Chapter 11
Analysis of Group Conversations: Modeling Social Verticality

Oya Aran and Daniel Gatica-Perez

11.1 Introduction

Social interaction is a fundamental aspect of human life and is also a key research area in psychology and cognitive science. Social psychologists have been researching the dimensions of social interaction for decades and found out that a variety of social communicative cues strongly determine social behavior and interaction outcomes. Many of these cues are consciously produced, in the form of spoken language. However, besides the spoken words, human interaction also involves nonverbal elements, which are extensively and often unconsciously used in human communication. The nonverbal information is conveyed as wordless messages, in parallel to the spoken words, through aural cues (voice quality, speaking style, intonation) and also through visual cues (gestures, body posture, facial expression, and gaze) [31]. These cues can be used to predict human behavior, personality, and social relations. It has been shown that, in many social situations, humans can correctly interpret the nonverbal cues and can predict behavioral outcomes with high accuracy, when exposed to short segments or "thin slices" of expressive behavior [1]. The length of these thin slices can change from a few seconds to several minutes depending on different situations.

Computational analysis of social interaction, in particular of face-to-face group conversations is an emerging field of research in several communities such as human–computer interaction, machine learning, speech and language processing, and computer vision [20, 38]. Close connection with other disciplines including psychology and linguistics also exist in order to understand what kind of verbal

O. Aran (✉) · D. Gatica-Perez
Idiap Research Institute, Martigny, Switzerland
e-mail: oya.aran@idiap.ch

D. Gatica-Perez
e-mail: gatica@idiap.ch

A.A. Salah, T. Gevers (eds.), *Computer Analysis of Human Behavior*,
DOI 10.1007/978-0-85729-994-9_11,

and nonverbal signals are used in diverse social situations to infer human behavior. The ultimate aim is to develop computational systems that can automatically infer human behavior by observing a group conversation via sensing devices such as cameras and microphones. Besides the value for several social sciences, the motivation behind the research on automatic sensing, analysis, and interpretation of social behavior has several dimensions. These systems could open doors to a number of relevant applications that support interaction and communication. These include tools that improve collective decision making and that support self-assessment, training, and education, with possible example applications such as automatic meeting evaluators, trainers, and automatic personal coaches for self learning. Moreover, not only supporting human interaction, these systems can also support a natural human–robot or human–computer interaction, by designing socially aware systems [38], i.e. by enabling a robot to understand the social context around it and to act accordingly.

In this chapter we focus on one aspect of social relations and interactions: social verticality. Social verticality is one of the many dimensions of human relations and refers to the structure of interpersonal relations positioned in a low-to-high continuum, stating a kind of social hierarchy among people [22]. It relates to power, status, dominance, leadership, and other related concepts. The vertical dimension is in contrast to the horizontal dimension, which is the affective and socio-emotional dimension that describes the emotional closeness of human relations. Instead, vertical dimension describes how each person is positioned in the group, e.g. as higher status/lower status. We present computational models for the analysis of social verticality through nonverbal cues in small groups.

The next section gives definitions and a brief summary of the psychological and cognitive aspects of the display and perception social verticality during human interactions. Section 11.3 describes computational methods and Sect. 11.4 presents four case studies. A summary of the chapter, acknowledgments, end-of-chapter questions and a small glossary can be found in the remaining sections.

11.2 Social Verticality in Human Interaction and Nonverbal Behavior

Social verticality constructs, such as power, status, and dominance, are related to each other with important differences in their definitions. For example, power indicates “the capacity to produce intended effects, and in particular, the ability to influence the behavior of another person” [17] (p. 208), without implying any respect or prestige. As power is defined as an ability, it is not always exercised. On the other hand, dominance can be defined as “a personality trait involving the motive to control others, the self perception of oneself as controlling others, and/or as a behavioral outcome with a success on controlling others” [22] (p. 898), and as a result, it is “necessarily manifest” [17] (p. 208). Dominance can be seen as a “behavioral manifestation of the relational construct of power” [17] (p. 208).

Status relates to both power and dominance and is defined as an "ascribed or achieved quality implying respect and privilege, does not necessarily include the ability to control others" [22] (p. 898). Leadership is another related construct, which can be defined as the ability of motivating a group of people to pursue a common goal. Thus, leadership is related to the end result, not just a manifested act. Among various types of leadership types, "emergent leadership", for example, is where the leader arises from a group of equal status people [46].

Dominance is one of the fundamental dimensions of social interaction. It is signaled via both verbal and nonverbal cues. The nonverbal cues include vocalic ones such as speaking time [45], loudness, pitch, vocal control, turns, and interruptions [17] and kinesic ones such as gesturing, posture, facial expressions, and gaze [16]. Dominant people are in general more active both vocally and kinesically, with an impression of relaxation and confidence [22]. It has been shown that they also have a higher visual dominance ratio (looking-while speaking to looking-while-listening ratio), i.e. they look at others more while speaking and less while listening [16].

In a study that investigated the relationship between the leadership style and sociable and aggressive dominance, it is found out that there is a higher correlation between leadership and sociable dominance [28]. Sociably dominant people look at others more while speaking and use more gestures, receiving more frequent and longer-lasting glances from the group; whereas aggressively dominant people interrupt more, and they look at others less while listening.

Social verticality in a group can also be defined by roles that constitute a hierarchy-like structure, such as physician/patient, manager/employee, teacher/learner, interviewer/interviewee, where one part has more expertise than the other, in terms of knowledge or rank. In [45], it is shown that the association between speaking time and dominance is higher for both dominant and high status people. It is important to note that not all the role distributions of a group present a vertical relationship, i.e. a hierarchical relation. In this chapter, we mainly refer to the roles that have a vertical dimension. We also consider the roles that are defined based on the psychological behavior (i.e. functional roles) of the participants in a group that partly shows a hierarchical structure.

The relationship between the vertical constructs and personality traits is also of interest to social psychologists. Personality is defined as a collection of consistent behavioral and emotional traits that characterize a person [18]. While personality addresses stable and consistent behavior of a person, the social verticality constructs address the behavior of a person in a group which may not be consistent across time, relations and situations [22]. Nevertheless, verticality constructs are closely related with the personality traits of the individuals, as personality strongly influences verticality in a social relationship of peers. For example, it is shown that cognitive ability and the personality traits of extroversion and openness to experience are predictive of emergent leadership behavior [29].

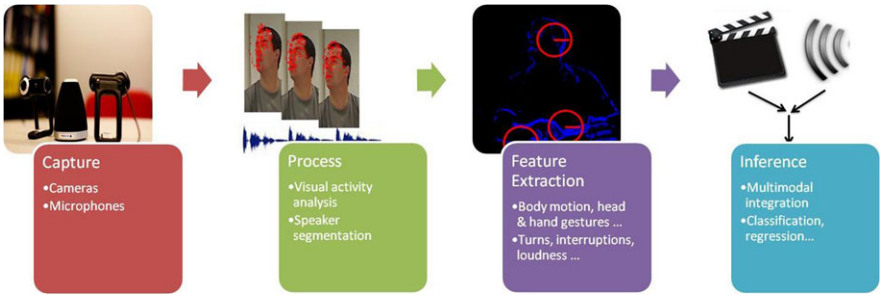

Fig. 11.1 The functional blocks of social interaction analysis

11.3 Automatic Analysis of Social Verticality from Nonverbal Features

The aim of the automatic analysis of group interaction is to infer and possibly predict aspects of the underlying social context, including both individual attributes and interactions with other people in the group. As a specific case, the concept of social verticality imposes a social hierarchy on the group and requires a special interest.

In this chapter we concentrate on three social constructs of social verticality, dominance, roles, and leadership, and explain the functional blocks, as shown in Fig. 11.1, that are required for analysis. Table 11.1 presents an overview of recent works selected from the literature and Table 11.2 summarizes several datasets used for social verticality analysis.

In the next sections, we mainly focus on audio and visual sensors as the two primary sources of information. A brief discussion on other sensors to capture social behavior is given in Sect. 11.3.2.3.

11.3.1 Processing

Although it is known that nonverbal communication involves both the aural and visual modalities, most initial works in the literature on computational analysis of group interactions have largely focused on audio features. This is partly related to the data capture technology. Reasonable quality audio capture systems were available earlier than video capture systems. Overall, video capture is more problematic than audio capture; video is quite sensitive to environmental conditions, and requires adequate resolution and frame rates. A second dimension is related to the challenges of video processing in natural conversations. One last dimension is related to privacy. People are in general less willing to have their video recorded compared to audio.

