Modeling Vocal Interaction for Text-Independent Participant Characterization in Multi-Party Conversation

Kornel Laskowski
Cognitive Systems Labs
Universität Karlsruhe
Karlsruhe, Germany
kornel@ira.uka.de

Mari Ostendorf
Dept. of Electrical Engineering
University of Washington
Seattle WA, USA
mo@ee.washington.edu

Tanja Schultz
Cognitive Systems Labs
Universität Karlsruhe
Karlsruhe, Germany
tanja@ira.uka.de

Abstract

An important task in automatic conversation understanding is the inference of social structure governing participant behavior. We explore the dependence between several social dimensions, including assigned role, gender, and seniority, and a set of low-level features descriptive of talkspurt deployment in a multiparticipant context. Experiments conducted on two large, publicly available meeting corpora suggest that our features are quite useful in predicting these dimensions, excepting gender. The classification experiments we present exhibit a relative error rate reduction of 37% to 67% compared to choosing the majority class.

1 Introduction

An important task in automatic conversation understanding is the inference of social structure governing participant behavior; in many conversations, the maintenance or expression of that structure is an implicit goal, and may be more important than the propositional content of what is said.

There are many social dimensions along which participants may differ (Berger, Rosenholtz and Zelditch, 1980). Research in social psychology has shown that such differences among participants entail systematic differences in observed turn-taking and floor-control patterns (e.g. (Bales, 1950), (Tannen, 1996), (Carletta, Garrod and Fraser-Krauss, 1998)), and that participant types are not independent of the types and sizes of conversations in which they appear. In the present work, we consider the dimensions of assigned role, gender, and seniority level. We explore the predictability of these dimensions from a set of low-level speech activity features, namely the *probabilities* of initiating and continuing talkspurts in specific multiparticipant contexts, estimated from entire conversations. For our purposes, talkspurts (Norwine and Murphy, 1938) are contiguous intervals of speech, with internal pauses no longer than 0.3 seconds. Features derived from talkspurts are not only easier to compute than higher-level lexical, prosodic, or dialogue act features, they are also applicable to scenarios in which only privacy-sensitive data (Wyatt et al, 2007) is available. At the current time, relatively little is known about the predictive power of talkspurt timing in the context of large multi-party corpora.

As stated, our primary goal is to quantify the dependence between specific types of speech activity features and specific social dimensions; however, doing so offers several additional benefits. Most importantly, the existence of significant dependence would suggest that multiparticipant speech activity detectors (Laskowski, Fügen and Schultz, 2007) relying on models conditioned on such attributes may outperform those relying on general models. Furthermore, conversational dialogue systems deployed in multi-party scenarios may be perceived as more human-like, by humans, if their talkspurt deployment strategies are tailored to the personalities they are designed to embody.

Computational work which is most similar to that presented here includes the inference of static dominance (Rienks and Heylen, 2005) and influence (Rienks et al., 2006) rankings. In that work, the authors employed several speech activity features differing from ours in temporal scale and normaliza-

Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue, pages 148–155,
Columbus, June 2008.

tion. Notably, their features are not probabilities which are directly employable in a speech activity detection system. In addition, several higher-level features were included, such as topic changes, participant roles, and rates of phenomena such as turns and interruptions, and these were shown to yield the most robust performance. Our aim is also similar to that in (Vinciarelli, 2007) on radio shows, where the proposed approach relies on the relatively fixed temporal structure of production broadcasts, a property which is absent in spontaneous conversation. Although (Vinciarelli, 2007) also performs single-channel speaker diarization, he does not explore behavior during vocalization overlap.

Aside from the above, the focus of the majority of existing research characterizing participants is the detection of dynamic rather than static roles (i.e. (Banerjee and Rudnicky, 2004), (Zancanaro et al, 2006), (Rienks et al., 2006)). From a mathematical perspective, the research presented here is a continuation of our earlier work on meeting types (Laskowski, Ostendorf and Schultz, 2007), and we rely on much of that material in the presentation which follows.

2 Characterizing Participants

Importantly, we characterize participants in entire *groups*, rather than characterizing each participant independently. Doing so allows us to apply constraints on the group as a whole, eliminating the need for hypothesis recombination (in the event that more than one participant is assigned a role which was meant to be unique). Additionally, treating groups holistically allows for modeling the interactions between specific pairs of participant types.

