Recognition of Dialogue Acts in Multiparty Meetings Using a Switching DBN

Alfred Dielmann and Steve Renals, *Member, IEEE*

***Abstract*—This paper is concerned with the automatic recognition of dialogue acts (DAs) in multiparty conversational speech. We present a joint generative model for DA recognition in which segmentation and classification of DAs are carried out in parallel. Our approach to DA recognition is based on a switching dynamic Bayesian network (DBN) architecture. This generative approach models a set of features, related to lexical content and prosody, and incorporates a weighted interpolated factored language model. The switching DBN coordinates the recognition process by integrating the component models. The factored language model, which is estimated from multiple conversational data corpora, is used in conjunction with additional task-specific language models. In conjunction with this joint generative model, we have also investigated the use of a discriminative approach, based on conditional random fields, to perform a reclassification of the segmented DAs. We have carried out experiments on the AMI corpus of multimodal meeting recordings, using both manually transcribed speech, and the output of an automatic speech recognizer, and using different configurations of the generative model. Our results indicate that the system performs well both on reference and fully automatic transcriptions. A further significant improvement in recognition accuracy is obtained by the application of the discriminative reranking approach based on conditional random fields.**

***Index Terms*—AMI corpus, conditional random field (CRF), dialogue act (DA), dynamic Bayesian network (DBN), interpolated factored language model (FLM).**

I. Introduction

DIALOGUE acts (DAs) form a useful level of representation for the interpretation of conversations. A DA is a construct that describes the role that an utterance plays in a conversation and provides a bridge between an orthographic word-level transcription and a richer representation of the discourse. A conversation may be segmented into a sequence of DAs, with each DA assigned a label that describes the function played by that utterance within the conversation. DA labels may incorporate syntactic, semantic, and pragmatic factors: in addition to providing information about the structure of a dialogue and the course of a conversation, DAs are also able to capture, at a coarse level, individual speaker attitudes and intentions, their interaction role, and their level of involvement.

Manuscript received July 31, 2007; revised February 19, 2008. This work was supported in part by the European Union 6th FWP IST Integrated Project AMI (FP6-506811) and AMIDA (FP6-033812, publication AMIDA-33). This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Helen Meng.

The authors are with the Centre for Speech Technology Research (CSTR), University of Edinburgh, Informatics Forum, Edinburgh EH8 9AB, U.K. (e-mail: a.dielmann@ed.ac.uk; s.renals@ed.ac.uk).

Digital Object Identifier 10.1109/TASL.2008.922463

Multiparty meetings have been intensively researched over the past several years, with a growing focus on how a meeting may be automatically analyzed and interpreted in terms of the group discourse and interaction. Example applications have included automatic summarization [1], topic segmentation and labeling [2], [3], group action detection [4]–[6], participant influence [7], and dialog structure annotation [8]. The reliable recognition of the DA sequence in a meeting, and the resulting knowledge of the discourse structure, plays an important role in the development of such applications.

In this paper, we present a flexible trainable approach for the automatic recognition of DAs in meetings, based on a switching dynamic Bayesian network (DBN) model, a factored language model, and discriminative reranking. We present results on the AMI meeting corpus, in which we compare DA recognition accuracy on manual and automatic meeting transcriptions, and compare the effect of the different components of the overall approach.

The DA recognition task comprises two related subtasks: segmentation, and classification or tagging. These tasks may be performed jointly or sequentially. In a sequential approach the conversation is first segmented into unlabeled DA segments, then each detected segment is tagged with a DA label. The joint approach performs both tasks concurrently, detecting DA segment boundaries and assigning labels in a single step. The joint approach is able to examine multiple segmentation and classification hypotheses in parallel, whereas only the most likely segmentation is supplied to the DA classifier in a sequential approach. The joint approach is potentially capable of greater accuracy, since it is able to explore a wider search space, but the optimization problem can be more challenging. In a sequential system, the two subtasks can be optimized independently.

We present an approach to DA recognition that takes advantage of both techniques by employing a joint generative infrastructure followed by a discriminative classifier. Both system components make use of supervised learning from manually annotated data. The joint recognition is coordinated by a switching DBN which integrates a discourse language model, six lexical and prosodic features, and two factored language models trained on the orthographic transcriptions. The recognized sequence of DA units is then reclassified using a conditional random field DA tagger trained using the lexical content and a set of discrete features.

We have performed tagging, segmentation, and recognition experiments using the joint generative approach on unseen meetings with three different modelling configurations, based on both manual and automatic speech recognition (ASR) transcriptions. We demonstrate in additional experiments, that the accuracy of DA recognition using this joint approach can be further improved through discriminative postprocessing.

II. Multiparty Conversational Data Resources

In our main experiments, we have used the AMI meeting corpus [9], which is a multimodal collection of annotated meeting recordings. It consists of about 100 h of meetings collected in three instrumented meeting rooms. About two thirds of the corpus consists of meetings elicited using a scenario in which four meeting participants, playing different roles in a team, take a product development project from beginning to completion. The scenario portion of the corpus consists of a number of meeting series, with four meetings per series. Each series of four meetings involves the same four participant roles, and comprises project kickoff, functional design, conceptual design, and detailed design meetings. The remaining meetings in the corpus, "nonscenario," are naturally occurring meetings, with three to five participants.

The aim of the corpus collection was to obtain a multimodal record of the complete communicative interaction between the meeting participants. To this end, the meeting rooms were instrumented with a set of synchronized recording devices, including lapel and headset microphones for each participant, an eight-element circular microphone array, six video cameras (four close-up and two room-view), capture devices for the whiteboard and data projector, and digital pens to capture the handwritten notes of each participant. The corpus has been manually annotated at several levels, including orthographic transcriptions, various linguistic phenomena including DAs, head and hand movements, and focus of attention.[1] The DA annotation scheme for the AMI corpus, outlined in Table I, is based around a categorization tailored for group decision making, and consists of six broad categories and a total of 15 DA classes. Each DA unit is assigned to a single class, corresponding to the speaker's intent for the utterance. The distribution of the DA classes, shown in Table I, is rather imbalanced, with over 60% of DAs corresponding to one of the three most frequent classes (inform, backchannel or assess). Over half the DA classes account for less than 10% of the observed DAs.

We performed our experiments on the 138 meetings that form the scenario subset of the AMI corpus, following the subdivision into training, development, and test sets suggested in the corpus documentation. There were 98 meetings in the training set, 20 in the development set, and 20 in the test set.

We have used two further corpora in this work: the ICSI meetings corpus [10] and the Fisher corpus [11]. The ICSI meetings corpus consists of 72 h of naturally occurring research group meetings at the International Computer Science Institute in Berkeley, CA, during the years 2000–2002, recorded using close-talking microphones worn by each participant (in addition, there were also four tabletop microphones). The ICSI corpus has been orthographically transcribed and annotated in terms of DAs. However, the DA annotation scheme is different to the one used for the AMI corpus—it is not possible to test a DA recognition system developed on the AMI data on the ICSI corpus or vice-versa. The ICSI corpus was annotated according to the Meeting Recorder Dialog Act (MRDA) scheme, which utilizes 11 generic tags and 40 specific subtags resulting in more than 1000 unique DA labels [12]. This large set of DA classes may be transformed (by rule) to a set of five broad DA classes: statements (52.2% of annotated DAs), questions (6.2%), disruptions (12.9%), fillers (10.3%), and backchannels (12.3%). It is not feasible to build a mapping between the ICSI and AMI DA classes.

[1]The annotated corpus is freely available from http://corpus.amiproject.org

TABLE I
Six Broad Categories and 15 Specialized DA Classes Used in the AMI Corpus DA Annotation Scheme, With the Percentage of DAs in Each Class

Category	DA class	Proportion %
Information exchange	*inform*	26.6
	elicit inform	3.4
Individual or group action	*suggest*	7.5
	offer	1.2
	elicit offer or suggestion	0.5
Comment on previous discussion	*assess*	16.7
	elicit assessment	1.7
	comment about understanding	1.8
	elicit comment understanding	0.2
Social function	*be positive*	1.8
	be negative	0.1
No speaker intention	*backchannel*	17.6
	stall	6.3
	fragment	13.0
Other	*other*	1.8

The Fisher corpus consists of more than 16 000 English telephone conversations on a wide range of elicited topics, resulting in about 2000 hours of recorded speech, which has been orthographically transcribed. Although it is not possible to use these corpora directly as training data for DA recognition (using the AMI corpus annotation scheme) they represent valuable additional sources of transcribed conversational data. The Fisher corpus was of particular utility, since it contains over 10 million words, making it an order of magnitude larger than the AMI and ICSI corpora.

