The CALO Meeting Assistant System

Gokhan Tur, *Senior Member, IEEE*, Andreas Stolcke, *Senior Member, IEEE*, Lynn Voss, Stanley Peters, Dilek Hakkani-Tür, *Senior Member, IEEE*, John Dowding, Benoit Favre, Raquel Fernández, Matthew Frampton, Mike Frandsen, Clint Frederickson, Martin Graciarena, Donald Kintzing, Kyle Leveque, Shane Mason, John Niekrasz, Matthew Purver, Korbinian Riedhammer, Elizabeth Shriberg, Jing Tien, Dimitra Vergyri, *Member, IEEE*, and Fan Yang

***Abstract*—The CALO Meeting Assistant (MA) provides for distributed meeting capture, annotation, automatic transcription and semantic analysis of multiparty meetings, and is part of the larger CALO personal assistant system. This paper presents the CALO-MA architecture and its speech recognition and understanding components, which include real-time and offline speech transcription, dialog act segmentation and tagging, topic identification and segmentation, question-answer pair identification, action item recognition, decision extraction, and summarization.**

***Index Terms*—Multiparty meetings processing, speech recognition, spoken language understanding.**

I. Introduction

In most organizations, staff spend many hours each week in meetings, and technological advances have made it possible to routinely record and store meeting data. Consequently, automatic means of transcribing and understanding meetings would greatly increase productivity of both meeting participants and nonparticipants. The meeting domain has a large number of subdomains including judicial and legislative proceedings, lectures, seminars, board meetings, and a variety of less formal group meeting types. All these meeting types could benefit immensely from the development of automatic speech recognition (ASR), understanding, and information extraction technologies that could be linked with a variety of online information systems.

In this paper we present the meeting recognition and understanding system for the CALO Meeting Assistant (CALO-MA) project. CALO-MA is an automatic agent that assists meeting participants, and is part of the larger CALO [1] effort to build a "Cognitive Assistant that Learns and Organizes" funded under the "Perceptive Assistant that Learns" (PAL) program [2] of the DARPA. The focus of CALO in general is "learning in the wild", or continuous improvement of the system's abilities as a result of system use.

Significant anecdotal evidence suggests that companies have collected a wide range of meeting data over the years. Broadcast data and recorded conferences are also available. Further, public data such as council meetings and government proceedings is often accessible. However, little of the data is useful for research purposes. First, privacy and competitive advantage requirements preclude the use of most business meeting data. Privacy, copyright, and signal quality bar the use of most other types of "found" data as well. Rather, collection with the specific intent of providing a basis for research is required.

Projects initiated at CMU [3] and ICSI [4] in the late 1990s and early 2000s collected substantial meeting corpora and investigated many of the standard speech processing tasks on this genre. Subsequently, several large, interdisciplinary, and multisite government-funded research projects have investigated meetings of various kinds. The AMI (Augmented Multiparty Interaction) Consortium [5] project concentrates on conference-room meetings with small numbers of participants, similar to the CALO-MA system. The CHIL (Computers in the Human Interaction Loop) project [6] collected a series of lectures dominated by a single presenter with shorter question/answer portions, as well as some "interactive" lectures involving smaller groups. AMI and CHIL also produced corpora of time-synchronized media, generally including close-talking and far-field microphones, microphone arrays, individual and room-view video cameras, and output from slide projectors and electronic whiteboards.

Starting in 2002, the annual NIST Rich Transcription (RT) Evaluations [7] have become a driving force for research in meeting processing technology, with substantial performance improvements in recent years. In order to promote robustness and domain independence, the NIST evaluations cover several meeting genres and topics, ranging from largely open-ended, interactive chit-chat, to topic-focused project meetings and technical seminars dominated by lecture-style presentations. However, NIST evaluates only the speech recognition and speaker diarization systems, with a focus on recognition from multiple distant table-top microphones. Higher level semantic understanding tasks ranging from dialog act tagging to summarization are only indirectly evaluated in the framework of larger meeting processing projects.

In the following sections we discuss the speech-based component technologies contributing to CALO-MA, including speech recognition, dialog act segmentation and tagging, topic

Manuscript received March 03, 2009; revised October 25, 2009. This work was supported in part by the DARPA CALO funding (FA8750-07-D-0185, Delivery Order 0004), in part by the European Union IST Integrated Project AMIDA FP6-506811, and in part by the Swiss National Science Foundation through NCCR's IM2 project. The associate editor coordinating the review of this manuscript and approving it for publication was XXXXX XXXXXX.

G. Tur, A. Stolcke, L. Voss, M. Frandsen, C. Frederickson, M. Graciarena, D. Kintzing, K. Leveque, S. Mason, E. Shriberg, J. Tien, D. Vergyri, and F. Yang are with SRI International, Menlo Park, CA 94025 USA.

S. Peters, R. Fernandez, J. Niekrasz, M. Purver, J. Dowding, and M. Frampton are with CSLI, Stanford University, Stanford, CA 94305 USA

D. Hakkani-Tür, B. Favre, and K. Riedhammer are with ICSI, Berkeley, CA 94704 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2038810

- *John Smith*: so we need to arrange an office for joe browning (statement/all)
- *Kathy Brown*: are there special requirements (question/John)
- *Cindy Green*: when is he co- (disruption/John)
- *John Smith*: yes (affirmation/Kathy) // there are (statement/Kathy)
- *John Smith*: we want him to be close to you (statement/Kathy)
- *Kathy Brown*: okay (agreement/John) // I'll talk to the secretary (commitment/John)
- *Cindy Green*: hold on (floor grabber/all) // wh- when is he coming (question/John)
- *John Smith*: next monday (statement/Cindy)
- *Cindy Green*: uh-huh (backchannel/all)

Action Item: Arrangement of Joe's office location
Owner: Kathy

Decision: Location of Joe's office to be close to Kathy

Summary:

- *John Smith*: so we need to arrange an office for joe browning (statement/all)
- *John Smith*: we want him to be close to you (statement/Kathy)

Fig. 1. An example of meeting data. Dialog act tags and addressed persons are shown in parantheses. This meeting data has one action item and one decision. A brief extractive summary corresponding to this meeting data follows.

segmentation and identification, action item and decision detection, and summarization. We conclude by pointing out research challenges and directions for future work. This paper significantly extends the previous IEEE SLT workshop paper [8] with much more detailed task descriptions, literature surveys, and thorough analyses.

II. CALO-MA Framework

A. Task and Corpora

Speech and language processing technology has advanced such that many types of meeting information can be detected and evaluated—including dialog acts, topics, and action items. For example, Fig. 1 presents an imagined excerpt from a meeting. The speakers and the words spoken are transcribed, along with the dialog acts (listed in parentheses). Dialog act boundaries in a single turn are separated by // tokens. In an agenda-driven meeting, each agenda item can be considered a separate topic. The example shown discusses a particular agenda item (*Arrangements for Joe Browning*). It also contains discussions about action items, due dates, and assignees. Automatically extracting this information from the signals would provide significant advantages in applications ranging from meeting browsing and search to summarization, minutes generation, and automated meeting assistants.

Apart from being highly usable in its present form, the CALO-MA system presents an experimentation platform to support ongoing research in natural language and speech processing technologies. The nature of multiparty interactions and the extreme variability found in meeting genres make this one of the most challenging domains for speech and natural language processing today.

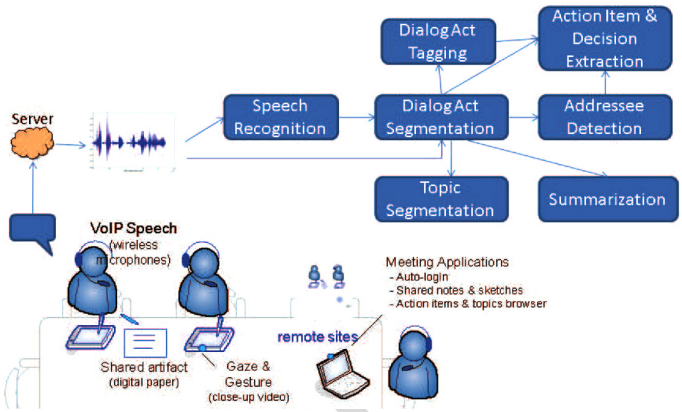

Fig. 2. The CALO-MA conceptual framework.

Fig. 2 presents the overall CALO-MA framework. CALO-MA supports multiparty meetings with a variety of information capture and annotation tools. Meetings are recorded via client software running on participants' laptop computers. The system is aware of each participant's identity. Meetings may be geographically distributed as long as a broadband Internet connection to the server is available (a phone-based interface is being developed as well). The client software captures the participants' audio signals, as well as optional handwriting recorded by digital pens. During the meeting, the participants have a real-time transcript available to which annotations may be attached. Real-time chat via keyboard input is also supported. All interactions are logged in a database, and at the conclusion of the meeting various further automatic annotation and interpretation technologies are initiated, for later browsing via a web-based interface.

The speech utterances are delivered to the server which performs real-time and offline tasks. First, the utterance is recognized, and segmented into sentences. This is the primary input for most of the following tasks. Then the sentences are assigned to dialog act tags and addressee information, which are used to improve action item and decision extraction components. Finally, the meeting is segmented into topically coherent segments and summarized according to parameters set by the user.

The CALO-MA system is used for collecting eight sequences of five meetings, each about 40 minutes long on average. The experimental results provided in the following sections either use CALO, ICSI, and/or AMI meetings corpora.

B. Meeting Capture

An early goal of the CALO-MA project was to allow lightweight data capture. Because of this, highly instrumented rooms were avoided in favor of running on each individual's Java Runtime enabled computer. Meeting participants can attend meetings by using a desktop or laptop running Windows® XP/Vista, Linux, or Mac OS X Leopard. Servers for data transport, data processing, and meeting data browsing run on Windows and Linux environments. If scaling is an issue, additional servers can be integrated into the framework to load balance the various tasks. New efforts will allow participants to conference into a meeting via a bridge between the data transport server and the public switched telephone network (PSTN).

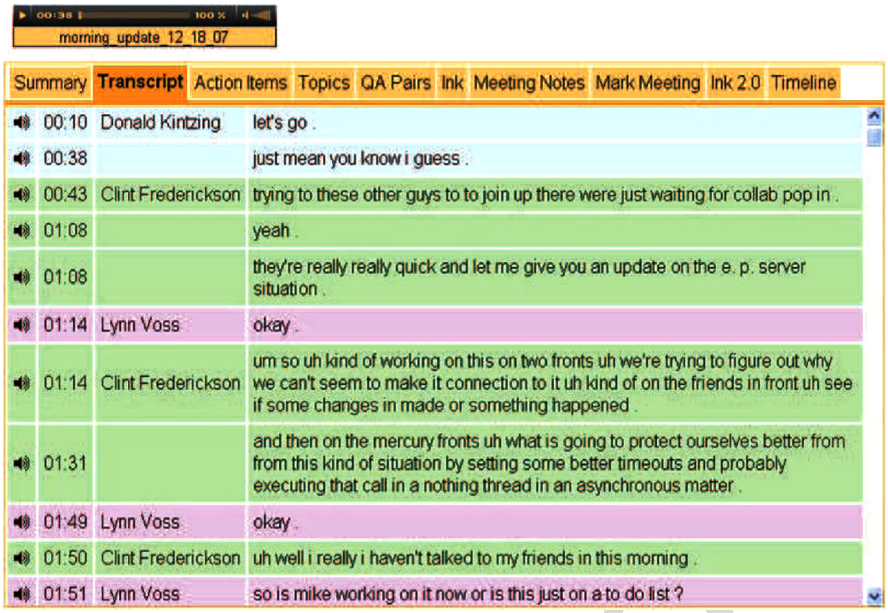

Fig. 3. Snapshot from the CALO-MA offline meeting browser.

During a meeting, client software sends Voice over Internet Protocol (VoIP) compressed audio data to the server either when energy thresholds are met or when a hold-to-talk mechanism is enabled. The data transport server splits the audio: sending one stream to meeting data processing agents for preprocessing and processing the data. Any processing agents that operate in real-time send their data back to the data transport server that relays the data back to the meeting participants.

C. Integration With Other CALO Components

Both during the live meeting and at any time after the meeting, the meeting data transport server makes available all meeting data to interested parties using XML-RPC interfaces. This allows both local and distributed users and processing agents to access the data in a language-neutral way. Meeting processing agents that are order dependent register with a meeting post processor framework to ensure that processing order is enforced (e.g., speech transcription, prosodic feature detection, dialog act recognition, action item detection, decision detection, topic boundary detection, meeting summarization, and email notification to meeting participants) and processing load is balanced.

Any CALO components outside the meeting processing framework (including the meeting browser) can send XML-RPC queries to the meeting data transport server. Those components can then perform further integration with user desktop data to facilitate additional machine learning (a focus of many other CALO processes) or present other visualizations of the data to the user.

D. Meeting Browser

After the meeting has been fully processed, email is sent out to all meeting participants. This email includes a static version of the meeting data and a link to a website where the data can be browsed dynamically from any Internet-enabled device as shown in Fig. 3. Once connected to the browser, the user can select a meeting to review and browse any of the data: both user-generated (e.g., shared files and notes) and auto-generated (e.g., detected action items and summaries). As all data is time stamped, a user can click on any data element and bring up the corresponding section of the transcript to read what was being discussed at that time. To overcome any speech transcription errors, all transcript segments can be selected for streaming audio playback. We are currently working on a framework that will allow the users to correct transcription errors.

III. Speech Recognition

A crucial first step toward understanding meetings is transcription of speech to text (STT). The NIST RT evaluations have driven the research in this field, starting out with roundtable, or "conference," meetings and recently adding other meeting genres such as lectures (mainly one person speaking) and "coffee breaks" (informal discussions following lectures). The best meeting recognition systems typically make use of the full arsenal of state-of-the-art STT techniques employed in recognizing other kinds of speech. Here, we give a brief summary with special emphasis on the approaches that deal specifically with meeting data.