For each conversation or meeting[1] of K participants, we compute a feature vector $\mathbf{F}$, in which all one-participant and two-participant speech activity features are found in a particular order, typically imposed by microphone channel or seating assignment (the specific features are described in Section 4). The goal is to find the most likely group assignment of participant labels that account for the observed $\mathbf{F}$. In (Laskowski, Ostendorf and Schultz, 2007), it was shown that meeting types in a large meeting corpus can be successfully inferred from $\mathbf{F}$ using this approach; here, we employ the same framework to classify participant types in the K-length vector $\mathbf{g}$, for the group as a whole:

$$\begin{aligned} \mathbf{g}^* &= \arg\max_{\mathbf{g}\in\mathcal{G}} P(\mathbf{g}\,|\,\mathbf{F}) \\ &= \arg\max_{\mathbf{g}\in\mathcal{G}} \underbrace{P(\mathbf{g})}_{\text{MM}} \underbrace{P(\mathbf{F}\,|\,\mathbf{g})}_{\text{BM}}, \end{aligned} \quad (1)$$

where MM and BM are the membership and behavior models, respectively, and $\mathcal{G}$ is the set of all possible assignments of $\mathbf{g}$.

In the remainder of this section, we define the participant characteristics we explore, which include assigned role, gender, and seniority. We treat these as separate tasks, applying the same classification framework. We also show how our definitions provide search space constraints on Equation 1.

2.1 Conversations with Unique Roles

Given a meeting of K participants, we consider a set of roles $\mathcal{R} = \{R_1, R_2, \cdots, R_K\}$ and assign to each participant k, $1 \leq k \leq K$, exactly one role in $\mathcal{R}$. An example group assignment is the vector $\mathbf{r}_1$ of length K, where $\mathbf{r}_1[k] = R_k$. The set $\mathbb{R}$ of group assignment alternatives $\mathbf{r} \in \mathbb{R}$ is given by permutations $\alpha : \mathbb{R} \mapsto \mathbb{R}$, where $\alpha \in \mathbb{S}_K$, the *symmetric group on K symbols*[2]. The number of elements in $\mathbb{R}$ is identically the number of unique permutations in $\mathbb{S}_K$, a quantity known as its *order* $|\mathbb{S}_K| = K!$.

To identify the most likely group assignment $\mathbf{r}^* = \alpha^*(\mathbf{r}_1)$ given the set $\mathbf{F}$ of observables, we iterate over the $K!$ elements of $\mathbb{S}_K$ using

$$\alpha^* = \arg\max_{\alpha\in\mathbb{S}_K} P(\mathbf{F}\,|\,\alpha(\mathbf{r}_1)) \;, \quad (2)$$

where we have elided the prior $P(\alpha)$ assuming that it is uniform. Following the application of Equation 2, the most likely role of participant k is given by $\alpha^*(\mathbf{r}_1)[k]$.

Alternately, we may be interested in identifying only a subset of the roles in $\mathcal{R}$, namely a leader, or a manager. In this case, participant roles are drawn from $\mathcal{L} = \{L, \neg L\}$, under the constraint that exactly one participant is assigned the role L. The set $\mathbb{L}$ of

[1] "Conversation" and "meeting" will be used interchangeably in the current work.

[2] For an overview of group theoretic notions and notation, we refer the reader to (Rotman, 1995).

alternative group assignments has K indicator vector members $\mathbf{l}_j$, $1 \leq j \leq K$, where $\mathbf{l}_j[k]$ is L for $k = j$ and $\neg L$ otherwise.[3] We iterate over the indicator vectors to obtain

$$j^* = \underset{j \in \{1, \cdots, K\}}{\arg\max} \; P(\mathbf{F} \mid \mathbf{l}_j) \; , \tag{3}$$

assuming uniform priors $P(\mathbf{l}_j)$. Following the application of Equation 3, j^* is the index of the most likely L participant.

We note that this framework for unique role classification is applicable to classifying unique ranks, without first having to collapse them into non-unique rank classes as was necessary in (Rienks et al., 2006).

