AUTOMATIC SOCIAL ROLE RECOGNITION IN PROFESSIONAL MEETINGS USING CONDITIONAL RANDOM FIELDS

Ashtosh Sapru[1,2] and Hervé Bourlard[1,2]

[1] Idiap Research Institute, 1920, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
ashtosh.sapru@idiap.ch,herve.bourlard@idiap.ch

ABSTRACT

Social roles characterize relation between participants in a conversation and, in turn, influence their interaction patterns. This paper investigates automatic social role recognition in professional meetings using a completely discriminative framework based on conditional random fields. We present a novel approach which combines information from multiple layers of data. The conversation layer models the influence of social roles on turn taking patterns of participants present in multiparty interactions. A conditional random field augmented with hidden state sequences is used to estimate the posterior distribution of social roles in this layer. The other novelty of our approach consists in modeling statistical dependencies between roles across adjacent segments of meeting. The posterior distribution estimated in conversation layer is combined with role transition information to improve the model. Experiments conducted on more than 40 hours of data reveal that the proposed approach reaches a recognition accuracy of 67% in classifying four social roles using information from conversation layer. Moreover, recognition accuracy increases to 70% when information from multiple layers is taken into consideration.

Index Terms— Social Role Labeling, Turn Taking, Conditional random fields, Lexical and Prosody.

1. INTRODUCTION

Analyzing spoken documents in terms of speaker role information is useful for enriching the content description of multimedia data. It can be used in applications like information retrieval, enhancing multimedia content browsing and allowing summarization of multimedia documents [1]. Speaker roles are stable behavioral patterns in an audio recording and the problem of role recognition consists in assigning a label, i.e., a role to each of the speakers. Automatic labeling of speaker roles has been widely studied in case of Broadcast News (BN) recordings [2, 3, 4]. These roles are imposed from the news format and relate to the task each participant performs in the conversation like anchorman, journalists, interviewees, etc. In the last few years automatic role recognition has also been investigated for meeting recordings and broadcast conversations. Typical roles in these studies can vary with environment and applications such as project manager in AMI corpus [5], student, faculty member in ICSI corpus [6]. Common features used in these studies extract relevant information from conversation features, lexical features, prosody and dialog act tags [2, 3, 4].

For the studies mentioned above participants role is formal and considered to remain constant over the duration of entire audio recording. Other role coding schemes have also been proposed in literature which put roles in a more dynamic setting, such as socio-emotional roles (here after referred to as social roles) [7, 8, 9, 10]. Social roles describe relation between conversation participants and their roles "*oriented towards functioning of group as a group*". Social roles are useful to characterize the dynamics of the conversation, i.e., the interaction between the participants and can be generalized across any type of conversation. They are also related to phenomena studied in meetings like social dominance, engagement and also hot-spots [11].

Automatic social role recognition was first investigated in meetings recorded for problem solving sessions [7]. Their approach was based on applying a support vector machine classifier to discriminate between social roles using simple speech activity features. Other studies have also investigated social role recognition in professional meetings (AMI corpus). The research in [12] revealed that automatically extracted subjectivity features from lexical and prosodic cues are correlated with social roles. More recently, in [10] a multiclass boosting classifier was used to integrate evidence from several information streams i.e. speech activity, dialog act tags, lexical and prosody for social role recognition. Investigations in this work also highlighted that some social roles are more correlated with lexical content and dialog act tags. The generative classifiers for social role recognition based on Hidden Markov Model (HMM) were considered in [9, 8]. While the work in [8] used speech activity and video features, in [9] the generative framework was used to combine prosody and turn duration. Discriminative approaches [10, 7] do not consider the sequential nature of conversations while generative approaches [9] model sequential information under the assumption that multistream observations are conditionally independent. Furthermore, previous studies in social role recognition have relied on limited amount of data for implementing automatic role recognition systems and it needs to be investigated how these algorithms scale on larger datasets.

In this paper we investigate social role recognition on a much larger database containing 128 different speakers for a total of more than 40 hours of speaker data. The social roles of participants within a segment of meeting (a relatively short windowed chunk of meeting recording) are considered as fixed. Our role recognition system is based on a completely discriminative framework which employs conditional random fields (CRF) [13] on multiple layers of data. The conversation layer is an abstraction to model social role of a participant within a meeting segment. In this work we model conversation layer using a hidden conditional random field (HCRF). HCRFs have earlier been successfully used for phone classification [14] and gesture recognition [15] where they show superior performance when compared with HMMs.