At the front end, these techniques include speaker-level vocal tract length normalization, cepstral feature normalization, heteroscedastic linear discriminant feature transforms, and nonlinear discriminant transforms effected by multilayer perceptrons (MLPs). Hidden Markov model (HMM) acoustic models based on clustered Gaussian mixtures are trained using discriminative criteria such as minimum phone error (MPE) and/or a related feature-level transform (fMPE). An interesting challenge for acoustic modeling is that only relatively small amounts of actual meeting data (about 200 hours) are publicly available, compared to thousands of hours for other domains. This has engendered much research in techniques to adapt models and data from other domains for this task. For example, discriminative versions of Bayesian maximum *a posteriori* adaption (MAP) are used for Gaussian training and fMPE transform estimation

and feature estimation MLPs that were pretrained on large background corpora are retargeted to the meeting domain by limited retraining [9]. Feature transforms are also used to bridge differences in signal bandwidth between background and target data [10]. All state-of-the-art systems proceed in batch mode, decoding meetings in their entirety multiple times for the purpose of unsupervised acoustic adaptation (using maximum likelihood linear regression (MLLR)), and also for the purpose of combining multiple hypothesis streams, often based on subsystems that differ in the features or models used so as to generate complementary information. For example, a system might recognize speech based on both Mel cepstral coefficients and perceptual linear prediction cepstrum, and combine the results.

Recognizers also use large n-gram language models drawn from a range of corpora: telephone speech for conversational speaking style, technical proceedings for coverage of lecture topics, broadcast transcripts and news texts for general topic coverage, as well as smaller amounts of actual meeting transcripts available from the research projects mentioned earlier. Data are also culled from the World Wide Web using targeted search to find conversational-style transcripts as well as relevant subject matter. Source-specific component language models (LMs) are then trained and interpolated with weights optimized to maximize likelihood on representative sample data.

Even with close-talking microphones, crosstalk between channels (especially with lapel-type microphones) can be a significant problem since words from the "wrong" speakers end up being inserted into a neighboring speaker's transcript. This problem has been addressed with echo-cancellation type algorithms or cross-channel features that allow crosstalk to be suppressed during speech/nonspeech segmentation. Word error rates (WERs) on recent NIST evaluation data are in the 20% to 30% range for close-talking microphones. In the CALO-MA system, the audio stream from each meeting participant is transcribed into text by using two separate recognition systems. A real-time recognizer generates "live" transcripts with 5 to 15 seconds of latency for immediate display (and possible interactive annotation) in the CALO-MA user interface. Once the meeting is concluded, a second, offline recognition system generates a more accurate transcript for later browsing and serves as the input to the higher-level processing step described in the following sections.

The offline recognition system is a modified version of the SRI-ICSI NIST meeting recognizer [9]. It performs a total of seven recognition passes, including acoustic adaptation and language model rescoring, in about 4.2 times real-time (on a 4-core 2.6 GHz Opteron server). The real-time recognition systems consists of an online speech detector, causal feature normalization and acoustic adaptation steps, and a sub-real-time trigram decoder. On a test set where the offline recognizer achieves a word error rate (WER) of 26.0%, the real-time recognizer obtains 39.7% on the CALO corpus. We have also demonstrated the use of unsupervised adaptation methods for about 10% relatively better recognition using the recognition outputs of previous meetings [11]. Recent work includes exploiting user feedback for language model adaptation in speech recognition, by allowing users to modify the meeting transcript from the meeting browser [12].

IV. Dialog Act Segmentation

Output from a standard speech recognition system typically consists of an unstructured stream of words lacking punctuation, capitalization, or formatting. Sentence segmentation for speech enriches the output of standard speech recognizers with this information. This is important for the readability of the meetings in the CALO-MA offline meeting browser and the following processes which use sentences as the processing units, such as action item extraction or summarization.

Previous work on sentence segmentation used lexical and prosodic features from news broadcasts and spontaneous telephone conversations [13]. Work on multiparty meetings has been more recent (e.g., [14] and [15]). In the meetings domain, what constitutes a sentential unit (called as a dialog act unit) is defined by the DAMSL (Dialog Act Markup in Several Layers) [16] and MRDA (Meeting Recorder Dialog Act) [17] standards as explained in the next section.

For dialog act segmentation, similar to the approaches taken for sentence segmentation, the CALO-MA system exploits lexical and prosodic information (such as the use of pause duration [15] and others [18]). Dialog act segmentation is treated as a binary boundary classification problem where the goal is finding the most likely word boundary tag sequence, $T = t_1, \ldots, t_n$, given the features, $F = f_1, \ldots, f_n$ for n words:

$$\text{argmax}_T P(T \mid F)$$

To this end, for CALO-MA, we use hybrid models combining both generative and discriminative classification models. As the generative model, we use the hidden event language model, as introduced by [19]. In this approach, sentence boundaries are treated as the hidden events and the above optimization is simply done by the Viterbi algorithm using only lexical features, i.e., language model. Later, a discriminative classification approach is used to build hybrid models to improve this approach by using additional prosodic features [13]. The posterior probabilities obtained from the classifier are simply converted to state observation likelihoods by dividing to their priors following the well-known Bayes rule:

$$\text{argmax}_T \frac{P(T \mid F)}{P(T)} = \text{argmax}_T P(F \mid T).$$

For the ICSI corpus, using only lexical or prosodic information with manual transcriptions resulted in around 48% NIST error rate, which is the number of erroneous boundaries divided by the number of sentences (i.e., a baseline of 100% error rate). Using the hybrid approach to combine these information sources resulted in 33% NIST error rate, a significant improvement. The performance drops by 20%–25% relatively when ASR output is used instead, where the WER is around 35%. For the CALO corpus, using only lexical information resulted in 57% NIST error rate, and this was reduced to 39% using the hybrid approach with manual transcriptions.

With the advances in discriminative classification algorithms, other researchers also tried using Conditional Random Fields (CRFs) [20], Boosting [21], and hybrid approaches using Boosting and Maximum Entropy classification algorithms [22].

Our recent research has focused on model adaptation methods for improving dialog act segmentation for meetings using spontaneous telephone conversations, and speaker-specific prosodic [18] and lexical modeling [21].

In order to exploit the sentence boundary tagged meeting corpora as obtained from other projects such as ICSI and AMI, we also proposed model adaptation [21] and semi-supervised learning techniques, such as co-training [23] and co-adaptation [24], for this task. Model adaptation reduced the NIST error rate for the CALO corpus to 30%.

V. Dialog Act Tagging

A dialog act is a primitive abstraction or an approximate representation of the illocutionary force of an utterance, such as *question* or *backchannel*. Dialog acts are designed to be task independent. The main goal of dialog acts is to provide a basis for further discourse analysis and understanding.

For CALO-MA, dialog acts are very useful for most of the following processes, such as using action motivators for action item detection or using question/statement pairs for addressee detection. Note that dialog acts can be organized in a hierarchical fashion. For instance, statements can be further subcategorized as *command* or *agreement*. Depending on the task, which will use the DA tags, the granularity of the tags is determined. Furthermore, dialog act tags can be used for correct punctuation such as period versus question marks.

The communicative speech act theory goes back to the 1960s, and there are a number of contemporary dialog act sets in the literature, such as DAMSL [16] and MRDA [17], as mentioned in the previous section. DAMSL focuses on providing multiple layers of dialog act markup. Each layer allows multiple communicative functions of an utterance to be labeled. The Forward Communicative Functions consist of a taxonomy in a style similar to the actions of traditional speech act theory. The Backward Communicative Functions indicate how the current utterance relates to the previous dialog, such as accepting a proposal confirming understanding or answering a question. Utterance features include information about an utterance's form and content such as whether an utterance concerns the communication process itself or deals with the subject at hand. The latter popular dialog act tag annotation scheme, MRDA, focuses on multiparty meetings. While similar to DAMSL, one big difference is that it includes a set of labels for floor management mechanisms, such as *floor grabbing* and *holding*, which are common in meetings. In total it has 11 general (such as question) and 39 specific (such as yes/no question) dialog act tags.

Dialog act tagging is generally framed as an utterance classification problem [25], ([26], among others). The basic approach as taken by [26] is to treat each sentence independently and to employ lexical features in classifiers. Additional features such as prosodic cues have also been successfully used for tagging dialog acts using multilayer perceptrons [27]. The approach taken by [25] is more complex and classifies dialog acts based on lexical, collocational, and prosodic cues, as well as on the discourse coherence of the dialog act sequence. The dialog model is based on treating the discourse structure of a conversation as an HMM and the individual dialog acts as observations emanating from the model states. Constraints on the likely sequence of dialog acts are modeled via a dialog act n-gram. The statistical dialog act grammar is combined with word n-grams, decision trees, and neural networks modeling the idiosyncratic lexical and prosodic manifestations of each dialog act. Note the similarity of this approach with the hybrid dialog act segmentation method described above. There are also more recent studies performing joint dialog act segmentation and tagging [28], [29].

For the CALO-MA project, dialog act tagging is framed as an utterance classification problem using Boosting. More specifically, we built three different taggers.

1) For capturing high-level dialog act tags (statement, question, disruption, floor mechanism, and backchannel): To build this model, we used only lexical features; Using the ICSI corpus, the classification error rate was found to be 22% using manual transcriptions, where the baseline is 42% using the majority class.
2) For detecting action motivators since they are shown to help action item extraction [30]: For this, we considered only suggestion, command, and commitment dialog act tags using only lexical features using manual transcriptions; The performance was 35% F-score where the baseline was 6% by marking all sentences as action motivators.
3) For detecting agreement and disagreement dialog act tags for single-word utterances, such as *yeah* or *okay*: For this task we used prosodic and contextual information using manual transcriptions, which resulted in a performance of 61% compared to the baseline of 36% F-score.

VI. Topic Identification and Segmentation

Identifying topic structure provides a user with the basic information of *what* people talked about *when*. This information can be a useful end product in its own right: user studies show that people ask general questions like *"What was discussed at the meeting?"* as well as more specific ones such as *"What did X say about topic Y?"* [31]. It can also feed into further processing, enabling topic-based summarization, browsing, and retrieval. Topic modeling can be seen as two subtasks:

- *segmentation*, dividing the speech data into topically coherent units (the *"when"* question);
- *identification*, extracting some representation of the topics discussed therein (the *"what"*).

While both tasks have been widely studied for broadcast news (see, e.g., [32]–[34]), the meeting domain poses further challenges and opportunities. Meetings can be much harder to segment accurately than news broadcasts, as they are typically more coherent overall and have less sharp topic boundaries: discussion often moves naturally from one subject to another. In fact, even humans find segmenting meetings hard: [35] found that annotators asked to mark topic shifts over the open-domain ICSI Meeting Corpus did not agree well with each other at all, especially with fine-grained notions of topic; and although [36] did achieve reasonable agreement with coarser-grained topics, even then some meetings were problematic. On the other hand, meetings may have an agenda and other observable topic-related behavior such as note taking, which may provide helpful independent information (and [37] found that inter-annotator agreement could be much improved by providing such information).

The segmentation problem has received more attention, with typical lexical cohesion based approaches focusing on changes in lexical distribution (following text-based methods such as TextTiling [38])—the essential insight being that topic shifts tend to change the vocabulary used, which can be detected by looking for minima in some lexical cohesion metric. Reference [36], for example, used a variant that pays particular attention to chains of repeated terms, an approach followed by [39] and [40], while [41] stuck closer to the original TextTiling approach. Various measures of segmentation accuracy exist; one of the more common is P_k, which gives the likelihood that a segmentation disagrees with the gold standard about whether an arbitrary two points in the dialogue are separated by a topic shift (better segmentation accuracy therefore corresponds to a lower P_k—see [32]). Reference [36]'s essentially unsupervised approach gives P_k between 0.26 and 0.32 on the ICSI Corpus; supervised discriminative approaches can improve this, with [42] achieving 0.22.

Of course, there is more to meeting dialog than the words it contains, and segmentation may be improved by looking beyond lexical cohesion to features of the interaction itself and the behavior of the participants. Reference [37], for example, provided meeting participants with a note-taking tool that allows agenda topics to be marked, and use their interaction with that tool as implicit supervision. We cannot always assume such detailed information is available, however—nor on the existence of an agenda—but simpler features can also help. Reference [36] found that features such as changes in speaker activity, amounts of silence and overlapping speech, and the presence of certain cue phrases were all indicative of topic shifts, and adding them to their approach improved their segmentation accuracy significantly. Reference [43] found that similar features also gave some improvement with their supervised approach, although [39] found this only to be true for coarse-grained topic shifts (corresponding in many cases to changes in the activity or state of the meeting, such as introductions or closing review), and that detection of finer-grained shifts in subject matter showed no improvement.

The identification problem can be approached as a separate step after segmentation: [40] showed some success in using supervised discriminative techniques to classify topic segments according to a known list of existing topics, achieving F-scores around 50%. However, there may be reason to treat the two as joint problems: segmentation can depend on the topics of interest. Reference [37], for example, showed improvement over a baseline lexical cohesion segmentation method by incorporating some knowledge of agenda items and their related words. Reference [44] investigated the use of Latent Semantic Analysis, learning vector-space models of topics and using them as the basis for segmentation, but accuracy was low.

Instead, in CALO-MA, we therefore use a generative topic model with a variant of Latent Dirichlet Allocation [45] to learn models of the topics automatically, without supervision, while simultaneously producing a segmentation of the meeting [46]. Topics are modeled as probability distributions over words, and topically coherent meeting segments are taken to be generated by fixed weighted mixtures of a set of underlying topics. Meetings are assumed to have a Markov structure, with each utterance being generated by the same topic mixture as its predecessor, unless separated by a topic shift, when a new mixture is chosen. By using Bayesian inference, we can estimate not only the underlying word distributions (the topics) but the most likely position of the shifts (the segmentation). The segmentation is then used in the system to help users browse meetings, with the word distributions providing associated keyword lists and word clouds for display. Similarity between distributions can also be used to query for related topics between meetings.

