk in no specific order.

The meeting data we used was loosely structured. The meetings follow an agenda consisting of meeting items. The items of a meeting generally have a single main presenter, followed by questions and discussion. As noted before, finding this structure serves two purposes: 1. it allows the creation of an outline of the meeting, that reflects the partitioning of the meeting into differing subject matter; 2. it allows us to identify salient and synoptic portions of the meeting.

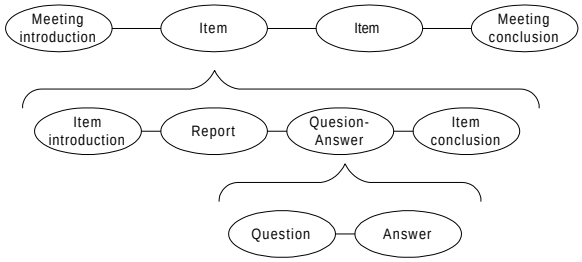

Figure 1. Hierarchical structure of a meeting.

2.2. Organization of the paper:

Speaker identification, word spotting and prosody analysis all serve as inputs to the structure analysis. We begin by presenting an overview of the structure analysis. We then describe the subsystems, i.e. the Speaker Identification, Word Spotting and Prosody Analysis subcomponents. Then we discuss details of structure analysis and present early results. Finally we describe an interface that is used both for controlling the analysis and editing the results as well as for browsing the meeting.

3. Overview of Structure Analysis

Before discussing the component systems, we present an overview of the complete system in order to motivate the relevant issues when discussing each of the subsystems.

The structure analysis is comprised of three subsystems, (see Figure 2). The audio of a recorded meeting is fed through the feature extraction stage, which calculates spectral features and prosodic features (pitch and energy). Using these features, we perform speaker identification, word spotting and prosody analysis. The output of the speaker identification is used as a basis for observation intervals. The output of the three component analysis is fused into one observation for each observation interval. Detection of structural elements of the meeting is based on this observation sequence.

The fused observation vector includes verbal, prosodic, utterance and silence duration, and speaker sequence information necessary for producing a detailed analysis of the discourse.

4. Speaker Identification

The identity of the speaker is a crucial component in the current framework. Interaction of the speakers is represented by the length of the turns and the sequence of the

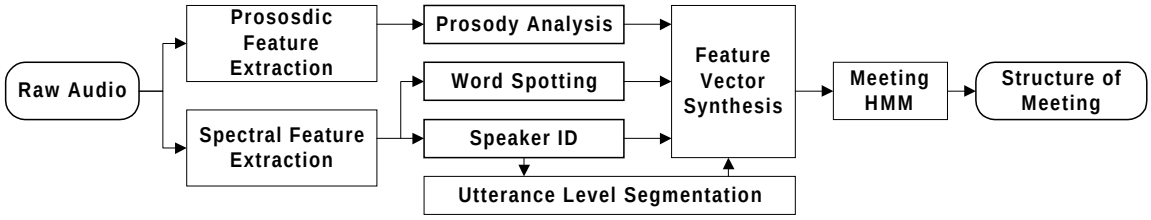

Figure 2. Overview of Structure Analysis.

speakers. The duration of silent segments (i.e. pauses) is also important for finding the boundaries between meeting components.

Much work has been done on speaker identification and verification [7]. Traditionally, the work has focused on identifying a speaker using a sentence or paragraph of speech, for security applications etc. The current task differs from the traditional one in that we want to identify a relatively small set of speakers at each time instant, with high resolution and accuracy. Similar work has been done [4], where speakers were identified in a meeting scenario. Our approach differs in the use of a constraint on the minimum length of an utterance and in that the user is included in the loop to achieve a more accurate segmentation.

The speaker identification is based on Gaussian mixture models. Each speaker is modeled by a mixture of 15 Gaussians. The features used are the traditional cepstrum based features used in speech recognition systems. In order to impose minimum length constrains, 7 Gaussians are strung together in a Hidden Markov Model. The resulting minimum segment length of 70ms corresponds to the length of a short syllable. This approach allows for constraining the minimum length of a segment while retaining finer resolution in determining transitions between speakers or between speech and silence.

The procedure for producing the speaker identification is a process of iteration. If models are available for the speakers from a previous meeting, they are used to produce an initial labeling. A revised labeling is then produced by correcting portions with errors as well as copying portions with correct labels. The models are updated based on this revised labeling.

This process is easy due to the interface that allows the user to easily navigate and correct errors in the automatic segmentation. An additional aid is a colored bar that represents the confidence of the automatic segmentation for each segment (see Figure 6). A low score is represented by red. This allows the user to locate mislabelings quickly.