It is also important to note that social interaction capture systems should not use sensors that affect the interaction. The subjects should be able to act naturally without any distraction caused by the sensors that are used. For instance, to record audio,

Table 11.1 Related works on the analysis of social verticality

	Task	Audio features	Video features	Fusion	Inference
Dominance					
[24]	most/least dominant person	speaking turn	NA	NA	rule-based
[2]	most/least dominant person	speaking turn	visual activity, audio-visual activity	score level	rule-based
[12]	correlations with dominance	prosodic	NA	NA	correlations
[27]	most/least dominant person	speaking turn, prosodic	visual activity	feature level	rule-based supervised -SVM
[26]	most dominant/ high status person	speaking turn	visual activity, visual attention	NA	rule based
[42]	dominance level (as high, normal, low)	speaking turn (manual)	NA	NA	supervised -SVM
Roles					
[41]	role patterns	speaking turn	NA	NA	rule-based supervised -Influence Model
[43]	role recognition	speaking turn	NA	feature level	supervised -MAP, ML est. -Simulated ann.
[19]	role recognition	speaking turn, verbal	NA	score level	supervised -ML estimate -Boosting
[15]	role recognition	speaking turn	visual activity	feature level	supervised -SVM -HMM -Influence Model
[50]	role recognition	speaking turn	visual activity	feature level	supervised -SVM
[6]	role identification	speaking turn, verbal	NA	NA	supervised -Boosting
Leadership					
[44]	emergent leadership	speaking turn	NA	feature level	rule-based
[48]	leadership	prosodic	visual activity, gestures	feature level	rule-based
[25]	group conversational patterns	speaking turn	NA	feature level	unsupervised -topic models

Table 11.2 Selected databases used for social verticality analysis

Dataset Name	Task	Details	Length	References
DOME (part 1,2)	Dominance	Meetings, a subset of the AMI corpus, publicly available	∼10 hours	[2, 3]
DOME (part 1)	Dominance	Meetings, a subset of the AMI corpus, publicly available	∼5 hours	[12, 24, 26, 27]
–	Dominance, Roles	Meetings from "The Apprentice" TV show	90 minutes	[41]
–	Roles	News, talk shows from Swiss radio	∼46 hours	[43]
AMI	Roles	Meetings, publicly available	∼46 hours	[19, 43]
Survival	Roles	Meetings on the mission survival task	∼5 hours	[15, 40, 50]
ELEA	Leadership	Meetings of newly formed groups	∼10 hours	[44]

a distant microphone array device should be preferred to head-set microphones that are attached to the people. Regarding the cameras, while it is true that people might be distracted or feel self-conscious at the beginning of an interaction, when they are in a natural environment and focused on an engaging task, they rapidly tend to forget about the presence of cameras and act naturally.

11.3.1.1 Audio Processing

The key audio processing step is to segment each speaker's data in the conversation such that it allows robust further processing to extract features for each of the participants separately. This process is called speaker diarization, a combination of speaker segmentation (finding speaker change points) and speaker clustering (grouping segments for each speaker). Speaker diarization is an active topic, not only in social interaction analysis, but in speech processing in general. Here we refer to its applications in social interaction analysis very briefly.

As the audio capture methodology, one can use different setups ranging from one single microphone [24], to microphone arrays [44] and head-set microphones [2, 27]. Each of these setups have different noise levels and require different levels of processing for speaker diarization. Head-set microphones provide lower levels of noise, however, there is still a need for speaker diarization, since voices of other participants also exist in the recordings. In [24], the authors used a single audio source and investigated the performance of speaker diarization and dominance estimation under different conditions. Their results show that dominance estimation is robust to diarization noise in the single audio source case. For recording three-four people meetings, Sanchez-Cortes et al. [44] used a commercial microphone array,

which provides the speaker diarization output along with the audio recordings. The speaker diarization output is used for estimating the emergent leader in the group.

11.3.1.2 Video Processing

Once the camera input is received, the visual activity of each participant in the meeting needs to be processed. The level of processing depends on the features that will be extracted and ranges from face detection to motion detection, from face tracking to skin blob tracking and body parts tracking (see Chap. 3 for more details).

Face detection algorithms in general are designed to detect either frontal or profile faces. Their performance is affected especially if there are out-of-plane rotations. During group conversations, even if the participants are sitting around a table and are stationary, as a part of the interaction, they may frequently move their heads, and gaze at each other. Thus, a face detection algorithm alone is not sufficient to extract the positions of the faces during an interaction. Face tracking algorithms (using Kalman filter, particle filter, etc.) could be applied for better results. On top of face tracking, if the body movements, such as hand gestures and body postures, are also of interest, advanced techniques to track body parts or just the skin-colored parts can be applied. A lower level of processing, such as motion detection, can be applied if the interest is on the general visual activity of the participant and not on the individual body parts' activity. More detail on these techniques is given in Sect. 11.3.2.2.

11.3.2 Feature Extraction

The descriptors of a social interaction and social behavior can be categorized with respect to the sensor used, but also with respect to whether the feature is extracted from a single participant's activity or from an interaction that involves more than one person. We will call the former type of features as "independent features" and the latter as "relational features". Moreover, one can extract features that represent the overall interaction of the whole meeting, instead of the participants one by one, which we refer as "meeting features". The following sections present a sensor based categorization of both independent, relational, and meeting features.

The features presented below are generally extracted from thin slices of meetings, summarizing independent and group behavior for that meeting segment. Research in social psychology has shown that by examining only brief observations of expressive behavior, humans are able to predict behavioral outcomes [1]. The current research in computational social behavior analysis uses these conclusions and investigates whether computational methods are also able to predict similar outcomes by applying thin slice based processing. The length of the slices and how these slices are to be processed should be determined with respect to the

task and are open questions. The whole meeting duration can be used as one single segment, or the entire segment can be described as an accumulation of shorter slices.

11.3.2.1 Audio Nonverbal Features

In this section, we present the audio nonverbal features in two groups: speaking turn features and prosodic features. The term "speaking turn features" refers to audio nonverbal features that are extracted based on the speaking status of the participants and their turn taking behavior. The prosodic features on the other hand, represent the rhythm, stress, and intonation of speech.

Speaking Turn Features

Speaking turn features are frequently used in social interaction analysis for two main reasons. First, given the speaker diarization output, they are easy to calculate, with very low computational complexity. Second, despite their simplicity, they are very successful in many social tasks, supported by both social psychology and social computing research [20, 45].

Speaking turn features can be categorized as independent and relational. Independent features describe the speaking activity for one participant. These include speaking length, number of speaking turns, turn duration statistics (average, min, max, etc.). A turn is defined as one continuous segment where a participant starts and ends her/his speech. The relational features describe the interaction of one participant with other participants in the group. These include interruptions, overlapped speech, turn taking order (i.e. who speaks after whom) and also centrality features. Centrality features are relational features that represent the relative position of each participant in the group. We can represent the interaction of a group as a graph, by taking the nodes as the participants and the edges as the indication of how one person relates to others. The edges can be connected to several other relational features, such as "who interrupts whom", "who speaks after whom", etc. One can assign weights to the edges, representing the strength of the relation. Based on the definition of the relational features, the edges can be directed, in which case they are called arcs. Once the graph is formed, centrality measures can be calculated in various ways, such as indegree and outdegree of each node, closeness to other nodes (with weights defined as distances), etc.

The interaction patterns between participants can also be defined via Social Affiliation Networks (SAN) and used to represent the relationships between the roles [43]. A SAN is a graph that encodes "who interacts with whom and when". The two kinds of nodes in a SAN refer to the actors and the events, respectively. In [43], the events are defined as the segments from the recordings, and the participants are linked to the events if they talk during the corresponding segment. The assumption in this representation is that if the roles influence the structure of the interaction, similar interaction patterns should correspond to the same roles.

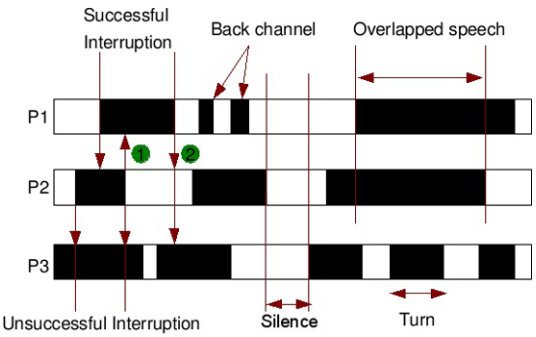

Fig. 11.2 Speaking turn audio nonverbal features

Speaking turn features that describe the whole meeting include the amounts of silence, overlapped speech and non-overlapped speech, among others. Accumulated statistics of all participants can also be used as meeting features (total number of turns, interruptions, etc.).

