Recovering Participant Identities in Meetings from a Probabilistic Description of Vocal Interaction

Kornel Laskowski and Tanja Schultz

Cognitive Systems Lab, Universität Karlsruhe, Karlsruhe, Germany
Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, USA
kornel@cs.cmu.edu

Abstract

An important decision in the design of automatic conversation understanding systems is the level at which information streams representing specific participants are merged. In the current work, we explore participant-dependence of low-level interactive aspects of conversation, namely the observed contextual preferences for talkspurt deployment. We argue that strong participant-dependence at this level gives cause for merging participant streams as early as possible. We demonstrate that our probabilistic description of talkspurt deployment preferences is strongly participant-dependent, and frequently predictive of participant identity.

Index Terms: Vocal interaction, Automatic conversation understanding, Meetings.

1. Introduction

An important decision in the design of automatic multiparticipant conversation understanding systems pertains to the level at which information streams representing specific participants are merged. The majority of such systems, to date, have delayed fusion until as late as possible; with very few exceptions, speech/non-speech segmentation, automatic laughter detection, automatic speech recognition, and automatic punctuation are performed independently for each participant.

A growing body of evidence suggests that modeling interactive aspects of conversation as early as possible may be beneficial, as stream fusion provides a context for the activity of each participant. For example, in speech/non-speech segmentation, it has recently been shown that modeling contiguous intervals of speech, or talkspurts [5], *simultaneously* for all participants makes it possible to impose conversation- and participant-independent overlap constraints [6], and thereby largely eliminate the problem of crosstalk observed in close-talk microphone recordings of meetings.

Although adaptation to specific participants (and conversations) represents an opportunity for improved system performance, very little is known about how participant-specific the observed contextual preferences for vocal activity deployment actually are. In meetings, work on dominance classification [9], influence ranking [10], meeting type classification [1], and role and seniority classification [2] has shown that systematic differences in vocal activity deployment do exist across classes of conversations and participants. In the current work, we hypothesize that contextual preferences for vocal activity deployment actually vary quite significantly across *specific* participants, more so than they do within each participant's repertoire. This would make such preferences predictive of participant identity, suggesting that low-level processing systems stand to benefit from adaptation passes prior to final decoding.

The goal of the current work is to explore whether talkspurt deployment timing differentiates between participants. To this end, we present experiments in which participant identity in unseen meeting data is recovered given only the parameters of a probabilistic model of talkspurt deployment. Although our motivation is not competitive speaker identification, these features may, in the multiparticipant context, be complementary to standard acoustic, prosodic, lexical, and semantic features typically computed for this task. Our results indicate that when the participants to a conversation are known, the channels corresponding to each participant can be correctly identified in the majority of cases. When the identities of the participants are not known ahead of time, the observed channel activity can be correctly attributed to specific participants in over a third of the cases. We note additionally that the proposed framework makes it possible to assess the extent to which classification of participants into equivalence classes, as in our earlier related work [2], relies on the detection of specific participants rather than of classes.

2. Data

The data used in the current work is the same as that used in [2], namely the ICSI Meeting Corpus [3] meetings of type

$$u \quad \in \quad \mathcal{U} \quad \equiv \quad \{\texttt{Bed}, \texttt{Bmr}, \texttt{Bro}\} \ , \tag{1}$$

representing longitudinal recordings of three research groups at ICSI. Each of the 67 meetings in this subset is identified by a string consisting of the type u, and a numerical identifier d. As in [2], we have divided them into: ICSITRAINSET, consisting of the 33 meetings for which $d \bmod 4 \in \{1, 2\}$; ICSIDEVSET, consisting of the 18 meetings for which $d \bmod 4 \equiv 3$; and ICSIEVALSET, consisting of the 16 meetings for which $d \bmod 4 \equiv 0$. The three sets are not disjoint in participants, and the number of instrumented participants K varies from meeting to meeting, between 3 and 9.

We use and contrast three separate multiparticipant vocal activity segmentations. All three segmentations are binary, in that at any point in time t each participant k is considered to be either vocalizing or not vocalizing. The first segmentation, $\mathcal{S}$, consists of all talkspurts and is constructed from the forced-alignment lexical item endpoints found in the ICSI MRDA Corpus [4]; inter-item gaps shorter than 0.3 seconds are bridged. The second segmentation, $\mathcal{S} - \mathcal{B}$, is constructed in the same way, but only non-backchannel lexical items are considered. Finally, the third segmentation $\mathcal{L}$ of laugh bouts is as described in [7]. Each of the three segmentations is discretized [6] using a particular frame step ΔT and frame size T_S.