In comparison to CRF which associates a label sequence with

an observation sequence, HCRF associates a single class label with an observation sequence. The association between class label and observation sequence is mediated using a hidden layer of variables which capture latent structure in data. In the social role classification task, observation sequence in the conversation layer is collected from multiple information streams including prosody, word usage and duration, produced at multiple times due to turn taking in conversations. HCRF estimates posterior probability of social role from observation sequence for each participant in a windowed segment of recording. The posterior distribution estimated at conversation layer is, in turn, used to model the segment layer. The segment layer investigates statistical dependencies between social roles of a participant in adjacent meeting segments. A CRF is used to model complete role sequence for each participant over the length of meeting recording, where observation sequence comes from role posteriors estimated in the conversation layer.

2. DATA AND ANNOTATION

The AMI Meeting Corpus is a collection of meetings captured in specially instrumented meeting rooms, which record the audio and video for each meeting participant. The corpus contains both scenario and non-scenario meetings. In the scenario meetings, four participants play the role of a design team composed of *Project Manager (PM), Marketing Expert (ME), User Interface Designer (UI),* and *Industrial Designer (ID)* tasked with designing a new remote control. A subset of 59 meetings from the scenario portion of AMI Meeting Corpus containing 128 different speakers (84 male and 44 female participants) is selected from the entire corpus. Subsequently each meeting was segmented into short clips (with a minimum duration of 20 seconds) based on presence of long pauses i.e. pauses longer than 1 second. Within each such meeting segment social role of the participant is assumed to remain constant. From each meeting a total duration of approximately 12 minutes long audio/video data was selected. Meeting segments are resampled so as to cover the entire length of recording comprising various parts of meeting such as openings, presentation, discussion and conclusions.

Since social roles are subjective labels and require human annotators, the annotation scheme was implemented as follows. Each annotator is asked to view and listen the entire video segment and tasked with assigning a speaker to role mapping based on a list of specified guidelines. These guidelines define a set of acts and behaviors that characterize each social role and is summarized in the following: *Protagonist* - a speaker that takes the floor, drives the conversation, asserts its authority and assume a personal perspective; *Supporter* - a speaker that shows a cooperative attitude demonstrating attention and acceptance providing technical and relational support; *Neutral* - a speaker that passively accepts other speaker's ideas; *Gatekeeper* - a speaker that acts like group moderator, mediates and encourage the communication; *Attacker* - a speaker who deflates the status of others, express disapproval and attacks other speakers. At least 10 annotators were asked to label each video clip.

Figure 1 shows the distribution of roles over all the meeting segments present in the data set. It can be seen that the neutral role has been labeled most often by annotators. This is followed by supporter, gatekeeper and protagonist. Comparatively the attacker role has received the fewest labels as observed by multiple annotators. A reason for this distribution may be due to collaborative nature of AMI meetings. The reliability of labeling scheme as measured through Fliess's kappa shows a value 0.5 which is considered to have moderate agreement according to Landis and Koch's criterion [7]. In terms of inter annotator agreement we find that neutral label is most reliable one as measured through category wise κ statistic with a value of 0.7. The intermediate level of agreement is present for supporter 0.36 and gatekeeper 0.38 labels. This is followed by the protagonist role which shows a fair level of agreement with a κ value of 0.29. One difference from the earlier studies [9, 10] is the higher percentage of gatekeeper role. A reason for this behavior can be explained from Figure 2 which reveals that annotators were more likely to associate the role of PM with gatekeeper compared to other formal roles.

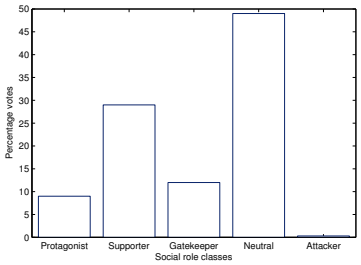

Fig. 1. Social role distribution in the annotated corpus. The vertical axis represents percentage votes for each class as labeled by multiple annotators.