Segmentation performance is competitive with that of an unsupervised lexical cohesion approach (P_k between 0.27 and 0.33 on the ICSI Meeting Corpus) and is more robust to ASR errors, showing little if any reduction in accuracy. The word distributions simultaneously learned (the topic identification models) rate well for coherence with human judges, when presented with lists of their top most distinctive keywords. Incorporating non-lexical discourse features into the model is also possible, and [47] shows that this can further improve segmentation accuracy, reducing P_k in the ICSI corpus for a fully unsupervised model from 0.32 to 0.26.

VII. Action Item and Decision Extraction

Among the most commonly requested outputs from meetings (according to user studies [31], [48]) are lists of the decisions made, and the tasks or action items people were assigned (*action items* are publicly agreed commitments to perform a given task). Since CALO is a personal intelligent assistant, for CALO-MA, keeping track of action items and decisions have special importance. The CALO meetings are also designed to cover many action items, such as organizing an office for a new employee in the example of Fig. 2. Again, we can split the problem into two subtasks:

- *detection* of the task or decision discussion;
- *summarization* or *extraction* of some concise descriptive representation (for action items, typically the task itself together with the due date and responsible party; for decisions, the issue involved and the resolved course of action).

Related work on action item detection from email text approaches it as a binary classification problem, and has shown reasonable performance [49]–[51]: F-scores around 80% are achieved on the task of classifying messages as containing action items or not, and 60% to 70% when classifying individual sentences.

However, applying a similar approach to meeting dialog shows mixed results. Some success has been shown in detecting decision-making utterances in meetings in a constrained domain [52], [53]; features used for classification include lexical cues (words and phrases), prosodic (pitch and intensity), semantic (dialog act tags, temporal expressions) and contextual (relative position within the meeting). [53] achieve F-scores of 60% to 70% for the task of detecting decision-making utterances from within a manually selected summary set. On the other hand, when the task is to detect utterances from within an entire meeting, and when the domain is less constrained, accuracy seems to suffer significantly: [54] achieved F-scores only around 30% when detecting action item utterances over the ICSI Meeting Corpus using similar features.

The reason for this may lie in the nature of dialog: whereas tasks or decisions in text tend to be contained within individual sentences, this is seldom true in speech. Tasks are defined incrementally, and commitment to them is established through interaction between the people concerned; cues to their detection can therefore lie as much in the discourse structure itself as in the content of its constituent sentences. CALO-MA therefore takes a structural approach to detection: utterances are first classified according to their role in the commitment process (e.g., task definition, agreement, acceptance of responsibility, issue under discussion, decision made) using a suite of binary SVM classifiers, one for each possible utterance role, and then action item or decision discussions are detected from patterns of these roles using a binary classifier or a probabilistic graphical model. This structural approach significantly improves detection performance. The detectors used in CALO-MA are trained on mutiparty meeting data from the AMI Meeting Corpus. On manual transcripts, the detectors achieve F-scores around 45% for action items [55] and 60% for decisions [56]. This is a significant improvement over the baseline results obtained with non-structured detectors trained on the same data, which achieve 37% and 50% F-scores, respectively. When ASR output is used there is a drop in detection performance, but this is still above the baseline. A real-time decision detector does not perform significantly worse than the offline version [57]. Here, the detector runs at regular and frequent intervals during the meeting. It reprocesses recent utterances in case a decision discussion straddles these and brand new utterances, and it merges overlapping hypothesized decision discussions, and removes duplicates.

Once the relevant utterances or areas of discussion have been detected, we must turn to the summarization or extraction problem, but this has received less attention so far. On email text, [49] used a parsing-based approach, building logical form representations from the related sentences and then generating descriptions via a realizer. With spoken language and ASR output, the parsing problem is of course more difficult, but in CALO-MA we investigated a similar (although slightly shallower) approach: a robust parser is used to extract candidate fragments from a word confusion network classified as task- or decision-related [55], [58]. These are then ranked by a regression model learned from supervised training data (as we explain below, this ranking allows the meeting browser to display several hypotheses to the user). Results were encouraging for extracting due dates, but task descriptions themselves are more problematic, often requiring deeper linguistic processing such as anaphora and ellipsis resolution. Identifying the responsible party requires a slightly different approach: mention of the person's name is rare, it is usually expressed via "I" or "you" rather than a full name, so parsing or entity extraction cannot get us very far. Much more common are the cases of speakers volunteering themselves, or asking for their addressee's commitment, so the task becomes one of speaker and/or addressee identification as explained in the next section.

In CALO-MA, the user can access the summaries extracted from the detected decisions and action items via the meeting browser. The browser presents the extracted information in a convenient and intuitive manner and, most importantly, allows the user to make modifications or corrections when the generated output falls short of the mark. The hypotheses corresponding to properties of action items and decisions—such as their descriptions, timeframes, or the decisions made—are highlighted at various degrees of illumination, according to the level of confidence given to each hypothesis by the classifiers. A user can click on the correct hypothesis, edit the proposed text, add action items to a to-do list, or delete an erroneous action item or decision discussion altogether. Any of these actions will feed back to the detection and extraction models, which can be retrained on the basis of this feedback.

VIII. Reference and Addressee Resolution

An important intermediate step in the analysis of meeting conversations is to determine the entities and individuals to which the participants are speaking, listening and referring. This means predicting individuals' focus of attention, identifying the addressees of each utterance, and resolving any linguistic or gestural references to individuals or present objects. In the CALO-MA system, one particular concern is the word "you," which can refer to a single individual, a group, or can be generic, referring to nobody in particular. As action items are often assigned to "you," the system must determine referentiality and (if applicable) the actual addressee reference in order to determine the owner of an action item.

Recent research in this area has shown the importance of multimodality—that is, of visual as well as linguistic information. For example, [59] used a combination of lexical features of the utterance (e.g., personal, possessive, and indefinite pronouns, and participant names) and manually annotated gaze features for each participant in order to detect addressee(s) in four-person meetings using Bayesian Networks. Here, using only utterance features gave 53% accuracy, speaker gaze 62%, all participants' gaze, 66%, and their combination, 71%.

In the CALO-MA project, our approach to automatically resolving occurrences of *you* is dividing the problem into three tasks [60], [61]: 1) distinguish between generic versus referential *you* (GVR); 2) referential singular versus plurals (RSVP); and 3) identify the individual addressee for the referential singulars (IA). Our experimental data-set comes from the AMI corpus and is composed of around 1000 utterances which contain the word *you*. We experimented with Bayesian Networks, using linguistic and visual features, both manually annotated and fully automatic. For the former, features are derived from manual transcripts and AMI Focus of Attention (FOA) annotations,[1] while for the latter, they are generated from ASR transcripts and with a six degree-of-freedom head tracker.

For each *you*-utterance, we computed visual features to indicate at which target each participant's gaze was directed the longest during different periods of time. The target could be any of the other participants, or the white-board/projector screen at the front of the meeting room, while the different time periods included each third of the utterance, the utterance as a whole, and the periods from 2 seconds before until 2 seconds after the start time of the word *you*. A further feature indicated with

[1]A description of the FOA labeling scheme is available from the AMI Meeting Corpus website: http://corpus.amiproject.org/documentations/guidelines-1

whom the speaker spent most time sharing a mutual gaze over the utterance as a whole.

Our generic features include firstly, features which encode structural, durational, lexical and shallow syntactic patterns of the *you*-utterance. Second, there are Backward Looking (BL)/ Forward Looking (FL) features, which express the similarity or distance (e.g., ratio of common words, time separation) between the *you*-utterance and the previous/next utterance by each non-speaker. Others include the BL/FL speaker order and the number of speakers in the previous/next five utterances. Finally, for the manual systems, we also use the AMI dialogue acts of the *you*-utterances, and of the BL/FL utterances.

Our most recent results are as follows: in a tenfold cross-validation using manual features, the system achieves accuracy scores of 88%, 87%, and 82% in the GVR, RSVP and IA tasks, respectively, or 75% on the (five-way) combination of all three. A fully automatic system gives accuracies of 83%, 87%, and 77%, (all higher than majority class baselines, $p < 0.05$). Taking away FL features (as required for a fully online system) causes a fairly large performance drop in the IA task—9% for the manual system, and 8% for the automatic—but less in the other two. Although at this point the actual CALO-MA system is not able to process visual information, our experiments show that visual features produce a statistically significant improvement in the IA and RSVP tasks. The speaker's visual features are most predictive in the IA task, and it seems that when listeners look at the white-board/projector screen, this is indicative of a referential plural. Of the linguistic features, sentential, especially those concerning lexical properties help in the GVR and RSVP tasks. Fewer speaker changes correlate more with plural than singular referential and in the IA task, FL/BL speaker order is predictive. As for dialogue acts, in the GVR tasks, a *you* in a question is more likely to be referential, and in the RSVP task, questions are more likely to have an individual addressee, and statements, plural addressees.

IX. Summarization

A recent interest for CALO-MA is summarizing meetings. The goal of summarization is to create a shortened version of a text or speech while keeping important points. While textual document summarization is a well-studied topic, speech summarization (and in particular meeting summarization) is an emerging research area, and apparently very different from text or broadcast news summarization. The aim is basically filtering out the unimportant chit-chat from contentful discussions. While hot-spot detection, action item extraction, dialog act tagging, and topic segmentation and detection methods can be used to improve summarization, there are also preliminary studies using lexical, acoustic, prosodic, and contextual information.

In text or broadcast news summarization, the dominant approach is extractive summarization where "important" sentences are concatenated to produce a summary. For meeting summarization, it is not clear what constitutes an important utterance. In an earlier study [62], the sentences having the highest number of frequent content words are considered to be important. Using the advances in written and spoken document extractive summarization [63], some recent studies focused on feature-based classification approaches [64], while others mainly used maximum marginal relevance (MMR) [65] for meeting summarization [64], [66]. MMR iteratively selects utterances most relevant to a given query, which is expected to encode the user's information need, while trying to avoid utterances redundant to the already-selected ones. Due to the lack of a query, the common approach for meetings has been to use the centroid vector of the meeting as the query [64].

In CALO-MA, our summarization work mainly focused on investigating the boundaries of extractive meeting summarization in terms of different evaluation measures [67]. The most widely used is ROUGE [68], a metric that compares the produced summary against a set of reference summaries using word n-gram overlaps. We proposed to compute a simple baseline for summarization that consists in selecting the longest utterances in the meeting, which is more challenging to beat than the random baseline which selects random utterances. We also proposed a method to compute "oracle" summaries that extracts the set of sentences maximizing the ROUGE performance measure. For example, on the ICSI meeting corpus selecting the longest sentences yields a ROUGE-1 score of 0.15 (all scores are obtained on manual transcriptions), the oracle performs at 0.31 and a one of the most popular method for summarization, MMR, performs at 0.17. Improvements over the MMR system using keyphrases instead of words to represent the information increases ROUGE-1 to 0.20 [69] and a different model maximizing information recall (presented in [70]) performs at 0.23. Nevertheless, we observed that even the oracle summaries did not match the human capability for abstraction because they tend to stack up many unrelated facts. Hence, another trend is to use the sentences selected in the summaries as starting point for browsing the meetings. This helps users recontextualize the information and improve their ability to locate information as shown by [71]. To this end, in [69], we proposed a user interface for improving the capture of a user's information need by presenting automatically extracted keyphrases that can be refined and used to generate summaries for meeting browsing.

X. Conclusion and Future Work

We have presented a system for automatic processing of tasks involving multiparty meetings. Progress in these tasks, from low-level transcription to higher-level shallow understanding functions, such as action item extraction and summarization, has a potentially enormous impact on human productivity in many professional settings. However, there are practical and technical difficulties. In practice, people are not used to instrumented (virtual) meeting rooms. Technically, most higher level semantic understanding tasks are only vaguely defined and the annotator agreements are still very low. User feedback with support for adaptive training is critical for customizing the applications for individual use.

Further integration of these tasks and multiple potential modalities, such as video, or digital pen and paper, is part of the future work. Furthermore, meta information such as project related documentation or emails may be exploited for better performance. Another interesting research direction would be processing aggregate of meetings, tracking the topics, participants, and action items.

REFERENCES

[1] "DARPA cognitive agent that learns and organizes (CALO) project." [Online]. Available: http://www.ai.sri.com/project/CALO

[2] "DARPA perceptive assistant that learns (PAL) program." [Online]. Available: http://www.darpa.mil/ipto/programs/pal/pal.asp

[3] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Proc. ICSLP*, Denver, CO, 2002.

[4] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI meeting project: Resources and research," in *Proc. ICASSP*, Montreal, QC, Canada, 2004.

[5] "Augmented multi-party interaction," [Online]. Available: http://www.amiproject.org

[6] "Computers in the human interaction loop," [Online]. Available: http://chil.server.de

[7] "Rich transcription evaluations," [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2007

[8] G. Tur, A. Stolcke, L. Voss, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Hakkani-Tür, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, S. Peters, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang, "The CALO meeting speech recognition and understanding system," in *Proc. IEEE/ACL SLT Workshop*, Goa, India, 2008.

[9] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grézl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," in *Proc. MLMI*, 2005.

[10] T. Hain, L. Burget, J. Dines, G. Garau, V. Wan, M. Karafiat, J. Vepa, and M. Lincoln, "The AMI system for the transcription of speech in meetings," in *Proc. ICASSP*, Honolulu, HI, 2007, pp. 357–360.

[11] G. Tur and A. Stolcke, "Unsupervised language model adaptation for meeting recognition," in *Proc. ICASSP*, Honolulu, HI, 2007, pp. 173–176.

[12] D. Vergri, A. Stolcke, and G. Tur, "Exploiting user feedback for language model adaptation in meeting recognition," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 4737–4740.

[13] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Commun.*, vol. 32, no. 1–2, pp. 127–154, 2000.