2.2 Conversations with Non-Unique Roles

The second type of inference we consider is for dimensions in which roles are not unique, i.e. where participants are in principle drawn independently from a set of alternatives. This naturally includes dimensions such as gender, seniority, age, etc.

As an example, we treat the case of gender. Participant genders are drawn independently from $\mathcal{H} = \{♀, ♂\}$. The set of group assignment alternatives $\mathbf{h}$ is given by the Cartesian product $\mathcal{H}^K$, of 2^K unique elements. We search for the most likely group assignment $\mathbf{h}^*$, given the observables $\mathbf{F}$, by iterating over these elements using

$$\mathbf{h}^* = \underset{\mathbf{h} \in \mathcal{H}^K}{\arg\max} \; P(\mathbf{h}) \; P(\mathbf{F} \mid \mathbf{h}) \; . \tag{4}$$

Once $\mathbf{h}^*$ is found, the gender of each participant k is available in $\mathbf{h}^*[k]$.

A similar scenario is found for seniority, when it is not uniquely ranked. We assume a set of N_S mutually exclusive seniority levels $S_i \in \mathcal{S} = \{S_1, S_2, \cdots, S_{N_S}\}$, $1 \leq i \leq N_S$. During search, each participant's seniority level is drawn independently from $\mathcal{S}$, leading to group assignments $\mathbf{s} \in \mathcal{S}^K$, of which there are N_S^K. As for gender, we iterate over these to find

$$\mathbf{s}^* = \underset{\mathbf{s} \in \mathcal{S}^K}{\arg\max} \; P(\mathbf{s}) \; P(\mathbf{F} \mid \mathbf{s}) \; . \tag{5}$$

The seniority of participant k, following the application of Equation 5, is $\mathbf{s}^*[k]$.

[3] For completeness, we note that each $\mathbf{l}_j$ corresponds to a permutation $\beta : \mathbb{L} \mapsto \mathbb{L}$ of $\mathbf{l}_1$, and that $\beta \in \langle \tau \rangle$, the *cyclic subgroup generated by* τ, where τ is the permutation $(1, 2, \cdots, K)$.

3 Data

In the current work, we use two different corpora of multi-party meetings. The first, the scenario subset of the AMI Meeting Corpus (Carletta, 2007), consists of meetings involving $K = 4$ participants who play different specialist roles in a product design team. We have observed the recommended division of this data into: AMITRAINSET of 98 meetings; AMIDEVSET of 20 meetings; and AMIEVALSET, also of 20 meetings. Although each participant takes part in approximately 4 meetings, the 3 sets are disjoint in participants. We use only the provided word alignments of these meetings. The corpus is accompanied by metadata which specifies the gender and assigned role of each participant.

The second corpus consists of the `Bed`, `Bmr`, and `Bro` meeting types in the ICSI Meeting Corpus (Janin et al., 2003). Each meeting is identified by one of $\{\texttt{Bed}, \texttt{Bmr}, \texttt{Bro}\}$, as well as a numerical identifier d. We have divided these meetings 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 $d \bmod 4 \equiv 0$. These three sets are not disjoint in participants, and the number of instrumented participants K varies from meeting to meeting, between 3 and 9. The corpus is accompanied by metadata specifying the gender, age, and education level of each participant. We use only the forced alignments of these meetings, available in the accompanying MRDA Corpus (Shriberg et al, 2004).

4 Features

Our observation space is the complete K-participant vocal interaction on-off pattern description for a meeting $\mathcal{C}$, a discretized version of which we denote as $\mathbf{q}_t \in \{0, 1\}^K$ for $1 \leq t \leq T$, where T is the duration of $\mathcal{C}$ in terms of the number of 100 ms frames. Details regarding the discretization (and subsequent feature computation) can be found in (Laskowski, Ostendorf and Schultz, 2007).

