Modeling Vocal Interaction for Text-Independent Classification of Conversation Type

Kornel Laskowski
interACT
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
interACT
Carnegie Mellon University
Pittsburgh PA, USA
tanja@cs.cmu.edu

Abstract

We describe a system for conversation type classification which relies exclusively on multi-participant vocal activity patterns. Using a variation on a well-studied model from stochastic dynamics, we extract features which represent the transition probabilities that characterize the evolution of participant interaction. We also show how vocal interaction can be modeled between specific participant pairs. We apply the proposed system to the task of classifying meeting types in a large multi-party meeting corpus, and achieve a three-way classification accuracy of 84%. This represents a relative error reduction of more than 50% over a baseline which uses only individual speaker times (i.e. no interaction dynamics). Random guessing on this data yields an accuracy of 43%.

1 Introduction

An important and frequently overlooked task in automatic conversation understanding is the characterization of conversation type. In particular, search and retrieval in multi-participant conversation corpora stands to benefit from indexing by broad conversational style, as tending towards one or more speech-exchange prototypes (Sacks et al, 1974) such as interactive seminar, debate, formal business meeting, or informal chat. Current state-of-the-art speech understanding systems are well-poised to tackle this problem through up-stream fusion of multiparticipant contributions, following automatic speech recognition and dialog act classification. Unfortunately, such reliance on lexical information limits the ultimate application of conversational style classification to only a handful of languages with well-developed lexical components, notably English.

In the current work, we attempt to address this limitation by characterizing conversations in terms of their patterns of on-off vocal activity, referred to as *vocal interaction* by the psycholinguistic community (Dabbs and Ruback, 1987). In doing so, we rely only on the joint multi-participant vocal activity segmentation of a conversation (Renals and Ellis, 2003), and ignore other features. The text-independent features we explore here can of course be combined with text-dependent cues, and prosodic and/or speaker cues, depending on the reliability of these components.

To the best of our knowledge, there is currently little if any work on the continuous modeling of vocal interaction for conversations with arbitrary numbers of participants. Some very recent research exists with goals related to those in this work, most frequently focusing on the classification of time-dependent, evolving phenomena. Examples include the recognition of meeting states and participant roles (Banerjee and Rudnicky, 2004), the detection of interaction groups in meetings (Brdiczka et al., 2005), the recognition of individual and group actions in meetings (McCowan et al, 2005), and the recognition of participant states (Zancanaro et al, 2006). Modeling multi-participant vocal interaction to improve vocal activity detection in meetings was first explored in (Laskowski and Schultz, 2006) and elaborated in (Laskowski and Schultz, 2007); it has

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 194–201,
Antwerp, September 2007.

since been explored for privacy-sensitive data collection in more general settings (Wyatt et al, 2007). The rare examples of time-independent characterization of conversations in their entirety, as pursued in the current work, include the detection of conversational pairs (Basu, 2002) and the classification of dominance in meetings (Rienks and Heylen, 2005).

We begin this paper by proposing a computational framework which allows for the modeling of interactions among specific participants. We propose several time-independent interaction features, together with a robust means for computing them. Finally, we apply the proposed text-independent classification system to the task of meeting type classification. Our results show that features extracted from the multi-participant segmentation of a conversation can be successfully used for classifying meeting type through the observed conversational style.

2 Bayesian Framework

We introduce the notion of a *group* of participants, which we denote as $\mathcal{G}$ and which we define to be a specific ordering of all $K \equiv \|\mathcal{G}\|$ participants in a particular conversation $\mathcal{C}$. Each conversation is of exactly one type $\mathcal{T}$, from among $N_{\mathcal{T}}$ possible types. Participants are drawn without replacement from a potentially unknown population $\mathcal{P}$, of size $\|\mathcal{P}\|$. In general, $\|\mathcal{P}\| > \|\mathcal{G}\|$.