Figure 11.2 shows the illustration of a conversation between three participants. Each line represents the timeline for one participant and black segments indicate that the participant is speaking. Each black segment is a turn. The overlapped speech and silence segments are indicated. The interruptions and backchannels are also represented. The automatic detection of successful interruptions can be done in several ways if the verbal information is to be omitted. One definition can be made with respect to the interruptee's point of view (indicated with 1 in Fig. 11.2): "P1 started speaking when P2 was already speaking and P2's turn ended before P1's". Another definition uses the interrupter's point of view (indicated with 2 in Fig. 11.2): "P1 started speaking when P2 was already speaking and when P1's turn ended P2 was not speaking anymore". If we follow the first definition, P1 successfully interrupts P3 as well, however, with the second definition, it is an unsuccessful interruption. By definition, an interruption occurs between two participants. In that case, to calculate the number of interruptions for one person, all possible pairings with that person should be considered. As an alternative, interruptions that affect the whole group can be extracted [2].

All of these features need to be normalized with respect to the meeting duration (and with respect to the number of participants) before they are used in inference.

Prosodic Features

Other than speaking turn features, prosodic nonverbal features are also indicators of social behavior and used recently in several tasks such as dominance estimation. Features like pitch, energy, rhythm, or spectral features like formants, bandwidths, spectrum intensity can be extracted as independent features for each participant. Recently Charfuelan et al. [12] investigated the correlations between various prosodic features and dominance. Their results show that the most dominant person tends to speak louder and the least dominant person tends to speak softer than average.

In [48], prosodic features from audio such as loudness and spectral features are fused with visual features to estimate leadership in musical performances. In [26, 27], the speaking energy is used as a prosodic feature, together with other speaking turn features, for dominance estimation tasks. However, the results show that the speaking turn audio features, such as total speaking length and number of turns, can outperform speaking energy in predictive power.

11.3.2.2 Visual Nonverbal Features

Visual Activity

Most of the works in the nonverbal communication literature extract low-level visual features based on global image motion or geometric image primitives. In part, this approach can be feasible as there are no clearly defined hand shapes and hand trajectories for the visual nonverbal features of social verticality. Image and motion based approaches either assume that the background is stationary and any detected motion will indicate participant's visual activity, or find the skin-colored regions or faces and calculate the motion for these parts only.

In [27], two types of visual information, extracted from the compressed domain [49], are used for modeling dominance: the motion vector magnitude and the residual coding bit rate. While the motion vector magnitude reflects the global motion, the residual coding bit rate provides the local motion, such as lip movement on the face or finger movement on the hands. These two types of information can be used as indicators of the visual activity of the participant, either alone, or as a combination. In [2, 27], by thresholding the motion information, the authors extracted a binary vector in which zeros indicate no-motion segments and ones indicate the segments with motion. A number of higher level visual features are extracted from this binary vector, including the length, turns and interruptions of visual activity, similar to the audio speaking turn features explained in Sect. 11.3.2.1. Moreover, audio-visual versions of these features can also be extracted, by looking at the joint speaking and visual activity behavior. For example, in [2], the authors extracted visual attention features while speaking to estimate the most/least dominant persons in a group.

Another method is to use the motion history templates [9] for the detection and understanding of visual activity. In [13], motion history images are calculated for skin-colored regions and the amount of fidgeting, defined as "a condition of restlessness as manifested by nervous moments", is measured, by applying empirically determined thresholds. These features are used for the recognition of functional roles [15, 40, 50].

Visual Attention and Gaze

When and how much people look at each other during a conversation is a clear indicator of many social constructs. For example, dominant people often look at

others more while speaking and less while listening. And the ratio between these two measures, defined as the "Visual Dominance Ratio (VDR)", is considered as one of the classic measures of dominance [16]. Moreover, receiving more visual attention from other participants is an indicator of dominance.

For the automatic estimation of gaze and visual Focus of Attention (FOA), one can either use eye gaze or head pose. Although eye gaze is a more reliable source, with the current technology, this can be only implemented via eye trackers or high resolution cameras focused on each participants face area. Alternatively, the head pose can be estimated as an approximation to the actual eye gaze [5, 21, 36]. In a natural conversation environment, the focus target needs to be defined. Other than the participants in the conversation, there can be other targets such as the table, laptops, presentation screen, board, etc. In [5], an input–output hidden Markov model is used to detect the FOA of the group participants. In their work, the authors propose to recognize the FOA of all participants jointly and introduce a context dependent interaction model. Their model achieves around 10% performance increase when compared to using independent models for each participant. More details on FOA estimation can be found in Chap. 4.

Once the FOA of participants for each time frame is extracted, several measures can be defined as indicators of dominance. These include received visual attention, given visual attention (looking at others), and the VDR [23, 26]. These features, which are initially defined for dyadic conversations, can be generalized to the multi-party case by accumulating all possible pairwise participant combinations. It is important to note that VDR is by definition a multimodal cue, as it considers the FOA of a participant with respect to speaking status. For VDR, one needs to calculate "looking-while-speaking" and "looking-while-listening" measures. The "looking-while-speaking" case is trivial, however, "looking-while-listening" can be defined as "looking-while-not-speaking" or "looking-while-someone-else-is-speaking". In [26], the authors define two variants of VDR following these two definitions, and use them for dominance and status estimation.

Gestures and Facial Expressions

Despite the progress in computer vision to analyze structured gestures (e.g. hand gesture recognition, sign language recognition, gait recognition, etc.), the use of more accurate models of visual nonverbal communication has been largely unexplored. The main challenge is the lack of clearly defined gestures for visual nonverbal features. Another challenge is the uncontrolled experimental setup. Contrary to the natural conversational environment that is required for social interaction analysis, most of the developed gesture recognition algorithms require controlled environments and restrict people to perform gestures in a certain way.

To the authors' knowledge, the use of specific hand gestures, other than extracting general hand activity, has not been applied to the automatic analysis of social verticality. In one study [48], expressive gestures for musical performance, such as motion fluency, impulsiveness, directness, are used as features for leadership estimation. Head and body gestures have been used in related tasks in social interaction

analysis. While most of the works focus on the face area and head gestures, there are a few studies that also use the body motion. In [10], for estimating the participant status, the authors extract features from the face area and estimate yes-no head movements and also the global body movement. In [37], for an addressing task, to respond to questions such as "who responds to whom, when and how?", the authors extract features from the head area using a discrete wavelet transform to estimate head gestures such as nodding, shaking and tilt. A magnetic sensor is used in this study to capture the head motion.

Facial expressions are also strong indicators of social behavior. Despite progress on the automatic analysis of facial expressions, a widely studied topic in recent years, they are not as widely used in social interaction analysis. The main reason behind this is that facial expression analysis requires high resolution recordings of the facial region. However, most of the databases used for social interaction analysis use upper body or full body recordings of the participants and the captured facial region in these recordings are not good enough to perform high-level automatic expression analysis. Among the few works, in [32], the participants' smiling status is extracted during an interaction.

11.3.2.3 Other Sensors

The increasing use of mobile devices in people's daily lives introduced the opportunity to researchers to use these devices as capture devices. Mobile devices can record vast amounts of data from people's daily interaction via built-in audio and video sensors, but also via other sensors such as accelerometers to measure body movement, bluetooth or radio signals to measure proximity between two devices, and several others [34, 39].

The main challenge of using mobile devices for social behavior capture is limited computational resources. To overcome this difficulty, special equipment to collect social behavioral data can be developed. The sociometric badge is an example of such devices: it collects and analyzes social behavioral data. It allows voice capture, infrared (IR) transmission and reception and is capable of extracting features that can further be used for social behavior analysis [30], in real-time.

11.3.3 Inference

This section presents four groups of methods that have been used to infer social verticality:

1. *Rule-based approaches*
2. *Unsupervised approaches*
3. *Supervised approaches*
4. *Temporal approaches*

In the first approach, a decision with regards to the social verticality concept, for instance dominance, is made via rules defined using expert knowledge, without the need of training data. In the second and third approaches, unsupervised or supervised machine learning techniques are used, respectively. Their main difference is in the availability of labeled training data. Finally, in the fourth approach, the entire temporal dynamics of interaction is taken into account.

11.3.3.1 Rule-Based Approaches

For some social behavioral concepts, it is shown that people use several nonverbal cues more frequently or less frequently, with respect to other people in the group. For example, according to social psychology, dominant people often speak more, move more, or grab the floor more often [17, 22], so if someone speaks the most or moves the most, he/she is more likely to be perceived as dominant over the other people in the meeting. Following this information, one can assume that the nonverbal cues defined above are positively correlated with dominance and define a rule-based estimator on each related nonverbal feature [2, 27]. Similarly, other rule-based estimators can be defined for other social tasks, based on sociological and psychological aspects. Here we give an example of a rule-based estimator for dominance.