III. Joint DA Recognition System

We have developed a joint approach to DA recognition based on a switching DBN generative model. The observed features that are generated by this model are the words spoken by the meeting participants, together with a set of word-based prosodic features related to timing, intonation, and energy. The mapping from DA labels to word sequences was modeled using a factored language model (FLM) and an interpolated FLM. The probability of observing a certain sequence of DA labels (discourse model) was represented through a simple trigram language model over DAs. The set of continuous word-based prosodic features was integrated into the recognizer using a Gaussian mixture model (GMM). The overall recognition process is actively controlled by a switching DBN which integrates information derived from words, prosodic features, and language models. Section III-A outlines the use of an automatic

speech recognizer to produce a transcription, and the extraction of the prosodic features. Sections III-B and III-C discuss the factored language models and the switching DBN model that underlie the DA recognition system.

A. Feature Extraction

We have used two sets of features in the DA recognition system: the transcription of the spoken words obtained using an ASR system (Section III-A1) and the continuous prosodic features (Section III-A2).

1) Speech Recognition: Fully automatic DA recognition requires speech recognition. The AMI corpus has been manually transcribed at the word level, as well as being processed by an ASR system, thus enabling us to assess the robustness of the DA recognition system to speech recognition errors.

Large vocabulary continuous speech recognition (LVCSR) of conversational speech is a significant research domain, and the recognition of speech in meetings has been intensively studied and evaluated in recent years.[2] Automatic transcriptions of the AMI meeting corpus were obtained using the AMI–ASR system [13]. This LVCSR system is based on decision tree clustered crossword triphone hidden Markov models and a trigram language model. For the multiparty meeting domain, the front end was enhanced using acoustic echo cancellation, and the perceptual linear prediction acoustic features were processed using heteroscedastic linear discriminant analysis. The acoustic feature space was normalized by speaker, using vocal tract length normalization, and the model space was adapted using maximum-likelihood linear regression.

The meeting domain acoustic models were trained on the AMI corpus data. To recognize the complete corpus, a fivefold cross-validation was employed using equal splits of the corpus. Two transcription versions were generated in each case: a fully automatic one achieved by applying the full system to automatically segmented audio files; and a semiautomatic transcription obtained using a manual segmentation into utterances. The "manual segmentation" system also used a simpler ASR component, in which speaker adaptation was not used. The fully automatic system resulted in an overall word error rate (WER) of about 36%; the simpler system, using manual segmentation, resulted in a WER of about 39%. In both cases the system operated on signals recorded from the close-talking microphones.

The automatic DA recognition experiments performed on the AMI corpus (Section V-B) compared both transcription versions. The speaker adapted "automatic segmentation" ASR output offers an overall improvement in terms of WER compared with the "manual segmentation" ASR output. However, entire utterances may be deleted by the automatic acoustic segmentation, and consequently whole DA segments are irredeemably lost (Section IV). Moreover, the word boundary times of the "manual segmentation" ASR output, are more accurate, compared with the reference orthographic transcription, since they cannot cross the manually annotated utterance boundaries. Accurately timed word boundaries are desirable for the extraction of prosodic features at the word level and are also required to evaluate segmentation into DAs.

[2]NIST rich transcription meeting recognition evaluation available at http://www.nist.gov/speech/

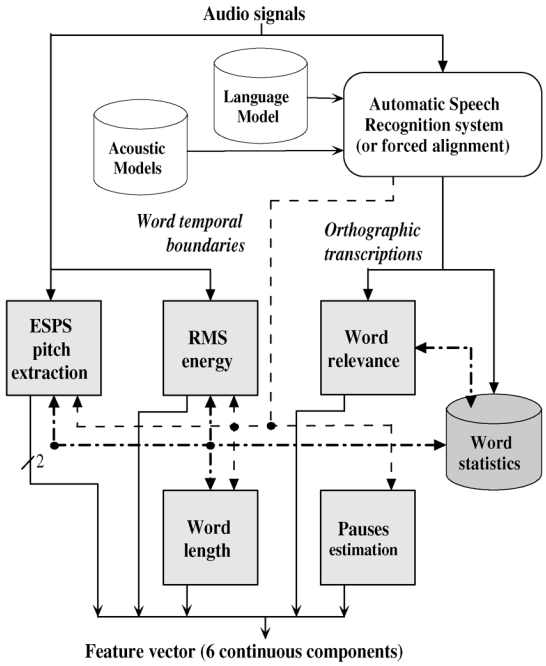

Fig. 1. Data-flow of the automatic speech transcription and feature extraction process.

Although both ASR versions offer valuable insights during the evaluation of our system, the "automatic segmentation" ASR output represents the main test condition since it does not require any manual intervention.

2) Prosodic Features: Six continuous prosodic features were extracted for each word, using the audio signal and the transcription (Fig. 1): mean and variance of the fundamental frequency (F0), mean energy, word duration, pause duration, and word relevance. For the reference transcription, the times of word boundaries were obtained using a forced alignment against the audio. For the ASR transcriptions, the word boundary timings were output as part of the recognition process. The F0 tracks were estimated using ESPS $\mathrm{get\_}f0$ [14], and the mean and variance were computed. The mean pitch was also normalized by speaker and by the average pitch for that term, with the objective of having a speaker-independent measure able to highlight content words with a significant pitch shift. A similar normalization technique was applied during RMS energy estimation with the aim of compensating for different channel gains and to highlight emphasized words. Word duration was "term normalized," being thus divided by the average word duration for that term, in order to highlight words which last more (or less) than the usual occurrences of that term. Unit duration, pitch, and energy were assigned to words which appear only once in the training set and to out-of-vocabulary words observed during testing but absent from the training set. Interword pauses were also estimated from the word boundary times. Pauses are often associated with speaker turn alternations and other relevant changes in the conversational process such as topic shifts, and it is known that they provide a valuable cue for DA segmentation [15], [16]. Word relevance was estimated as the ratio between local term frequency within the current conversation and absolute term frequency across the whole meetings collection, thus assigning high scores to globally infrequent terms which occur frequently in the current conversation.

B. Interpolated Factored Language Models

Conventional language models construct a joint probability distribution over word sequences $P(w_1, \ldots, w_n)$, which is factorized as a product of conditional probabilities $P(w_t \mid w_{t-1}, w_{t-2}, \ldots, w_{t-k})$. This concept can be generalized by replacing words $w_1, \ldots, w_n$ with bundles of factors $\mathbf{v_1}, \ldots, \mathbf{v_n}$, to construct a factored language model (FLM) [17]. Each factor bundle, $\mathbf{v}_t \equiv \{v_t^0, v_t^1, \ldots, v_t^k\}$, is a vector whose components are factors such as word identity, part of speech tag, word stem, and enclosing dialogue act label. Conventional LMs can be interpreted as a special case of FLMs with a single factor, the actual words: $\mathbf{v}_t \equiv \{w_t\}$. Word identities are usually included in the collection of factors employed in an FLM. The smoothing and discounting techniques used for conventional LMs may be applied to FLMs, with the added flexibility of choosing which factor to drop when constructing simpler models for interpolation or backoff. Moreover, it is possible to drop more than one factor at a time and to follow multiple concurrent backoff paths using *generalised parallel backoff* [17]. FLMs have an increased number of degrees of freedom, compared with conventional LMs, and it is possible to choose the factor set, the number of backoff steps, the backoff topology, and the discounting method associated to each backoff step.

We use FLMs to map word sequences into DA units, and we are primarily interested in evaluating these models in terms of DA labeling accuracy, rather than perplexity. It is possible to select the optimal FLM topology automatically [18], and we experimented with a simple search algorithm that randomly sampled the search space. The resulting models tended to employ a large number of factors (seven or more), implying many backoff steps. These automatically discovered topologies resulted in a slightly improved DA tagging accuracy (up to 2% absolute) when compared to manually developed FLMs, but the more intricate structure requires a more elaborate DBN infrastructure and substantially increases computational cost. In order to reach a tradeoff between simplicity, cost, and accuracy, we decided to employ a simpler FLM topology with three factors (and two backoff steps). Although this topology was initially designed by hand, it was also discovered by the automatic search procedure (with an improved set of discounting parameters).