$\mathcal{G}\,[k]$, for $1 \leq k \leq K$, is an attribute of the kth participant; k represents a particular cardinal ordering of participants in group $\mathcal{G}$, which is immutable for the duration of a meeting (in this work, k is the channel number). $\mathcal{G}$ may be unique in $\mathcal{P}$, i.e. it may represent a specific participant; alternately, it may represent a category of participant, such as age group, social standing, or vocalizing time rank. When participants are unique in $\mathcal{P}$, the number of unique groups $N_{\mathcal{G}} = \|\mathcal{P}\|!/\left(\|\mathcal{P}\| - \|\mathcal{G}\|\right)!$ is simply the number of permutations of $\|\mathcal{P}\|$ taken $\|\mathcal{G}\|$ at a time.

Our observation space is the complete vocal interaction on-off pattern description for conversation $\mathcal{C}$, a discretized version of which we denote as $\mathbf{q}_t$ for $1 \leq t \leq T$, where T is the duration of the conversation. Our goal in the present work is to extract from $\mathbf{q} = \{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_T\}$ a feature vector $\mathbf{F} \equiv f(\mathbf{q})$ which will discriminate among the $N_{\mathcal{T}}$ different conversation types under study.

We classify the type $\mathcal{T}$ of conversation $\mathcal{C}$, given observations $\mathbf{F}$, using:

$$\begin{aligned}
\mathcal{T}^* &= \arg\max_{\mathcal{T}} P\left(\mathcal{T} \,|\, \mathbf{F}\right) \\
&= \arg\max_{\mathcal{T}} \sum_{\mathcal{G}} P\left(\mathcal{G}, \mathcal{T}, \mathbf{F}\right) \\
&= \arg\max_{\mathcal{T}} \sum_{\mathcal{G}} P\left(\mathcal{T}\right) \times \qquad (1) \\
&\qquad \underbrace{P\left(\mathcal{G} \,|\, \mathcal{T}\right)}_{\substack{\text{Membership} \\ \text{Model}}} \times \underbrace{P\left(\mathbf{F} \,|\, \mathcal{G}, \mathcal{T}\right)}_{\substack{\text{Behavior} \\ \text{Model}}} .
\end{aligned}$$

The behavior model in Equation 1 is responsible for the likelihood of $\mathbf{F}$, describing the behavior of the participants of $\mathcal{G}$ during a conversation of type $\mathcal{T}$. The membership model provides a prior distribution for participant presence in conversations of type $\mathcal{T}$.

3 Vocal Interaction Features

We propose to extract interactional aspects of multiparticipant conversations by studying the presence of vocal activity for all participants at a fixed analysis frame rate. After some limited initial experimentation, we have chosen to use a frame shift of 100 ms. We consider two mutually exclusive vocal activity states, vocalizing ($\mathcal{V}$) and not vocalizing (i.e. silent, $\mathcal{N}$). Figure 1 graphically depicts the discretization of a multichannel segmentation, which allows us to treat a particular conversation as the output of a simple Markov process $\mathbf{q}$ over an alphabet of 2^K symbols, with

$$\mathbf{q}_t \;\in\; \mathbf{\Psi} \times \mathbf{\Psi} \times \mathbf{\Psi} \times \ldots \times \mathbf{\Psi} \qquad (2)$$

of K products, where $\mathbf{\Psi} \equiv \{\mathcal{N}, \mathcal{V}\}$, and t is the time index of the frame.

3.1 Feature Design

In the current work, we assume $\mathbf{q}$ to be a first-order Markov process which can be described by symbol transition probabilities

$$a_{ij} \;=\; P\left(\mathbf{q}_{t+1} = \mathbf{S}_j \,|\, \mathbf{q}_t = \mathbf{S}_i\right) . \qquad (3)$$

Furthermore, we assume that participants behave independently of each other, given their immediately preceding joint vocal activities,

$$a_{ij} \;=\; \prod_{k=1}^{K} P\left(\mathbf{q}_{t+1}\,[k] = \mathbf{S}_j\,[k] \,|\, \mathbf{q}_t = \mathbf{S}_i\right) . \qquad (4)$$