We compute from $\mathbf{q}_t$ the following features[4] which are the elements of $\mathbf{F}$: f_k^{VI}, the probabil-

[4] Feature type superscripts indicate talkspurt initiation (I) or continuation (C), for either single-participant vocalization (V) or vocalization overlap (O).

ity 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 features, which are time-independent probabilities, are estimated using a variant of the Ising model (cf. (Laskowski, Ostendorf and Schultz, 2007)). Additionally, we compute a feature f_k^V, the probability that participant k vocalizes at time t, and single-participant averages of the two-participant features: $\langle f_{k,j}^{OI} \rangle_j$, $\langle f_{j,k}^{OI} \rangle_j$, $\langle f_{k,j}^{OC} \rangle_j$, and $\langle f_{j,k}^{OC} \rangle_j$. The complete feature vector for a conversation of K participants then consists of $7K$ one-participant features, and $2(K^2-K)$ two-participant features.

We note that multiple phenomena contribute to the overlap features. The features $f_{k,j}^{OI}$ are based on counts from interruptions, backchannels, and precise floor handoffs. The features $f_{k,j}^{OC}$ are based on counts from interruptions, attempts to hold the floor, and backchannels. Both feature types also contain counts incurred during schism, when the conversation splits into two sub-conversations.

5 Models

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 and that each is described by its own univariate Gaussian model $N(\mu, \sigma^2)$. These parameters are maximum likelihood estimates from the f_k and $f_{k,j}$ values in a training set of conversations. In most of these experiments, where the number of classes is small, no parameter smoothing is needed.

For the cases where the group prior is not uniform and participant types are not unique, the membership model assumes independent participant types and has the general form

$$P(\mathbf{g}) = \prod_{k=1}^{K} P(\mathbf{g}[k]), \tag{6}$$

where $P(\mathbf{g}[k])$ is the probability that the k-th participant is type $\mathbf{g}[k]$. This model is used for gender ($P(\mathbf{h})$) and seniority ($P(\mathbf{s})$). The probabilities of specific types are maximum likelihood estimates from the training data.

6 Assigned Role Classification

6.1 Classifying Unique Roles

For unique role classification, we use the AMI Meeting Corpus. All meetings consist of $K=4$ participants, and each participant is assigned one of four roles: project manager (PM), marketing expert (ME), user interface designer (UI), or industrial designer (ID).

As mentioned in Section 2.1, classifying the unique role of all participants, jointly, involves enumerating over the possible permutations of $\{\mathrm{PM}, \mathrm{ME}, \mathrm{UI}, \mathrm{ID}\}$. We use AMITRAINSET to train the behavior model, and then classify AMIDEVSET using Equation 2, one feature type at a time, to identify the best 3 feature types for this task; development experiments suggest that classification rates level off after a small handful of the best performing feature types is included. Those feature types were found to be f_k^{VI}, $\langle f_{k,j}^{OI} \rangle_j$, and $f_{k,j}^{OI}$, capturing the probability of initiating a talkspurt in silence, of initiating a talkspurt when someone else is speaking, and of initiating a talkspurt when a participant in a specific other role is speaking, respectively. On AMIEVALSET, these feature types lead to single-feature-type 4-way classification rates of 41%, 29%, and 53%, respectively. When all three types are used together ($3K+K^2$ features in total), the rate is 53%. Accuracy when all feature types are used is 46%, indicating that some feature types are detrimental to this task.

The confusion matrix for classification using the three best feature types is shown in Table 1. The matrix shows that association between the reference assignment of PM, as well as of UI, and the hypothesized assignment based on the three feature types mentioned is statistically significant. On the other hand, assignment of ID and ME does not deviate significantly from chance.

6.2 Finding the Manager

Using the same data as above, we explore the simplified task of finding a specific participant type. We

Ref	Hyp			
	ID	ME	PM	UI
ID	**8**	6	4	2
ME	5	**8**	4	3
PM	3	4	++**12**	− 1
UI	4	2	−− 0	++**14**

Table 1: Confusion matrix for role classification on AMIEVALSET; reference assignment is found in the rows, hypothesized assignment in columns. Correctly classified roles, along the diagonal, are highlighted in bold. Statistical significance of association at the $p < 0.005$ level per class, using a 2×2 χ^2-test, is shown using "++" and "−−", for above chance and below chance values, respectively; the same is true of "+" and "−", for significance at the $0.005 \leq p < 0.05$ level.