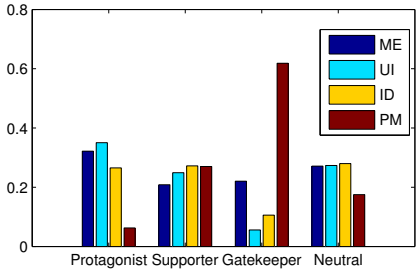

Fig. 2. Social role distribution conditioned on formal role that the speaker has in the meeting.

3. FEATURE EXTRACTION

Audio from the independent headset microphones (IHM) is processed through a speech segmentation system [16] for obtaining estimated speech/non-speech boundaries for each meeting participant. The output of speech/non speech system for each speaker is a sequence of speech and silence regions in time which arise due to turn taking in conversations. However, since meeting conversations involve multiple speakers, some activity regions (speech overlaps) will have more than one participant speaking simultaneously. Also silence regions corresponding to each participant can take multiple meanings. Silence due to conversation floor changes or whenever speakers pause to take breathe. On the other hand silence regions can simply be the listening silence from the perspective of some speakers when other speaker(s) is/are speaking.

Each participant's sequence of speech silence regions are tagged with one of the turn taking states defined as: talkspurts (TS) - a region of speech when only a single speaker speaks; pauses (PA) - regions when all the speakers are silent; overlaps (OV) - regions where multiple speakers are speaking simultaneously; listening silence (LS) - regions from perspective of current participant when some other speaker is speaking. We hypothesize that social role influence the distribution of turn taking states. For example, it is more

likely that a participant with a more active role will grab the conversation floor after a pause. Similarly, the participant's role is expected to affect whether it keeps control of conversation after a speech overlap or not. The participant's interaction record is represented by $\{q_n\}_{n=1}^{N}$, where $q_n \in \{PA, OV, LS, TS\}$, and N is the length of sequence under consideration. Also associated with each state q_n is its duration d_n. Each of these regions is smoothed using a minimum duration criterion. Furthermore we also extract prosodic and lexical features aligned with TS and OV regions. Prosodic feature vector X_n^p is represented using measures like mean F0 mean,max,min and slope, mean energy and speech rate (see [10] for details). Lexical features X_n^l are words corresponding to speaker utterances including backchannels.

4. ROLE RECOGNITION

A meeting $\mathbf{M}$ in the corpus can be seen as a sequence of windowed segments $\{\mathbf{M_1}, \mathbf{M_2}, ..., \mathbf{M_K}\}$ arranged from start to end of meeting. The participants in each meeting segment have a social role assigned to them as a result of annotation process. While social role of a participant changes across segments, within each segment the role of the participant is fixed. The feature extraction step associates with some participant $\mathcal{P}$ in $\mathbf{M_k}$ a conversation sequence $C_k^{\mathcal{P}} = \{(q_n, d_n, X_n^p, X_n^l)\}_{n=1}^{N_k}$, where N_k represents the length of conversation sequence in segment k. The problem of automatic role recognition is defined as assigning a role sequence $\tilde{\mathbf{R}} = \{R_1, R_2, ..., R_K\}^{\mathcal{P}}$ for every participant $\mathcal{P}$ present in the meeting, where R_k comes from a finite set of possible social roles $\mathcal{R}$. In a probabilistic framework, this assignment satisfies the following equation:

$$\tilde{\mathbf{R}} = \underset{\{R_1,...,R_K\} \in \mathcal{R}^K}{\arg\max}\, p(\{R_1, ..., R_K\}^{\mathcal{P}} | \{C_1, ..., C_K\}^{\mathcal{P}}) \quad (1)$$

In this work, the information present in each conversation sequence C_k, is coded in terms of probability vector ϕ_k which is estimated from a separate model and we work under the assumption $p(\mathbf{R}|\{C_k\}^{\mathcal{P}}, \{\phi_k\}^{\mathcal{P}}) = p(\mathbf{R}|\{\phi_k\}^{\mathcal{P}})$. In the following we provide details about estimation of ϕ_k and its use in role recognition system.