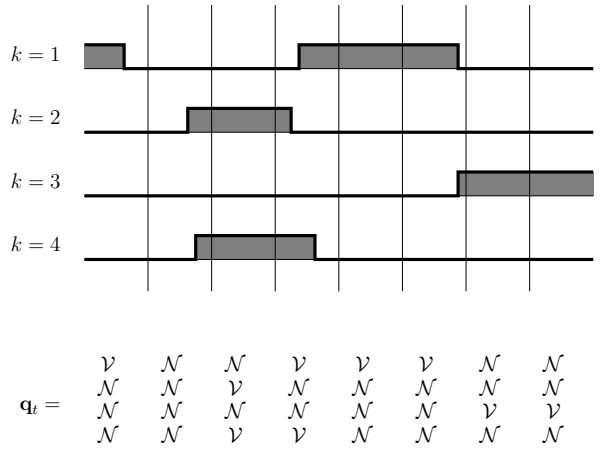

Figure 1: Discretization of multichannel segmentation references by assigning $\mathcal{V}$ for participant k at time t if that participant vocalizes for more than 50% of the duration of the frame centered at t, and $\mathcal{N}$ otherwise.

We propose to characterize the vocal behavior of participants over the entire course of conversation $\mathcal{C}$ using a subset of the probabilities a_{ij}. The features we explore, shown in Equations 5 to 8, represent the probability that participant k initiates vocalization during silence (VI), the probability that participant k continues non-overlapped vocalization (VC), the probability that participant k initiates overlap (OI) while only participant j vocalizes, and the probability that participant k continues vocalizing in overlap (OC) with participant j only, respectively. For this work, we neglect cases where more than one participant (other than j) is vocalizing at time t before participant k starts vocalizing, since such instances are rare.

The probabilities in Equations 5 to 8 can be estimated directly using a maximum likelihood (ML) criterion by accumulating bigram counts matching the event classes in each equation. For simplicity, we set the probabilities for which the conditioning context is never observed to 0.5.

In characterizing an entire conversational group of K participants, the feature vector $\mathbf{F}$ consists of K one-participant features of type f_k^{VI} and K one-participant features of type f_k^{VC}, as well as $K^2 - K$ two-participant features of type $f_{k,j}^{OI}$ and $K^2 - K$ two-participant features of type $f_{k,j}^{O}C$. This results in a total of $N_{\mathbf{F}} = 2K^2$ features per conversation; we note that conversations vary in the participant number K and therefore in their feature vector size.

3.2 Feature Estimation using the Ising Model

We contrast ML estimation of features with estimation which relies on a particular form of parameter tying, under an asymmetric infinite-range variant of the Ising model (Glauber, 1963). Canonically, the Ising model is used to study an ensembles emergent macroscopic properties, which are due to the microscopic interactions among its very large number of binary particles; we apply it here to study the emergent vocal interaction patterns of K participants. The modified Ising model is easily implemented as a single-layer neural network (Hertz et al., 1991) of K input units, K output units, and a sigmoid transfer function,

$$y_k(\mathbf{x}) \;=\; \frac{1}{1 + e^{-\beta\left(\sum_{j=1}^{K} w_{k,j} x_j + b_k\right)}} \quad , \quad (9)$$

where β is a parameter which is inversely proportional to the pseudo-temperature; we set it here to unity for convenience. x_j are the elements of vector $\mathbf{x}$, $w_{k,j}$ are the elements of a weight matrix $\mathbf{W} \in \Re^{K \times K}$, and b_k are the elements of a bias vector $\mathbf{b} \in \Re^{K}$. We show this network in Figure 2. When presented with an input vector $\mathbf{q}_t$, the network produces at each output unit the quantity

$$P(\,\mathbf{q}_{t+1}[k] = \mathcal{V} \,|\, \mathbf{q}_t = \mathbf{S}_i\,) \;=\; y_k(\mathbf{S}_i) \;. \quad (10)$$

In computing $y_k(\mathbf{S}_i)$, $\mathcal{V}$ and $\mathcal{N}$ are mapped to 1 and 0, respectively.