Once the models have been trained on a selected database, they reflect both the characteristics of the speaker and the channel. The characteristics of the channel are effected by the acoustic characteristics of a room, the position of the speaker in the room, as well at the microphone. This makes the recognition much more accurate than if channel independent models were used.

An accuracy of 98.2% was achieved in one iteration in our initial experiment.

5. Prosody Analysis

Prosody conveys a variety of nonverbal information, e.g. mood and attitude. It also conveys information about emphatic and contrastive focus [8] as well as segmental information at a word and sentence level. At the discourse level it conveys structural information [2].

We use prosody to locate emphatic portions of the meeting. This is based on the pitch and energy contours of the speech.

Three HMM models were used for emphasis detection, one for high emphasis, one for moderate emphasis and one for un-emphasized speech. Each model had 3 states of 3 Gaussian mixtures. The high and moderate emphasis models were left to right, while the no-emphasis model also had a transition from the third to the first state.

The pitch and amplitude were first calculated for 30ms frames at 10ms steps. In order to introduce temporal context, derivatives were calculated by first order regression analysis of the pitch and energy contours. This provided the mean and slope of the contours at 100ms intervals.

Using this method, the accuracy of the automatic detection for *high-emphasis* was 73.8% and 85.6% for *no-emphasis*.

In order to estimate the emphasis of an utterance, we find by the ratio of the length of the emphasized sub-segments of an utterance to the total length of the utterance. This information can be displayed as a colored bar, parallel to the speaker ID information in the user interface. This gives the user visual information for quickly finding the emphasized portions of the dialog, which often correspond to salient portions.

6. Word Spotting

The current work circumvents the need for a complete transcription of the meeting. We use word spotting technology that has been shown to be more robust than large vocabulary speech recognition. Instead of producing a complete transcription, we find a small set of words that relate to the structure of the meeting. Word spotting can also be used to locate salient concepts such as dates and monetary units.

A great deal of work has been done on word spotting [9]. The difference between word spotting and Large Vocabulary Automatic Speech Recognition is that in the former, only a small set of words are identified. By reducing the number of words and by introducing stricter verification measures, a higher level of accuracy can be achieved. The word spotting system we used is based on a system developed at Bell Laboratories.

Our system is comprised of two components: a candidate generation stage, followed by a verification stage. For each utterance, HTK[1] was used to find candidates for the position of each word within a segment. The verification is based on the use of phone anti-models [9].

We used a set of 1117 context dependent biphone models and 41 context independent anti-models. Word models were produced by concatenating the appropriate phone models according to the phonetic transcription of the word. In some cases we used multiple pronunciations found by examining the actual phone recognition of the system.

6.1. Structural keywords and keyphrases

The word spotting system provides information about the presence of 38 *structural-keywords* and *keyphrases*[2]. Words such as "start" and "meeting" have a high correlation to the meeting introduction, whereas "conclude", "adjourn" and "thanks" have a high correlation to the meeting conclusion, especially when uttered by the chairperson. Words such as "so" and "OK" in combination with relatively long silences often correlate with the end of a topic. We found that the word "basically" often marks synoptic speech.

The correlation of the words to the discourse elements and discourse boundaries are determined by finding the conditional probability of the discourse state, given the structural keyword, as estimated by frequency counts.

The set of words can be expanded for other types of meetings. Since the framework is probabilistic, the correlation strengths are learned and not determined from heuristics.

[1] Hidden Markov Model Toolkit from Entropic

[2] Similar terms used in other contexts: cue phrases, clue words, discourse markers, discourse particles

7. Hierarchical HMM based meeting model

Given the data produced by the speaker identification, word spotting and prosody analysis, we propose the use of a hidden Markov model to merge these information sources to find the structure. The states of the hidden Markov model capture the statistics of a structural components, e.g. a long turn will probably be part of a monolog. The transition probabilities allow us to model the sequential order of the elements of the structure, e.g. that the meeting introduction precedes the first agenda item, and that each agenda item often has a monolog followed by a discussion. It also allows the modeling of interaction between participants such as question-answer units. In the current system each discourse state is modeled with two states in the HMM, corresponding to utterance and silence. A natural extension of this is to use 3 pairs of utterance/silence states. This would allow for the modeling of phenomena such as longer pauses towards the end of a monolog or dialog. This also allows words that mark discourse boundaries to be associated with beginning or ending states.

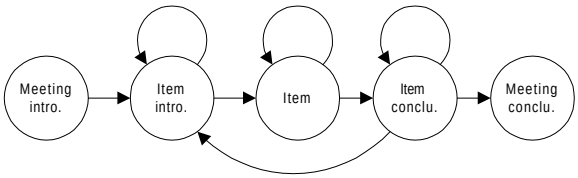

Figure 3. Meeting model.