The probability vector ϕ_k needs to code all the available social role information in the conversation layer. In that case the optimal coding would be represented in terms of conditional probability of participants current role given the conversation sequence C_k, i.e., $\phi_k = \{p(R_k|C_k)\} \forall R_k \in \mathcal{R}$, here we have dropped the index $\mathcal{P}$ for notational convenience. The present work estimates $\{p(R_k|C_k)\}$ using a chain structured HCRF, which models the conditional probability distribution as:

$$P(R_k|\lambda, C_k) = \frac{\sum_{\mathbf{h} \in R_k} \exp(\lambda . f(R_k, \mathbf{h}, C_k))}{Z(C_k, \lambda)} \quad (2)$$

Here λ is a parameter vector and $f(R_k, \mathbf{h}, C_k)$ is a feature vector. Following [14], we refer to f as feature vector of sufficient statistics which are extracted from conversation sequence C_k. The term $Z(C_k, \lambda)$ is called partition function and acts as a normalization constant to ensure an appropriately normalized probability distribution.

$$Z(C_k, \lambda) = \sum_{R_k} \sum_{\mathbf{h} \in R_k} \exp(\lambda . f(R_k, \mathbf{h}, C_k)) \quad (3)$$

The choice of sufficient statistics vector f determines the dependencies modeled by the HCRF. In this work we have consider the following components for f,

$$\begin{aligned}
f^{ULM}(R_k, \mathbf{h}, C_k) &= \delta(R_k = \hat{R}) \\
f^{Tr}(R_k, \mathbf{h}, C_k) &= \sum_t \delta(h_t = h, h_{t-1} = \hat{h}) \\
f^{Occ}(R_k, \mathbf{h}, C_k) &= \sum_t \delta(h_t = h) \\
f_{q0}^{Obs}(R_k, \mathbf{h}, C_k) &= \delta(h_0 = h, q_0 = q) \\
f_{\hat{q}q}^{Obs}(R_k, \mathbf{h}, C_k) &= \sum_t \delta(h_t = h, q_t = q, q_{t-1} = \hat{q}) \\
f_{qx}^{Obs}(R_k, \mathbf{h}, C_k) &= \sum_t \delta(h_t = h, q_t = q) x_t \\
f_{ql}^{Obs}(R_k, \mathbf{h}, C_k) &= \sum_t \delta(h_t = h, q_t = q, x_t^{lex} = w)
\end{aligned} \quad (4)$$

where f^{ULM} is 1 when current role label is $\hat{r}$, 0 otherwise. The transition features f^{Tr} count the number of times a specific transition $\hat{h}h$ occurs in $\mathbf{h}$ and f^{Occ} counts the occurrence of state h. f_{q0}^{Obs} triggers when conversation label at the start takes a values q. This is expected to represent which roles are more likely to start the conversation. $f_{\hat{q}q}^{Obs}$ represents the number of times conversation state changes from $\hat{q}$ to q and represents information whether a role is likely to grab conversation floor after a pause or an overlap. f_{qx}^{Obs} are continuous features which represent duration and prosody information for a given conversation state while f_{ql}^{Obs} counts the occurrence of word w in conversation. The topology of HCRF model forms a Markov chain as can be seen from the feature terms in sufficient statistics which depend on at most a pair of adjacent nodes. This makes it possible to use efficient algorithms like Forward Backward and Viterbi decoding similar to HMMs.

To estimate the parameters of model λ we train the model given a set of $\{R_i, C_i\}_{i=1}^{T_N}$, where T_N is the total number of conversation sequences available for training. The parameter vector λ is found by maximizing the conditional log likelihood of training data,

$$L(\lambda) = \sum_i \log p(R_i|C_i; \lambda) \quad (5)$$

The objective function can be maximized using an iterative algorithm like stochastic gradient descent or second order approaches like L-BFGS. In this work we have used L-BFGS algorithm as it is a scalable with low memory requirements and has been applied successfully for training CRFs [14].

The trained HCRF model is used to estimate $\phi_k = \{P(R_k/C_k)\}$, which serve as features in role modeling at the segment layer. We hypothesize that social roles of a participant are correlated across adjacent meeting segments. This information is used in a linear chain CRF which estimates the role sequence of a participant for the segments of meeting recording. The model is described as,

$$p(\mathbf{R}|\{\phi_k\}^{\mathcal{P}}, \theta) = \frac{\sum_k \sum_j exp(\theta_j . g_j(R_{k-1}, R_k, \phi_k))}{Z(\{\phi_k\}, \theta)} \quad (6)$$

Similar to Equation 3, $Z(\{\phi_k\}, \theta)$ is the normalization constant and θ is the parameter vector for CRF model. The parameters are trained in a similar fashion, i.e., by maximizing the conditional log likelihood of the role sequence given the posterior probability estimate sequence, as described in case of HCRF model.