The FLM that we used for the DA recognition task was based on three factors: the word identity w_t, the dialogue act label d_t associated to each word w_t, and the relative word position n_t in the context of the DA unit. The word sequence probability was modeled using a product of word bigrams conditioned also on word position and DA label $P(w_t \mid w_{t-1}, n_t, d_t)$. The model was smoothed using two backoff steps and Kneser–Ney discounting. w_{t-1} was the first term to be dropped leading to a unigram like term $P(w_t \mid n_t, d_t)$. In the case of a subsequent backoff, the DA label factor d_t was the next term to be dropped, leading to $P(w_t \mid n_t)$. The FLM was estimated using the training subset of the AMI scenario meeting data outlined in Section II (470 000 words and a dictionary of about 9000 unique terms).

FLMs with the same topology may be interpolated, similarly to word-based n-grams. This enables the construction of combined models, whose component FLMs are trained using different data resources. We built FLMs for DA recognition using the ICSI meetings corpus and the Fisher corpus of conversational telephone speech, in addition to an FLM built on the target AMI corpus, integrating them into a single interpolated factored language model.

The AMI meetings corpus has a size of 0.97 million words in total, with about 0.47 million words in our training set of 98 meetings. The ICSI corpus, which is from a similar domain, contains 0.74 million words. The Fisher corpus, which is based on two party telephone conversations is much larger, containing 10.62 million words. Building an interpolated FLM from these data sources, enriches the baseline FLM trained on AMI meetings only, by extending the vocabulary and thus reducing the out-of-vocabulary, and by improving the n-gram counts with word sequences that are not observed in the AMI training data-set alone. However, neither the ICSI or Fisher corpora are annotated using the AMI DA annotation scheme. (The ICSI corpus has been annotated for DAs, but using a different and incompatible scheme.) In the absence of compatible DA annotations, both the ICSI and FISHER corpora were duplicated 15 times when training the FLMs, labeling every sentence with all the 15 possible DA labels in the AMI DA annotation scheme. FLMs trained on artificially duplicated data are obviously not discriminative in a DA classification task, but they are able to enhance the dictionary and n-gram counts of the resulting interpolated FLM.

As will be discussed in Section V, the use of an interpolated FLM provides an improvement in DA segmentation at the price of slightly reduced DA classification accuracy. To address this, we conducted experiments with a hybrid approach in which the baseline FLM trained on the AMI data is combined with an interpolated FLM at the sequence decoding level by maximizing the product of the joint probabilities associated to the two concurrent FLMs.

C. Switching DBN Architecture

In a DA recognition system, segmentation and classification are strongly related—the output of the DA classifier is dependent on the optimal placement of the DA unit boundaries, and the placement of the DA boundaries depends on the labels assigned to the DAs. In this paper, we treat the segmentation and classification problems jointly and the process is coordinated by a switching DBN model [19], implemented using the Graphical Model ToolKit (GMTK) [20].

Fig. 2 depicts the switching DBN model [21]. The transcribed words are represented as the sequence of discrete observable nodes $W_0, \ldots, W_{t-1}, W_t$. The FLM and interpolated FLM outlined in the previous section are depicted using dotted arcs, and each word is observed twice: once for the baseline FLM and once for the interpolated FLM. The relative position of each word W_t in the current DA unit DA_t^0 is represented by the discrete node N_t. N_t relies on a bounded word counter C_t, which is incremented at every word encountered in the current DA unit. After each block of five words, C_t is reset to zero, and N_t is incremented, thus indicating to which "block of five words" the current word W_t belongs to

$$\begin{aligned} &\text{if } C_{t-1} < 4: \quad C_t := C_{t-1} + 1 \\ &\text{if } C_{t-1} = 4: \quad C_t := 0 \quad N_t := N_{t-1} + 1. \end{aligned} \tag{1}$$

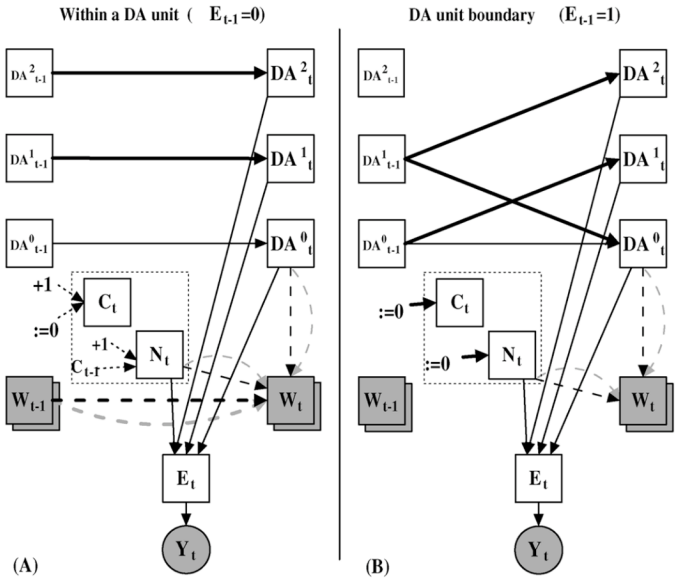

Fig. 2. Switching dynamic Bayesian network model for the joint dialogue act recognition task. (A) *Intra-DA* topology adopted within a DA unit. (B) *Inter-DA* topology used at DA boundaries. The model switches between the two operating conditions (topologies) according to the state of the DA boundary detector node E. Square nodes represent discrete random variables; round nodes are continuous variables. Shaded nodes represent observable features; unshaded nodes are hidden variables. Plain arcs visually encode statistical dependences between random variables and dotted arcs highlight the dependences implied by FLMs.

The final length of an automatically detected DA unit is not known *a priori* and is only available at the end of the DA recognition process; therefore, it is impractical to estimate word position features normalized for DA length.

The DA recognition history is represented by the current and the two previous DA labeling hypotheses, $\mathrm{DA}_t^0, \mathrm{DA}_t^1$ and DA_t^2. This history is needed by the DA boundary detector, the hidden binary variable E_t. E_t is the principal switching variable in the model, switching from zero to one when a boundary between two disjoint DA units is detected. In the absence of a DA boundary $(E_{t-1} = 0)$ the DBN assumes the *intra-DA* topology shown in Fig. 2(A); when a boundary is likely to be present $(E_{t-1} = 1)$, the model adopts the alternative *inter*-DA topology depicted in Fig. 2(B).

The dependency of the observable prosodic feature vectors Y_t on E_t is modeled using a Gaussian mixture model (GMM) with n components

$$P(Y_t = y \,|\, E_t = i) = \sum_{j=1}^{n} C(i,j) N(y; \mu_{i,j}, \Sigma_{i,j}) \qquad (2)$$

where $N(y; \mu_{i,j}, \Sigma_{i,j})$ is a Gaussian density with mean $\mu_{i,j}$ and covariance $\Sigma_{i,j}$, evaluated at y. $C(i,j)$ is the conditional prior weight of each mixture component j, and the optimal number of mixture components n for each state $i = [0,1]$ is automatically selected during training [20]. The GMM relates the six-dimensional prosodic features to the two discrete states of E_t, thus helping to predict the DA segmentation.

The cardinalities of the discrete random variables reflect the function they serve in the model, thus: $|E_t| = 2, |C_t| = 5, |\mathrm{DA}_t^0| = |\mathrm{DA}_t^1| = |\mathrm{DA}_t^2| = 15$, and W_t has as many states as the number of words in the dictionary. Since the vast majority of the DA units have fewer than 75 words, the word block counter cardinality has been constrained to $|N_t| = 15$.

The intra-DA topology used within a DA unit [Fig. 2(A)] accumulates the joint probability for a sequence of $k+1$ words $W_{t-k}, \ldots, W_t$ as the product of a FLM and a weighted interpolated FLM given the current DA label hypothesis DA_t^0 and the deterministic counter nodes N_t and C_t. The two language model probabilities (FLM and interpolated FLM) are combined by using an equally weighted stream weighting combination

$$\begin{aligned} &P(W_{t-k}, \ldots, W_t \,|\, \mathrm{DA}^0) \\ &\quad = \prod_{i=t-k}^{t} \{P_{\mathrm{IFLM}}(W_i \,|\, W_{i-1}, N_i, \mathrm{DA}^0) \\ &\qquad \cdot P_{\mathrm{FLM}}(W_i \,|\, W_{i-1}, N_i, \mathrm{DA}^0)\} \end{aligned} \qquad (3)$$

where $P(W_{t-k}, \ldots, W_t \,|\, \mathrm{DA}^0)$ represents the joint probability for the observed utterance $W_{t-k}, \ldots, W_t$, given the current DA classification hypothesis DA^0; P_{FLM} and P_{IFLM} are the probabilities, respectively, provided by the baseline and the interpolated FLMs.