[14] J. Kolar, E. Shriberg, and Y. Liu, "Using prosody for automatic sentence segmentation of multi-party meetings," in *Proc. Int. Conf. Text, Speech, Dialogue (TSD)*, Czech Republic, 2006.

[15] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005, pp. 1061–1064.

[16] M. Core and J. Allen, "Coding dialogs with the DAMSL annotation scheme," in *Proc. Working Notes AAAI Fall Symp. Commun. Action in Humans Mach.*, Cambridge, MA, Nov. 1997.

[17] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SigDial Workshop*, Boston, MA, May 2004.

[18] J. Kolar, Y. Liu, and E. Shriberg, "Speaker adaptation of language models for automatic dialog act segmentation of meetings," in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[19] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proc. ICASSP*, Atlanta, GA, May 1996, pp. 405–408.

[20] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Using conditional random fields for sentence boundary detection in speech," in *Proc. ACL*, Ann Arbor, MI, 2005.

[21] S. Cuendet, D. Hakkani-Tür, and G. Tur, "Model adaptation for sentence segmentation from speech," in *Proc. IEEE/ACL SLT Workshop*, Aruba, 2006, pp. 102–105.

[22] M. Zimmerman, D. Hakkani-Tür, J. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu, "The ICSI+ multilingual sentence segmentation system," in *Proc. ICSLP*, Pittsburg, PA, 2006.

[23] U. Guz, D. Hakkani-Tür, S. Cuendet, and G. Tur, "Co-training using prosodic and lexical information for sentence segmentation," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.

[24] G. Tur, "Co-adaptation: Adaptive co-training for semi-supervised learning," in *Proc. ICASSP*, Taipei, Taiwan, 2009.

[25] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Comput. Linguist.*, vol. 26, no. 3, pp. 339–373, 2000.

[26] G. Tur, U. Guz, and D. Hakkani-Tür, "Model adaptation for dialog act tagging," in *Proc. IEEE/ACL SLT Workshop*, 2006, pp. 94–97.

[27] M. Mast, R. Kompe, S. Harbeck, A. Kiessling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, and V. Warnke, "Dialog act classification with the help of prosody," in *Proc. ICSLP*, Philadelphia, PA, Oct. 1996, pp. 1732–1735.

[28] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "Toward joint segmentation and classification of dialog acts in multiparty meetings," in *Proc. MLMI*, Edinburgh, U.K., July 2005.

[29] V. Warnke, R. Kompe, H. Niemann, and E. Nöth, "Integrated dialog act segmentation and classification using prosodic features and language models," in *Proc. Eurospeech*, Rhodes, Greece, Sep. 1997.

[30] F. Yang, G. Tur, and E. Shriberg, "Exploiting dialog act tagging and prosodic information for action item identification," in *Proc. ICASSP*, Las Vegas, NV, 2008, pp. 4941–4944.

[31] A. Lisowska, "Multimodal interface design for the multimodal meeting domain: preliminary indications from a query analysis study," ISSCO, Univ. of Geneva, Tech. Rep. IM2.MDM-11, Nov. 2003.

[32] D. Beeferman, A. Berger, and J. D. Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, no. 1–3, pp. 177–210, 1999.

[33] J. Reynar, "Statistical models for topic segmentation," in *Proc. ACL*, 1999, pp. 357–364.

[34] G. Tur, D. Hakkani-Tür, A. Stolcke, and E. Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Comput. Linguist.*, vol. 27, no. 1, pp. 31–57, 2001.

[35] A. Gruenstein, J. Niekrasz, and M. Purver, , L. Dybkjaer and W. Minker, Eds., "Meeting structure annotation: Annotations collected with a general purpose toolkit," in *Recent Trends in Discourse and Dialogue*. Berlin/Heidelberg: Springer-Verlag, 2007, Text, Speech and Language Technology..

[36] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proc. ACL*, 2003.

[37] S. Banerjee and A. Rudnicky, "Segmenting meetings into agenda items by extracting implicit supervision from human note-taking," in *Proc. IUI*, Honolulu, HI, Jan. 2007, ACM.

[38] M. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, 1997.

[39] P.-Y. Hsueh, J. Moore, and S. Renals, "Automatic segmentation of multiparty dialogue," in *Proc. EACL*, 2006.

[40] P.-Y. Hsueh and J. Moore, "Automatic topic segmentation and labeling in multiparty dialogue," in *Proc. 1st IEEE/ACM Workshop Spoken Lang. Technol. (SLT)*, Palm Beach, Aruba, 2006.

[41] S. Banerjee and A. Rudnicky, "A TextTiling based approach to topic boundary detection in meetings," in *Proc. ICSLP*, Pittsburgh, PA, Sep. 2006.

[42] M. Georgescul, A. Clark, and S. Armstrong, "Word distributions for thematic segmentation in a support vector machine approach," in *Proc. CoNLL*, New York, Jun. 2006, pp. 101–108.

[43] M. Georgescul, A. Clark, and S. Armstrong, "Exploiting structural meeting-specific features for topic segmentation," in *Actes de la 14 ème Conf. sur le Traitement Automatique des Langues Naturelles*, Toulouse, France, Jun. 2007, Association pour le Traitement Automatique des Langues.

[44] A. Popescu-Belis, A. Clark, M. Georgescul, D. Lalanne, and S. Zufferey, , S. Bengio and H. Bourlard, Eds., "Shallow dialogue processing using machine learning algorithms (or not)," in *MLMI, Revised Selected Papers*. New York: Springer, 2005, vol. 3361, Lecture Notes in Computer Science, pp. 277–290.

[45] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[46] M. Purver, K. Körding, T. Griffiths, and J. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *Proc. COLING-ACL*, Sydney, Australia, Jul. 2006, pp. 17–24.

[47] M. Dowman, V. Savova, T. L. Griffiths, K. P. Körding, J. B. Tenenbaum, and M. Purver, "A probabilistic model of meetings that combines words and discourse features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 7, pp. 1238–1248, Sep. 2008.

[48] S. Banerjee, C. Rosé, and A. Rudnicky, "The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing," in *Proc. CHI*, 2005.

[49] S. Corston-Oliver, E. Ringger, M. Gamon, and R. Campbell, "Task-focused summarization of email," in *Proc. ACL Workshop Text Summarization Branches Out*, 2004.

[50] P. N. Bennett and J. Carbonell, "Detecting action-items in e-mail," in *Proc. ACM SIGIR*, Salvador, Brazil, Aug. 2005.

[51] P. N. Bennett and J. G. Carbonell, "Combining probability-based rankers for action-item detection," in *Proc. HLT/NAACL*, Rochester, NY, Apr. 2007.
[52] A. Verbree, R. Rienks, and D. Heylen, "First steps towards the automatic construction of argument-diagrams from real discussions," in *Proc. 1st Int. Conf. Comput. Models of Argument, September 11 2006, Frontiers Artif. Intell. Applicat.*, 2006, vol. 144, pp. 183–194, IOS press.
[53] P.-Y. Hsueh and J. Moore, "What decisions have you made?: Automatic decision detection in meeting conversations," in *Proc. NAACL/HLT*, Rochester, NY, 2007.
[54] W. Morgan, P.-C. Chang, S. Gupta, and J. M. Brenier, "Automatically detecting action items in audio meeting recordings," in *Proc. SIGdial Workshop Discourse and Dialogue*, Sydney, Australia, Jul. 2006.
[55] M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, S. Noorbaloochi, and S. Peters, "Detecting and summarizing action items in multi-party dialogue," in *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, Sep. 2007.
[56] R. Fernández, M. Frampton, P. Ehlen, M. Purver, and S. Peters, "Modelling and detecting decisions in multi-party dialogue," in *Proc. 9th SIGdial Workshop Discourse and Dialogue*, Columbus, OH, 2008.
[57] M. Frampton, J. Huang, T. H. Bui, and S. Peters, "Real-time decision detection in multi-party dialogue," in *Proc. EMNLP*, Singapore, Aug. 2009.
[58] R. Fernández, M. Frampton, J. Dowding, A. Adukuzhiyil, P. Ehlen, and S. Peters, "Identifying relevant phrases to summarize decisions in spoken meetings," in *Proc. Interspeech*, Brisbane, Australia, 2008.
[59] N. Jovanovic, R. op den Akker, and A. Nijholt, "Addressee identification in face-to-face meetings," in *Proc. EACL*, Trento, Italy, 2006, pp. 169–176.
[60] M. Frampton, R. Fernández, P. Ehlen, M. Christoudias, T. Darrell, and S. Peters, "Who is "you"? Combining linguistic and gaze features to resolve second-person references in dialogue," in *Proc. EACL*, 2009.
[61] M. Purver, R. Fernández, M. Frampton, and S. Peters, "Cascaded lexicalised classifiers for second-person reference resolution," in *Proc. SIGDIAL Meeting Discourse and Dialogue*, London, U.K., 2009.
[62] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, "Meeting browser: Tracking and summarizing meetings," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, Jun. 1998.
[63] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005.
[64] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005.
[65] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. ACM SIGIR*, Melbourne, Australia, 1998.
[66] S. Xie and Y. Liu, "Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization," in *Proc. ICASSP*, Las Vegas, NV, 2008, pp. 4985–4988.
[67] K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tür, "Packing the meeting summarization knapsack," in *Proc. Interspeech*, Brisbane, Australia, 2008.
[68] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Text Summarization Workshop*, 2004.
[69] K. Riedhammer, B. Favre, and D. Hakkani-Tür, "A keyphrase based approach to interactive meeting summarization," in *Proc. IEEE/ACL SLT Workshop*, Goa, India, 2008.
[70] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tür, "A global optimization framework for meeting summarization," in *Proc. IEEE ICASSP*, Taipei, Taiwan, 2009, pp. 4769–4772.
[71] G. Murray, T. Kleinbauer, P. Poller, S. Renals, J. Kilgour, and T. Becker, "Extrinsic summarization evaluation: A decision audit task," in *Proc. MLMI*, Utrecht, The Netherlands, 2008.

Gokhan Tur (M'01–SM'05) is currently with the Speech Technology and Research Lab of SRI International, Menlo Park, CA. From 2001 to 2005, he was with AT&T Labs-Research, Florham Park, NJ.

Dr. Tur is an Associate Editor of the IEEE Transactions on Speech and Audio Processing and was a member of IEEE SPS Speech and Language Technical Committee (SLTC).

Andreas Stolcke (M'96–SM'05) received the Ph.D. degree in computer science from the University of California, Berkeley, in 1994.

He is a Senior Research Engineer at the Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, and at the International Computer Science Institute (ICSI), Berkeley, CA.

Lynn Voss received the MBA degree from University of Phoenix, Phoenix, CA

He is with the Engineering and Systems Division (ESD), SRI International, Menlo Park, CA. He is the project manager for the CALO-MA project interacting with both the research and engineering teams.

Stanley Peters is the Director of the Center for the Study of Language and Information (CSLI), Stanford, CA, and a Professor at the Linguistics Department at Stanford University.

Dilek Hakkani-Tür (S'00–M'01–SM'05) is a Senior Research Scientist at the ICSI, Berkeley, CA. From 2001 to 2006, she was with the AT&T Labs-Research, Florham Park, NJ.

Dr. Hakkani-Tür was an Associate Editor of the IEEE Transactions on Speech and Audio Processing and is a member of the IEEE SPS SLTC.

John Dowding is with CSLI, Stanford University, Stanford, CA, and NASA.

Benoit Favre received the Ph.D. degree from the University of Avignon, Avignon, France, in 2007.

Until 2009, he was a Postdoctoral Researcher at ICSI, Berkeley, CA. He is currently a Research Engineer with LIUM in France.

Raquel Fernández received the Ph.D. degree from the University of Potsdam, Potsdam, Germany.

Until 2009, she was a Postdoctoral Researcher at CSLI, Stanford, University, Stanford, CA. She is currently with University of Amsterdam, Amsterdam, The Netherlands.

Matthew Frampton received the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K.

Since 2007 he has been an Engineering Research Associate at CSLI, Stanford, University, Stanford, CA.

Mike Frandsen is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Clint Frederickson is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Martin Graciarena received the Ph.D. degree from the University of Buenos Aires, Buenos Aires, Argentina, in 2009.

Since 1999, he has been with the STAR Lab, SRI International, Menlo Park, CA.

Donald Kintzing is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Kyle Leveque is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Shane Mason is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

John Niekrasz is currently a Research Fellow with the University of Edinburgh, Edinburgh, U.K. He was with CSLI, Stanford, University, Stanford, CA, when this work was done.

Matthew Purver is currently a Senior Research Fellow with Queen Mary University of London, London, U.K.

Korbinian Riedhammer is currently with the University of Erlangen, Erlangen, Germany. He was with ICSI, Berkeley, CA when this work was done.

Elizabeth Shriberg received the Ph.D. degree from the University of California, Berkeley, in 1994.

She is a Senior Researcher at both the Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, and the ICSI, Berkeley, CA.

Jing Tien is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Dimitra Vergyri (M'06) received the Ph.D. degree from Johns Hopkins University, Baltimore, MD, in 2000.

She is currently with the STAR Lab, SRI International, Menlo Park, CA.

Dr. Vergyri is an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.

Fan Yang received the Ph.D. degree from the Oregon Graduate Institute, Beaverton, in 2008.

He is currently with Next IT Corp. He was with SRI International, Menlo Park, CA, when this work was done.