The network is characterized by the parameters $\mathbf{W}$ and $\mathbf{b}$, which can be learned from $\mathbf{q}_t$, $1 \leq t \leq T$, using a standard first-order or second-order gradient descent technique, for example. At each time frame, the current $\mathbf{q}_t$ binary vector can be used as a "pattern", with the subsequent $\mathbf{q}_{t+1}$ binary vector as the "target"; there are a total of $T-1$ such pattern-target pairs. The appropriate objective function for outputs representing multiple (conditionally) independent attributes is the binomial error (Bishop, 1995). To distinguish from features estimated using ML, as described in the previous section, we henceforth refer to features estimated using the Ising model as "NN".

$$f_k^{VI} = P(\mathbf{q}_{t+1}[k] = \mathcal{V} \mid \mathbf{q}_t[i] = \mathcal{N} \quad \forall \quad 1 \leq i \leq K) , \tag{5}$$

$$f_k^{VC} = P(\mathbf{q}_{t+1}[k] = \mathcal{V} \mid \mathbf{q}_t[k] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N} \quad \forall \quad i \neq k, 1 \leq i \leq K) , \tag{6}$$

$$f_{k,j}^{OI} = P(\mathbf{q}_{t+1}[k] = \mathcal{V} \mid \mathbf{q}_t[j] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N} \quad \forall \quad i \neq j, 1 \leq i \leq K) , \quad j \neq k , \tag{7}$$

$$f_{k,j}^{OC} = P(\mathbf{q}_{t+1}[k] = \mathcal{V} \mid \mathbf{q}_t[k] = \mathbf{q}_t[j] = \mathcal{V}, \mathbf{q}_t[i] = \mathcal{N} \quad \forall \quad i \neq j, i \neq k, 1 \leq i \leq K) , \quad j \neq k . \tag{8}$$

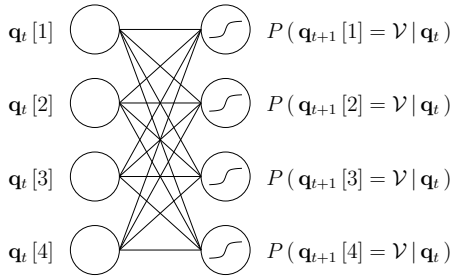

Figure 2: Infinite-range Ising model for predicting conditionally independent probabilities of activation at time $t+1$ given activations at time t, for a conversation with four participants; for clarity, bias connections are elided.

In closing this section, we note that the proposed interaction features have a particularly prosaic form under this model, when $\mathcal{N} = 0$ and $\mathcal{V} = 1$:

$$f_k^{VI} = \frac{1}{1 + e^{-b_k}} , \tag{11}$$

$$f_k^{VC} = \frac{1}{1 + e^{-b_k - w_{k,k}}} , \tag{12}$$

$$f_{k,j}^{OI} = \frac{1}{1 + e^{-b_k - w_{k,j}}} , \tag{13}$$

$$f_{k,j}^{OC} = \frac{1}{1 + e^{-b_k - w_{k,j} - w_{k,k}}} . \tag{14}$$

Furthermore, the total number of parameters to be estimated from segmentation data is $K(K+1)$, rather than $2K^2$ for the bigram ML model.

4 Modeling Groups

In this section we describe the structure, parameter estimation, and probability evaluation for the membership and the behavior models as introduced in Equation 1.

4.1 Behavior Model

We assume conditional independence among the elements of the feature vector $\mathbf{F}$,

$$\mathbf{F} = \bigcup_{k=1}^{K} \left\{ f_k^{VI}, f_k^{VC}, \bigcup_{j \neq k}^{K} \left\{ f_{k,j}^{OI}, f_{k,j}^{OC} \right\} \right\} , \tag{15}$$

such that

$$P(\mathbf{F} \mid \mathcal{G}, \mathcal{T}) = \tag{16}$$
$$\prod_{k=1}^{K} P\left(f_k^{VI} \mid \theta^{VI}_{\mathcal{T},\mathcal{G}[k]} \right) P\left(f_k^{VC} \mid \theta^{VC}_{\mathcal{T},\mathcal{G}[k]} \right) \times$$
$$\prod_{j \neq k}^{K} P\left(f_{k,j}^{OI} \mid \theta^{OI}_{\mathcal{T},\mathcal{G}[k],\mathcal{G}[j]} \right) P\left(f_{k,j}^{OC} \mid \theta^{OC}_{\mathcal{T},\mathcal{G}[k],\mathcal{G}[j]} \right) .$$

In the above, each θ represents a single one-dimensional Gaussian mean μ and variance Σ pair. These parameters are maximum likelihood estimates from the f_k and $f_{k,j}$ values in a training set of conversations, smoothed towards their global values.