To estimate the most dominant person in meeting i, using feature f, the rule is defined as:

$$MD_i = \arg\max_p \left(f_p^i\right), \quad p \in \{1, 2, \ldots, P\}, \tag{11.1}$$

where p is the participant number, f_p^i is the value of the feature for that participant in meeting i, and P is the number of participants. The least dominant person can be estimated similarly using the following rule:

$$LD_i = \arg\min_p \left(f_p^i\right), \quad p \in \{1, 2, \ldots, P\}. \tag{11.2}$$

The main advantage of these rule based estimators is that they do not require any training data and they are very fast to compute. On the other hand, the major disadvantage is that they only allow the use of a single feature and cannot directly utilize the power of combining multiple features. By definition, the rule-based estimator is limited to a single feature. In the next section, we explain several approaches to perform fusion using the rule-based estimator.

Multimodal Fusion via Rule-Based Estimator

Although speaking length alone is a good estimator of dominance, there are other displays of dominance as well, such as the visual activity, which provides complementary information. Thus, different features representing different aspects of dominance could be fused together to obtain a better estimator. We can define a

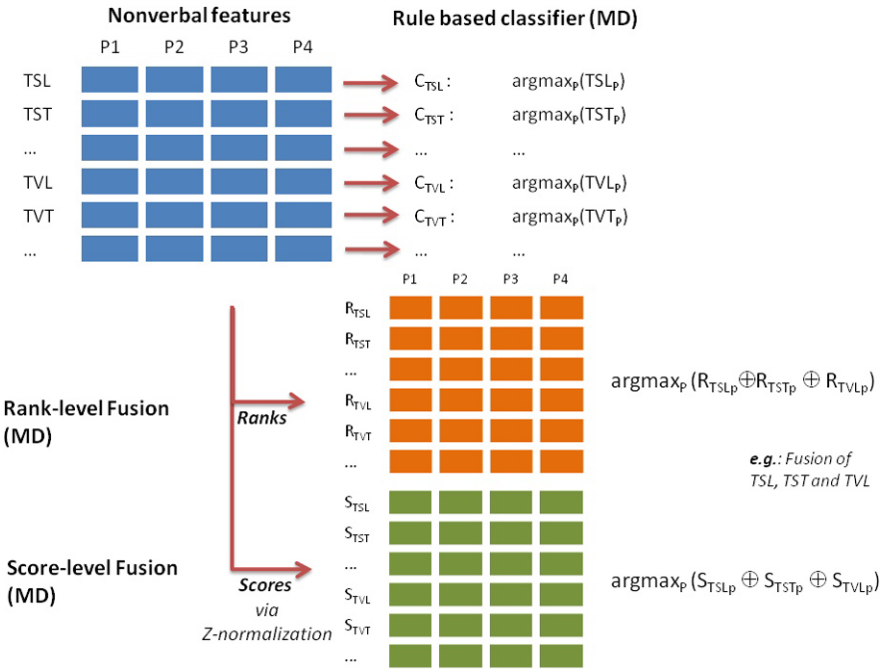

Fig. 11.3 Dominance estimation with rule based classifier and rank and score level fusion. The rank and score information is calculated from the extracted nonverbal features for each participant

rule-based estimator on each feature as an independent classifier and apply fixed combination rules on the decisions of these classifiers. Two different fusion architectures are presented in this section: score level and rank level fusion [33]. An overview of the fusion architectures is shown in Fig. 11.3.

For *Score Level Fusion*, each classifier should provide scores, representing the support of the classifier for each class. The scores of each classifier are then combined by simple arithmetic combination rules such as sum, product, etc. The scores to be combined should be in the same range, so a score normalization should be performed prior to fusion.

As the actual feature values are positively correlated with dominance, they can be used as the scores of the rule-based classifier, defined on that nonverbal feature, following a normalization step. z-normalization can be used to normalize the features for each meeting:

$$\hat{f}_p^i = \left(f_p^i - \mu_{fi}\right)/(\sigma_{fi}), \quad \forall p \in 1, \ldots, P, \tag{11.3}$$

where $\hat{f}_p^i$ and f_p^i are the values of the feature f for participant p in meeting i, z-normalized and prior to normalization, respectively. μ_{fi} and σ_{fi} are the mean and the standard deviation over all participants. The score level fusion can then be performed by using an arithmetic combination rule. For meeting i, this would mean

using feature combination $\mathcal{C}$, combining the scores for each participant (e.g. with sum rule) and selecting the participant with the highest total score:

$$S_i^{\mathcal{C}} = \arg\max_p \left(\sum_{f \in \mathcal{C}} \hat{f}_p^i \right), \quad \mathcal{C} \subseteq \mathcal{F}, \tag{11.4}$$

where $\mathcal{F}$ is the set of all features.

Rank Level Fusion is a direct extension of the rule-based estimator. Instead of selecting the participant with the maximum feature value, the participants are ranked and the rank information is used to fuse different estimators based on different features. The ranks for each participant are summed up and the one with the highest total rank is selected as the most dominant. For meeting i, using feature combination $\mathcal{C}$, the most dominant participant is selected by

$$R_i^{\mathcal{C}} = \arg\max_p \left(\sum_{f \in \mathcal{C}} r_{f_p}^i \right), \quad \mathcal{C} \subseteq \mathcal{F}, \tag{11.5}$$

where $r_{f_p}^i$ is the rank of participant p using feature f in meeting i. In case of ties, the selection can be performed based on the z-normalized scores.

11.3.3.2 Unsupervised Approaches

Unsupervised approaches can be applied to analyze social behavior data to discover and to differentiate patterns of certain behavior types. The motivation behind using unsupervised approaches is that they decrease the dependency to labeled training data. Given the difficulty of collecting data annotations for social interactions, this is a huge opportunity. Moreover, for problems with none or vague class descriptions, unsupervised approaches provide better models. Although there is always a trade-off between the performance and the amount of labeled training data, efficient unsupervised (or semi-supervised) techniques can be developed that would result in low performance degradation by using none or a very small amount of training data, when compared to using huge amounts of data.

Among the many diverse methods, we present here topic models, in particular Latent Dirichlet Allocation (LDA) [8], as an example model to discover social patterns. Topic models are probabilistic generative models that are proposed to analyze the content of documents. Although topic models were originally used in text modeling, they are capable of modeling any collection of discrete data. The patterns are discovered based on word co-occurrence. In topic models, each document is viewed as a mixture of topics, where topics are distributions over words. A word is defined as a basic unit of the discrete data. The probability of a word w in a document, assuming the document is generated from a convex combination of T topics, is given as

$$p(w_i) = \sum_{t=1}^{T} p(w_i | z_i = t) p(z_i = t), \tag{11.6}$$

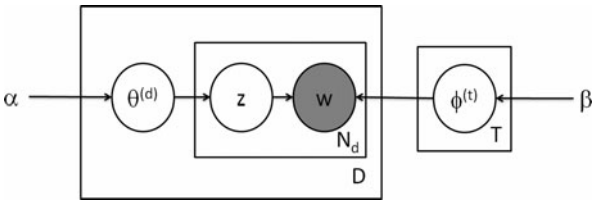

Fig. 11.4 Latent Dirichlet allocation model.

where z_i is a latent variable indicating the topics from which the word w_i may be drawn.

Assume that a document d is a bag of N_d words, and a corpus is a collection of D documents, with a total of N words (i.e. $N = \sum N_d$), and a vocabulary size of V. Let $\phi^{(t)} = p(w|z = t)$ refer to the multinomial word distribution for each topic t, and $\theta^{(d)} = p(z)$ refer to the topic distribution for each document. ϕ indicates which words are important for each topic, and θ indicates which topics are important for each document.

LDA [8] model assumes a Dirichlet prior both on the topic and word distributions ($p(\theta)$ and $p(\phi)$ are Dirichlet with hyperparameters α and β, respectively) to provide a complete generative model for documents. As the Dirichlet distribution is a conjugate prior to the multinomial, its usage simplifies the statistical inference problem and allows variational inference methods to be used. Then, the joint distribution of the set of all words in a given document is given by

$$p(z, w, \theta, \phi|\alpha, \beta) = \prod_{i=1}^{N} p(w_i|z_i, \phi)p(z_i|\theta)p(\theta|\alpha)p(\phi|\beta), \tag{11.7}$$

where z_i is the topic assignment of the ith word w_i.

The graphical model for LDA is shown in Fig. 11.4. The shaded variables indicate the observed variables, whereas the unshaded ones indicate the unobserved/latent variables. In the case of LDA, words are the only observed variables.

The objective of LDA inference is to estimate the word distribution for each topic $\phi^{(t)} = p(w|z = t)$, and the topic distribution for each document $\theta^{(d)} = p(z)$, given a training corpus and the parameters α, β, and T. The posterior distribution over z for a given document can be calculated by marginalizing over θ and ϕ, using Gibbs sampling. More details of Gibbs sampling for LDA inference can be found in [47].