The CALO Meeting Assistant System

Gokhan Tur, *Senior Member, IEEE*, Andreas Stolcke, *Senior Member, IEEE*, Lynn Voss, Stanley Peters, Dilek Hakkani-Tür, *Senior Member, IEEE*, John Dowding, Benoit Favre, Raquel Fernández, Matthew Frampton, Mike Frandsen, Clint Frederickson, Martin Graciarena, Donald Kintzing, Kyle Leveque, Shane Mason, John Niekrasz, Matthew Purver, Korbinian Riedhammer, Elizabeth Shriberg, Jing Tien, Dimitra Vergyri, *Member, IEEE*, and Fan Yang

***Abstract*—The CALO Meeting Assistant (MA) provides for distributed meeting capture, annotation, automatic transcription and semantic analysis of multiparty meetings, and is part of the larger CALO personal assistant system. This paper presents the CALO-MA architecture and its speech recognition and understanding components, which include real-time and offline speech transcription, dialog act segmentation and tagging, topic identification and segmentation, question-answer pair identification, action item recognition, decision extraction, and summarization.**

***Index Terms*—Multiparty meetings processing, speech recognition, spoken language understanding.**

I. Introduction

In most organizations, staff spend many hours each week in meetings, and technological advances have made it possible to routinely record and store meeting data. Consequently, automatic means of transcribing and understanding meetings would greatly increase productivity of both meeting participants and nonparticipants. The meeting domain has a large number of subdomains including judicial and legislative proceedings, lectures, seminars, board meetings, and a variety of less formal group meeting types. All these meeting types could benefit immensely from the development of automatic speech recognition (ASR), understanding, and information extraction technologies that could be linked with a variety of online information systems.

In this paper we present the meeting recognition and understanding system for the CALO Meeting Assistant (CALO-MA) project. CALO-MA is an automatic agent that assists meeting participants, and is part of the larger CALO [1] effort to build a "Cognitive Assistant that Learns and Organizes" funded under the "Perceptive Assistant that Learns" (PAL) program [2] of the DARPA. The focus of CALO in general is "learning in the wild", or continuous improvement of the system's abilities as a result of system use.

Significant anecdotal evidence suggests that companies have collected a wide range of meeting data over the years. Broadcast data and recorded conferences are also available. Further, public data such as council meetings and government proceedings is often accessible. However, little of the data is useful for research purposes. First, privacy and competitive advantage requirements preclude the use of most business meeting data. Privacy, copyright, and signal quality bar the use of most other types of "found" data as well. Rather, collection with the specific intent of providing a basis for research is required.

Projects initiated at CMU [3] and ICSI [4] in the late 1990s and early 2000s collected substantial meeting corpora and investigated many of the standard speech processing tasks on this genre. Subsequently, several large, interdisciplinary, and multisite government-funded research projects have investigated meetings of various kinds. The AMI (Augmented Multiparty Interaction) Consortium [5] project concentrates on conference-room meetings with small numbers of participants, similar to the CALO-MA system. The CHIL (Computers in the Human Interaction Loop) project [6] collected a series of lectures dominated by a single presenter with shorter question/answer portions, as well as some "interactive" lectures involving smaller groups. AMI and CHIL also produced corpora of time-synchronized media, generally including close-talking and far-field microphones, microphone arrays, individual and room-view video cameras, and output from slide projectors and electronic whiteboards.

Starting in 2002, the annual NIST Rich Transcription (RT) Evaluations [7] have become a driving force for research in meeting processing technology, with substantial performance improvements in recent years. In order to promote robustness and domain independence, the NIST evaluations cover several meeting genres and topics, ranging from largely open-ended, interactive chit-chat, to topic-focused project meetings and technical seminars dominated by lecture-style presentations. However, NIST evaluates only the speech recognition and speaker diarization systems, with a focus on recognition from multiple distant table-top microphones. Higher level semantic understanding tasks ranging from dialog act tagging to summarization are only indirectly evaluated in the framework of larger meeting processing projects.

In the following sections we discuss the speech-based component technologies contributing to CALO-MA, including speech recognition, dialog act segmentation and tagging, topic

Manuscript received March 03, 2009; revised October 25, 2009. This work was supported in part by the DARPA CALO funding (FA8750-07-D-0185, Delivery Order 0004), in part by the European Union IST Integrated Project AMIDA FP6-506811, and in part by the Swiss National Science Foundation through NCCR's IM2 project. The associate editor coordinating the review of this manuscript and approving it for publication was XXXXX XXXXXX.

G. Tur, A. Stolcke, L. Voss, M. Frandsen, C. Frederickson, M. Graciarena, D. Kintzing, K. Leveque, S. Mason, E. Shriberg, J. Tien, D. Vergyri, and F. Yang are with SRI International, Menlo Park, CA 94025 USA.

S. Peters, R. Fernandez, J. Niekrasz, M. Purver, J. Dowding, and M. Frampton are with CSLI, Stanford University, Stanford, CA 94305 USA

D. Hakkani-Tür, B. Favre, and K. Riedhammer are with ICSI, Berkeley, CA 94704 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2009.2038810

- *John Smith*: so we need to arrange an office for joe browning (statement/all)
- *Kathy Brown*: are there special requirements (question/John)
- *Cindy Green*: when is he co- (disruption/John)
- *John Smith*: yes (affirmation/Kathy) // there are (statement/Kathy)
- *John Smith*: we want him to be close to you (statement/Kathy)
- *Kathy Brown*: okay (agreement/John) // I'll talk to the secretary (commitment/John)
- *Cindy Green*: hold on (floor grabber/all) // wh- when is he coming (question/John)
- *John Smith*: next monday (statement/Cindy)
- *Cindy Green*: uh-huh (backchannel/all)

Action Item: Arrangement of Joe's office location
Owner: Kathy

Decision: Location of Joe's office to be close to Kathy

Summary:

- *John Smith*: so we need to arrange an office for joe browning (statement/all)
- *John Smith*: we want him to be close to you (statement/Kathy)

Fig. 1. An example of meeting data. Dialog act tags and addressed persons are shown in parantheses. This meeting data has one action item and one decision. A brief extractive summary corresponding to this meeting data follows.

segmentation and identification, action item and decision detection, and summarization. We conclude by pointing out research challenges and directions for future work. This paper significantly extends the previous IEEE SLT workshop paper [8] with much more detailed task descriptions, literature surveys, and thorough analyses.

II. CALO-MA Framework

A. Task and Corpora

Speech and language processing technology has advanced such that many types of meeting information can be detected and evaluated—including dialog acts, topics, and action items. For example, Fig. 1 presents an imagined excerpt from a meeting. The speakers and the words spoken are transcribed, along with the dialog acts (listed in parentheses). Dialog act boundaries in a single turn are separated by // tokens. In an agenda-driven meeting, each agenda item can be considered a separate topic. The example shown discusses a particular agenda item (*Arrangements for Joe Browning*). It also contains discussions about action items, due dates, and assignees. Automatically extracting this information from the signals would provide significant advantages in applications ranging from meeting browsing and search to summarization, minutes generation, and automated meeting assistants.

Apart from being highly usable in its present form, the CALO-MA system presents an experimentation platform to support ongoing research in natural language and speech processing technologies. The nature of multiparty interactions and the extreme variability found in meeting genres make this one of the most challenging domains for speech and natural language processing today.

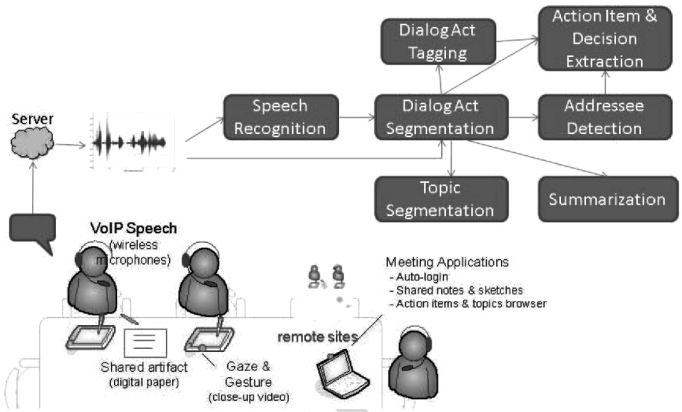

Fig. 2. The CALO-MA conceptual framework.

Fig. 2 presents the overall CALO-MA framework. CALO-MA supports multiparty meetings with a variety of information capture and annotation tools. Meetings are recorded via client software running on participants' laptop computers. The system is aware of each participant's identity. Meetings may be geographically distributed as long as a broadband Internet connection to the server is available (a phone-based interface is being developed as well). The client software captures the participants' audio signals, as well as optional handwriting recorded by digital pens. During the meeting, the participants have a real-time transcript available to which annotations may be attached. Real-time chat via keyboard input is also supported. All interactions are logged in a database, and at the conclusion of the meeting various further automatic annotation and interpretation technologies are initiated, for later browsing via a web-based interface.

The speech utterances are delivered to the server which performs real-time and offline tasks. First, the utterance is recognized, and segmented into sentences. This is the primary input for most of the following tasks. Then the sentences are assigned to dialog act tags and addressee information, which are used to improve action item and decision extraction components. Finally, the meeting is segmented into topically coherent segments and summarized according to parameters set by the user.

The CALO-MA system is used for collecting eight sequences of five meetings, each about 40 minutes long on average. The experimental results provided in the following sections either use CALO, ICSI, and/or AMI meetings corpora.

B. Meeting Capture

An early goal of the CALO-MA project was to allow lightweight data capture. Because of this, highly instrumented rooms were avoided in favor of running on each individual's Java Runtime enabled computer. Meeting participants can attend meetings by using a desktop or laptop running Windows® XP/Vista, Linux, or Mac OS X Leopard. Servers for data transport, data processing, and meeting data browsing run on Windows and Linux environments. If scaling is an issue, additional servers can be integrated into the framework to load balance the various tasks. New efforts will allow participants to conference into a meeting via a bridge between the data transport server and the public switched telephone network (PSTN).

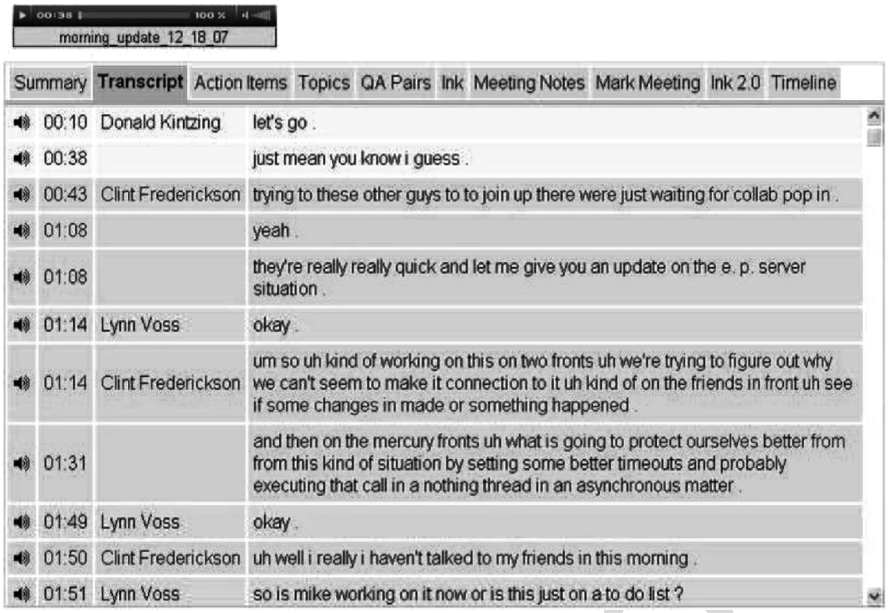

Fig. 3. Snapshot from the CALO-MA offline meeting browser.

During a meeting, client software sends Voice over Internet Protocol (VoIP) compressed audio data to the server either when energy thresholds are met or when a hold-to-talk mechanism is enabled. The data transport server splits the audio: sending one stream to meeting data processing agents for preprocessing and processing the data. Any processing agents that operate in real-time send their data back to the data transport server that relays the data back to the meeting participants.

C. Integration With Other CALO Components

Both during the live meeting and at any time after the meeting, the meeting data transport server makes available all meeting data to interested parties using XML-RPC interfaces. This allows both local and distributed users and processing agents to access the data in a language-neutral way. Meeting processing agents that are order dependent register with a meeting post processor framework to ensure that processing order is enforced (e.g., speech transcription, prosodic feature detection, dialog act recognition, action item detection, decision detection, topic boundary detection, meeting summarization, and email notification to meeting participants) and processing load is balanced.

Any CALO components outside the meeting processing framework (including the meeting browser) can send XML-RPC queries to the meeting data transport server. Those components can then perform further integration with user desktop data to facilitate additional machine learning (a focus of many other CALO processes) or present other visualizations of the data to the user.

D. Meeting Browser

After the meeting has been fully processed, email is sent out to all meeting participants. This email includes a static version of the meeting data and a link to a website where the data can be browsed dynamically from any Internet-enabled device as shown in Fig. 3. Once connected to the browser, the user can select a meeting to review and browse any of the data: both user-generated (e.g., shared files and notes) and auto-generated (e.g., detected action items and summaries). As all data is time stamped, a user can click on any data element and bring up the corresponding section of the transcript to read what was being discussed at that time. To overcome any speech transcription errors, all transcript segments can be selected for streaming audio playback. We are currently working on a framework that will allow the users to correct transcription errors.

III. Speech Recognition

A crucial first step toward understanding meetings is transcription of speech to text (STT). The NIST RT evaluations have driven the research in this field, starting out with roundtable, or "conference," meetings and recently adding other meeting genres such as lectures (mainly one person speaking) and "coffee breaks" (informal discussions following lectures). The best meeting recognition systems typically make use of the full arsenal of state-of-the-art STT techniques employed in recognizing other kinds of speech. Here, we give a brief summary with special emphasis on the approaches that deal specifically with meeting data.