4.2 Membership Model

Equation 1 allows for the inclusion of a prior probability on the presence and arrangement of participants with respect to channels. Although participants may have tendencies to sit in close proximity to certain other participants, we ignore channel preference in the current work. We employ the simple membership model

$$P(\mathcal{G} \mid \mathcal{T}) = \frac{1}{Z_{\mathcal{G}}} \prod_{k=1}^{K} P(\mathcal{G}[k] \mid \mathcal{T}) , \tag{17}$$

where $Z_{\mathcal{G}}$ is a normalization constant which ensures that $\sum_{N_{\mathcal{G}}} P(\mathcal{G} \mid \mathcal{T}) = 1$. We set each factor $P(\mathcal{G}[k] \mid \mathcal{T})$ to the ML estimate for participant $\mathcal{G}[k]$ in the training data. For example, if $\mathcal{G}[k]$ represents an identifier unique in $\mathcal{P}$, i.e. a name, then $P(\mathcal{G}[k] \mid \mathcal{T})$ is simply the proportion of meetings

of type $\mathcal{T}$ attended by the participant with that name. To allow the model to hypothesize rarely observed participants in the training material, we set this probability no lower than 0.1, a factor selected empirically without extensive validation.

4.3 Search

Equation 1 calls for the exhaustive enumeration of all possible groups $\mathcal{G}$. As mentioned in Section 2, there are $N_{\mathcal{G}} = \|\mathcal{P}\|! / (\|\mathcal{P}\| - \|\mathcal{G}\|)!$ different groups, which may make such enumeration intractable. Since we are not interested in automatically classifying participants, clustering participants in the training material and thereby reducing $\|\mathcal{P}\|!$ offers a simple means of limiting the magnitude of $N_{\mathcal{G}}$.

In the current work, we choose to cluster participants by training models not for specific participants, but for participant rank in terms of vocalizing time proportion. This makes the attribute $\mathcal{G}[k]$ unique in $\mathcal{G}$ rather than in $\mathcal{P}$. For each training conversation, we rank participants in terms of the overall proportion of time spent in state $\mathcal{V}$, in descending order, such that participant rank 1 refers to that participant who vocalizes most often during the conversation in question. This form of clustering also eliminates the problem of estimating models for specific participants which appear in only a handful of conversations.

Since a test conversation of K participants contains participant ranks $\{1, 2, \ldots, K\}$ and no others, the enumeration of $N_{\mathcal{G}}$ unique participant groups $\mathcal{G}$ in Equation 1 is replaced by an enumaration of $K! = \|\mathcal{G}\|!$ unique rank groups. However, we note that under this simplification, the membership model has only a small impact.

5 Classification Experiments

5.1 Data

In our experiments, we use the ICSI Meeting Corpus (Janin et al., 2003), consisting of 75 unscripted, naturally occurring multi-party meetings. There are 3 aspects which make this corpus attractive for the current work. First, it is larger than most multi-party conversation corpora. This is important because, in our framework, each meeting represents one data point. Second, meeting participants are drawn from a pool of 52 speakers, several of whom occur in more than one meeting type. Finally, meetings are not fixed in participant number, allowing us to demonstrate the generalization of our methods to arbitrary conversational group sizes.

$\mathcal{T}$	#	$\|\mathcal{P}\|$	$\|\mathcal{G}\|$		
			mod	min	max
Bed	15	13	6	4	7
Bmr	29	15	7	3	9
Bro	23	10	6	4	8

Table 1: Characteristics of the three ICSI meeting types considered: number of meetings (#); size of population from which participants are drawn ($\|\mathcal{P}\|$); mode (mod), minimum (min) and maximum (max) number of participants ($\|\mathcal{G}\|$) per meeting type $\mathcal{T}$.