As an example of the application of topic models to social verticality problems, we present a case study in Sect. 11.4.4 [25]. In this work, analogous to the bag-of-words approach in a text collection, bag-of-nonverbal patterns are defined to represent the group nonverbal behavior, for modeling group conversational patterns. In this context, the documents are the meetings, the topics are the conversational patterns, and the words are low-level nonverbal features, calculated from thin slices of small group meetings.

11.3.3.3 Supervised Approaches

Supervised approaches, including support vector machines, boosting methods, and naive Bayes, are frequently used in tasks like role recognition [6, 15, 19, 50] and dominance estimation [27, 42]. The details of these models are not given in this chapter, as they are well known models. Interested readers may refer to above mentioned references. In this section, we focus on two issues of using supervised models for social verticality problems. The first is on how to formulate the given problem as a supervised learning task, and the second is on how to obtain reliable labels for using during training from noisy and subjective annotations.

Depending on the task, the supervised learning problem can be formulated as a regression problem (e.g. if the leadership score of a participant is in question), as a binary classification problem (e.g. whether the person is dominant or not), or as a multiclass classification problem (e.g. assigning a role to each participant, among multiple role definitions). From the supervised learning point of view, one interesting problem is the estimation of the most dominant (or similarly the least dominant) person in a meeting. The trained supervised model needs to select exactly one participant from among the all participants in the meeting. In [27], the authors employed a binary classification approach to discriminate between the 'most' dominant participants and the rest, in each meeting. They trained a two-class SVM, and for each test participant in a meeting, the SVM scores were calculated with respect to the distance to the class boundary. With this formulation, the participant that has the highest score receives the 'most dominant' label, generating exactly one most dominant person per meeting. An alternative approach would be to define the problem as a regression problem and assign a dominance score to each participant. Then, the participant receiving the highest score could be selected as the most dominant person.

One of the challenges of using supervised models in social verticality problems is the need for a labeled training dataset, as obtaining these labels is not trivial for most of the social behavior estimation tasks, for which there is no "true label". As a result, the labels need to be collected from human annotators. However, when the question at hand is the existence of a social construct, even human judgments can differ, given the fact that a single correct answer does not necessarily exist. To cope with this variability, multiple annotators are used to annotate social behavior data. A common approach is to use majority agreement of annotators as the ground-truth labels. However, majority agreement has its disadvantages. It discards data points for which the annotators do not have an agreement. Furthermore, it weighs each annotator equally, without considering their different levels of expertise. Other than using majority voting, several other approaches in diverse domains are proposed to model multiple human judgments to estimate the underlying true label. In the field of social computing, as the only example so far, Chittaranjan et al. proposed an Expectation-Maximization (EM) based approach that uses annotations, and also the annotator confidences to model the ground truth [14].

11.3.3.4 Temporal Modeling

Instead of modeling a meeting as a whole, modeling the temporal evolution of the interaction in the meeting could reveal further properties of the interaction, enabling a better analysis. Dynamic Bayesian networks, in particular Hidden Markov Models (HMM) and its variants, are the most popular temporal models used in interaction analysis (see Chap. 2).

A straightforward idea is to model each participant in a meeting with one HMM and then to compute all the combinations of interacting states between these chains. However, this approach results in a high number of states, exponential with the number of chains. An alternative approach would be to use coupled HMMs or N-chain coupled HMMs. However both approaches require large number of parameters,

As an alternative to these models, in [4, 7], the *influence model* is presented and used for analyzing the interaction in groups and various social constructs such as roles [15, 41], and dominance [7].

The influence model is proposed as a generative model for describing the connections between many Markov chains. The parametrization of the model allows the representation of the influence of each chain on the others. The advantage of the influence model with respect to HMMs or coupled HMMs is that it models interacting chains while still keeping the model tractable. Figure 11.5 shows the graphical models of coupled HMM and the influence model.

The graphical model for the influence model and for the generalized N-chain coupled HMM are identical, with one very important simplification [7]. In the influence model, the probability of being at state i at time t is approximated by the pairwise conditional probabilities instead of modeling them jointly:

$$P\left(S_t^i | S_{t-1}^1, \ldots, S_{t-1}^N\right) = \sum_j \alpha_{ij} P\left(S_t^i | S_{t-1}^j\right), \tag{11.8}$$

where α_{ij} indicates the influence of chain i on chain j. Although this pairwise modeling limits the capability of the model, it allows tractability and scalability. The details of the model and the EM algorithm for learning influence model parameters can be found in [7].

11.4 Case Studies

This section presents example studies on automatic analysis of social verticality, for four different social constructs: dominance, emergent leadership, roles, and leadership styles.

11.4.1 Dominance Estimation

In this section we report a study that explores ways to combine audio and visual information to estimate the most and least dominant person in small group interactions. More details of this work can be found in [2].

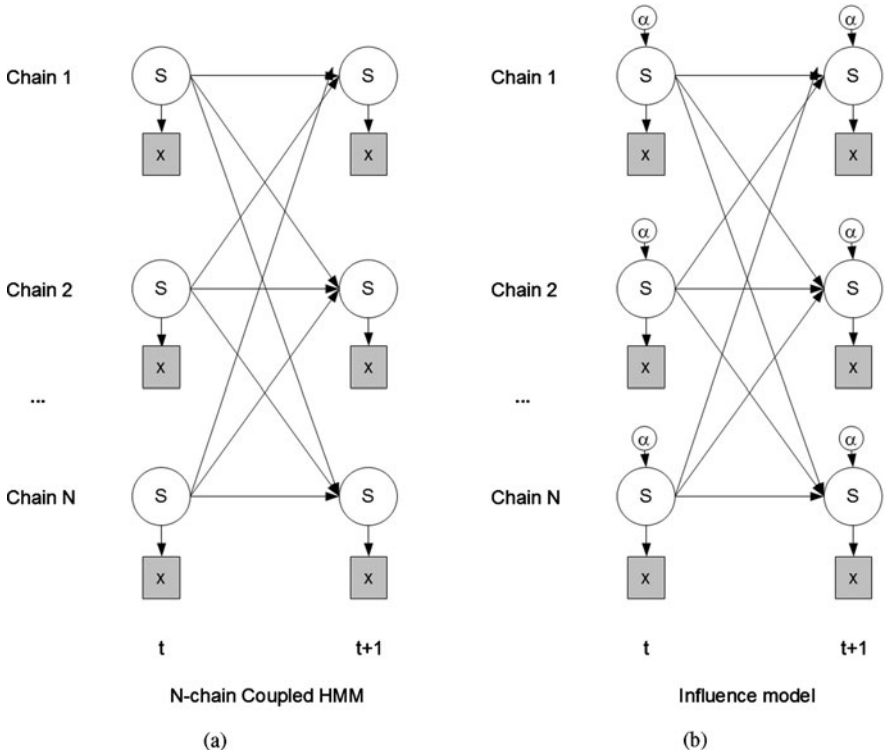

Fig. 11.5 Graphical models for (**a**) coupled HMM, and (**b**) influence model with hidden states

11.4.1.1 Task Definition and Data

Given a meeting, a small group conversation, this study focuses on finding the Most Dominant (MD) and Least Dominant (LD) participants in the group.

The data used in this study are publicly available as the DOME corpus [3]. The DOME corpus contains 125 five-minute meeting segments selected from the Augmented Multi-party Interaction (AMI) corpus [11]. Each meeting has four participants, and is recorded with multiple cameras and microphones. The total length of the DOME corpus corresponds to more than 10 hours of recordings. Each meeting segment in the DOME corpus is annotated by three annotators. The annotators ranked the participants according to their level of perceived dominance. Then the agreement (full and majority agreement on most and least dominant person) between the annotators for each meeting is assessed. Following this procedure, two annotated meeting datasets for each task are obtained (see Table 11.3).

Table 11.3 Number of meetings with full and majority agreement in DOME corpus

	Full	Maj		Full	Maj
Most dominant	67	121	Least dominant	71	117

11.4.1.2 Features and Model

Social psychology research states that dominance is displayed via audio nonverbal cues such as the speaking time, number of turns and interruptions, pitch, as well as visual cues such as visual activity, expressions and gaze [22, 31]. Based on these studies, several audio and visual features can be extracted as descriptors of some of the above cues.

For audio nonverbal features, speaking turn features such as speaking time, number of turns and interruptions are considered. The audio recordings from the close-talk microphones are processed for each participant and their speech activity, in the form of binary speaking status, is extracted. The following speaking turn features are used in this study: Total Speaking Length (TSL), Total Speaking Turns (TST), TST without Short Utterances (TSTwoSU), Total Successful Interruptions (TSI), and Average Speaker Turn Duration (AST).

Visual activity based nonverbal features are extracted from the close-up camera that captures the face and the upper body of each participant. The amount of motion in the skin-colored regions are calculated using compressed domain processing (see Sect. 11.3.2.2), in the form of binary visual activity information for each participant. Visual activity (-V-) equivalents of the above given audio features are extracted as visual nonverbal features.