The absence of a DA boundary implies that the DA recognition history remains unaltered; hence, the content of DA_{t-1}^1 needs to be cloned into DA_t^1, and similarly $\mathrm{DA}_t^2 := \mathrm{DA}_{t-1}^2$. Since the word sequence $W_{t-k}, \ldots, W_t$ has been generated by the same DA unit with label DA_t^0, and no DA boundaries have been spotted between time $t-k$ and time t, it follows that $\mathrm{DA}_{t-k}^j = \cdots = \mathrm{DA}_{t-1}^j = \mathrm{DA}_t^j$ for $j = [0,2]$.

If a DA boundary is hypothesized $(E_{t-1} = 1)$, then the model switches to the inter DA topology [Fig. 2(B)], which integrates the probability from the 3-gram discourse LM into the overall recognition process and starts the evaluation of a new DA unit, reinitializing the counter nodes: $C_t = 0, N_t = 0$. The DA recognition history is updated and a new set of DA classification hypotheses DA_t^0, for the next DA unit beginning with W_t, is generated following the 3-gram discourse language model $P(\mathrm{DA}_t^0 \,|\, \mathrm{DA}_{t-1}^1, \mathrm{DA}_{t-1}^2)$.

When $t = 0$ a slightly modified intra-DA topology $(E_{-1} = 0)$ needs to be adopted, with both the DA recognition history and the counter nodes initialized to zero $(\mathrm{DA}_0^1 = \mathrm{DA}_0^2 = 0, C_0 = 0, N_0 = 0)$.

Segmentation and classification are carried out concurrently. The classification process accounts for the joint probability of the transcription $W_{t-k}, \ldots, W_t$ accumulated by the two concurrent FLMs given the current classification hypothesis DA_t^0, the probability of DA_t^0 given the two previously recognized DA units, and the segmentation hypothesis (a DA unit starting at time $t-k$ and ending at time t). Several alternative segmentation hypotheses are generated, with the probability of each segmentation combining the likelihood of generating the observed prosodic feature vectors Y_t and the likelihood of the DA unit generating the observed words $W_{t-k}, \ldots, W_t$. A pruned Viterbi decoding is used to find the most likely sequence of labeled DA segments.[3]

[3]The decoding runtime for this model is about 10 times slower than realtime on a 3–GHz P4 equipped with 1 GB of RAM.

TABLE II
DA SEGMENTATION AND *RECOGNITION* EVALUATION METRICS

	Normalized by:		
	DA boundaries	Words	DA units
Tolerant:	NIST-SU	Boundary	
1 matching boundary	*NIST-SU*		
Rigorous:		Strict	DSER
2 matching boundaries		*Strict*	*DER*

Since this approach cannot generate a DA segmentation without an associated DA labeling hypothesis, the segmentation accuracy is assessed by ignoring the recognized DA labels. Classification of the DA units for a reference segmentation can be achieved by constraining the state of the boundary detector nodes E.

IV. EVALUATION

DA tagging accuracy can be easily evaluated by scoring the automatic DA classification output on a test set against the corresponding reference DA annotation. The percentage of correctly classified DA units, or its complement the classification error rate, is a standard metric for the DA tagging task, along with class-based precision and recall measures [22].

The evaluation of DA segmentation accuracy is less straightforward. The concept of a "correct" DA segmentation is not unequivocally defined, since it may be in terms of the overall sequence of DA units, or may demand precise timing of the DA boundaries. Moreover, a segmentation metric may be expressed and normalized in terms of DA units, DA boundaries, or words. A number of different metrics have been proposed, each offering a different perspective on the task of DA segmentation. In this paper, we report our results using four previously defined metrics: the NIST Sentence like Unit (NIST-SU), Strict, and Boundary metrics [15], and the DA Segmentation Error Rate (DSER) metric [23], [16]. These metrics are summarized in Table II.

According to the Strict and DSER metrics a DA unit has been correctly detected only when both boundaries are correctly located and no other boundaries fall within the detected unit; the NIST-SU and Boundary metrics focus on individual boundaries, rather than on DA units, and are thus more tolerant. The NIST-SU metric scores the sum of missed DA boundaries and false alarms divided by the number of reference DA boundaries. In case of a high number of insertions (false alarms), the NIST-SU metric can assume values well above 100% [16]. The Boundary metric has the same numerator as the NIST-SU metric (missed boundaries + insertions) but is normalized by the total number of nonboundaries in the reference, which is equivalent to the number of reference words. Since there are usually many more reference words than segmentation errors, this metric tends to be skewed toward very low error rates. The DSER metric is the complement of the percentage of correctly detected DA units; similarly the Strict metric can be defined as the percentage of words belonging to incorrectly segmented units. The Strict metric is a severe metric heavily influenced by the length of DA units in terms of words.

Since the DA recognition task combines segmentation and tagging, it is possible to translate most of the segmentation metrics into recognition metrics by requiring that the detected DA unit labels match the reference annotation. Therefore, the NIST-SU, Strict, and DSER (usually referred as DA error rate or DER in the recognition task) metrics can be easily adapted to the recognition task by adding the constraint that wrongly labeled units will be scored as errors even if their boundaries are a perfect match. This added requirement implies that these recognition metrics will result in error rates at least as great as their segmentation counterparts. The Boundary segmentation metric is an exception, since it is translated into the Lenient recognition metric [15], which is defined as the percentage of correctly classified words independent of the segmentation. Since it is focused exclusively on tagging accuracy, this metric should be regarded as a DA classification metric rather than a genuine recognition metric.

The reference DA annotation is produced in terms of the manually transcribed word sequence. When processing ASR output, the DA tags will be applied to a different word sequence, owing to ASR errors. Since a manual re-annotation of the ASR output would be extremely expensive, we have adopted the evaluation scheme proposed by Ang *et al.* [15]: ASR words are mapped into the manually annotated segments according to their midpoint $0.5 * (\text{word\_start\_time} + \text{word\_end\_time})$, thus inheriting their reference DA labels. Because of ASR deletions and the time-based alignment, several DA units will be empty. As we have adopted a word-based approach, these lost segments cannot be successfully recognized and will be reported as errors by every segmentation/recognition metric. Conversely, on a pure DA tagging evaluation task, empty segments will be scored as if they were tagged with a randomly drawn label, thus reducing the biasing effect of words and utterances deleted by the ASR system.

V. EXPERIMENTS

We have used the switching DBN model for tagging, segmentation, and recognition of DAs in the ICSI and AMI meeting corpora, using the three language model configurations described in Section III-B: FLM, interpolated FLM, and a hybrid in which the interpolated FLM is focused on segmentation and the baseline FLM is focused on tagging. These experiments extend our previously published results in which an early version of the switching DBN model, without the use of interpolated FLMs, was used for DA recognition on the ICSI meetings corpus [24], and experiments on the AMI corpus using manual transcriptions only [21]. Our initial experiments, applying the complete framework to the five DA ICSI task, validates the methodology on an established task, forming the base for our investigations on the novel 15 DA AMI task.