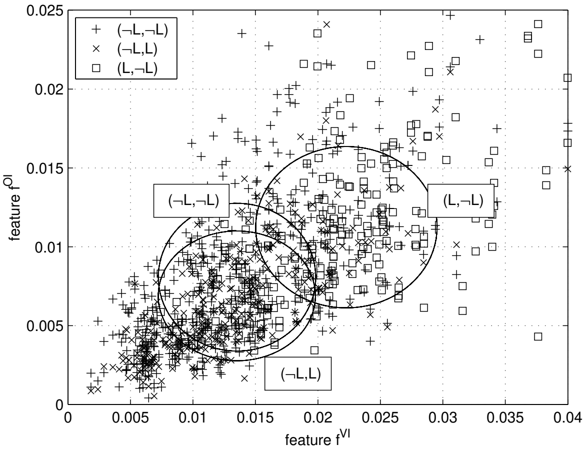

Figure 1: Distribution of $(f_k^{VI}, f_{k,j}^{OI})$ pairs for each of $(\neg L, \neg L)$, $(\neg L, L)$, and $(L, \neg L)$. Ellipses are centered on AMITRAINSET means and encompass one standard deviation.

equate the project manager role with L, and the remaining roles with $\neg L$. This is justified by the AMI meeting scenario, in which participant groups take a product design from start to prototype, and in which the project manager is expected to make the group run smoothly.

The behavior model, trained on AMITRAINSET, is applied using Equation 3 to determine the most likely index j^* of the leader L, given the observed $\mathbf{F}$, from among the $K = 4$ alternatives. To select the best 3 feature types, we once again use AMIDEVSET; these turn out to be the same as those for role classification, namely f_k^{VI}, $\langle f_{k,j}^{OI} \rangle_j$, and $f_{k,j}^{OI}$. Using these three feature types individually, we are able to identify the leader PM in 12 of the 20 meetings in AMIEVALSET. When all three are used together, the identification rate is 60%. However, when all feature types are used, the identification rate climbs to 75%. Since all participants are equally likely to be the leader, the baseline for comparison is random guessing (25% accuracy).

Figure 1 shows the distribution of two of the selected features, f_k^{VI} and $f_{k,j}^{OI}$, for the data in AMITRAINSET; we also show the first standard deviation of the single-Gaussian diagonal-covariance models induced. We first note that f_k^{VI} and $f_{k,j}^{OI}$ are correlated, i.e. that the probability of beginning a talkspurt in silence is correlated with the probability of beginning a talkspurt when someone else is speaking. L consistently begins more talkspurts, both in silence and during other people's speech. It is also interesting that $\neg L$ is slightly less likely to initiate a talkspurt when L is already speaking than when another $\neg L$ is. This suggests that $\neg L$ participants consistently observe the L-status of the already speaking party when contemplating talkspurt production. Finally, we note that neither the probability of continuing a talkspurt f_k^{VC} (related to talkspurt duration) nor f_k^{V} (related to overall amount of talk) are by themselves good $L/\neg L$ discriminators.

7 Gender Classification

Gender classification is an example of a task with a Cartesian search space. For these experiments, we use the AMI Meeting Corpus and the ICSI Meeting Corpus. In both corpora, gender is encoded in the first letter of each participant's unique identifier. The ratio of male to female occurrences is $2:1$ in AMITRAINSET, and $4:1$ in ICSITRAINSET. Choosing the majority class leads to gender classification rates of 65% and 81% on AMIEVALSET and ICSIEVALSET, respectively.

We enumerate alternative group assignments using Equation 4. Somewhat surprisingly, no single feature type leads to AMIEVALSET or ICSIEVALSET classification rates higher than those obtained by hypothesizing all participants to be male. On AMIDEVSET, one feature type ($f_{k,j}^{OI}$) yields negligibly better accuracy, but does not generalize to the corresponding evaluation data. Furthermore, the association between reference gender labels and hypothesized gender labels, on both evaluation sets, does not appear to be statistically significant at the $p < 0.05$ level. This finding that males and females do not differ significantly in their deployment of talkspurts is likely a consequence of the social structure of the particular groups studied. The fact that AMI roles are acted may also have an effect.