At the front end, these techniques include speaker-level vocal tract length normalization, cepstral feature normalization, heteroscedastic linear discriminant feature transforms, and nonlinear discriminant transforms effected by multilayer perceptrons (MLPs). Hidden Markov model (HMM) acoustic models based on clustered Gaussian mixtures are trained using discriminative criteria such as minimum phone error (MPE) and/or a related feature-level transform (fMPE). An interesting challenge for acoustic modeling is that only relatively small amounts of actual meeting data (about 200 hours) are publicly available, compared to thousands of hours for other domains. This has engendered much research in techniques to adapt models and data from other domains for this task. For example, discriminative versions of Bayesian maximum *a posteriori* adaption (MAP) are used for Gaussian training and fMPE transform estimation

and feature estimation MLPs that were pretrained on large background corpora are retargeted to the meeting domain by limited retraining [9]. Feature transforms are also used to bridge differences in signal bandwidth between background and target data [10]. All state-of-the-art systems proceed in batch mode, decoding meetings in their entirety multiple times for the purpose of unsupervised acoustic adaptation (using maximum likelihood linear regression (MLLR)), and also for the purpose of combining multiple hypothesis streams, often based on subsystems that differ in the features or models used so as to generate complementary information. For example, a system might recognize speech based on both Mel cepstral coefficients and perceptual linear prediction cepstrum, and combine the results.

Recognizers also use large n-gram language models drawn from a range of corpora: telephone speech for conversational speaking style, technical proceedings for coverage of lecture topics, broadcast transcripts and news texts for general topic coverage, as well as smaller amounts of actual meeting transcripts available from the research projects mentioned earlier. Data are also culled from the World Wide Web using targeted search to find conversational-style transcripts as well as relevant subject matter. Source-specific component language models (LMs) are then trained and interpolated with weights optimized to maximize likelihood on representative sample data.

Even with close-talking microphones, crosstalk between channels (especially with lapel-type microphones) can be a significant problem since words from the "wrong" speakers end up being inserted into a neighboring speaker's transcript. This problem has been addressed with echo-cancellation type algorithms or cross-channel features that allow crosstalk to be suppressed during speech/nonspeech segmentation. Word error rates (WERs) on recent NIST evaluation data are in the 20% to 30% range for close-talking microphones. In the CALO-MA system, the audio stream from each meeting participant is transcribed into text by using two separate recognition systems. A real-time recognizer generates "live" transcripts with 5 to 15 seconds of latency for immediate display (and possible interactive annotation) in the CALO-MA user interface. Once the meeting is concluded, a second, offline recognition system generates a more accurate transcript for later browsing and serves as the input to the higher-level processing step described in the following sections.

The offline recognition system is a modified version of the SRI-ICSI NIST meeting recognizer [9]. It performs a total of seven recognition passes, including acoustic adaptation and language model rescoring, in about 4.2 times real-time (on a 4-core 2.6 GHz Opteron server). The real-time recognition systems consists of an online speech detector, causal feature normalization and acoustic adaptation steps, and a sub-real-time trigram decoder. On a test set where the offline recognizer achieves a word error rate (WER) of 26.0%, the real-time recognizer obtains 39.7% on the CALO corpus. We have also demonstrated the use of unsupervised adaptation methods for about 10% relatively better recognition using the recognition outputs of previous meetings [11]. Recent work includes exploiting user feedback for language model adaptation in speech recognition, by allowing users to modify the meeting transcript from the meeting browser [12].

IV. Dialog Act Segmentation

Output from a standard speech recognition system typically consists of an unstructured stream of words lacking punctuation, capitalization, or formatting. Sentence segmentation for speech enriches the output of standard speech recognizers with this information. This is important for the readability of the meetings in the CALO-MA offline meeting browser and the following processes which use sentences as the processing units, such as action item extraction or summarization.

Previous work on sentence segmentation used lexical and prosodic features from news broadcasts and spontaneous telephone conversations [13]. Work on multiparty meetings has been more recent (e.g., [14] and [15]). In the meetings domain, what constitutes a sentential unit (called as a dialog act unit) is defined by the DAMSL (Dialog Act Markup in Several Layers) [16] and MRDA (Meeting Recorder Dialog Act) [17] standards as explained in the next section.

For dialog act segmentation, similar to the approaches taken for sentence segmentation, the CALO-MA system exploits lexical and prosodic information (such as the use of pause duration [15] and others [18]). Dialog act segmentation is treated as a binary boundary classification problem where the goal is finding the most likely word boundary tag sequence, $T = t_1, \ldots, t_n$, given the features, $F = f_1, \ldots, f_n$ for n words:

$$\text{argmax}_T P(T \mid F)$$

To this end, for CALO-MA, we use hybrid models combining both generative and discriminative classification models. As the generative model, we use the hidden event language model, as introduced by [19]. In this approach, sentence boundaries are treated as the hidden events and the above optimization is simply done by the Viterbi algorithm using only lexical features, i.e., language model. Later, a discriminative classification approach is used to build hybrid models to improve this approach by using additional prosodic features [13]. The posterior probabilities obtained from the classifier are simply converted to state observation likelihoods by dividing to their priors following the well-known Bayes rule:

$$\text{argmax}_T \frac{P(T \mid F)}{P(T)} = \text{argmax}_T P(F \mid T).$$

For the ICSI corpus, using only lexical or prosodic information with manual transcriptions resulted in around 48% NIST error rate, which is the number of erroneous boundaries divided by the number of sentences (i.e., a baseline of 100% error rate). Using the hybrid approach to combine these information sources resulted in 33% NIST error rate, a significant improvement. The performance drops by 20%–25% relatively when ASR output is used instead, where the WER is around 35%. For the CALO corpus, using only lexical information resulted in 57% NIST error rate, and this was reduced to 39% using the hybrid approach with manual transcriptions.

With the advances in discriminative classification algorithms, other researchers also tried using Conditional Random Fields (CRFs) [20], Boosting [21], and hybrid approaches using Boosting and Maximum Entropy classification algorithms [22].

Our recent research has focused on model adaptation methods for improving dialog act segmentation for meetings using spontaneous telephone conversations, and speaker-specific prosodic [18] and lexical modeling [21].

In order to exploit the sentence boundary tagged meeting corpora as obtained from other projects such as ICSI and AMI, we also proposed model adaptation [21] and semi-supervised learning techniques, such as co-training [23] and co-adaptation [24], for this task. Model adaptation reduced the NIST error rate for the CALO corpus to 30%.

V. Dialog Act Tagging

A dialog act is a primitive abstraction or an approximate representation of the illocutionary force of an utterance, such as *question* or *backchannel*. Dialog acts are designed to be task independent. The main goal of dialog acts is to provide a basis for further discourse analysis and understanding.

For CALO-MA, dialog acts are very useful for most of the following processes, such as using action motivators for action item detection or using question/statement pairs for addressee detection. Note that dialog acts can be organized in a hierarchical fashion. For instance, statements can be further subcategorized as *command* or *agreement*. Depending on the task, which will use the DA tags, the granularity of the tags is determined. Furthermore, dialog act tags can be used for correct punctuation such as period versus question marks.

The communicative speech act theory goes back to the 1960s, and there are a number of contemporary dialog act sets in the literature, such as DAMSL [16] and MRDA [17], as mentioned in the previous section. DAMSL focuses on providing multiple layers of dialog act markup. Each layer allows multiple communicative functions of an utterance to be labeled. The Forward Communicative Functions consist of a taxonomy in a style similar to the actions of traditional speech act theory. The Backward Communicative Functions indicate how the current utterance relates to the previous dialog, such as accepting a proposal confirming understanding or answering a question. Utterance features include information about an utterance's form and content such as whether an utterance concerns the communication process itself or deals with the subject at hand. The latter popular dialog act tag annotation scheme, MRDA, focuses on multiparty meetings. While similar to DAMSL, one big difference is that it includes a set of labels for floor management mechanisms, such as *floor grabbing* and *holding*, which are common in meetings. In total it has 11 general (such as question) and 39 specific (such as yes/no question) dialog act tags.

Dialog act tagging is generally framed as an utterance classification problem [25], ([26], among others). The basic approach as taken by [26] is to treat each sentence independently and to employ lexical features in classifiers. Additional features such as prosodic cues have also been successfully used for tagging dialog acts using multilayer perceptrons [27]. The approach taken by [25] is more complex and classifies dialog acts based on lexical, collocational, and prosodic cues, as well as on the discourse coherence of the dialog act sequence. The dialog model is based on treating the discourse structure of a conversation as an HMM and the individual dialog acts as observations emanating from the model states. Constraints on the likely sequence of dialog acts are modeled via a dialog act n-gram. The statistical dialog act grammar is combined with word n-grams, decision trees, and neural networks modeling the idiosyncratic lexical and prosodic manifestations of each dialog act. Note the similarity of this approach with the hybrid dialog act segmentation method described above. There are also more recent studies performing joint dialog act segmentation and tagging [28], [29].

For the CALO-MA project, dialog act tagging is framed as an utterance classification problem using Boosting. More specifically, we built three different taggers.

1) For capturing high-level dialog act tags (statement, question, disruption, floor mechanism, and backchannel): To build this model, we used only lexical features; Using the ICSI corpus, the classification error rate was found to be 22% using manual transcriptions, where the baseline is 42% using the majority class.
2) For detecting action motivators since they are shown to help action item extraction [30]: For this, we considered only suggestion, command, and commitment dialog act tags using only lexical features using manual transcriptions; The performance was 35% F-score where the baseline was 6% by marking all sentences as action motivators.
3) For detecting agreement and disagreement dialog act tags for single-word utterances, such as *yeah* or *okay*: For this task we used prosodic and contextual information using manual transcriptions, which resulted in a performance of 61% compared to the baseline of 36% F-score.

VI. Topic Identification and Segmentation

Identifying topic structure provides a user with the basic information of *what* people talked about *when*. This information can be a useful end product in its own right: user studies show that people ask general questions like *"What was discussed at the meeting?"* as well as more specific ones such as *"What did X say about topic Y?"* [31]. It can also feed into further processing, enabling topic-based summarization, browsing, and retrieval. Topic modeling can be seen as two subtasks:

- *segmentation*, dividing the speech data into topically coherent units (the *"when"* question);
- *identification*, extracting some representation of the topics discussed therein (the *"what"*).

While both tasks have been widely studied for broadcast news (see, e.g., [32]–[34]), the meeting domain poses further challenges and opportunities. Meetings can be much harder to segment accurately than news broadcasts, as they are typically more coherent overall and have less sharp topic boundaries: discussion often moves naturally from one subject to another. In fact, even humans find segmenting meetings hard: [35] found that annotators asked to mark topic shifts over the open-domain ICSI Meeting Corpus did not agree well with each other at all, especially with fine-grained notions of topic; and although [36] did achieve reasonable agreement with coarser-grained topics, even then some meetings were problematic. On the other hand, meetings may have an agenda and other observable topic-related behavior such as note taking, which may provide helpful independent information (and [37] found that inter-annotator agreement could be much improved by providing such information).

The segmentation problem has received more attention, with typical lexical cohesion based approaches focusing on changes in lexical distribution (following text-based methods such as TextTiling [38])—the essential insight being that topic shifts tend to change the vocabulary used, which can be detected by looking for minima in some lexical cohesion metric. Reference [36], for example, used a variant that pays particular attention to chains of repeated terms, an approach followed by [39] and [40], while [41] stuck closer to the original TextTiling approach. Various measures of segmentation accuracy exist; one of the more common is P_k, which gives the likelihood that a segmentation disagrees with the gold standard about whether an arbitrary two points in the dialogue are separated by a topic shift (better segmentation accuracy therefore corresponds to a lower P_k—see [32]). Reference [36]'s essentially unsupervised approach gives P_k between 0.26 and 0.32 on the ICSI Corpus; supervised discriminative approaches can improve this, with [42] achieving 0.22.

Of course, there is more to meeting dialog than the words it contains, and segmentation may be improved by looking beyond lexical cohesion to features of the interaction itself and the behavior of the participants. Reference [37], for example, provided meeting participants with a note-taking tool that allows agenda topics to be marked, and use their interaction with that tool as implicit supervision. We cannot always assume such detailed information is available, however—nor on the existence of an agenda—but simpler features can also help. Reference [36] found that features such as changes in speaker activity, amounts of silence and overlapping speech, and the presence of certain cue phrases were all indicative of topic shifts, and adding them to their approach improved their segmentation accuracy significantly. Reference [43] found that similar features also gave some improvement with their supervised approach, although [39] found this only to be true for coarse-grained topic shifts (corresponding in many cases to changes in the activity or state of the meeting, such as introductions or closing review), and that detection of finer-grained shifts in subject matter showed no improvement.

The identification problem can be approached as a separate step after segmentation: [40] showed some success in using supervised discriminative techniques to classify topic segments according to a known list of existing topics, achieving F-scores around 50%. However, there may be reason to treat the two as joint problems: segmentation can depend on the topics of interest. Reference [37], for example, showed improvement over a baseline lexical cohesion segmentation method by incorporating some knowledge of agenda items and their related words. Reference [44] investigated the use of Latent Semantic Analysis, learning vector-space models of topics and using them as the basis for segmentation, but accuracy was low.

Instead, in CALO-MA, we therefore use a generative topic model with a variant of Latent Dirichlet Allocation [45] to learn models of the topics automatically, without supervision, while simultaneously producing a segmentation of the meeting [46]. Topics are modeled as probability distributions over words, and topically coherent meeting segments are taken to be generated by fixed weighted mixtures of a set of underlying topics. Meetings are assumed to have a Markov structure, with each utterance being generated by the same topic mixture as its predecessor, unless separated by a topic shift, when a new mixture is chosen. By using Bayesian inference, we can estimate not only the underlying word distributions (the topics) but the most likely position of the shifts (the segmentation). The segmentation is then used in the system to help users browse meetings, with the word distributions providing associated keyword lists and word clouds for display. Similarity between distributions can also be used to query for related topics between meetings.

Segmentation performance is competitive with that of an unsupervised lexical cohesion approach (P_k between 0.27 and 0.33 on the ICSI Meeting Corpus) and is more robust to ASR errors, showing little if any reduction in accuracy. The word distributions simultaneously learned (the topic identification models) rate well for coherence with human judges, when presented with lists of their top most distinctive keywords. Incorporating non-lexical discourse features into the model is also possible, and [47] shows that this can further improve segmentation accuracy, reducing P_k in the ICSI corpus for a fully unsupervised model from 0.32 to 0.26.