5. EXPERIMENTS

For evaluation of proposed method experiments were conducted using repeated cross-validation, wherein one set of meetings (all but

Table 1. Per role F-measure, Precision and Recalls obtained in recognizing social roles for the three considered models. Asterisk besides the accuracy shows that improvement compared to baseline is statistically significant according to paired *t* test with rejection of null hypothesis at 1%

	Per-role F-measure (Precision/Recall)				Accuracy
Model	Protagonist	Supporter	Gatekeeper	Neutral	
Baseline	0.45 (0.42/0.48)	0.72 (0.84/0.63)	0.53 (0.52/0.55)	0.52 (0.39/0.81)	0.61
HCRF	0.51 (0.47/0.57)	0.74 (0.77/0.72)	0.56 (0.55/0.57)	0.69 (0.69/0.69)	0.67*
CRF	**0.58 (0.54/0.63)**	**0.76 (0.78/0.74)**	**0.62 (0.63/0.62)**	**0.70 (0.70/0.70)**	**0.70***

two) was kept for training/tuning the model parameters, while a distinct set (remaining two meetings) was used for evaluation. The partition of meetings was done keeping in view that participants with same speaker identity do not appear in both training and test set. The ground truth for participant role labels was derived by majority voting. An initial filtering was done to consider only those meeting segments where a participant is active, also a few meeting segments, were majority voting resulted in participant having an attacker role label were not considered (very few labels, see Figure 1).

During the feature extraction process, the fundamental frequency (F0) is computed from the headset microphones using 30ms long windows shifted by 10ms. All the lexical information was extracted using output of AMI-ASR system [17]. The word error rate for the system is less than 25%. During training of HCRF and CRF models a regularization term was added to the objective function to avoid overfitting. This term corresponds to using a Gaussian prior with variance σ^2, i.e., $P(\lambda) \propto exp(\frac{||\lambda||^2}{2\sigma^2})$. The tuning parameters were also selected by evaluations on a randomly sampled portion of training data. All the models were evaluated on a separated test set and performance measured in terms of recognition accuracy and F-measure/Precision/Recall.

The baseline model is based upon the work presented in [10] which predicts a role label for each speaker turn observed in the meeting segment. The evidence from duration, lexical and prosody is combined using multiclass Boosting algorithm. Each speaker has a unique role within the meeting segment, however since the baseline system predicts labels at the turn level, we need to map turn labels for a speaker to a single role. This mapping is done by summing the duration for every turn in the segment which corresponds to a specific role label, thereby calculating the total duration for each role. The segment level predicted role is allocated as the label with the maximum duration.

Table 1 compares the performance of the baseline model, HCRF model and CRF trained on HCRF posteriors. Also reported are performance figures for different roles. It can be seen that baseline model doesn't perform as well as other models especially in recognition of neutral and protagonist roles. However, compared to baseline model, HCRF improves performance over all roles. The absolute improvement in overall recognition accuracy between the two models being 6%. In terms of individual roles, table numbers reveal that supporter achieves the highest F-measure 0.75 while protagonist achieves the lowest F-measure 0.49. The low performance for protagonist can be related to it also being the role with a lower κ score. The final model which uses the statistical dependencies between roles of adjacent meeting segments in a CRF shows the best performance 70%. The improvements come from both gatekeeper and protagonist roles. This behavior can be explained from Figure 3 which shows the normalized histogram of role distribution for a segment conditioned on the role for previous segment. From the figure it can be seen both gatekeeper and protagonist roles are more likely to continue across adjacent segments. Finally, the influence of hidden layer on the performance of system is shown in Figure 4. It can be seen that performance of the model depends on the number of hidden states. The best performance is obtained in the case of $4-5$ hidden states after which it saturates.