8 Seniority Classification

As a second example of non-unique roles, we attempt to classify participant seniority. For these experiments, we use the ICSI Meeting corpus, in which each participant's education level appears as an optional, self-reported attribute. We have manually clustered these attributes into $N_S = 3$ mutually exclusive seniority categories.[5] Each participant's seniority is drawn independently from $\mathcal{S} = \{\text{GRAD}, \text{PHD}, \text{PROF}\}$; a breakdown for ICSITRAINSET is shown in Table 2. Choosing the majority class ($P(\text{PHD}) = 0.444$ on ICSITRAINSET) yields a classification accuracy of 45% on ICSIEVALSET. We note that in this data, education level is closely correlated with age group.

Seniority	Number of		
	spkrs	occur	meets
GRAD	15	81	33
PHD	13	87	29
PROF	3	28	28
all	31	196	33

Table 2: Breakdown by seniority $\mathcal{S}$ in ICSITRAINSET by the number of unique participants (spkrs), the number of occurrences (occur), and the number of meetings (meets) in which each seniority occurs.

8.1 Classifying Participant Types Independently of Conversation Types

We first treat the problem of classifying participant seniority levels independently of the type of conversation being studied. We identify the most likely seniority assignment for all participants using Equation 5. The best three feature types, determined using ICSIDEVSET, are f_k^V, $f_{k,j}^{OI}$, and $f_{k,j}^{OC}$ (representing the probability of speaking, of beginning a talkspurt when a specific seniority participant is already speaking, and of continuing a talkspurt when a specific seniority participant is speaking), yielding single-feature-type classification rates of 52%, 59%, and 59%, respectively. When used together, these three feature types produce the confusion matrix shown in Table 3 and a rate of 61%, better than when all feature types are used (58%). This represents a 28% relative error reduction over chance. As can be seen in the table, association between the reference and hypothesized seniority assignments is statistically significant on unseen data. It is also evident that confusion between GRAD and PROF is lower than between more proximate seniority levels.

Ref	Hyp		
	GRAD	PHD	PROF
GRAD	++**11**	26	3
PHD	− 2	++**41**	− 3
PROF	0	−− 6	++**10**

Table 3: Confusion matrix for seniority classification on ICSIEVALSET; reference assignment is found in the rows, hypothesized assignment in columns. Highlighting and use of "++", "+", "−", and "−−" as in Table 1.

Figure 2 shows the distribution of $(f_k^V, f_{k,j}^{OC})$ pairs in ICSITRAINSET, together with the first standard deviation, for each combination of the already speaking seniority participant and the seniority participant initiating a new talkspurt (except for (PROF, PROF), since there is at most one PROF in each ICSITRAINSET meeting).

As is clear from the figure, PROF participants in this data talk more than either of the two other seniority types. The figure also demonstrates a difference of behavior during speech overlap. The four ellipses describing GRAD behavior when overlapping with any of the other three classes, as well as PHD behavior when overlapping with GRAD participants, are relatively broad and indicate the absence of strong tendency or preference. However, PHD participants are more likely to continue vocalizing in overlap with other PHD participants, and even more likely to continue through overlap with PROF partic-

[5] GRAD includes "Grad", as well as "Undergrad", "B.A.", and "Finished BA in 2001", due to their small number of exemplars; PHD includes "PhD" and "Postdoc"; and PROF includes "Professor" only.

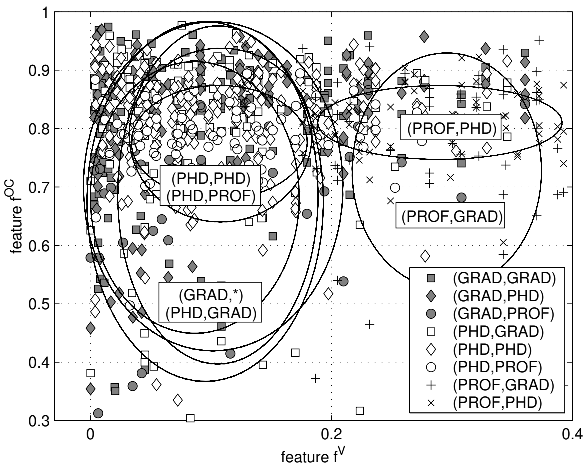

Figure 2: Distribution of $(f^V_k, f^{OC}_{k,j})$ feature value pairs for each of the (k, j) participant pairs (GRAD, GRAD), (GRAD, PHD), (GRAD, PROF), (PHD, GRAD), (PHD, PHD), (PHD, PROF), (PROF, GRAD), and (PROF, PHD). Ellipses are centered on ICSITRAINSET means and encompass one standard deviation.