A. Joint DA Recognition of ICSI Meetings

We performed DA tagging, segmentation and recognition on the ICSI meeting corpus, using the reference manual transcriptions. These experiments used the ICSI DA annotation scheme based on the five broad DA categories described in Section II. In order to facilitate comparison with the existing literature, we used the subdivision of the ICSI corpus defined by Ang *et al.* [15]. The results obtained using the three language model configurations are reported in Table III: the baseline FLM model

TABLE III
DA TAGGING, SEGMENTATION, AND RECOGNITION ERROR RATES (%) ON THE ICSI MEETING CORPUS USING A DICTIONARY OF FIVE BROAD DA CLASSES; RESULTS ARE REPORTED ON THREE DIFFERENT FLM SETUPS (BASELINE FLM, INTERPOLATED FLM, AND HYBRID FLM+IFLM) USING REFERENCE MANUAL TRANSCRIPTIONS

		Reference transcription		
Task	Metric	FLM	iFLM	Hybrid
TAG.	100 - %Correct	**24.0**	38.8	25.2
S	NIST-SU	35.6	**30.5**	32.0
E	DSER	48.9	**27.9**	27.8
G	Strict	56.5	**50.3**	52.3
M.	Boundary	5.5	**4.7**	4.9
R	NIST-SU	56.8	67.9	**59.5**
E	DER	61.4	57.9	**47.4**
C.	Strict	64.7	66.4	**62.7**
	Lenient	19.7	30.3	**20.9**

[24]; a novel weighted interpolated FLM trained on ICSI, AMI, and FISHER data (AMI and FISHER were duplicated five times, one .for each DA class); and a *hybrid* combination of the two FLMs

The results on the ICSI corpus indicate that the baseline FLM offers the best tagging performance; adoption of an interpolated FLM improves the segmentation accuracy at the cost of tagging. An effective tradeoff between DA tagging and segmentation, required for DA recognition, was obtained using the *hybrid* configuration (baseline FLM and interpolated FLM used in conjunction). In Section VII, we compare these results with the state-of-the-art results reported on this task [16].

B. Joint DA Recognition of AMI Meetings

We performed more extensive experiments using the switching DBN model and the three system configurations on the AMI meeting corpus. Each of these systems was run on three transcription conditions: manual reference transcription, ASR with manual utterance segmentation, and ASR with automatic utterance segmentation. As discussed in Section II, the AMI meeting corpus uses a set of fifteen DA classes, in contrast to the five broad DA classes used on the ICSI corpus, thus results for the two corpora are not directly comparable.

Error rates for the DA tagging, segmentation, and recognition tasks, using the three system configurations and the three transcription conditions are shown in Table IV. The three system configurations are as follows:

- *FLM*: simple FLM trained only on the AMI training set;
- *iFLM*: weighted interpolated FLM trained on AMI (relative combination weight of about 58.5%), ICSI (2.7%), and FISHER (38.8%) conversational data;
- *Hybrid*: *iFLM* and *FLM* combined at the decoding level.

These three systems were each run on three transcription conditions, described in Section III-A1:

- *Manual* hand transcription (WER: 0%);

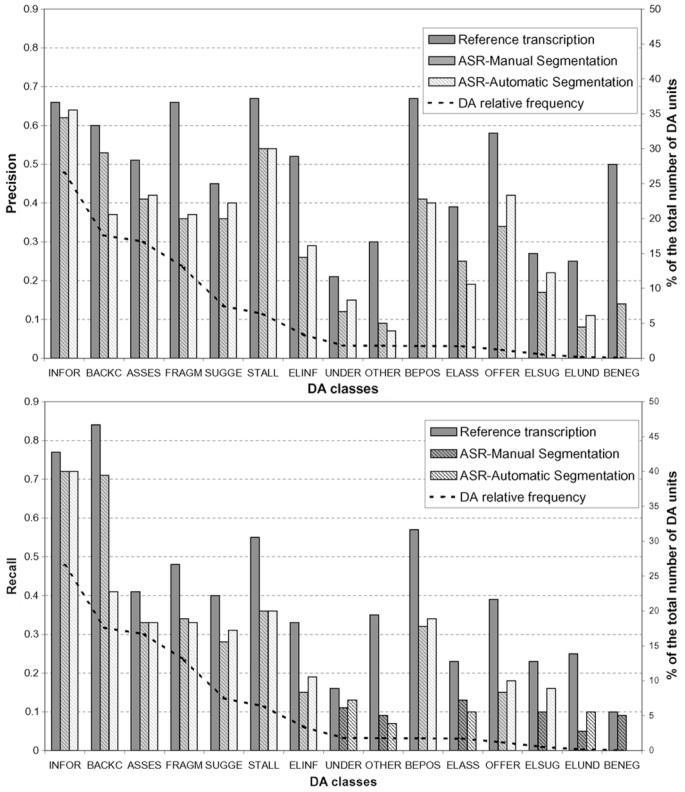

Fig. 3. DA class-based precision/recall metrics for the automatic DA tagging task on reference orthographic annotation and two versions of the ASR output. The 15 classes are sorted by their relative frequency in the AMI corpus.

- *ASR_AS* ASR with automatic segmentation: fully automatic system from ASR preprocessing up to DA segmentation and recognition (WER: 36%; 12.8% of DAs lost due to ASR deletions);
- *ASR_MS* ASR with manual segmentation: non-speaker adapted ASR with manual utterance segmentation (WER: 39%; 5.8% of DAs lost due to ASR deletions).

Although *ASR_MS* has a higher word error rate, the manual segmentation results in fewer complete DAs being deleted. Most of the deleted DA segments are very short, typically backchannels or fragments; an example of this is visible at the bottom of Fig. 4.

The *FLM* system has a classification error rate of about 10% absolute lower than the *iFLM* system for the tagging task, which uses a predefined segmentation. This is to be expected, since the additional data sources used in the *iFLM* system, the Fisher, and ICSI corpora do not have DA tags corresponding to the AMI scheme (Section III-B). Thus, although these additional data sources extend the vocabulary and n-gram counts, they are unable to provide information to help discriminate between DA classes. The trigram discourse model contributes to these results by about 7.0% absolute: DA tagging experiments using the *FLM* system without the discourse trigram, resulted in classification error rates of 47.7%, 57.5%, and 59.7%, respectively, for the *manual*, *ASR_MS*, and *ASR_AS* transcriptions.

Precision and recall of DA tagging is shown by class in Fig. 3. This graph indicates that DA tagging accuracy is influenced by the imbalanced distribution of DA labels. Not surprisingly, the classifier performs better on the two most frequent classes, *inform* and *backchannel*. However very infrequent classes such

TABLE IV
DA TAGGING, SEGMENTATION, AND RECOGNITION ERROR RATES (%) ON THE AMI MEETING CORPUS; RESULTS ARE REPORTED ON THREE DIFFERENT FLM SETUPS (BASELINE FLM, INTERPOLATED FLM, AND HYBRID FLM+IFLM) BOTH ON REFERENCE MANUAL TRANSCRIPTIONS AND ON TWO ASR OUTPUTS

		Reference transcription			ASR manual segmentation			ASR automatic segmentation		
Task	Metric	FLM	iFLM	Hybrid	FLM	iFLM	Hybrid	FLM	iFLM	Hybrid
TAG.	100 - %Correct	**40.9**	51.4	42.8	**50.7**	61.2	53.0	**52.7**	61.9	54.8
S	NIST-SU	70.7	**20.4**	25.6	77.6	**26.5**	34.1	102.6	**30.7**	34.0
E	DSER	78.0	**12.8**	17.0	85.5	**17.0**	22.8	94.2	**23.2**	25.8
G	Strict	74.4	**28.5**	36.9	81.8	**29.4**	39.5	91.5	**26.9**	33.7
M.	Boundary	10.8	**3.1**	3.9	12.8	**4.4**	5.6	16.7	**5.0**	5.5
R	NIST-SU	93.1	73.6	**71.3**	98.3	85.3	**85.9**	114.8	84.0	**81.2**
E	DER	85.5	57.0	**51.9**	91.7	67.0	**62.5**	96.5	68.6	**64.1**
C.	Strict	83.2	64.4	**62.1**	89.2	70.7	**68.5**	94.5	68.3	**64.7**
	Lenient	40.9	51.8	**42.2**	43.8	59.0	**48.3**	43.4	57.1	**46.9**

as *be-positive* and *offer* have good recall and precision scores, suggesting that even if rare, they can be well modeled and discriminated.

For the DA segmentation task, Table IV indicates that the *iFLM* system results in much lower errors, by a factor of three, compared with the basic *FLM* approach. In this case, the reduced discrimination of the *iFLM* system is outweighed by the extended dictionary and larger language model, obtained from the additional ICSI and Fisher corpora.

Since DA recognition needs both accurate segmentation and classification, we combined the *FLM* and *iFLM*, resulting in a hybrid approach which combines the two models at the decoding level. The segmentation error rates of the *hybrid* system are slightly higher than those provided by the *iFLM* approach, and the tagging error rate is slightly higher than the *FLM* approach, but on the joint recognition task, which involves both classification and segmentation, the *hybrid* provides the lowest errors.