VII. Action Item and Decision Extraction

Among the most commonly requested outputs from meetings (according to user studies [31], [48]) are lists of the decisions made, and the tasks or action items people were assigned (*action items* are publicly agreed commitments to perform a given task). Since CALO is a personal intelligent assistant, for CALO-MA, keeping track of action items and decisions have special importance. The CALO meetings are also designed to cover many action items, such as organizing an office for a new employee in the example of Fig. 2. Again, we can split the problem into two subtasks:

- *detection* of the task or decision discussion;
- *summarization* or *extraction* of some concise descriptive representation (for action items, typically the task itself together with the due date and responsible party; for decisions, the issue involved and the resolved course of action).

Related work on action item detection from email text approaches it as a binary classification problem, and has shown reasonable performance [49]–[51]: F-scores around 80% are achieved on the task of classifying messages as containing action items or not, and 60% to 70% when classifying individual sentences.

However, applying a similar approach to meeting dialog shows mixed results. Some success has been shown in detecting decision-making utterances in meetings in a constrained domain [52], [53]; features used for classification include lexical cues (words and phrases), prosodic (pitch and intensity), semantic (dialog act tags, temporal expressions) and contextual (relative position within the meeting). [53] achieve F-scores of 60% to 70% for the task of detecting decision-making utterances from within a manually selected summary set. On the other hand, when the task is to detect utterances from within an entire meeting, and when the domain is less constrained, accuracy seems to suffer significantly: [54] achieved F-scores only around 30% when detecting action item utterances over the ICSI Meeting Corpus using similar features.

The reason for this may lie in the nature of dialog: whereas tasks or decisions in text tend to be contained within individual sentences, this is seldom true in speech. Tasks are defined incrementally, and commitment to them is established through interaction between the people concerned; cues to their detection can therefore lie as much in the discourse structure itself as in the content of its constituent sentences. CALO-MA therefore takes a structural approach to detection: utterances are first classified according to their role in the commitment process (e.g., task definition, agreement, acceptance of responsibility, issue under discussion, decision made) using a suite of binary SVM classifiers, one for each possible utterance role, and then action item or decision discussions are detected from patterns of these roles using a binary classifier or a probabilistic graphical model. This structural approach significantly improves detection performance. The detectors used in CALO-MA are trained on mutiparty meeting data from the AMI Meeting Corpus. On manual transcripts, the detectors achieve F-scores around 45% for action items [55] and 60% for decisions [56]. This is a significant improvement over the baseline results obtained with non-structured detectors trained on the same data, which achieve 37% and 50% F-scores, respectively. When ASR output is used there is a drop in detection performance, but this is still above the baseline. A real-time decision detector does not perform significantly worse than the offline version [57]. Here, the detector runs at regular and frequent intervals during the meeting. It reprocesses recent utterances in case a decision discussion straddles these and brand new utterances, and it merges overlapping hypothesized decision discussions, and removes duplicates.

Once the relevant utterances or areas of discussion have been detected, we must turn to the summarization or extraction problem, but this has received less attention so far. On email text, [49] used a parsing-based approach, building logical form representations from the related sentences and then generating descriptions via a realizer. With spoken language and ASR output, the parsing problem is of course more difficult, but in CALO-MA we investigated a similar (although slightly shallower) approach: a robust parser is used to extract candidate fragments from a word confusion network classified as task- or decision-related [55], [58]. These are then ranked by a regression model learned from supervised training data (as we explain below, this ranking allows the meeting browser to display several hypotheses to the user). Results were encouraging for extracting due dates, but task descriptions themselves are more problematic, often requiring deeper linguistic processing such as anaphora and ellipsis resolution. Identifying the responsible party requires a slightly different approach: mention of the person's name is rare, it is usually expressed via "I" or "you" rather than a full name, so parsing or entity extraction cannot get us very far. Much more common are the cases of speakers volunteering themselves, or asking for their addressee's commitment, so the task becomes one of speaker and/or addressee identification as explained in the next section.

In CALO-MA, the user can access the summaries extracted from the detected decisions and action items via the meeting browser. The browser presents the extracted information in a convenient and intuitive manner and, most importantly, allows the user to make modifications or corrections when the generated output falls short of the mark. The hypotheses corresponding to properties of action items and decisions—such as their descriptions, timeframes, or the decisions made—are highlighted at various degrees of illumination, according to the level of confidence given to each hypothesis by the classifiers. A user can click on the correct hypothesis, edit the proposed text, add action items to a to-do list, or delete an erroneous action item or decision discussion altogether. Any of these actions will feed back to the detection and extraction models, which can be retrained on the basis of this feedback.

VIII. Reference and Addressee Resolution

An important intermediate step in the analysis of meeting conversations is to determine the entities and individuals to which the participants are speaking, listening and referring. This means predicting individuals' focus of attention, identifying the addressees of each utterance, and resolving any linguistic or gestural references to individuals or present objects. In the CALO-MA system, one particular concern is the word "you," which can refer to a single individual, a group, or can be generic, referring to nobody in particular. As action items are often assigned to "you," the system must determine referentiality and (if applicable) the actual addressee reference in order to determine the owner of an action item.

Recent research in this area has shown the importance of multimodality—that is, of visual as well as linguistic information. For example, [59] used a combination of lexical features of the utterance (e.g., personal, possessive, and indefinite pronouns, and participant names) and manually annotated gaze features for each participant in order to detect addressee(s) in four-person meetings using Bayesian Networks. Here, using only utterance features gave 53% accuracy, speaker gaze 62%, all participants' gaze, 66%, and their combination, 71%.

In the CALO-MA project, our approach to automatically resolving occurrences of *you* is dividing the problem into three tasks [60], [61]: 1) distinguish between generic versus referential *you* (GVR); 2) referential singular versus plurals (RSVP); and 3) identify the individual addressee for the referential singulars (IA). Our experimental data-set comes from the AMI corpus and is composed of around 1000 utterances which contain the word *you*. We experimented with Bayesian Networks, using linguistic and visual features, both manually annotated and fully automatic. For the former, features are derived from manual transcripts and AMI Focus of Attention (FOA) annotations,[1] while for the latter, they are generated from ASR transcripts and with a six degree-of-freedom head tracker.

For each *you*-utterance, we computed visual features to indicate at which target each participant's gaze was directed the longest during different periods of time. The target could be any of the other participants, or the white-board/projector screen at the front of the meeting room, while the different time periods included each third of the utterance, the utterance as a whole, and the periods from 2 seconds before until 2 seconds after the start time of the word *you*. A further feature indicated with

[1]A description of the FOA labeling scheme is available from the AMI Meeting Corpus website: http://corpus.amiproject.org/documentations/guidelines-1

whom the speaker spent most time sharing a mutual gaze over the utterance as a whole.

Our generic features include firstly, features which encode structural, durational, lexical and shallow syntactic patterns of the *you*-utterance. Second, there are Backward Looking (BL)/ Forward Looking (FL) features, which express the similarity or distance (e.g., ratio of common words, time separation) between the *you*-utterance and the previous/next utterance by each non-speaker. Others include the BL/FL speaker order and the number of speakers in the previous/next five utterances. Finally, for the manual systems, we also use the AMI dialogue acts of the *you*-utterances, and of the BL/FL utterances.

Our most recent results are as follows: in a tenfold cross-validation using manual features, the system achieves accuracy scores of 88%, 87%, and 82% in the GVR, RSVP and IA tasks, respectively, or 75% on the (five-way) combination of all three. A fully automatic system gives accuracies of 83%, 87%, and 77%, (all higher than majority class baselines, $p < 0.05$). Taking away FL features (as required for a fully online system) causes a fairly large performance drop in the IA task—9% for the manual system, and 8% for the automatic—but less in the other two. Although at this point the actual CALO-MA system is not able to process visual information, our experiments show that visual features produce a statistically significant improvement in the IA and RSVP tasks. The speaker's visual features are most predictive in the IA task, and it seems that when listeners look at the white-board/projector screen, this is indicative of a referential plural. Of the linguistic features, sentential, especially those concerning lexical properties help in the GVR and RSVP tasks. Fewer speaker changes correlate more with plural than singular referential and in the IA task, FL/BL speaker order is predictive. As for dialogue acts, in the GVR tasks, a *you* in a question is more likely to be referential, and in the RSVP task, questions are more likely to have an individual addressee, and statements, plural addressees.

IX. Summarization

A recent interest for CALO-MA is summarizing meetings. The goal of summarization is to create a shortened version of a text or speech while keeping important points. While textual document summarization is a well-studied topic, speech summarization (and in particular meeting summarization) is an emerging research area, and apparently very different from text or broadcast news summarization. The aim is basically filtering out the unimportant chit-chat from contentful discussions. While hot-spot detection, action item extraction, dialog act tagging, and topic segmentation and detection methods can be used to improve summarization, there are also preliminary studies using lexical, acoustic, prosodic, and contextual information.

In text or broadcast news summarization, the dominant approach is extractive summarization where "important" sentences are concatenated to produce a summary. For meeting summarization, it is not clear what constitutes an important utterance. In an earlier study [62], the sentences having the highest number of frequent content words are considered to be important. Using the advances in written and spoken document extractive summarization [63], some recent studies focused on feature-based classification approaches [64], while others mainly used maximum marginal relevance (MMR) [65] for meeting summarization [64], [66]. MMR iteratively selects utterances most relevant to a given query, which is expected to encode the user's information need, while trying to avoid utterances redundant to the already-selected ones. Due to the lack of a query, the common approach for meetings has been to use the centroid vector of the meeting as the query [64].

In CALO-MA, our summarization work mainly focused on investigating the boundaries of extractive meeting summarization in terms of different evaluation measures [67]. The most widely used is ROUGE [68], a metric that compares the produced summary against a set of reference summaries using word n-gram overlaps. We proposed to compute a simple baseline for summarization that consists in selecting the longest utterances in the meeting, which is more challenging to beat than the random baseline which selects random utterances. We also proposed a method to compute "oracle" summaries that extracts the set of sentences maximizing the ROUGE performance measure. For example, on the ICSI meeting corpus selecting the longest sentences yields a ROUGE-1 score of 0.15 (all scores are obtained on manual transcriptions), the oracle performs at 0.31 and a one of the most popular method for summarization, MMR, performs at 0.17. Improvements over the MMR system using keyphrases instead of words to represent the information increases ROUGE-1 to 0.20 [69] and a different model maximizing information recall (presented in [70]) performs at 0.23. Nevertheless, we observed that even the oracle summaries did not match the human capability for abstraction because they tend to stack up many unrelated facts. Hence, another trend is to use the sentences selected in the summaries as starting point for browsing the meetings. This helps users recontextualize the information and improve their ability to locate information as shown by [71]. To this end, in [69], we proposed a user interface for improving the capture of a user's information need by presenting automatically extracted keyphrases that can be refined and used to generate summaries for meeting browsing.

X. Conclusion and Future Work

We have presented a system for automatic processing of tasks involving multiparty meetings. Progress in these tasks, from low-level transcription to higher-level shallow understanding functions, such as action item extraction and summarization, has a potentially enormous impact on human productivity in many professional settings. However, there are practical and technical difficulties. In practice, people are not used to instrumented (virtual) meeting rooms. Technically, most higher level semantic understanding tasks are only vaguely defined and the annotator agreements are still very low. User feedback with support for adaptive training is critical for customizing the applications for individual use.

Further integration of these tasks and multiple potential modalities, such as video, or digital pen and paper, is part of the future work. Furthermore, meta information such as project related documentation or emails may be exploited for better performance. Another interesting research direction would be processing aggregate of meetings, tracking the topics, participants, and action items.

REFERENCES

[1] "DARPA cognitive agent that learns and organizes (CALO) project." [Online]. Available: http://www.ai.sri.com/project/CALO

[2] "DARPA perceptive assistant that learns (PAL) program." [Online]. Available: http://www.darpa.mil/ipto/programs/pal/pal.asp

[3] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Proc. ICSLP*, Denver, CO, 2002.

[4] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI meeting project: Resources and research," in *Proc. ICASSP*, Montreal, QC, Canada, 2004.

[5] "Augmented multi-party interaction," [Online]. Available: http://www.amiproject.org

[6] "Computers in the human interaction loop," [Online]. Available: http://chil.server.de

[7] "Rich transcription evaluations," [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2007

[8] G. Tur, A. Stolcke, L. Voss, J. Dowding, B. Favre, R. Fernandez, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena, D. Hakkani-Tür, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, S. Peters, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang, "The CALO meeting speech recognition and understanding system," in *Proc. IEEE/ACL SLT Workshop*, Goa, India, 2008.

[9] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grézl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," in *Proc. MLMI*, 2005.

[10] T. Hain, L. Burget, J. Dines, G. Garau, V. Wan, M. Karafiat, J. Vepa, and M. Lincoln, "The AMI system for the transcription of speech in meetings," in *Proc. ICASSP*, Honolulu, HI, 2007, pp. 357–360.

[11] G. Tur and A. Stolcke, "Unsupervised language model adaptation for meeting recognition," in *Proc. ICASSP*, Honolulu, HI, 2007, pp. 173–176.

[12] D. Vergri, A. Stolcke, and G. Tur, "Exploiting user feedback for language model adaptation in meeting recognition," in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 4737–4740.

[13] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tur, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Commun.*, vol. 32, no. 1–2, pp. 127–154, 2000.

[14] J. Kolar, E. Shriberg, and Y. Liu, "Using prosody for automatic sentence segmentation of multi-party meetings," in *Proc. Int. Conf. Text, Speech, Dialogue (TSD)*, Czech Republic, 2006.

[15] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005, pp. 1061–1064.

[16] M. Core and J. Allen, "Coding dialogs with the DAMSL annotation scheme," in *Proc. Working Notes AAAI Fall Symp. Commun. Action in Humans Mach.*, Cambridge, MA, Nov. 1997.

[17] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. SigDial Workshop*, Boston, MA, May 2004.