67 of the meetings in the corpus are of one of three distinct meeting types, Bed, Bmr, and Bro, representing different projects, with different purposes for holding meetings. This is reflected in differences between patterns of vocal interaction; for example, Bmr meetings consist of more free-form discussion, presumably among peers, than either Bed or Bro meeting types. In contrast, the latter two types exhibit more asymmetry in participant roles than do Bmr meetings, and therefore the more easily inferable social structure. Furthermore, there are three speakers in the corpus which attend both Bro and Bmr meeting types, and one speaker which attends both Bed and Bmr meeting types; Bro and Bed types, however, have disjoint attendee subpopulations. A participant which appears in multiple meeting types may affect the overall interaction styles of the two types to be less distinct. This is especially true if he or she attends the majority of meetings of both types, as is the case for two of the participants which attend both Bmr and Bro meetings.

We present several additional characteristics of these three meeting types in Table 1. We ignore the remaining 8 meetings in the corpus, representing types of which there are too few exemplars for modeling. As Table 1 shows, the prior distribution over the 3 considered types is such that random guessing yields a 43% three-way classification accuracy.

We obtain the vocal interaction record $\mathbf{q} =$

$\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_T\}$ for each of the 67 meetings by discretizing their reference segmentations. The latter were produced by: (1) generating a talk spurt segmentation through forced alignment of transcribed words (available as part of the ICSI MRDA Corpus (Shriberg et al, 2004)), and bridging of inter-word gaps shorter than 0.3 s; (2) inferring a segmentation for transcribed laughter from the forced alignment of surrounding words, and manually segmenting isolated bouts of laughter (as described in (Laskowski and Burger, 2007)); and (3) merging the talk spurt and laugh bout segmentations. Fully automatic inference of the vocal interaction record, from audio, is beyond the scope of the current work.

5.2 Baseline Performance

To assess the difficulty of the problem, we propose a baseline which relies only on the proportion of vocalizing time, f_k^T, for each participant k. This is a frequently studied quantity for describing conversational style (Burger et al., 2002) and for assessing the performance of speaker diarization systems (Jin et al., 2004) (Mirghafori and Wooters, 2006).

The classification accuracy of the baseline, using the framework described by Equations 1, 16 and 17, is 65.7%. This performance is achieved with leave-one-out classification, using 66 meetings for training and one for testing, 67 times. The accuracy figures in this and in the subsequent section should be treated as estimates on a development set; since the longitudinal nature of the ICSI corpus is relatively unique, it is has not been possible to construct a fair evaluation set without significantly depleting the amount of training material.

We note that, as mentioned in Subsection 4.3, the membership model has negligible impact when participant vocalizing rank is used as the clustering criterion during training. This condition identically affects all of the experiments which follow, allowing for an unbiased comparison of the proposed vocal interaction features.

5.3 Feature Comparison

We present several leave-one-out experiments in order to evaluate the utility of each of the VI, VC, OI, and OC feature types separately, without f_k^T, estimating them from the multichannel reference segmentation for each meeting using both maximum likelihood (column 2), and the proposed neural network model (column 4). The results show that classification using ML-estimated single-participant features f_k^{VI} and f_k^{VC} outperforms classification using NN-estimated features. However, NN estimation outperform ML estimation when it comes to the two-participant features $f_{k,j}^{OI}$ and $f_{k,j}^{OC}$. This result is not surprising, since vocalization in overlap is much more rare than vocalizing alone, rendering maximum likelihood estimation of overlap behavior uncompetitive without additional smoothing.

Feature(s)	ML Estimation		NN Estimation	
	w/o f_k^T	w/ f_k^T	w/o f_k^T	w/ f_k^T
baseline	—	65.7	—	65.7
f_k^{VI}	59.7	67.2	56.7	65.7
f_k^{VC}	62.7	77.6	56.7	71.6
$\langle f_{k,j}^{OI} \rangle_j$	35.8	52.2	64.2	67.2
$\langle f_{k,j}^{OC} \rangle_j$	53.7	67.2	64.2	80.6
$f_{k,j}^{OI}$	41.8	46.3	67.2	64.2
$f_{k,j}^{OC}$	61.2	68.7	73.1	79.1
all	61.2	64.2	74.6	82.1
opt	—	—	74.6	83.6

Table 2: Leave-one-out meeting type classification accuracy using various feature combinations within the proposed Bayesian framework. "opt" consists of the features f_k^{VI}, $f_{k,j}^{OI}$, and $f_{k,j}^{OC}$.