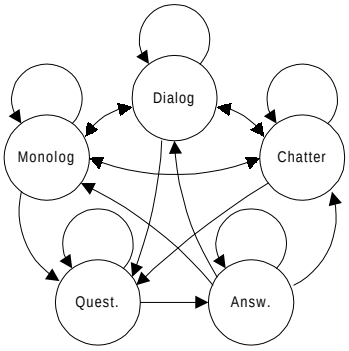

Figure 4. Item model.

7.1. Structural elements

The elements of a meeting form a hierarchical structure. The elements that we attempt to recognize are *meeting introduction*, *item introduction*, *monolog*, *discussion*, *question*, *answer*, *item conclusion* and *meeting conclusion*. These can be represented in a hierarchical fashion (see Fig-

ure 1). We also include a *Chatter* state that represents portions of a meeting where people talk or laugh at the same time, or in small groups. This happens before and after the meeting and sometimes also between other meeting states when the main speaker or speakers are being established.

7.2. Observations

The units produced in the speaker identification stage are used as the "time basis" for the observation sequence. The segments correspond either to utterances or to silence between utterances. By using the segmentation produced by the speaker ID, instead of a regularly paced segmentation, durations can be encoded in the feature vector. This allows the durations of the utterances to be modeled by the Gaussian mixtures of the state, rather than the transition probabilities. Transition probabilities lead to an exponential distribution, which is inadequate for modeling durations of speech segments.

For each segment, corresponding to silence or an utterance, the parameters that are calculated from the speaker ID data are: 1. *speech or silence*, 2. *chairperson weighting factor*, 3. *length of utterance*, 4. *length of turn*[3], 5. *number of intermittent speakers since last turn* and 6. *time since last turn*.

These features encode information on the segmental length and the speaker sequence. We must differentiate between the chairperson and the other participants, since the chairperson directs the meeting and utterances spoken by him/her are more relevant to the meeting structure. The *chairperson weighting* component of the feature vector encodes this information. The identity of the other participants does not matter for the purpose of determining structure; however, their local sequence is important. This information is represented by the *number of intermittent speakers since last turn* and *time since last turn* components.

The length of an utterance and length of a turn are important in discriminating between discourse states (see Figure 5), especially monolog and dialog. This information is represented by the *length of turn* and *length of utterance* components.

7.3. Results

We present early results, using only the utterance/silence duration and sequence information derived from the speaker identification subsystem. We recorded 4 meetings of 5 to 7 people, and lengths of 15 to 45 minutes. These meetings contained 1 to 3 agenda items. The recordings were done using two microphones mounted in the center of the board room table. The speaker segments were found as described

[3]Turns are defined as segments of uninterrupted utterances by one speaker.

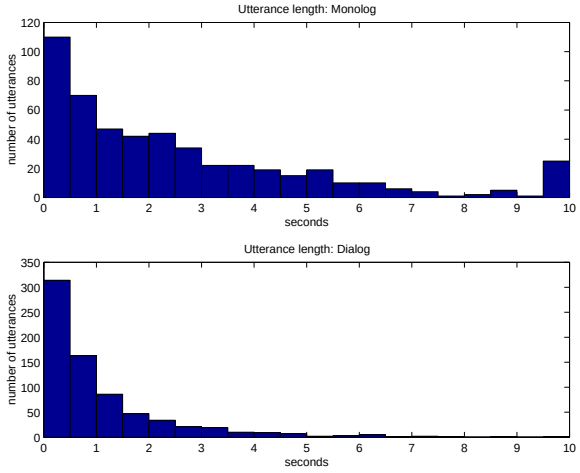

Figure 5. Histograms of utterance length in monolog (upper) and dialog (lower) discourse states.

Table 1. Confusion matrix for discourse type detection.

	Monolog	Dialog	Chatter
Monolog	60.9%	39.1%	0.0%
Dialog	3.4%	95.0%	1.6%
Chatter	7.1%	68.2%	24.7%

in the Speaker ID section. Observations were generated as described in the above section and analyzed using the framework. The discourse models were trained using 3 of the meetings and tested on one meeting. Since the results reported here do not include the word spotting input, we report only on the capability of the system to identify discourse of type *monolog*, *dialog* and *chatter*. The confusion matrix is shown in Table 1

These results show the capability of the system to recognize types of discourse from segmental length and speaker sequence alone. *Monolog* and *dialog* are well identified, but *chatter* is confused with *dialog*.

8. The user interface

The aim of this research is to provide the user with tools for quickly grasping the content of spoken interactions. An important aspect of this is giving the user easy-to-use tools for fast and efficient access to the original recording as well tools for extracting structure and highlighting salient portions. In addition, the tools should allow the user to edit

the output of the analysis and add textual descriptors of the material.

We stress the importance of keeping the user "in the loop", and allowing him/her to easily verify and correct errors in the automatic analysis.