Compared with the reference transcription, the automatically produced transcriptions, *ASR_AS* and *ASR_MS*, result in increased error rates for DA tagging, segmentation, and recognition. For tagging, the *ASR_AS* system results in an increased error of about 11% absolute, similar to that recorded on the ICSI tagging task [24]. Since the automatic DA segmentation strongly relies on the lexical content, a similar degradation can also be observed on DA segmentation metrics. The *iFLM* and *Hybrid* test conditions are less severely affected, suggesting that the larger language model results in a greater tolerance toward ASR inaccuracies. The full DA recognition task, representing a tradeoff between segmentation and classification, leads to an increase in the NIST-SU recognition metric by about 10% on *iFLM* and *Hybrid* setups and by 20% on the baseline *FLM* experiment.

However, the 12% of segments that are deleted in the *ASR_AS* transcription have an effect on the DA recognition results. In order to quantify this degradation, we compared the *ASR_AS* with the *ASR_MS* transcription which has an increased overall WER, but a reduced number of utterance deletions. Despite its higher WER, *ASR_MS* performs slightly better than *ASR_AS* on the isolated DA tagging task, although the lenient metric suggests that the situation is actually inverted when the DA classification is carried out as part of the joint DA recognition. Because of the lower number of deleted segments, *ASR_MS* outperforms *ASR_AS* on the DA segmentation subtask using both the *FLM* and *iFLM* systems. A similar discourse applies to the overall recognition performances on the baseline *FLM* setup. Thanks to the more ASR tolerant interpolated FLM and to the improved *ASR_AS* transcription quality, which leads to better dynamic classification performances (Lenient metric), *ASR_AS* offers a slightly improved DA recognition over *ASR_MS* on both *iFLM* and *Hybrid* setups.

An example of the automatic DA recognition output is shown in Fig. 4. The reference manually annotated DA units (bold text) have been aligned to the automatic DA recognizer output produced using both the reference transcription (plain text) and the *ASR_AS* output (italic text). An excerpt rich in interactions has been chosen for this example even if this often results in more ASR errors, because of overlapping speech and crosstalk between microphones, and thus in a lower DA recognition accuracy.

The switching DBN architecture generates both word sequences, using language models, and sequences of continuous prosodic features (using GMMs). We have performed a set of experiments to analyze the effect of the prosodic features. Table V gives tagging, segmentation, and recognition results for the *manual* and *ASR_AS* transcriptions, using a model that does not include the continuous prosodic features. The prosodic features do not contribute to the tagging task; hence, the results in this case are unchanged. For the segmentation and recognition tasks, it can be seen that removing the prosodic features results in a substantial increase in all the error rates, with the exception of the lenient error metric.

VI. DISCRIMINATIVE RECLASSIFICATION OF JOINT RECOGNITION OUTPUT

The use of static discriminative classifiers to rerank the output of sequential generative models has proven to be an effective technique in domains such as probabilistic parsing [25] and statistical machine translation [26]. Discriminative approaches have also been used to correct (or validate) the ASR transcription produced by a generative HMM system. Support vector ma-

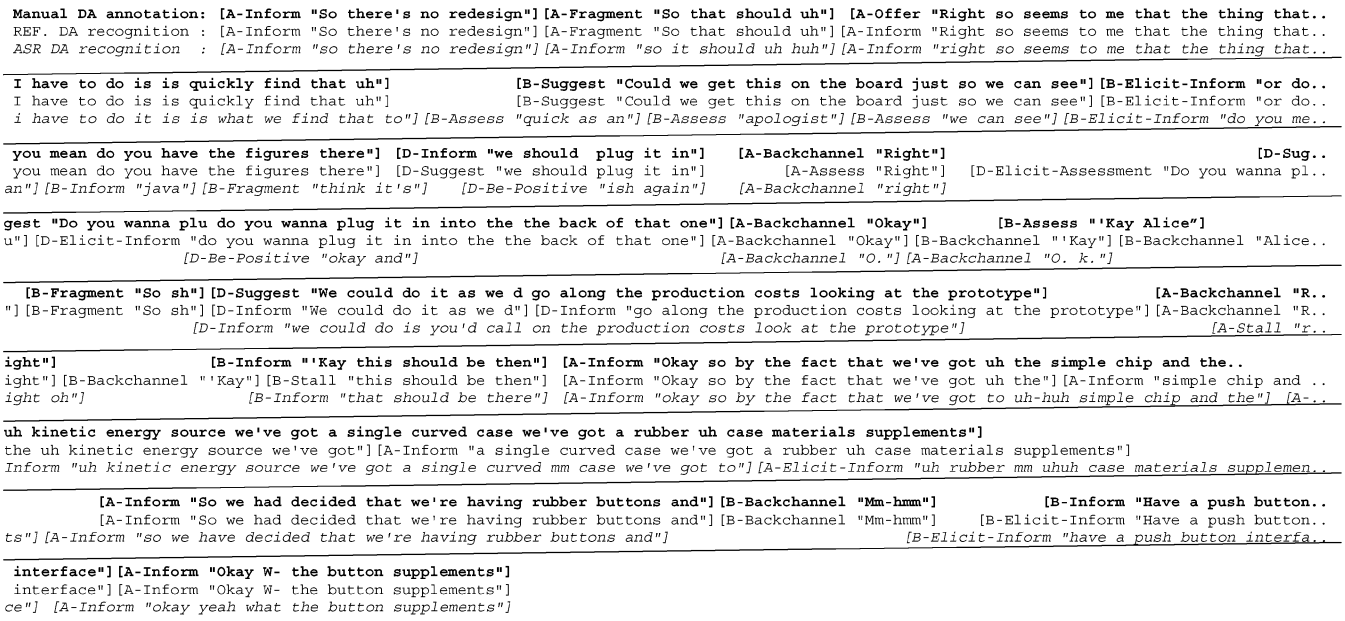

Fig. 4. Manually annotated DA units in bold (first row), and the automatic DA recognizer output obtained applying the switching DBN model with a *Hybrid* FLM configuration to the manual reference (second row) and the automatic ASR_AS transcriptions (third row, italic font). The DA segments have been specified using the following format: [speaker label—DA label "utterance"] where the four interacting speakers have been represented through the capital letters A, B, C, and D.

TABLE V
DA TAGGING, SEGMENTATION, AND RECOGNITION ERROR RATES (%) ON THE AMI MEETING CORPUS WITHOUT THE USE OF CONTINUOUS PROSODIC FEATURES; RESULTS ARE REPORTED ON THREE FLM SETUPS BOTH ON REFERENCE AND FULLY AUTOMATIC ASR TRANSCRIPTIONS

		Reference transcription			ASR automatic segm.		
Task	Metric	FLM	iFLM	Hybr.	FLM	iFLM	Hybr.
Tag.	CER	40.9	51.4	42.8	52.7	61.9	54.8
S	NIST-SU	88.5	31.9	51.8	103.0	45.6	70.9
E	DSER	79.6	24.5	36.0	99.7	47.8	62.1
G	Strict	82.7	50.7	63.2	88.6	51.2	67.5
M.	Boundary	13.5	4.9	7.9	16.8	7.4	11.5
R	NIST-SU	109.2	85.4	102.0	120.6	99.2	123.4
E	DER	86.3	61.8	61.7	104.8	85.3	87.1
C.	Strict	88.0	74.8	77.1	92.9	78.4	82.3
	Lenient	40.6	51.4	44.0	43.0	55.9	49.7

chines trained on features related to the acoustics are used in [27] to disambiguate confusable word pairs. In another application of static reranking of LVCSR n-best hypotheses, additional phonetic, lexical, syntactic, and semantic knowledge were used to discriminate between multiple recognition hypotheses [28].

This is an attractive approach for several reasons. First, since it is a postprocessing method, it may be applied to any preexisting system leaving it unaltered. Second, directly discriminant approaches explicitly optimize an error rate criterion, while exploiting temporal boundaries and recognition candidates estimated by the generative model. Finally, it is possible to add features to the joint recognition system, with the possibility of lower computational overhead.