3. Framework

The task in the current work is to hypothesize specific participant identities for each of K unknown participants in a particular meeting. Inference is based on the K observed vocal activity sequences. We denote these K sequences, jointly and in parallel, as the *vocal interaction* [8] record of a meeting, since for each participant k, $1 \leq k \leq K$, the remaining $K-1$ vocal activity sequences comprise the interactive context.

We treat the entirety of each meeting as a single occurrence of each of its K participants. Many ICSI meeting participants participate in only a small handful of meetings, and, for these, robust models of vocal activity behavior cannot be inferred. We therefore limit ourselves to identifying only those participants which occur in ICSITRAINSET 7 or more times; there are 14 such participants. We map all other participants to the class OTHER. Hypothesized participant identities are drawn from the set $\mathcal{G} = \{S_1, S_2, \cdots, S_{14}, \text{OTHER}\}$, where all identities except OTHER are unique within each conversation of interest.

To enforce this constraint, we hypothesize all K participant identities simultaneously. The K-length vector $\mathbf{g}$ of participant identities is such that the identity of the kth participant (with k an arbitrary enumerator such as channel number) is found in $\mathbf{g}[k]$. The alternative multiparticipant assignments $\mathbf{g}$ form a closed set $\mathbb{G}$, whose number of elements is

$$|\mathbb{G}| = \sum_{j=0}^{K} \frac{K!}{(K-j)!\, j!} \cdot \frac{(|\mathcal{G}|-1)!}{(|\mathcal{G}|-1-j)!} \quad . \tag{2}$$

The first term in Equation 2 represents the number of combinations of j indices in $\mathbf{g}$ at which j non-OTHER participants are found; the second term represents the number of permutations of $|\mathcal{G}|-1$ non-OTHER participants, taken j at a time.

Given a vector $\mathbf{F}$ of observables for a single meeting, we seek the best *a posteriori* assignment $\mathbf{g}^*$ using

$$\begin{aligned}
\mathbf{g}^* &= \arg\max_{\mathbf{g} \in \mathbb{G}} P(\mathbf{g} \,|\, \mathbf{F}) \\
&= \arg\max_{\mathbf{g} \in \mathbb{G}} \sum_{u \in \mathcal{U}} P(u, \mathbf{g}, \mathbf{F}) \\
&= \arg\max_{\mathbf{g} \in \mathbb{G}} \sum_{u \in \mathcal{U}} P(u) \underbrace{P(\mathbf{g} \,|\, u)}_{\text{MM}} \underbrace{P(\mathbf{F} \,|\, \mathbf{g}, u)}_{\text{BM}} . \quad (3)
\end{aligned}$$

In the above, "MM" is the membership model and "BM" is the behavior model [1]. The MM provides the prior probability that the conversational group $\mathbf{g}$ hold a meeting of type u; the BM provides the likelihood that the observed features $\mathbf{F}$ are produced by group $\mathbf{g}$ in meeting type u.

4. Observables

For a particular meeting, each discretized multiparticipant segmentation, $\mathcal{S}$, $\mathcal{S}-\mathcal{B}$, or $\mathcal{L}$, is a matrix of T columns $\mathbf{q}_t$, $1 \leq t \leq T$, where T is the number of frames. The columns $\mathbf{q}_t$ are K-length vectors in $\{0,1\}^K$. We compute from each of these three matrices the following feature types: f_k^{VI}, the probability that participant k initiates vocalization at time t when no-one else was speaking at $t-1$; f_k^{VC}, the probability that participant k continues vocalization at time t when no-one else was speaking at $t-1$; $f_{k,j}^{OI}$, the probability that participant k initiates vocalization at time t when participant j was speaking at $t-1$; and $f_{k,j}^{OC}$ the probability that participant k continues vocalization at time t when participant j was speaking at $t-1$. Values of the feature types, which are time-independent probabilities, are estimated using an asymmetric infinite-range variant of the Ising model, as used in [1]; the model implements a particular type of parameter tying, reducing model complexity from $K \cdot 2^K$ to $K + K^2$ independent parameters. Additionally, we compute a feature type f_k^V, the time-independent probability that participant k vocalizes at any time.