ipants. A similar trend is apparent for PROF participants: the mean likelihood that they continue vocalizing in overlap with GRAD participants lies below $\mu - \sigma$ (bottom 17%) of their model with PHD participants. We believe that the senior researchers in this data are consciously minimizing their overlap with students, who talk less, to make it easier for the latter to speak up.

8.2 Conditioning on Conversation Type

We now repeat the experiments in the previous section, but condition the behavior and membership models on meeting type t:

$$\mathbf{s}^* = \arg\max_{\mathbf{s}\in\mathcal{S}^K} \sum_{t\in\mathcal{T}} P(t)\; P(\mathbf{s}\,|\,t)\; P(\mathbf{F}\,|\,\mathbf{s},\,t)\;, \qquad (7)$$

where $t \in \mathcal{T} = \{\texttt{Bed}, \texttt{Bmr}, \texttt{Bro}\}$.

Performance using maximum likelihood estimates for the behavior model $P(\mathbf{F}\,|\,\mathbf{s},\,t)$ results in a seniority classification rate on ICSIEVALSET of 61%, i.e. no improvement over conversation-type-independent classification. We suspect this is due to the smaller amounts of training material. To verify this assumption, we smooth the maximum likelihood estimates, $\mu_{S_i,t}, \sigma^2_{S_i,t}$, towards the maximum likelihood conversation-type-independent estimates, μ_{S_i}, σ_{S_i}, using

$$\hat{\mu}_{S_i,t} = \alpha\mu_{S_i,t} + (1-\alpha)\,\mu_{S_i}\;, \qquad (8)$$

$$\hat{\sigma}^2_{S_i,t} = \alpha\sigma_{S_i,t} + (1-\alpha)\,\sigma^2_{S_i}\;, \qquad (9)$$

where the value of $\alpha = 0.7$ was selected using ICSIDEVSET. This leads to a rate of 63% on ICSIEVALSET. Furthermore, if instead of estimating the prior on conversation type $P(t)$ from the training data, we use our meeting type estimates from (Laskowski, Ostendorf and Schultz, 2007), the classification rate increases to 67%. A control experiment in which the true type t_{test} of each test meeting is known, i.e. $P(t) = 1$ if $t_{test} = t$ and 0 otherwise, shows that the maximum accuracy achievable under optimal $P(t)$ estimation is 73%.

9 Conclusions

We have explored several socially meaningful partitions of participant populations in two large multiparty meeting corpora. These include assigned role, leadership (embodied by a manager position), gender, and seniority. Our proposed classifier, which can represent participants in groups rather than independently, is able to leverage the observed differences between specific pairs of participant classes. Using only low-level features capturing when participants choose to vocalize relative to one another, it attains relative error rate reductions on unseen data of 37%, 67%, and 40% over chance on classifying role, leadership, and seniority, respectively. We have also shown that the same classifier, using the same features, cannot discriminate between genders in either corpus.

A comparison of the proposed feature types and their performance on the tasks we have explored is shown in Table 4. Consistently, the most useful feature types appear to be the probability of initiating a talkspurt in silence, and the probability of initiating a talkspurt when a participant of a specific type is already speaking. Additionally, on the ICSI Meeting Corpus, the probability of speaking appears to be dependent on seniority, and the probability of continuing to vocalize in overlap with another participant appears to depend on the seniority of the latter. Finally, we note that, for seniority classification on the unseen ICSIEVALSET, the top 3 feature types outperform the best single feature type, indicating a degree of feature type complementarity; this is also true for L-detection on AMIEVALSET when all feature types, as opposed to the single best feature type, are used.