In addition to the two-participant interaction features $f_{k,j}^{OI}$ and $f_{k,j}^{OC}$ described in Section 3, we also show the performance of summary single participant features $\langle f_{k,j}^{OI} \rangle_j = \sum_{j=1}^{K} f_{k,j}^{OI}/K$ and $\langle f_{k,j}^{OC} \rangle_j = \sum_{j=1}^{K} f_{k,j}^{OI}/K$, which average the overlap behavior of participant k over the possible identities of the already vocalizing participant j. When these features are used alone, they are outperformed by the two-participant features. This suggests that average overlap behavior does not distinguish between the three meeting types as well as does the overlap interaction between participants of specific vocalizing time rank.

Columns 3 and 5 of Table 2 show the performance of the same 6 feature types, in combination with the f_k^T features. Due to space constraints, we mention only that most feature types appear to combine additively with f_k^T. We also show, in the last two lines

Estimated	Actual Type		
	Bed	Bmr	Bro
Bed	11	1	3
Bmr	2	26	1
Bro	3	1	19

Table 3: Confusion matrix among the three ICSI meeting types studied, for classification with NN-estimated "opt" feature set (f_k^{VI}, $f_{k,j}^{OI}$, and $f_{k,j}^{OC}$).

of the table, the performance of all feature types together, as well as of an "oracle" feature set derived using backward feature selection, by removing the worst performing feature one at a time from the "all" feature set. The best number achieved, 83.6%, was obtained using total vocalizing proportion f_k^T, NN-estimated single-participant f_k^{VI}, and NN-estimated two-participant features $f_{k,j}^{OI}$ and $f_{k,j}^{O}C$, which describe the overlap behavior of specific participant ranks with respect to specific other participant ranks. The accuracy represents a 52% relative reduction over the baseline (from 34.3% to 16.4%).

We show the confusion matrix of the "opt" NN-estimated feature set in Table 3. Although the amount of data is too small to draw statistically meaningful conclusions, the symmetrical misclassification of 3 Bro meetings as type Bed and 3 Bed meetings as type Bro suggests that in fact the Bro and Bed meeting types are more similar to each other than either is to the Bmr meeting type.

6 Conclusions

We have proposed a framework for the classification of conversational style in multi-participant conversation. The framework makes use of several novel elements. First, it relies exclusively on text-independent features, extracted from the multiparticipant vocal interaction patterns of a conversation; the technique is directly deployable for languages for which mature automatic speech recognition or dialog act classification infrastructure may be lacking. Second, we have made use of a well-studied model in stochastic dynamics, the Ising model, to improve estimates of the transition probabilities that describe the evolution of multiparticipant vocal interaction over the course of conversation. Third, we have introduced the concept of enumerable groups of participants, making it possible to include features which model the interaction between specific pairs of participants, for meetings with any number of participants. Finally, we have applied the framework to the task of classifying meeting types. Our experiments show that features describing the text-independent interaction between participants of specific vocalizing time rank, when used in conjunction with a feature which performs poorly on its own f_k^{VI}, lead to a relative error reduction of 52% over our baseline.

The key findings from the analysis of different interaction features are that having detailed 2-participant features is better than simply using the average for a given target speaker, and that using interaction features (conversation dynamics) gives better results than the static measure of relative speaking time. Of course, the best results are achieved with a combination of these types of features.

7 Future Work

In the future, we will apply the proposed classification system to automatically generated multichannel segmentation and alternatives to the Gaussian classifier. It may also be interesting to investigate separately representing different types of vocalization (e.g. speech vs. laughter) and features related to overlaps of more than two speakers.

For resource rich languages, meeting type can be classified using lexical features from speech recognition. However, if one is interested in detecting meeting type independent of content, the choice of word features needs to factor out topic. It would be interesting to assess the relative importance of words vs. interactions, and the degree to which they are complementary, in the topic-independent context.