[18] J. Kolar, Y. Liu, and E. Shriberg, "Speaker adaptation of language models for automatic dialog act segmentation of meetings," in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[19] A. Stolcke and E. Shriberg, "Statistical language modeling for speech disfluencies," in *Proc. ICASSP*, Atlanta, GA, May 1996, pp. 405–408.

[20] Y. Liu, A. Stolcke, E. Shriberg, and M. Harper, "Using conditional random fields for sentence boundary detection in speech," in *Proc. ACL*, Ann Arbor, MI, 2005.

[21] S. Cuendet, D. Hakkani-Tür, and G. Tur, "Model adaptation for sentence segmentation from speech," in *Proc. IEEE/ACL SLT Workshop*, Aruba, 2006, pp. 102–105.

[22] M. Zimmerman, D. Hakkani-Tür, J. Fung, N. Mirghafori, L. Gottlieb, E. Shriberg, and Y. Liu, "The ICSI+ multilingual sentence segmentation system," in *Proc. ICSLP*, Pittsburg, PA, 2006.

[23] U. Guz, D. Hakkani-Tür, S. Cuendet, and G. Tur, "Co-training using prosodic and lexical information for sentence segmentation," in *Proc. Interspeech*, Antwerp, Belgium, Aug. 2007.

[24] G. Tur, "Co-adaptation: Adaptive co-training for semi-supervised learning," in *Proc. ICASSP*, Taipei, Taiwan, 2009.

[25] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. van Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Comput. Linguist.*, vol. 26, no. 3, pp. 339–373, 2000.

[26] G. Tur, U. Guz, and D. Hakkani-Tür, "Model adaptation for dialog act tagging," in *Proc. IEEE/ACL SLT Workshop*, 2006, pp. 94–97.

[27] M. Mast, R. Kompe, S. Harbeck, A. Kiessling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, and V. Warnke, "Dialog act classification with the help of prosody," in *Proc. ICSLP*, Philadelphia, PA, Oct. 1996, pp. 1732–1735.

[28] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "Toward joint segmentation and classification of dialog acts in multiparty meetings," in *Proc. MLMI*, Edinburgh, U.K., July 2005.

[29] V. Warnke, R. Kompe, H. Niemann, and E. Nöth, "Integrated dialog act segmentation and classification using prosodic features and language models," in *Proc. Eurospeech*, Rhodes, Greece, Sep. 1997.

[30] F. Yang, G. Tur, and E. Shriberg, "Exploiting dialog act tagging and prosodic information for action item identification," in *Proc. ICASSP*, Las Vegas, NV, 2008, pp. 4941–4944.

[31] A. Lisowska, "Multimodal interface design for the multimodal meeting domain: preliminary indications from a query analysis study," ISSCO, Univ. of Geneva, Tech. Rep. IM2.MDM-11, Nov. 2003.

[32] D. Beeferman, A. Berger, and J. D. Lafferty, "Statistical models for text segmentation," *Mach. Learn.*, vol. 34, no. 1–3, pp. 177–210, 1999.

[33] J. Reynar, "Statistical models for topic segmentation," in *Proc. ACL*, 1999, pp. 357–364.

[34] G. Tur, D. Hakkani-Tür, A. Stolcke, and E. Shriberg, "Integrating prosodic and lexical cues for automatic topic segmentation," *Comput. Linguist.*, vol. 27, no. 1, pp. 31–57, 2001.

[35] A. Gruenstein, J. Niekrasz, and M. Purver, , L. Dybkjaer and W. Minker, Eds., "Meeting structure annotation: Annotations collected with a general purpose toolkit," in *Recent Trends in Discourse and Dialogue*. Berlin/Heidelberg: Springer-Verlag, 2007, Text, Speech and Language Technology..

[36] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proc. ACL*, 2003.

[37] S. Banerjee and A. Rudnicky, "Segmenting meetings into agenda items by extracting implicit supervision from human note-taking," in *Proc. IUI*, Honolulu, HI, Jan. 2007, ACM.

[38] M. Hearst, "TextTiling: Segmenting text into multi-paragraph subtopic passages," *Comput. Linguist.*, vol. 23, no. 1, pp. 33–64, 1997.

[39] P.-Y. Hsueh, J. Moore, and S. Renals, "Automatic segmentation of multiparty dialogue," in *Proc. EACL*, 2006.

[40] P.-Y. Hsueh and J. Moore, "Automatic topic segmentation and labeling in multiparty dialogue," in *Proc. 1st IEEE/ACM Workshop Spoken Lang. Technol. (SLT)*, Palm Beach, Aruba, 2006.

[41] S. Banerjee and A. Rudnicky, "A TextTiling based approach to topic boundary detection in meetings," in *Proc. ICSLP*, Pittsburgh, PA, Sep. 2006.

[42] M. Georgescul, A. Clark, and S. Armstrong, "Word distributions for thematic segmentation in a support vector machine approach," in *Proc. CoNLL*, New York, Jun. 2006, pp. 101–108.

[43] M. Georgescul, A. Clark, and S. Armstrong, "Exploiting structural meeting-specific features for topic segmentation," in *Actes de la 14 ème Conf. sur le Traitement Automatique des Langues Naturelles*, Toulouse, France, Jun. 2007, Association pour le Traitement Automatique des Langues.

[44] A. Popescu-Belis, A. Clark, M. Georgescul, D. Lalanne, and S. Zufferey, , S. Bengio and H. Bourlard, Eds., "Shallow dialogue processing using machine learning algorithms (or not)," in *MLMI, Revised Selected Papers*. New York: Springer, 2005, vol. 3361, Lecture Notes in Computer Science, pp. 277–290.

[45] D. M. Blei, A. Y. Ng, M. I. Jordan, and J. Lafferty, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[46] M. Purver, K. Körding, T. Griffiths, and J. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *Proc. COLING-ACL*, Sydney, Australia, Jul. 2006, pp. 17–24.

[47] M. Dowman, V. Savova, T. L. Griffiths, K. P. Körding, J. B. Tenenbaum, and M. Purver, "A probabilistic model of meetings that combines words and discourse features," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 7, pp. 1238–1248, Sep. 2008.

[48] S. Banerjee, C. Rosé, and A. Rudnicky, "The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing," in *Proc. CHI*, 2005.

[49] S. Corston-Oliver, E. Ringger, M. Gamon, and R. Campbell, "Task-focused summarization of email," in *Proc. ACL Workshop Text Summarization Branches Out*, 2004.

[50] P. N. Bennett and J. Carbonell, "Detecting action-items in e-mail," in *Proc. ACM SIGIR*, Salvador, Brazil, Aug. 2005.

[51] P. N. Bennett and J. G. Carbonell, "Combining probability-based rankers for action-item detection," in *Proc. HLT/NAACL*, Rochester, NY, Apr. 2007.
[52] A. Verbree, R. Rienks, and D. Heylen, "First steps towards the automatic construction of argument-diagrams from real discussions," in *Proc. 1st Int. Conf. Comput. Models of Argument, September 11 2006, Frontiers Artif. Intell. Applicat.*, 2006, vol. 144, pp. 183–194, IOS press.
[53] P.-Y. Hsueh and J. Moore, "What decisions have you made?: Automatic decision detection in meeting conversations," in *Proc. NAACL/HLT*, Rochester, NY, 2007.
[54] W. Morgan, P.-C. Chang, S. Gupta, and J. M. Brenier, "Automatically detecting action items in audio meeting recordings," in *Proc. SIGdial Workshop Discourse and Dialogue*, Sydney, Australia, Jul. 2006.
[55] M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, S. Noorbaloochi, and S. Peters, "Detecting and summarizing action items in multi-party dialogue," in *Proc. 8th SIGdial Workshop on Discourse and Dialogue*, Antwerp, Belgium, Sep. 2007.
[56] R. Fernández, M. Frampton, P. Ehlen, M. Purver, and S. Peters, "Modelling and detecting decisions in multi-party dialogue," in *Proc. 9th SIGdial Workshop Discourse and Dialogue*, Columbus, OH, 2008.
[57] M. Frampton, J. Huang, T. H. Bui, and S. Peters, "Real-time decision detection in multi-party dialogue," in *Proc. EMNLP*, Singapore, Aug. 2009.
[58] R. Fernández, M. Frampton, J. Dowding, A. Adukuzhiyil, P. Ehlen, and S. Peters, "Identifying relevant phrases to summarize decisions in spoken meetings," in *Proc. Interspeech*, Brisbane, Australia, 2008.
[59] N. Jovanovic, R. op den Akker, and A. Nijholt, "Addressee identification in face-to-face meetings," in *Proc. EACL*, Trento, Italy, 2006, pp. 169–176.
[60] M. Frampton, R. Fernández, P. Ehlen, M. Christoudias, T. Darrell, and S. Peters, "Who is "you"? Combining linguistic and gaze features to resolve second-person references in dialogue," in *Proc. EACL*, 2009.
[61] M. Purver, R. Fernández, M. Frampton, and S. Peters, "Cascaded lexicalised classifiers for second-person reference resolution," in *Proc. SIGDIAL Meeting Discourse and Dialogue*, London, U.K., 2009.
[62] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, "Meeting browser: Tracking and summarizing meetings," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, Jun. 1998.
[63] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005.
[64] G. Murray, S. Renals, and J. Carletta, "Extractive summarization of meeting recordings," in *Proc. Interspeech*, Lisbon, Portugal, Sep. 2005.
[65] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. ACM SIGIR*, Melbourne, Australia, 1998.
[66] S. Xie and Y. Liu, "Using corpus and knowledge-based similarity measure in maximum marginal relevance for meeting summarization," in *Proc. ICASSP*, Las Vegas, NV, 2008, pp. 4985–4988.
[67] K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tür, "Packing the meeting summarization knapsack," in *Proc. Interspeech*, Brisbane, Australia, 2008.
[68] C. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Text Summarization Workshop*, 2004.
[69] K. Riedhammer, B. Favre, and D. Hakkani-Tür, "A keyphrase based approach to interactive meeting summarization," in *Proc. IEEE/ACL SLT Workshop*, Goa, India, 2008.
[70] D. Gillick, K. Riedhammer, B. Favre, and D. Hakkani-Tür, "A global optimization framework for meeting summarization," in *Proc. IEEE ICASSP*, Taipei, Taiwan, 2009, pp. 4769–4772.
[71] G. Murray, T. Kleinbauer, P. Poller, S. Renals, J. Kilgour, and T. Becker, "Extrinsic summarization evaluation: A decision audit task," in *Proc. MLMI*, Utrecht, The Netherlands, 2008.

Gokhan Tur (M'01–SM'05) is currently with the Speech Technology and Research Lab of SRI International, Menlo Park, CA. From 2001 to 2005, he was with AT&T Labs-Research, Florham Park, NJ.

Dr. Tur is an Associate Editor of the IEEE Transactions on Speech and Audio Processing and was a member of IEEE SPS Speech and Language Technical Committee (SLTC).

Andreas Stolcke (M'96–SM'05) received the Ph.D. degree in computer science from the University of California, Berkeley, in 1994.

He is a Senior Research Engineer at the Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, and at the International Computer Science Institute (ICSI), Berkeley, CA.

Lynn Voss received the MBA degree from University of Phoenix, Phoenix, CA

He is with the Engineering and Systems Division (ESD), SRI International, Menlo Park, CA. He is the project manager for the CALO-MA project interacting with both the research and engineering teams.

Stanley Peters is the Director of the Center for the Study of Language and Information (CSLI), Stanford, CA, and a Professor at the Linguistics Department at Stanford University.

Dilek Hakkani-Tür (S'00–M'01–SM'05) is a Senior Research Scientist at the ICSI, Berkeley, CA. From 2001 to 2006, she was with the AT&T Labs-Research, Florham Park, NJ.

Dr. Hakkani-Tür was an Associate Editor of the IEEE Transactions on Speech and Audio Processing and is a member of the IEEE SPS SLTC.

John Dowding is with CSLI, Stanford University, Stanford, CA, and NASA.

Benoit Favre received the Ph.D. degree from the University of Avignon, Avignon, France, in 2007.

Until 2009, he was a Postdoctoral Researcher at ICSI, Berkeley, CA. He is currently a Research Engineer with LIUM in France.

Raquel Fernández received the Ph.D. degree from the University of Potsdam, Potsdam, Germany.

Until 2009, she was a Postdoctoral Researcher at CSLI, Stanford, University, Stanford, CA. She is currently with University of Amsterdam, Amsterdam, The Netherlands.

Matthew Frampton received the Ph.D. degree from the University of Edinburgh, Edinburgh, U.K.

Since 2007 he has been an Engineering Research Associate at CSLI, Stanford, University, Stanford, CA.

Mike Frandsen is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Clint Frederickson is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Martin Graciarena received the Ph.D. degree from the University of Buenos Aires, Buenos Aires, Argentina, in 2009.

Since 1999, he has been with the STAR Lab, SRI International, Menlo Park, CA.

Donald Kintzing is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Kyle Leveque is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Shane Mason is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

John Niekrasz is currently a Research Fellow with the University of Edinburgh, Edinburgh, U.K. He was with CSLI, Stanford, University, Stanford, CA, when this work was done.

Matthew Purver is currently a Senior Research Fellow with Queen Mary University of London, London, U.K.

Korbinian Riedhammer is currently with the University of Erlangen, Erlangen, Germany. He was with ICSI, Berkeley, CA when this work was done.

Elizabeth Shriberg received the Ph.D. degree from the University of California, Berkeley, in 1994.

She is a Senior Researcher at both the Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, and the ICSI, Berkeley, CA.

Jing Tien is with ESD at SRI International, Menlo Park, CA, working on software engineering and user interfaces.

Dimitra Vergyri (M'06) received the Ph.D. degree from Johns Hopkins University, Baltimore, MD, in 2000.

She is currently with the STAR Lab, SRI International, Menlo Park, CA.

Dr. Vergyri is an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.

Fan Yang received the Ph.D. degree from the Oregon Graduate Institute, Beaverton, in 2008.

He is currently with Next IT Corp. He was with SRI International, Menlo Park, CA, when this work was done.