5. Behavior Model

Since K may change from meeting to meeting, the size of the feature vector $\mathbf{F}$ must be considered variable. We therefore factor the behavior model, assuming that all features are mutually independent. Each feature is described by its own univariate Gaussian model $N(\mu, \sigma^2)$, whose parameters we compute using $\hat{\mu} = C^1/C^0$ and $\hat{\sigma}^2 = C^2/C^0 - \hat{\mu}^2$, where C^m is the zeroth, first, or second order ($m \in \{0,1,2\}$) cumulant. For one-participant feature types for participant $\xi \in \mathcal{G}$, and for two-participant feature types for participants ξ and ζ, these are given by

$$C^m_{u;\xi} = \sum_{r=1}^{R} \delta(u_r, u) \sum_{k=1}^{K_r} \delta(\mathbf{g}_r[k], \xi) \times (f_{r,k})^m , \tag{4}$$

$$\begin{aligned}
C^m_{u;\xi,\zeta} &= \sum_{r=1}^{R} \delta(u_r, u) \sum_{k=1}^{K_r} \delta(\mathbf{g}_r[k], \xi) \times \qquad (5) \\
&\qquad \sum_{j=1}^{K_r} \delta(\mathbf{g}_r[j], \zeta) \times (f_{r,k,j})^m \quad ,
\end{aligned}$$

respectively. Here, δ is the Kronecker delta, and r enumerates over the R meetings in the training corpus. u_r is the type of the rth meeting, and $f_{r,k}$ (and $f_{r,k,j}$) are the features from the kth (and jth) participant in the rth meeting.

Because certain participants and participant pairs in $\mathcal{G}$ may occur only rarely in the training data, we rely also on less specific cumulants. For one-participant feature types, this includes meeting-type-independent cumulants $C^m_{*;\xi} = \sum_u C^m_{u;\xi}$, meeting-type-specific but participant-independent cumulants $C^m_{u;*} = \sum_\xi C^m_{u;\xi}$, and meeting-type-independent and participant-independent cumulants $C^m_{*;*} = \sum_u \sum_\xi C^m_{u;\xi}$. For two-participant feature types, this also includes meeting-type-independent cumulants $C^m_{*;\xi,\zeta} = \sum_u C^m_{u;\xi,\zeta}$; meeting-type-specific but participant-independent cumulants $C^m_{u;\xi,*} = \sum_\zeta C^m_{u;\xi,\zeta}$, $C^m_{u;*,\zeta} = \sum_\xi C^m_{u;\xi,\zeta}$, and $C^m_{u;*,*} = \sum_\xi \sum_\zeta C^m_{u;\xi,\zeta}$; and meeting-type-independent and participant-independent cumulants $C^m_{*;\xi,*} = \sum_u \sum_\zeta C^m_{u;\xi,\zeta}$, $C^m_{*;*,\zeta} = \sum_u \sum_\xi C^m_{u;\xi,\zeta}$, and $C^m_{*;*,*} = \sum_u \sum_\xi \sum_\zeta C^m_{u;\xi,\zeta}$.

Maximum a posteriori (MAP) model estimates are computed by combining these cumulants using

$$\tilde{C}^m_{u;\xi} = C^m_{u;\xi} + \lambda_{*;\xi} C^m_{*;\xi} + \lambda_{u;*} C^m_{u;*} + \lambda_{*;*} C^m_{*;*} \,. \tag{6}$$

and

$$\begin{aligned}
\tilde{C}^m_{u;\xi,\zeta} &= C^m_{u;\xi,\zeta} + \lambda_{*;\xi,\zeta} C^m_{*;\xi,\zeta} + \lambda_{u;\xi,*} C^m_{u;\xi,*} \qquad (7) \\
&+ \lambda_{*;\xi,*} C^m_{*;\xi,*} + \lambda_{u;*,\zeta} C^m_{u;*,\zeta} + \lambda_{*;*,\zeta} C^m_{*;*,\zeta} \\
&+ \lambda_{u;*,*} C^m_{u;*,*} + \lambda_{*;*,*} C^m_{*;*,*} \quad .
\end{aligned}$$

In the current work, the λ interpolation factors are tuned using the development data; however, it should be noted that minimal effort has gone into optimizing these values.