Finally, another important future direction is the application of the techniques to the dual of Equation 5,

$$\begin{aligned} \mathcal{G}^* &= \arg\max_{\mathcal{G}} P(\mathcal{G} \mid \mathbf{F}) \\ &= \arg\max_{\mathcal{G}} \sum_{\mathcal{T}} P(\mathcal{G}, \mathcal{T}, \mathbf{F}) \\ &= \arg\max_{\mathcal{G}} \sum_{\mathcal{T}} P(\mathcal{T}) \times \\ &\qquad P(\mathcal{G} \mid \mathcal{T}) \times P(\mathbf{F} \mid \mathcal{G}, \mathcal{T}) \end{aligned} \tag{18}$$

namely the problem of jointly characterizing participants rather than conversations.

8 Acknowledgments

This work was partly supported by the European Union under the integrated project CHIL (IST-506909), Computers in the Human Interaction Loop, while Mari Ostendorf was a Visiting Professor at the University of Karlsruhe.

References

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. *Proceedings of INTERSPEECH*, Jeju Island, South Korea.

S. Basu 2002. Conversational Scene Analysis. doctoral thesis, MIT.

O. Brdiczka and J. Maisonnasse and P. Reignier. 2005. Automatic detection of interaction groups. *Proceedings of ICMI*, Trento, Italy.

S. Burger and V. MacLaren and H. Yu. 2002 The ISL Meeting Corpus: The impact of meeting type on speech style. *Proceedings of ICSLP*, Denver CO, USA, pp301–304.

J. Dabbs and R. Ruback. 1987. Dimensions of group process: Amount and structure of vocal interaction. *Advances in Experimental Psychology*, 20, pp123–169.

N. Fay and S. Garrod and J. Carletta. 2000. Group discussion as interactive dialogue or as serial monologue: The influence of group size. *Psychological Science*, 11(6), pp487–492.

R. Glauber. 1963. Time-dependent statistics of the Ising model. *Journal of Mathematical Physics*, 4(2), pp294–307.

J. Hertz and A. Krogh and R. Palmer. 1991. *Introduction to the Theory of Neural Compuations.* Addison-Wesley Longman.

A. Janin and D. Baron and J. Edwards and D. Ellis and D. Gelbart and N. Morgan and B. Peskin and T. Pfau and E. Shriberg and A. Stolcke and C. Wooters. 2003. The ICSI Meeting Corpus. *Proceedings of ICASSP*, Hong Kong, China, pp364–367.

Q. Jin and K. Laskowski and T. Schultz. 2004 Speaker segmentation and clustering in meetings. *Proceedings of ICASSP NIST RT-04s Spring Meeting Recognition Evaluation Workshop*, Montreal, Canada.

K. Laskowski and T. Schultz. 2006. Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings. *Proceedings of ICASSP*, Toulouse, France, pp993–996.

K. Laskowski and T. Schultz. 2007. Modeling vocal interaction for segmentation in meeting recognition. *Proceedings of MLMI (to appear)*, Brno, Czech Republic.

K. Laskowski and T. Schultz. 2007. Analysis of the occurrence of laughter in meetings. *Proceedings of INTERSPEECH (to appear)*, Antwerpen, Belgium.

I. McCowan and S. Bengio and D. Gatica-Perez and G. Lathout and M. Barnard and D. Zhang. 2005. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3), pp305–317.

N. Mirghafori and C. Wooters. 2006. Nuts and Flakes: A study of data characteristics in speaker diarization. *Proceedings ICASSP*, Toulouse, France, pp1017–1020.

S. Renals and D. Ellis. 2003. Audio information access from meeting rooms. *Proceedings ICASSP*, Hong Kong, China, pp744–747.

R. Rienks and D. Heylen. 2005. Dominance detection in meetings using easily obtainable features. *Proceedings MLMI*, Edinburgh, UK.

H. Sacks and E. Schegloff and G. Jefferson. 1974. A simplest semantics for the organization of turn-taking for conversation. *Language*, 50(4), pp696–735.

E. Shriberg and R. Dhillon and S. Bhagat and J. Ang and H. Carvey. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *Proceedings SIGdial*, Cambridge MA, USA, pp97–100.

D. Wyatt and J. Bilmes and T. Choudhury and H. Kautz. 2007. A privacy-sensitive approach to modeling multi-person conversations. *Proceedings IJCAI*, Hyderabad, India, pp1769–1775.

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