Furthermore, in addition to the above audio-only and video-only features, a set of multimodal features is defined, which represent the audio-visual (-AV-) activity jointly. The visual activity of the person is measured only while speaking, and audio visual equivalents of the audio-only and video-only features are extracted.

Dominance estimation is performed by a rule-based estimator. The fusion of audio and visual nonverbal features is done via rank and score level multimodal fusion, using the rule-based estimator. The details of these techniques are presented in Sect. 11.3.3.1.

11.4.1.3 Experiments and Results

The experiments are performed on Full and Maj datasets for MD and LD tasks on the DOME corpus (see Table 11.3). The accuracy is calculated as follows: it is assumed that the estimation is correct with weight one, if it matches the agreement. If there is a tie, and one of the tied results is correct, a weight is assigned, which is the reciprocal of the number of ties.

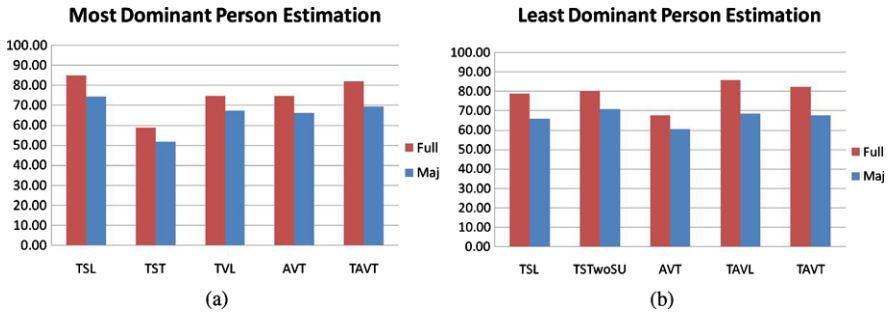

Fig. 11.6 Single feature accuracy for selected audio and visual nonverbal features for (**a**) MD task and (**b**) LD task

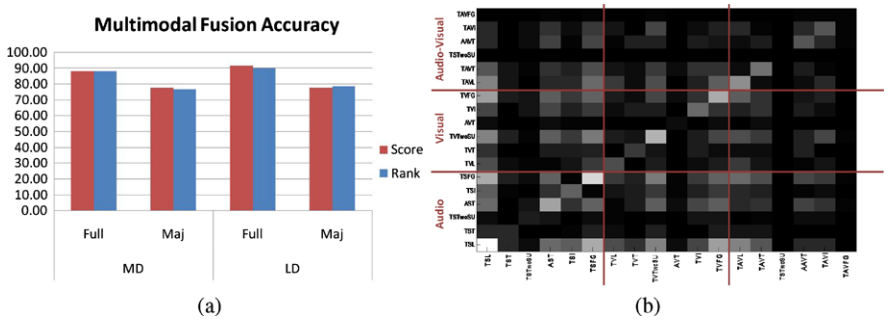

Fig. 11.7 (**a**) Multimodal fusion accuracy for MD and LD tasks. (**b**) Pairwise feature frequencies in best combinations for MD task. *Lighter colors* indicate higher frequency

The classification accuracies for selected single nonverbal features are shown in Fig. 11.6. For the MD task, the best results are obtained with TSL (85.07% and 74.38%) and for the LD task, with Total Audio-Visual Length (TAVL) (85.92%) and TSTwoSU (70.94%), on Full and Maj datasets, respectively.

To find the best combination of nonverbal features in multimodal fusion, an exhaustive search is performed: all feature combinations are evaluated and the best one that combines fewest number of features is reported. The classification accuracies for the best combinations are shown in Fig. 11.7(a). The results show that one can achieve ~3% increase on MD task and ~7% on LD task using rank or score level fusion. It is important to note that there is more than one combination that gives the highest result. Figure 11.7(b) shows the pairwise feature frequencies in best combinations for MD task.

As dominance is displayed multimodally, via audio and visual features, automatic methods should utilize multimodal fusion techniques for estimation of dominance. The results above show that the visual information is complementary to audio, and multimodal fusion is needed to achieve better performance.

11.4.2 Identifying Emergent Leaders

The second study focuses on automatically identifying the emergent leader in small groups. More details of this work can be found in [44].

11.4.2.1 Task Definition and Data

Two main questions are asked in the context of predicting emergent leadership in small groups based on automatic sensing. The first question deals with the existence of a correlation between how the emergent leader is perceived and her/his nonverbal behavior. The second question is whether one can predict emergent leadership using automatically extracted acoustic nonverbal features.

To study emergence of leadership, an audio-visual corpus is collected: The Emergent LEAder data corpus (ELEA) includes audio-visual recordings of groups performing a 'winter survival task' [29], and also questionnaires filled by each group member before and after the interaction. The winter survival task focuses on ranking a list of items in order to survive an airplane crash in winter [29]. The groups are composed of previously unacquainted people. The questionnaires ask participants about themselves and also about the other group members, to evaluate their leadership skills and personality. Several variables are computed from the questionnaires, indicating the perceived leadership, perceived dominance, dominance rank, and perceived competence.

11.4.2.2 Features and Model

The audio recordings of ELEA corpus are collected with a microphone array, which creates automatic speaker segmentations along with the audio recording. This results in a binary segmentation for each participant, indicating the binary speaking status. From this binary segmentation, speaking turn audio features are extracted as audio nonverbal features. The features include the speaking length (TSL), turns (TST and TSTf), average turn duration (TSTD) and interruptions (TSI and TSIf). For turns and interruptions, the filtered versions (TSTf, TSIf) consider only the turns longer than two seconds. Two different definitions of interruptions are used (TSI^1 and TSI^2) (see Sect. 11.3.2.1 and Fig. 11.2) for both the filtered and non-filtered versions. The rule-based estimator, presented in Sect. 11.3.3.1, is used for automatic identification of the emergent leader. The variables from questionnaires were used as a ground truth for evaluation purposes.

11.4.2.3 Experiments and Results

Correlation Between the Questionnaires and the Nonverbal Features Table 11.4 shows Pearson correlation values between questionnaire outputs and nonverbal features. There is a correlation between several nonverbal features and perceived leadership, suggesting that emergent leadership perception has a connection

Table 11.4 Correlation values between variables from questionnaires and nonverbal acoustic features

	Perc. leadership	Perc. dominance	Ranked dominance	Perc. competence
TSL	0.51	0.46	0.49	0.28
TSTD	0.44	0.39	0.40	0.19
TSTf	0.60	0.60	0.53	0.27
TSIf	0.62	0.60	0.54	0.26

Table 11.5 Accuracy (%) of individual features on predicting emergent leadership

	TSL	TSTD	TST	TSTf	TSI^1	$TSIf^1$	TSI^2	$TSIf^2$
Plead	60	70	35	65	50	65	55	70

to the person who talks the most, has more turns, and interrupts the most. Furthermore, several nonverbal features also have correlation with perceived or ranked dominance. The correlations with perceived competence is relatively low.

Automatic Inference Table 11.5 shows the accuracy using single features, where the best accuracy for perceived leadership is achieved using $TSIf^2$ and TSTD with 70%, followed by TSTf and $TSIf^1$ with 65%. The accuracy is calculated as in the previous case study: it is assumed that the estimation is correct with weight one, if it matches the agreement. If there is a tie, and one of the tied results is correct, a weight is assigned, which is the reciprocal of the number of ties.

Score level fusion (see Sect. 11.3.3.1) is applied to combine different acoustic nonverbal features. For the estimation of perceived leadership, a 10% increase in the accuracy is observed, achieving an accuracy of 80%, via the combination of TSTD and TSI features.

This study, summarized from [44], is a first attempt to automatically identify the emergent leader in small groups. Although the collected corpus is currently quite limited, several observations can be made. First there are correlations between the perceived leadership and automatically extracted acoustic nonverbal features. The emergent leader was perceived by his/her peers as a dominant person, who talks the most, and has more turns and interruptions. An accuracy up to 80% is obtained to identify the emergent leader using a combination of nonverbal features.

11.4.3 Recognizing Functional Roles

The third case study attempts to recognize functional roles in meetings using the influence model. More details can be found in [15, 40, 50].

11.4.3.1 Task Definition and Data

Given a small group meeting, the task in this study is to identify the role of each participant. The roles are defined based on the Functional Role Coding Scheme (FRCS), in two complementary areas. The *task area* includes the roles related to the tasks and the expertise of the meeting participants. These include 'orienter', 'giver', 'seeker', and 'follower' roles. The *socio-emotional area* roles, i.e. 'attacker', 'protagonist', 'supporter', and 'neutral', are related to the relationships of the group members.