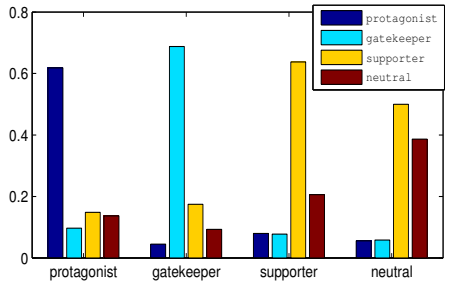

Fig. 3. Social role distribution conditioned on participants social role in the previous meeting segment.

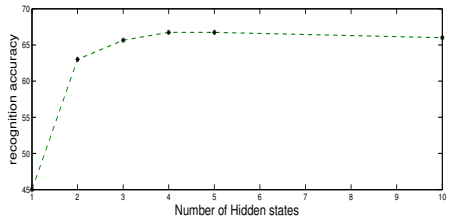

Fig. 4. Variation in social role recognition accuracy as the number of hidden states is increased in HCRF model.

6. CONCLUSION

In this paper, we investigated a completely discriminative framework for automatic social role recognition based on conditional random fields. The proposed approach models statistical dependencies at multiple layers of data. The conversation layer implements a HCRF to model turn taking patterns in meetings and combines duration, prosody and lexical information to reach an accuracy of 67% in classifying four social roles. The segment layer uses the posterior role distribution estimated at the conversation layer and statistical dependencies between roles to further increase accuracy to 70%. In summary proposed approach leads us to conclude that recognizing social roles requires extracting meaningful information at different layers of data. In future we plan to extend our study on other meeting environments.

7. ACKNOWLEDGEMENT

This work was funded by the Hasler Stiftung under SESAME grant, the EU NoE SSPNet, and the Swiss National Science Foundation NCCR IM2.

8. REFERENCES

[1] Gabriel Murray, Pei-Yun Hsueh, Simon Tucker, Jonathan Kilgour, Jean Carletta, Johanna Moore, and Steve Renals, "Automatic Segmentation and Summarization of Meeting Speech," *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2007.

[2] Barzilay R., Collins M., Hirschberg J., and Whittaker S., "The rules behind roles: Identifying speaker role in radio broadcasts," *Proceedings of AAAI*, 2000.

[3] Wang W., Yaman S., Precoda P., and Richey C., "Automatic identification of speaker role and agreement/disagreement in b roadcast conversation.," in *Proceedings of ICASSP*, 2011.

[4] Damnati G. and Charlet D., " Robust speaker turn role labeling of TV Broadcast News shows," *proceedings of ICASSP*, 2011.

[5] Salamin H. et al., "Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction," *IEEE Transactions on Multimedia*, vol. 11, November 2009.

[6] Laskowski K., Ostendorf M., and Schultz T., "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, 2008.

[7] Zancaro M. et al., "Automatic detection of group functional roles in face to face interactions," *Proceedings of ICMI*, 2006.

[8] Dong W. et al., "Using the influence model to recognize functional roles in meetings," *Proceedings of ICMI*, 2007.

[9] Valente F. and Vinciarelli A., "Language-independent socio-emotional role recognition in the ami meetings corpus," in *Proceedings of Interspeech*, 2011.

[10] Sapru A. and Valente F., "Automatic speaker role labeling in AMI meetings: recognition of formal and social roles," *Proceedings of Icassp*, 2012.

[11] Wrede D. and Shriberg E., "Spotting "hotspots" in meetings: Human judgments and prosodic cues," *Proc. Eurospeech*, 2003.

[12] Wilson T. et. al., "Using linguistic and vocal expressiveness in social role recognition," *Proceedings of the Conference on Intelligent User Interfaces(IUI)*, 2011.

[13] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

[14] Asela Gunawardana, Milind Mahajan, Alex Acero, and John C. Platt, "Hidden conditional random fields for phone classification," in *Proceedings of Interspeech*, 2005.

[15] Ariadna Quattoni, Michael Collins, and Trevor Darrell, "Conditional random fields for object recognition," in *In NIPS*, 2004.

[16] Hain, Vepa J., and J. T.Dines, "The segmentation of multi-channel meeting recordings for automatic speech recognition," *Proceedings of Interspeech*, 2006.

[17] Hain T., Wan V., Burget L., Karafiat M., J. Dines, Vepa J., Garau G., and Lincoln M., "The AMI System for the Transcription of Speech in Meetings.," *Proceedings of Icassp*, 2007.