The argumentation so far assumes a single feature vector $\mathbf{F}$, extracted from a binary vocal activity segmentation

$\mathcal{V} \in \{\mathcal{S}, \mathcal{S}-\mathcal{B}, \mathcal{L}\}$ using a particular *framing policy*; here, a policy consists of a specific *frame step* ΔT and *frame size* T_S. A feature family for a particular segmentation type and framing policy is denoted $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$; we assume such feature families to be independent, i.e. that $P(\mathbf{F}, \mathbf{F}', \cdots \,|\, \mathbf{g}, u) = P(\mathbf{F} \,|\, \mathbf{g}, u) \times P(\mathbf{F}' \,|\, \mathbf{g}, u) \times \cdots$.

6. Membership Model

The membership model used in the current work is identical to that in [1] and [2]. It assumes that participants attend meetings of specific type independently of other participants, and has the general form $P(\mathbf{g} \,|\, u) = \prod_{k=1}^{K} P(\mathbf{g}[k] \,|\, u)$, where $P(\mathbf{g}[k] \,|\, u)$ is the probability that the k-th participant has identity $\mathbf{g}[k]$, conditioned on the meeting type u. The probabilities are found using maximum likelihood estimation.

7. Search

As Equation 2 illustrates, the number of possible multiparticipant alternatives $\mathbf{g}$ can be intractably large. Our proposed greedy algorithm, which does not exhaustively iterate over $\mathbb{G}$, is shown below; at every point in the algorithm's execution, $\mathcal{G}'$ is the set of currently unhypothesized specific participants, and $\mathcal{I}$ is the set of indices in $\{1, 2, \cdots, K\}$ currently unoccupied by specific participants.

1. $\mathcal{G}' = \mathcal{G}$. $\mathcal{I} = \{1, 2, \cdots, K\}$. $\mathbf{g}[k] = $ OTHER, for all $1 \leq k \leq K$. Estimate u-conditioned MMs and BMs. Compute $LL = \text{score}(\mathbf{F}, \text{MM}, \text{BM})$. $LL^* = LL$.
2. While $\mathcal{I} \neq \emptyset$,
   (a) $g^* = \emptyset$. Set g to the first element of $\mathcal{G}'$.
   (b) Set i to the first element in $\mathcal{I}$.
   (c) $\mathbf{g}[i] = g$. Estimate u-conditioned MMs and BMs. Compute $LL = \text{score}(\mathbf{F}, \text{MM}, \text{BM})$. If $LL > LL^*$, $\mathbf{g}^* = \mathbf{g}$, $LL^* = LL$, $g^* = g$, and $i^* = i$.
   (d) $\mathbf{g}[i] = $ OTHER. Set i to the next element of $\mathcal{I}$. If $i \neq \emptyset$, return to Step 2c.
   (e) Set g to the next element of $\mathcal{G}'$. If $g \neq \emptyset$, return to Step 2b.
   (f) Remove g^* from $\mathcal{G}'$. Remove i^* from $\mathcal{I}$. Return to Step 2.

The algorithm aims to identify all K participants, one iteration at a time; $\text{score}(\mathbf{F}, \text{MM}, \text{BM})$ is the joint probability product in Equation 3. Step 1 hypothesizes a background participant model at each index $1 \leq i \leq K$. The algorithm then enumerates over the currently still unhypothesized specific participants $\mathcal{G}'$. Each such participant g is evaluated as being at each currently still unused index $i \in \mathcal{I}$. Once the first participant is located at his/her best index i^*, the algorithm proceeds to identify a next participant, by enumerating over all remaining participants and over all remaining vacant indices.

8. Experiments

We present two experiments; in the first, the identities of participants are known and only need to be attributed to specific observed participants; in the second, participants must first be drawn from a larger population. The numbers we present should be contrasted with majority class guessing. Always guessing the most probable single participant in ICSITRAINSET yields accuracies of 11.9% and 11.8% on ICSIDEVSET and ICSIEVALSET, respectively (always guessing OTHER yields 20.6% and 22.9%, respectively, but effectively fails to identify any specific participants).

8.1. Known group g

We first explore the performance of the behavior model, under the assumption of a perfect membership model. We do this by allowing both the type of meeting u and the identity of the K participants in $\mathbf{g}$ to be known in advance; the only task is to determine the correct permutatation α which maximizes the posterior probability of $\mathbf{g}$ given the observed $\mathbf{F}$:

$$\alpha^* = \arg\max_{\alpha \in \mathbb{S}_K} P(\mathbf{F} \,|\, \alpha(\mathbf{g}_0), u^*), \tag{8}$$

where $\mathbb{S}_K$ is the *symmetric group on K symbols*, ie. the space of all possible $K!$ permutations of K elements, and $\mathbf{g}_0$ is an arbitrary but fixed ordering of the correct K participant identities. u^* is the known type of the meeting. Following the application of Equation 8, $\mathbf{g}^* = \alpha^*(\mathbf{g}_0)$.