The Meeting Analysis and Editing Tool is shown in Figure 6. The timeline of the dialog is shown with vertical color-coded bars from top to bottom, as well as text labels. The right half of the frame shows the results of the speaker identification.

The left half of the frame shows the structure of the meeting. The meeting components are indented to represent the hierarchical structure of the meeting. When editing the results of the structure analysis, segments can easily be inserted or deleted, and aligned to the speaker identity/silence segmentation.

When browsing, points of interest can be played by clicking on the bars. The user is aided by the ability to play segments that correspond to utterances, turns and meeting components. The results of the emphasis evaluation can be displayed as a color coded bar, to the right of the speaker identity bar, in a similar manner to the confidence bar.

The interface allows the user to control the speaker identification, word spotting, and prosody analysis as well as the structure analysis, and edit the results.

In addition to the time line representation of a meeting, a textual representation, in the form of a hierarchical table of content, is of great value to the user (not shown).

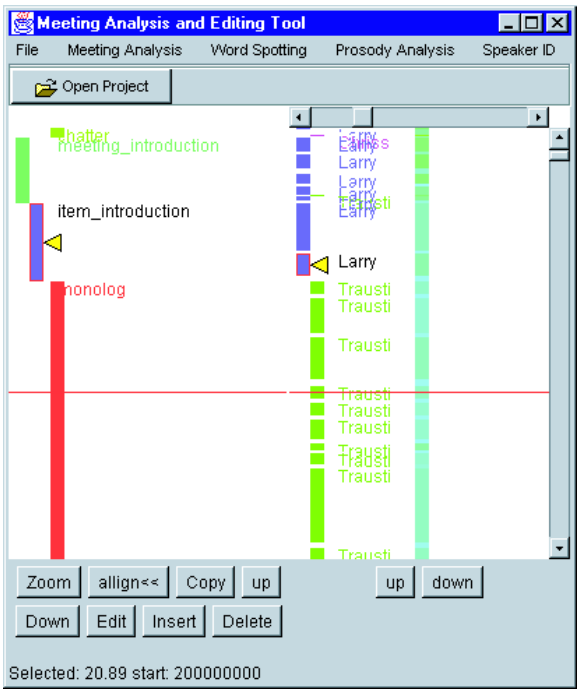

Figure 6. Meeting Analysis and Editing tool.

9. Conclusion and future work

In this paper we have presented a framework for merging various speech analysis technologies to produce an outline of a meeting. We have described a system that provides powerful analysis aids to the user and an interface for browsing spoken interactions.

We believe that the hidden Markov model provides a powerful framework for modeling the large scale structure of a meeting. This framework is readily applicable to other types of spoken interactions, such as lectures, seminars and interviews. By using a different set of preprocessing subsystems, we believe it is also applicable to other types of media that contains temporal structure such as television news and sporting coverage.

The next step is to incorporate the results of the word spotting subsystem. Future work includes expanding the types of prosodic patterns found in prosody analysis and improving the speaker identification.

References

[1] R. Kazman, R. Al-Halimi, W. Hunt, M. Mantei, "Four Paradigms for Indexing Video Conferences," *IEEE Multimedia Magazine*, 1996 pp. 63 – 73

[2] B. Arons, "Pitch-Based Emphsis Detection for Segmenting Speech Recordings", *Proc. of Int'l Conf. on Spoken Language Processing* September 1994, Vol.4, 1994, pp. 1931-1934

[3] F. R. Chen, M. Whithgott, "The Use of Emphasis to Automatically Summarize a Spoken Discourse," *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, 1992, vol. 1, p. 229-232.

[4] L. Wilcox, F. Chen, D. Kimber, V. Balasubramanian, "Segmentation of Speech Using Speaker Identification," *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, Vol. 1, 1994, pp. 161 – 164

[5] M. Das, S. Liou, "A New Hybrid Approach to Video Organization for Content-Based Indexing," *Int. Conf. on Multimedia Systems and Computing*, 1998 (Accidentally left out of proceedings)

[6] B.J. Grosz, C.L. Sidner "Attention, Intentions, and the Structure of Discourse," *Computational Linguistics*, Vol. 12, No. 3, July-September 1986 pp. 175 – 204

[7] H. Gish, M. Schmidt, "Text-Independent Speaker Identification," *IEEE Signal Processing Magazine*, October 1994 pp. 18 – 32

[8] J. Pierrehumbert, J. Hirshberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," Intentions in Communication, edited by Cohen, Morgan & Pollack MIT Press, 1989, pp. 270–323

[9] M.W. Koo, C. H. Lee, B. H. Juang "A New Hybrid Decoding Algorithm for Speech Recognition and Utterance Verification," *Proc. of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 303–310