The Mission Survival Corpus is used as the meeting corpus [40], which includes eight four-people meetings, recorded with microphones and cameras. The annotations for the roles are done by one annotator, by considering the participant's behavior every five seconds. As a result, instead of assigning one role for each participant for the entire meeting, a thin slice based approach is used. This coding scheme assumes that the participants can have different roles throughout the meeting.

11.4.3.2 Features and Model

As features, the authors use automatically extracted speech and visual activity features. The speech recorded from close-talk microphones is automatically segmented for each participant and speaking/non-speaking status is used as speech activity features. The number of simultaneous speakers is also used as a feature. As visual activity features, the amount of fidgeting (i.e. the amount of energy) for hands and body is used [13].

The influence model is proposed as a suitable approach to model the group interaction (see Sect. 11.3.3.4), as it can model complex and highly structured interacting processes. To model a meeting with the influence model, two processes per participant are used: one for the task roles, and another for the socio-emotional roles. The latent states of the models are the role classes.

11.4.3.3 Experiments and Results

The performance of the influence model is compared with two other models: SVM and HMM. For each of the models, the training is performed with half of the available meeting data, using two fold cross-validation. The feature vector for each participant is composed of all extracted audio-visual features. For SVM, the feature vectors of each participant is concatenated and a single feature vector is composed.

The role recognition accuracies for each model is presented in Table 11.6, as reported in [15]. The SVM suffers from the curse of dimensionality and overfitting. The influence model achieves the highest accuracy, as it handles the curse of dimensionality by modeling each participant with different processes. Although the HMM handles the curse of dimensionality using the same approach, as there is no interaction between the processes, the recognition accuracy is lower. Another advantage of using the influence model is its flexibility: it is adaptable to different-sized groups.

Table 11.6 Role recognition accuracies (%) of the Influence model, HMM, and SVM

	Task roles	Social roles	Overall
Influence model	75	75	75
HMM	60	70	65
SVM	–	–	70

11.4.4 Discovering Leadership Styles in Group Conversations

As the last example, we present a study that differs from the previously presented ones, in the sense that it aims to model the group as a whole, instead of modeling individuals. More details of this study can be found in [25].

11.4.4.1 Task Definition and Data

The addressed problem in this study is to automatically discover group conversational patterns from nonverbal features, extracted from brief observations of interaction. Specifically, following the definition in [35], the group conversations can be grouped in three categories: autocratic groups, in which the decisions are determined by the leader; participative groups, in which the leader encourages group discussion and decision making; and free-rein groups, in which the group has complete freedom to decide without leader participation. The study uses a subset of the AMI corpus [11], corresponding to 17 hours of meetings. Part of this subset is annotated by human annotators and used for assessment of the group conversation type.

11.4.4.2 Features and Model

In this study, a novel descriptor of interaction slices—a bag of group nonverbal patterns is described, which captures the behavior of the group as a whole, and its leader's position in the group. The discovery of group interaction patterns is done in an unsupervised way, using principled probabilistic topic modeling.

Analogous to how topics are inferred from a text collection, by representing documents in a corpus as histograms of words, group dynamics can be characterized via bag-of group nonverbal patterns (bag-of-NVPs). The bag-of-NVPs are produced from low-level speaking turn audio nonverbal features, calculated from thin slices of small group meetings. The low-level features include individual and group speaking features such as speaking length, turns, interruptions, backchannels, overlapped/non-overlapped speech, and silence. These low-level features are quantized to generate the bag-of-NVPs. There are two types of bag-of-NVPs: generic group patterns and leadership patterns. Generic group patterns describe the group as a whole without using the identity information. The leadership patterns describe the leader in the group. A diagram showing the features is given in Fig. 11.8(a). Once the bag-of-NVPs are produced, the mining of group patterns is done using

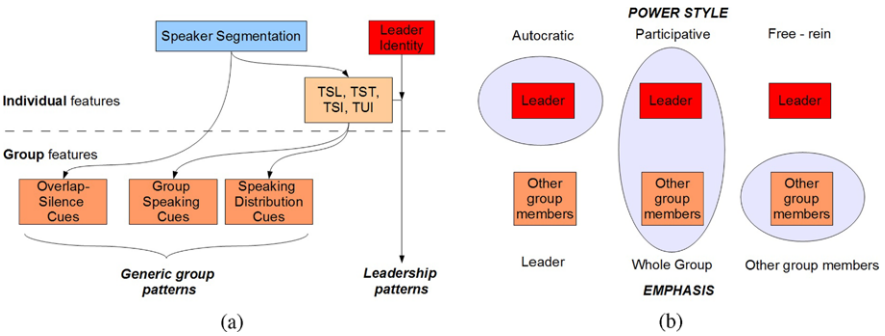

Fig. 11.8 (**a**) Extracted features to characterize individual and group behavior. (**b**) Leadership styles by Lewin et al. [35]

latent Dirichlet allocation topic model (see Sect. 11.3.3.2), to discover topics by considering the co-occurrence of word patterns.

11.4.4.3 Experiments and Results

The experiments are performed on a subset of meetings selected from the AMI corpus [11]. Effect of different time scales and different combinations of bag features are analyzed. The evaluation is done via comparison with human annotations.

On different time scales, the authors observed that the group interactions look more like a monologue at finer time scales (e.g., 1 minute) and like a discussion at coarser time scales (e.g., 5 minute). Also, successful interruptions are not very common at fine time scales.

The LDA based discovery approach is applied to discover three topics. The results on 5-minute scale show that the three discovered topics resemble three classic leadership styles of Lewin et al. [35], as illustrated in Fig. 11.8(b). In comparison to the human annotators, the accuracy of the model for autocratic, participative and free-rein classes are 62.5%, 100%, and 75%, respectively. This suggests that the discovered topics are indeed meaningful. The LDA experiments are repeated for a 2-minute scale as well. The distribution of the topics found with the 2-minute scale is more balanced than the topics at 5-minute scale, indicating that at longer time intervals, the interaction styles are captured more strongly.

11.5 Summary

This chapter focuses on the computational analysis of social verticality. Social verticality refers to the vertical dimension of social interactions, in which the participants of the group position themselves in a hierarchical-like structure. We presented a brief summary of main nonverbal features that humans display and perceive during social interactions that represent social verticality constructs such as dominance,

power, status, and leadership. As the main sources of these nonverbal features are audio and video, we described processing and feature extraction techniques for these modalities. Different inference approaches, such as rule-based, unsupervised, supervised, and temporal are also discussed with examples from the literature. We also present a non-exhaustive survey on the computational approaches for modeling social verticality. In the last section of this chapter we presented four case studies on dominance estimation, identifying emergent leadership, role recognition, and discovering leadership styles in group conversations as examples to the techniques discussed in the chapter.

The future dimensions of this field lie in all the functional blocks that are presented in the chapter, with the inference block being the core challenge. Developments on new sensor technologies will result in better capture of social behavior of humans. On top of this, the current research on tracking human movements should be further extended to cover human behavior in natural settings. Features that better represent nonverbal social behavior should be investigated in close contact with social psychology research. The inference models lie at the core of social behavior analysis. Flexible models that can handle dynamic groups with varying numbers of participants are needed, applicable to different settings to estimate and model social constructs that relate to individuals, as well as to their group behavior.

11.6 Questions

1. What is the difference between verbal and nonverbal communication?
2. What are the differences between power, status, and dominance?
3. What kind of audio nonverbal features can be extracted for social verticality analysis?
4. What kind of visual nonverbal features can be extracted for social verticality analysis?
5. What are the techniques that can be used to fuse different modalities for social verticality analysis?
6. What is thin slice based modeling?
7. Discuss what kind of models can be used in a meeting scenario or how the standard models can be modified when the number of participants in a group vary, i.e. the dataset contains meetings with different number of participants.
8. Explain the differences between the influence model and coupled HMM.

11.7 Glossary

- *Emergent leadership*: The leader who arises from an interacting group and has a base of power arising from followers rather than from a higher authority.
- *Influence model*: A representation to model the dynamics between interacting processes.

- *Thin slice*: The smallest segment from which, when exposed, humans can correctly predict behavioral outcomes with high accuracy by interpreting nonverbal cues.
- *Topic models*: A statistical model for discovering the hidden topics that occur in a collection of documents.
- *Visual dominance ratio*: Looking-while-speaking to looking-while-listening ratio.

Acknowledgements This work is supported by the EU FP7 Marie Curie Intra-European Fellowship project "Automatic Analysis of Group Conversations via Visual Cues in nonverbal Communication" (NOVICOM), and by the Swiss National Science Foundation under the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (IM2) and by the Sinergia project on "Sensing and Analysing Organizational Nonverbal Behaviour" (SONVB). The authors would like to thank Dinesh Jayagopi, Dairazalia Sanchez-Cortes, and Gokul Chittaranjan for their contributions to several studies presented in this chapter.