Table 1 shows the performance of each feature family $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$, computed using 6 different framing policies and 3 different binary segmentations, and using all feature types for all participants and participant pairs, on ICSIDEVSET. It appears that the $\mathcal{S}$ segmentation is the most informative, that removing backchannels from the $\mathcal{S}$ segmentation lowers performance slightly, and that the $\mathcal{L}$ segmentation, alone, leads to classification accuracies which are approximately 33% relative lower than accuracies obtained using the $\mathcal{S}$ segmentation.

ΔT (ms)	T_S (ms)	Segmentation Type $\mathcal{V}$ $\mathcal{S}$	$\mathcal{S}-\mathcal{B}$	$\mathcal{L}$
50	100	55.9	57.6	36.4
100	200	**60.2**	56.8	**42.4**
200	400	**60.2**	53.4	35.6
400	800	55.1	**58.5**	31.4
800	1600	47.5	47.5	38.1
1600	3200	54.2	56.8	32.2

Table 1: Identity classification accuracy using all feature types in each feature family $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$, on ICSIDEVSET, for 6 different framing policies and 3 different binary segmentations. For each test meeting, the meeting type and participant identities are known (but not attributed). Best performing $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$'s for each segmentation type are shown in bold.

In Table 2 we show the 5 complete feature families $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$ from Table 1 which, when combined, yield the highest identity classification accuracy on ICSIDEVSET. We show the performance of each feature family separately, all 5 feature families together, and, for each feature family, the performance of the other 4 feature families together. On ICSIDEVSET, performance using all 5 feature families is 9.3% higher than for the best feature family alone. We note that $\mathbf{F}^{\mathcal{L}}_{0.1/0.2}$, obtained using $\mathcal{L}$, improves the accuracy of classification based only on talkspurt production by 5.9%. The table also shows that performance on ICSIEVALSET is approximately 10% lower than on ICSIDEVSET. Furthermore, unlike for the latter, the feature family obtained using the laughter segmentation $\mathcal{L}$ lowers performance for ICSIEVALSET, from 54.9% to 53.9%.

We note that dropping the assumption that the test meeting type u is known during testing does not impact the results in Table 2, because meeting type u can always be correctly inferred from the (unordered) participant identities $\mathbf{g}_0$.

Feature family	icsiDevSet		icsiEvalSet	
	accur	compl	accur	compl
$\mathbf{F}^{\mathcal{S}}_{0.1/0.2}$	60.2	60.2	50.0	52.0
$\mathbf{F}^{\mathcal{S}}_{0.2/0.4}$	60.2	61.9	45.1	51.0
$\mathbf{F}^{\mathcal{S}}_{0.8/1.6}$	47.5	61.0	**54.9**	50.0
$\mathbf{F}^{\mathcal{S}-\mathcal{B}}_{0.2/0.4}$	53.4	59.3	48.0	50.0
$\mathbf{F}^{\mathcal{L}}_{0.1/0.2}$	42.4	63.6	29.4	**54.9**
all 5 $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$'s	**69.5**		53.9	

Table 2: Identity classification accuracy (%) using each of the best five feature families by themselves (accur), together (all 5 $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$'s), and leaving each of the five out, one at a time (compl), for both icsiDevSet and icsiEvalSet. For each test meeting, the meeting type and participant identities are known (but not attributed); the best-performing feature families are shown in bold.

8.2. Unknown group $\mathbf{g}$

Next, we drop the assumption that the (unordered) set of participants is known in advance. This exercises the membership model, as we now need to validate candidate $\mathbf{g}$'s, drawn from $\mathbb{G}$. Selection is performed using the algorithm described in Section 7, and the simple membership model from Section 6. Results are shown in Table 3.

Feature family	icsiDevSet		icsiEvalSet	
	accur	compl	accur	compl
$\mathbf{F}^{\mathcal{S}}_{0.1/0.2}$	39.0	30.5	27.5	33.3
$\mathbf{F}^{\mathcal{S}}_{0.2/0.4}$	30.5	35.6	20.6	**37.3**
$\mathbf{F}^{\mathcal{S}}_{0.8/1.6}$	16.9	31.4	28.4	32.4
$\mathbf{F}^{\mathcal{S}-\mathcal{B}}_{0.2/0.4}$	29.7	33.1	24.5	**37.3**
$\mathbf{F}^{\mathcal{L}}_{0.1/0.2}$	16.9	**40.7**	21.6	24.5
all 5 $\mathbf{F}^{\mathcal{V}}_{\Delta T/T_S}$'s	29.7		30.4	

Table 3: Identity classification accuracy (%) when meeting type and participant identities are not known; symbols as in Table 2.