We have applied discriminative reranking to automatic DA recognition, postprocessing the output of the *iFLM* system with

TABLE VI
DA RECOGNITION ERROR RATES (%) OF A CRF-BASED RECLASSIFICATION SYSTEM WITH AND WITHOUT THE USE OF DISCRETISED PROSODIC FEATURES; BEST PRIOR RECOGNITION PERFORMANCES USING THE *HYBRID* APPROACH HAVE BEEN REPORTED IN BRACKETS

Recognition metrics	Reference transcription	ASR manual segmentation	ASR automatic segmentation
NIST-SU	59.2 - 59.3 (71.3)	70.3 - 72.6 (85.9)	71.3 - 71.8 (81.2)
DER	46.7 - 46.7 (51.9)	56.1 - 58.0 (62.5)	59.7 - 60.0 (64.1)
Strict	54.2 - 54.5 (62.1)	59.3 - 61.2 (68.5)	57.4 - 58.2 (64.7)
Lenient	36.0 - 36.5 (42.2)	40.6 - 43.2 (48.3)	40.5 - 41.7 (46.9)

a static discriminative classifier based on conditional random fields [29]. CRFs are undirected graphical models frequently used with a simplified *linear chain* topology (first-order CRF) which can be interpreted as a generalization of HMMs. Since CRFs are trained to maximize the conditional likelihood of a given training sequence, rather than the joint likelihood, they offer improved discrimination and a better support of correlated features. Moreover, during CRF decoding, the classification decision is taken globally over the entire sequence and not locally on a single observation.

The linear chain CRF has been used to associate DA labels with the best segmentations provided by the switching DBN. The prosodic features that we used in the generative model (with the exception of F0 variance) were discretized and used in conjunction with the lexical information during the CRF relabeling process, implemented with CRF + +.[4]

Table VI reports the recognition performances after discriminative reclassification. The improvement is consistent on all the transcription conditions and on all the evaluation metrics, with reduction of 5%–12% absolute. This improvement is mainly due to the discriminative use of the lexical content;

[4][Online]. Available: http://crfpp.sourceforge.net/

prosodic features provide a marginal contribution of less than 0.5% on reference transcriptions, 2.6% on *ASR_MS*, and 1.2% on *ASR_AS*. This confirms that acoustics related features can help to discriminate between DA units with similar lexical realizations, but word identities play a more central role in DA classification. The experiments reported in Table V show that prosodic related features have a more substantial impact on the segmentation task, confirming the intuition behind exploiting the prosodic information in the switching DBN approach only for segmentation. This approximation also helped to reduce the model's complexity.

VII. Related Work

Most previous work concerned with DA modeling has focused on tagging presegmented DAs, rather than the overall recognition task which includes segmentation and tagging. Indeed, automatic linguistic segmentation [30], [31] is often regarded as a research problem itself.

The use of an HMM discourse model has underpinned most approaches to DA modeling, and a good overview of this approach is given by Stolcke *et al.* [32]. The discourse history is typically modeled using an n-gram over DAs, although approaches such as polygrams [33] have been tested. Lexical features have been widely used for DA tagging, via cue words or statistical language models, including approaches such as multiple parallel n-grams [34], hidden event language models [23], and factored language models [35]. The factored language model approach of Ji and Bilmes [35], the closest to the work reported here, presents a DA tagging approach for the ICSI corpus based on a switching DBN, using a set of 62 DA classes. Several authors have previously investigated the use of prosody to disambiguate between different DAs with a similar lexical realization [36], and investigated approaches to automatically select the most informative features [37].

More recently, there have been a number of conditional models applied to DA classification including support vector machines (SVMs) [38], [39] and maximum entropy classifiers [34], [15]. Features for these models include both lexical and prosodic cues, as well as contextual DA information [34]. As outlined in Section VI, generative and conditional approaches can also be combined. For example, Surendran *et al.* [40] integrated local discriminative SVM classifiers (using prosodic and lexical features) within the HMM discourse model by applying Viterbi decoding to class posterior probabilities estimated using the SVMs.

Automatic DA recognition, in which segmentation and tagging are combined, is less well investigated. An early system for the integrated joint DA segmentation and classification [33] employed a multilayer perceptron and a language model for segmentation, a polygram LM for DA classification, and a joint search algorithm to score multiple joint recognition hypotheses. More recently Ang *et al.* [15] have proposed a sequential approach to segment the ICSI meetings and label the detected units using five broad DA categories (statements, questions, backchannels, floorgrabbers, and disruptions). The segmentation algorithm is based on a hidden event language model (HE-LM) and a DA boundary detector based on interword pauses jointly combined through an HMM framework. A maximum entropy classifier trained on lexical, prosodic, and DA contextual features performs the final DA tagging.

We have reported some preliminary results using a joint DA recognition system on the ICSI meeting data [24], using a framework in which components such as the interpolated FLM were missing. The DA segmentation and recognition results on that system, are similar to those of Ang *et al.*, although using a discriminative MaxEnt DA classifier [15] resulted in a 5% lower error rate for the tagging task. In a later work, Zimmerman *et al.* [23] compared two joint approaches on the same experimental setup. An extended HE-LM able to predict not only DA boundaries but also the type of the DA, and a HMM recognizer inspired by HMM-based part of speech taggers, was trained on lexical features and compared using several of the metrics discussed in Section IV. The joint HE-LM system obtained lower recognition error rates than the HMM based DA recogniser, achieving performances closer to the discriminative sequential approach of [15]. A further extension of this joint HE-LM DA recognizer [16] included a discriminative maximum-entropy DA boundary detector and tagger trained on discretized interword pauses with a lexical context of four words. The weighed combination of the classification probabilities for both systems provides the most likely sequence of labeled DA units, which is able to outperform the sequential approach of [15]. Our results applying the switching DBN model to the ICSI task (Section V-A) compare favorably to this novel combined joint approach. Although for tagging the FLM is less accurate than a discriminative DA classifier [15], the situation is inverted on the DA segmentation task [16], thanks to the added capability to include additional in-domain data by adopting an interpolated FLM. Joint recognition experiments, reported in Section V-A, suggest that these two effects can be carefully balanced (hybrid approach), leading to a competitive DA recognizer which performs well in comparison with the state of the art [16].

VIII. Summary and Conclusion

We have presented a framework for the automatic recognition of dialogue acts in multiparty conversations. DA recognition experiments were carried out on the AMI meeting corpus using a dictionary of 15 DA classes tailored for decision making meetings, and on the ICSI corpus using a more generic set of five DA classes. The system that we have presented employs a generative probabilistic approach implemented through the integration of a heterogeneous set of technologies: six continuous prosodic features extracted from the lexical and prosodic content facilitate the segmentation process; a trigram discourse language model estimated from observed sequences of DAs; a factored language model interpolated using multiple conversational data resources, used in conjunction with a plain FLM trained solely on in-domain data; and a switching DBN model with two alternating topologies, which coordinates the joint DA segmentation and classification task by integrating the available resources. Multiple concurrent DA segmentation and classification hypotheses are evaluated by this joint DA recognizer, enabling the investigation of a larger search space compared with a two-step sequential segmentation-classification approach.

Three experimental systems were investigated based on a simple FLM, an interpolated FLM, or a hybrid using both. The

simple FLM trained only on data from the target corpus offers the most accurate DA classification. However, the interpolated FLM, thanks to its richer dictionary and language model, reduces the number of segmentation errors by a factor of 2–3, at the cost of a slightly degraded DA classification accuracy. A hybrid approach, using both FLMs, allows a tradeoff between segmentation and classification, to improve the overall recognition accuracy. Experiments using each of the three systems on hand-transcribed and two kinds of automatically transcribed data, showed that these systems generalize well to automatic imperfect transcriptions. A further significant improvement in the recognition accuracy, of 5%–12%, was obtained using a discriminative DA reclassification process based on conditional random fields.

The degradation when moving from manual transcriptions to the output of a speech recognizer is less than 15% absolute for most tasks and metrics. Our experiments indicate that it is possible to perform automatic segmentation into DA units with a relatively low error rate. However, the operations of tagging and recognition into 15 imbalanced DA categories have a relatively high error rate, even after discriminative reclassification, indicating that this remains a challenging task. As the first complete set of DA recognition experiments reported on the AMI meetings, this work can also provide a baseline reference system for future work on this corpus.

REFERENCES

[1] G. Murray, S. Renals, J. Moore, and J. Carletta, "Incorporating speaker and discourse features into speech summarization," in *Proc. HLT-NAACL*, Jun. 2006, pp. 367–374.

[2] M. Galley, K. R. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proc. 41st Annu. Meeting Assoc. Comput. Linguist. (ACL-03)*, Jul. 2003, pp. 562–569.

[3] P. Hsueh and J. Moore, "Automatic topic segmentation and labelling in multiparty dialogue," in *Proc. IEEE/ACL Workshop Spoken Lang. Technol.*, Dec. 2006, pp. 98–101.