Feature Type	AMI			ICSI		
	$\mathcal{R}$	$\mathcal{L}$	$\mathcal{H}$	$\mathcal{H}$	$\mathcal{S}$	$\mathcal{S}\|t^*$
f_k^V	44	—	—	—	*52	*57
f_k^{VI}	*41	*60	—	—	52	56
f_k^{VC}	34	—	—	—	—	62
$\langle f_{j,k}^{OI} \rangle_j$	44	—	—	—	47	56
$\langle f_{k,j}^{OI} \rangle_j$	*29	*60	—	—	49	59
$f_{k,j}^{OI}$	*53	*60	64	—	*59	*59
$\langle f_{j,k}^{OC} \rangle_j$	24	—	—	—	—	57
$\langle f_{k,j}^{OC} \rangle_j$	—	—	—	—	54	59
$f_{k,j}^{OC}$	—	—	—	—	*59	*63
top 3*	53	60	—	—	61	67
all	46	75	43	47	58	57
priors	25	25	65	81	45	45

Table 4: Comparative classification performance for 3 experiments on AMIEVALSET and 3 experiments on ICSIEVALSET, per feature type; $\mathcal{R}$, $\mathcal{L}$, $\mathcal{H}$, and $\mathcal{S}$ as defined in Section 2. Also shown is performance on the best three feature types (selected using development data) and all feature types, as well as that when choosing the majority class ("prior"), informed by training data priors; for $\mathcal{R}$ and $\mathcal{L}$ classification, "prior" performance is equal to random guessing. "—" indicates that a feature type, by itself, did not perform above the corresponding "prior" rate; top-3 feature type selection indicated by "*".

Our results not only suggest new, easy-to-compute, low-level features for the automatic classification of participants into socially meaningful types, but also offer scope for informing turn-taking or talkspurt-deployment policies in conversational agents deployed in multi-party settings. Additionally, they suggest that implicit models of certain equivalence classes may lead to improved performance on other tasks, such as multi-participant vocal activity detection.

Acknowledgments

We would like to thank Jean Carletta for helpful comments during the final preparation of this manuscript, and Liz Shriberg for access to the ICSI MRDA Corpus.

References

R. Bales. 1950. *Interaction Process Analysis.* Addison-Wesley Press, Inc.

S. Banerjee and A. Rudnicky. 2004. Using simple speech based features to detect the state of a meeting and the roles of the meeting participants. *Proc. INTERSPEECH*, pp.2189-2192.

J. Berger, S. Rosenholtz, M. Zelditch Jr. 1980. Status Organizing Processes. *Annual Review of Sociology*, **6**:479-508.

J. Carletta, S. Garrod, and H. Fraser-Krauss. 1998. Communication and placement of authority in workplace groups — The consequences for innovation. *Small Group Research*, **29**(5):531-559.

J. Carletta. 2007. Unleashing the killer corpus: Experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, **41**(2):181–190.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. *Proc. ICASSP*, pp.364–367.

K. Laskowski, M. Ostendorf, and T. Schultz. 2007. Modeling vocal interaction for text-independent classification of conversation type. *Proc. SIGdial*, pp.194-201.

K. Laskowski, C. Fügen, and T. Schultz. 2007. Simultaneous multispeaker segmentation for automatic meeting recognition. *Proc. EUSIPCO*, pp.1294-1298.

A. Norwine and O. Murphy. 1938. Characteristic time intervals in telephonic conversation. *Bell System Technical Journal*, **17**:281-291.

R. Rienks and D. Heylen. 2005. Dominance detection in meetings using easily obtainable features. *Proc. MLMI.*

R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post. 2006. Detection and application of influence rankings in small-group meetings. *Proc. ICMI.*

J. Rotman. 1995. *An Introduction to the Theory of Groups.* Springer-Verlag New York, Inc.

E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *Proc. SIGdial*, pp.97–100.

D. Tannen. 1996. *Gender & Discourse.* Oxford University Press, USA.

A. Vinciarelli. 2007. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Trans. Multimedia*, **9**(6):1215-1226.

D. Wyatt, J. Bilmes, T. Choudhury, and H. Kautz. 2007. A privacy-sensitive approach to modeling multi-person conversations. *Proc. IJCAI*, pp.1769–1775.

M. Zancanaro, B. Lepri, and F. Pianesi. 2006. Automatic detection of group functional roles in face to face interactions. *Proc. ICMI.*