As can be seen in the table, classification accuracies are significantly lower when the K group participants must be not only permuted into the correct arrangement but also first drawn from a population larger than K. When all 5 feature families are used, classification accuracy is reduced by 44% relative on icsiEvalSet; it is reduced by 32% relative when only the best performing feature family combinations from among those shown are used. These reductions are likely attributable to both the membership model and the search algorithm, and are the subject of ongoing analysis.

8.3. Effect of training set size

In order to gain insight into how performance varies with the size of the training data, we repeat the above experiment but use both icsiTrainSet and icsiDevSet for training, leading to a 50% relative increase in training corpus size.

Space constraints prohibit a detailed analysis of the resulting improvements on icsiEvalSet; the latter are in the range 0.0-3.0% absolute for every feature family in Table 3. Using all 5 feature families increases the classification accuracy by 3.9% absolute.

9. Conclusions

We have explored the extent to which specific participant identities can be recovered from only the vocal interaction record of meetings. Our experiments suggest that when it is known who is present, identities can be successfully attributed to participants in the majority of cases (53.9% in unseen data). When it is not known who is present, and identity hypotheses must be drawn from a larger set than present at the meeting, classification accuracies degrade significantly; in the data used here, classification rates are 20% lower than when participant identities are known and must only be assigned. However, even in this case, the numbers represent a 20% relative reduction in classification error over a baseline which always assigns the most frequent identity.

The presented system offers significant scope for feature type selection ([2]), joint membership modeling, and improvements to search. However, the ramifications of our current findings extend beyond participant identification. That participant identities can be recovered with above-chance accuracies using parametric models of vocal activity deployment indicates that applications relying on such models stand to gain significantly from participant-specific training, or participant adaptation. Example applications include speech/non-speech segmentation for meeting recognition, laughter detection for emotional valence classification, and text-independent backchannel detection.

10. Acknowledgments

We would like to thank Liz Shriberg for access to the ICSI MRDA Corpus.

11. References

[1] Laskowski, K., Ostendorf, M., and Schultz, T. “Modeling Vocal Interaction for Text-Independent Classification of Conversation Type”, Proc. SIGdial, Antwerpen, Belgium, pp.194–201, 2007.

[2] Laskowski, K., Ostendorf, M., and Schultz, T. “Modeling Vocal Interaction for Text-Independent Participant Characterization in Multi-Party Converation”, Proc. SIGdial, Columbus OH, USA, pp.148–155, 2008.

[3] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. “The ICSI Meeting Corpus”, Proc. of ICASSP, Hong Kong, China, pp.364–367, 2003.

[4] Shriberg, E., Dhillon, R., Bhagat, S., Ang, J., and Carvey, H. “The ICSI Meeting Recorder Dialog Act (MRDA) Corpus”, Proc. of SIGdial, Cambridge MA, USA, pp.97–100, 2004.

[5] Norwine, A. and Murphy, O. “Characteristic Time Intervals in Telephonic Conversation”, Bell System Technical Journal, 17:281–291, 1938.

[6] Laskowski, K. and Schultz, T. “Modeling Vocal Interaction for Segmentation in Meeting Recognition”, Proc. MLMI, Brno, Czech Republic, Springer LNCS 4892:259–270, 2007.

[7] Laskowski, K. and Burger, S. “Analysis of the Occurrence of Laughter in Meetings”, Proc. INTERSPEECH, Antwerpen, Belgium, pp.1258–1261, 2007.

[8] Dabbs, J. and Ruback, R. “Dimensions of group process: Amount and structure of vocal interaction”, Advances in Experimental Psychology, 20:123–169, 1987.

[9] Rienks, R. and Heylen, D. “Dominance detection in meetings using easily obtainable features”, Proc. of MLMI, Edinburgh, UK, 2005.

[10] Rienks, R., Zhang, D., Gatica-Perez, D., and Post, W. “Detection and application of influence rankings in small-group meetings”, Proc. of ICMI, Banff, Canada, 2006.