[4] M. Al-Hames, A. Dielmann, D. Gatica-Perez, S. Reiter, S. Renals, G. Rigoll, and D. Zhang, "Multimodal integration for meeting group action segmentation and recognition," in *Proc. Multimodal Interaction and Rel. Mach. Learn. Algorithms Workshop (MLMI-05)*, 2006, pp. 52–63.

[5] I. McCowan, D. Gatica-Perez, S. Bengio, and G. Lathoud, "Automatic analysis of multimodal group actions in meetings," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, pp. 305–317, Mar. 2005.

[6] A. Dielmann and S. Renals, "Automatic meeting segmentation using dynamic Bayesian networks," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 25–36, Jan. 2007.

[7] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, "Detection and application of influence rankings in small group meetings," in *Proc. Int. Conf. Multimodal Interfaces (ICMI-06)*, Nov. 2006, pp. 257–264.

[8] M. Purver, J. Niekrasz, and P. Ehlen, "Automatic annotation of dialogue structure from simple user interaction," in *Proc. Multimodal Interaction and Rel. Mach. Learn. Algorithms Workshop (MLMI-07)*, 2007, pp. 48–59.

[9] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Proc. Multimodal Interaction and Rel. Mach. Learn. Algorithms Workshop (MLMI-05)*, 2006, pp. 28–39.

[10] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. IEEE ICASSP*, Apr. 2003, vol. 1, pp. 364–367.

[11] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *Proc. LREC*, May 2004, pp. 69–71.

[12] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proc. HLT-NAACL SIGDIAL Workshop*, Apr.–May 2004, pp. 97–100.

[13] 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. IEEE ICASSP*, Apr. 2007, vol. 4, pp. 357–360.

[14] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. New York: Elsevier, 1995, pp. 495–518.

[15] J. Ang, Y. Liu, and E. Shriberg, "Automatic dialog act segmentation and classification in multiparty meetings," in *Proc. IEEE ICASSP*, Mar. 2005, vol. 11, pp. 1061–1064.

[16] M. Zimmermann, A. Stolcke, and E. Shriberg, "Joint segmentation and classification of dialog acts in multiparty meetings," in *Proc. IEEE ICASSP*, May 2006, vol. 1, pp. I-581–I-584.

[17] J. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proc. HLT-NAACL*, May 2003, vol. 2, pp. 4–6.

[18] K. Duh and K. Kirchhoff, "Automatic learning of language model structure," in *Proc. Int. Conf. Comput. Linguist. (COLING)*, Nov. 2004, article no.148.

[19] J. Bilmes, "Dynamic Bayesian multinets," in *Proc. Int. Conf. Uncertainty in Artif. Intell.*, Jun.–Jul. 2000.

[20] J. Bilmes and G. Zweig, "The Graphical model ToolKit: An open source software system for speech and time-series processing," in *Proc. IEEE ICASSP*, Jun. 2002, vol. 4, pp. 3916–3919.

[21] A. Dielmann and S. Renals, "DBN based joint dialogue act recognition of multiparty meetings," in *Proc. IEEE ICASSP*, Apr. 2007, vol. 4, pp. 133–136.

[22] S. Lesch, T. Kleinbauer, and J. Alexandersson, "A new metric for the evaluation of dialog act classification," in *Proc. Workshop Semantics and Pragmatics of Dialogue (SEMDIAL)*, Jun. 2005, pp. 143–146.

[23] M. Zimmermann, Y. Liu, E. Shriberg, and A. Stolcke, "Toward joint segmentation and classification of dialog acts in multiparty meetings," in *Proc. Multimodal Interaction and Rel. Mach. Learn. Algorithms Workshop (MLMI-05)*, 2006, pp. 187–193.

[24] A. Dielmann and S. Renals, "Multistream recognition of dialogue acts in meetings," in *Proc. Multimodal Interaction and Rel. Mach. Learn. Algorithms Workshop (MLMI-06)*, 2007, pp. 178–189.

[25] M. Collins and T. Koo, "Discriminative reranking for natural language parsing," *Computat. Linguist.*, vol. 31, no. 1, pp. 25–70, 2005.

[26] L. Shen, A. Sarkar, and F. Och, "Discriminative reranking for machine translation," in *Proc. HLT-NAACL*, May 2004, pp. 177–184.

[27] V. Venkataramani, S. Chakrabartty, and W. Byrne, "*Gini* support vector machines for segmental minimum Bayes risk decoding of continuous speech," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 423–442, Jul. 2007.

[28] M. Balakrishna, D. Moldovan, and E. Cave, "N-best list reranking using higher level phonetic, lexical, syntactic and semantic knowledge sources," in *Proc. IEEE ICASSP*, May 2006, vol. 1, pp. 413–416.

[29] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2001, pp. 282–289.

[30] A. Stolcke and E. Shriberg, "Automatic linguistic segmentation of conversational speech," in *Proc. ICSLP*, Oct. 1996, vol. 2, pp. 1005–1008.

[31] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Commun.*, vol. 32, pp. 127–154, Sep. 2000.

[32] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, D. Jurafsky, P. Taylor, R. Martin, C. V. 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.

[33] V. Warnke, R. Kompe, H. Niemann, and E. Nöth, "Integrated dialog act segmentation and classification using prosodic features and language models," in *Proc. Interspeech-Eurospeech*, Sep. 1997, vol. 1, pp. 207–210.

[34] A. Venkataraman, Y. Liu, and E. Shriberg, "Does active learning help automatic dialog act tagging in meeting data?," in *Proc. Interspeech-Eurospeech*, Sep. 2005, pp. 2777–2780.

[35] G. Ji and J. Bilmes, "Dialog act tagging using graphical models," in *Proc. IEEE ICASSP*, Mar. 2005, vol. 1, pp. 33–36.

[36] S. Bhagat, H. Carvey, and E. Shriberg, "Automatically generated prosodic cues to lexically ambiguous dialog acts in multiparty meetings," in *Proc. Int. Congr. Phonetic Sci.*, Aug. 2003, pp. 2961–2964.

[37] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, M. Meteer, and C. V. Ess-Dykema, "Can prosody aid the automatic classification of dialog acts in conversational speech?," *Lang. Speech*, no. 41, pp. 439–487, 1998.

[38] R. Fernandez and R. W. Picard, "Dialog act classification from prosodic features using support vector machines," in *Proc. Speech Prosody 2002*, Apr. 2002, pp. 291–294.

[39] Y. Liu, "Using SVM and error-correcting codes for multiclass dialog act classification in meeting corpus," in *Proc. Interspeech-ICSLP*, Sep. 2006, pp. 1938–1941.

[40] D. Surendran and G. A. Levow, "Dialog act tagging with support vector machines and hidden Markov models," in *Proc. Interspeech-ICSLP*, Sep. 2006, article no. 1831.

Alfred Dielmann received the Laurea degree in electronic engineering from the University of Cagliari, Cagliari, Italy, in 2002. He is currently pursuing the Ph.D. degree at the Centre for Speech Technology Research, School of Informatics, University of Edinburgh, Edinburgh, U.K.

From 2002 to 2003, he was a Graduate Research Assistant at the Speech and Hearing Research Group, Computer Science Department, University of Sheffield, Sheffield, U.K. Since October 2003, he has been a Research Associate at the Centre for Speech Technology Research, School of Informatics, University of Edinburgh. His research interests concern multimodal signal processing and machine learning, in particular probabilistic graphical models for multiparty interaction modeling and natural language processing.

Steve Renals (M'91) received the B.Sc. degree from the University of Sheffield, Sheffield, U.K., in 1986, and the M.Sc. and Ph.D. degrees from the University of Edinburgh, Edinburgh, U.K., in 1987 and 1991, respectively.

He is a Professor of speech technology and Director of the Center for Speech Technology Research at the University of Edinburgh. He held postdoctoral fellowships at the International Computer Science Institute, Berkeley, CA, (1991–1992) and at the University of Cambridge, Cambridge, U.K. (1992–1994). He was a Member of Academic Staff at the University of Sheffield for nine years as a Lecturer (1994–1999), then Reader (1999–2003). His main research areas are in acoustic and language modeling for speech recognition and audio and multimodal information access, and he has over 100 publications in these areas. Activities in the latter area include spoken document retrieval, speech summarization, and models that incorporate multiple modalities, particularly in the context of meetings.