References

1. Ambady, N., Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. Psychol. Bull. **111**, 256–274 (1992)
2. Aran, O., Gatica-Perez, D.: Fusing audio-visual nonverbal cues to detect dominant people in small group conversations. In: 20th International Conference on Pattern Recognition (ICPR), Istanbul, Turkey (2010)
3. Aran, O., Hung, H., Gatica-Perez, D.: A multimodal corpus for studying dominance in small group conversations. In: LREC Workshop on Multimodal Corpora, Malta (LREC MMC'10) (2010)
4. Asavathiratham, C., Roy, S., Lesieutre, B., Verghese, G.: The influence model. IEEE Control Syst. Mag. **21**, 52–64 (2001)
5. Ba, S.O., Odobez, J.-M.: Multiperson visual focus of attention from head pose and meeting contextual cues. IEEE Trans. Pattern Anal. Mach. Intell. **33**, 101–116 (2011)
6. Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S.: The rules behind roles: Identifying speaker role in radio broadcasts. In: 17th National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence, pp. 679–684. AAAI, Washington (2000)
7. Basu, S., Choudhury, T., Clarkson, B., Pentland, A.: Learning human interactions with the influence model. Technical report, MIT Media Lab, Cambridge, MA (2001)
8. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
9. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**, 257–267 (2001)
10. Campbell, N., Douxchamps, D.: Processing image and audio information for recognising discourse participation status through features of face and voice. In: INTERSPEECH 2007, pp. 730–733 (2007)
11. Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., Wellner, P.: The AMI meeting corpus: A pre-announcement. In: Workshop Mach. Learn. for Multimodal Interaction (MLMI'05), Edinburgh, UK, pp. 28–39 (2005)
12. Charfuelan, M., Schröder, M., Steiner, I.: Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings. In: Interspeech 2010, Makuhari, Japan, Sept. (2010)

13. Chippendale, P.: Towards automatic body language annotation. In: 7th International Conference on Automatic Face and Gesture Recognition (FG '06), Washington, DC (2006)
14. Chittaranjan, G., Aran, O., Gatica-Perez, D.: Exploiting observers' judgments for multimodal nonverbal group interaction analysis. In: 9th IEEE Conference on Automatic Face and Gesture Recognition, Santa Barbara, CA (2011)
15. Dong, W., Lepri, B., Cappelletti, A., Pentland, A.S., Pianesi, F., Zancanaro, M.: Using the influence model to recognize functional roles in meetings. In: 9th International Conference on Multimodal Interfaces (ICMI'07), pp. 271–278 (2007)
16. Dovidio, J.F., Ellyson, S.L.: Decoding visual dominance: Attributions of power based on relative percentages of looking while speaking and looking while listening. Soc. Psychol. Q. **45**(2), 106–113 (1982)
17. Dunbar, N.E., Burgoon, J.K.: Perceptions of power and interactional dominance in interpersonal relationships. J. Soc. Pers. Relatsh. **22**(2), 207–233 (2005)
18. Funder, D.C.: Personality. Annu. Rev. Psychol. **52**, 197–221 (2001)
19. Garg, N.P., Favre, S., Salamin, H., Hakkani Tür, D., Vinciarelli, A.: Role recognition for meeting participants: an approach based on lexical information and social network analysis. In: 16th ACM International Conference on Multimedia (MM'08), pp. 693–696 (2008)
20. Gatica-Perez, D.: Automatic nonverbal analysis of social interaction in small groups: a review. Image Vis. Comput. **27**(12), 1775–1787 (2009)
21. Gorga, S., Otsuka, K.: Conversation scene analysis based on dynamic bayesian network and image-based gaze detection. In: ICMI-MLMI 2010 (2010)
22. Hall, J.A., Coats, E.J., Smith LeBeau, L.: Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. Psychol. Bull. **1313**(6), 898–924 (2005)
23. Hung, H., Jayagopi, D.B., Ba, S., Gatica-Perez, D., Odobez, J.-M.: Investigating automatic dominance estimation in groups from visual attention and speaking activity. In: Int. Conf. on Multimodal Interfaces (ICMI), Chania, Greece (2008)
24. Hung, H., Huang, Y., Friedland, G., Gatica-Perez, D.: Estimating dominance in multi-party meetings using speaker diarization. IEEE Trans. Audio Speech Lang. Process. **19**(4), 847–860 (2011)
25. Jayagopi, D.B., Gatica-Perez, D.: Mining group nonverbal conversational patterns using probabilistic topic models. IEEE Trans. Multimed. (2010)
26. Jayagopi, D.B., Ba, S., Odobez, J.-M., Gatica-Perez, D.: Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In: Int. Conf. on Multimodal Interfaces (ICMI), Special Session on Social Signal Processing, Chania, Greece (2008)
27. Jayagopi, D.B., Hung, H., Yeo, C., Gatica-Perez, D.: Modeling dominance in group conversations from nonverbal activity cues. IEEE Trans. Audio Speech Lang. Process. **17**(3), 501–513 (2009). Special Issue on Multimodal Processing for Speech-based Interactions
28. Kalma, A.K., Visser, L., Peeters, A.: Sociable and aggressive dominance: Personality differences in leadership style? Leadersh. Q. **4**(1), 45–64 (1993)
29. Kickul, J., Neuman, G.: Emergent leadership behaviours: The function of personality and cognitive ability in determining teamwork performance and KSAs. J. Bus. Psychol. **15**(1) (2000)
30. Kim, T., Chang, A., Pentland, A.: Meeting mediator: Enhancing group collaboration with sociometric feedback. In: ACM Conference on Computer Supported Collaborative Work, San Diego, CA, pp. 457–466 (2008)
31. Knapp, M.L., Hall, J.A.: Nonverbal Communication in Human Interaction, 7th edn. Wadsworth, Belmont (2009)
32. Kumano, S., Otsuka, K., Mikami, D., Yamato, J.: Recognizing communicative facial expressions for discovering interpersonal emotions in group meetings. In: 11th International Conference on Multimodal interfaces (ICMI'09), ICMI-MLMI '09, pp. 99–106 (2009)
33. Kuncheva, L.I.: Combining Pattern Classifiers: Methods and Algorithms. Wiley, New York (2004)
34. Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. IEEE Commun. Mag. **48**, 140–150 (2010)

35. Lewin, K., Lippit, R., White, R.K.: Patterns of aggressive behavior in experimentally created social climates. J. Soc. Psychol. **10**, 271–301 (1939)
36. Otsuka, K., Yamato, J., Takemae, Y., Murase, H.: Conversation scene analysis with dynamic Bayesian network based on visual head tracking. In: ICME 2006 (2006)
37. Otsuka, K., Sawada, H., Yamato, J.: Automatic inference of cross-modal nonverbal interactions in multiparty conversations. In: ACM 9th Int. Conf. Multimodal Interfaces (ICMI2007), pp. 255–262 (2007)
38. Pentland, A.: Socially aware computation and communication. Computer **38**(3), 33–40 (2005)
39. Pentland, A.: Honest Signals: How They Shape Our World. MIT Press, Cambridge (2008)
40. Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A.: A multimodal annotated corpus of consensus decision making meetings. Lang. Resour. Eval. **41**, 409–429 (2007)
41. Raducanu, B., Gatica-Perez, D.: Inferring competitive role patterns in reality TV show through nonverbal analysis. Multimed. Tools Appl. 1–20 (2010)
42. Rienks, R.J., Heylen, D.: Automatic dominance detection in meetings using easily detectable features. In: Workshop Mach. Learn. for Multimodal Interaction (MLMI'05), Edinburgh, UK (2005)
43. Salamin, H., Favre, S., Vinciarelli, A.: Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. IEEE Trans. Multimed. **11**(7), 1373–1380 (2009)
44. Sanchez-Cortes, D., Aran, O., Schmid-Mast, M., Gatica-Perez, D.: Identifying emergent leadership in small groups using nonverbal communicative cues. In: 12th International Conference on Multimodal Interfaces (ICMI'10), Beijing, China (2010)
45. Schmid-Mast, M.: Dominance as expressed and inferred through speaking time: A meta-analysis. Hum. Commun. Res. **28**(3), 420–450 (2002)
46. Stein, R.T.: Identifying emergent leaders from verbal and nonverbal communications. J. Pers. Soc. Psychol. **32**(1), 125–135 (1975)
47. Steyvers, M., Griffiths, T.: Probabilistic Topic Models. Erlbaum, Hillsdale (2007)
48. Varni, G., Volpe, G., Camurri, A.: A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. IEEE Trans. Multimed. **12**(6), 576–590 (2010)
49. Yeo, C., Ahammad, P., Ramchandran, K., Sastry, S.S.: High-speed action recognition and localization in compressed domain videos. IEEE Trans. Circuits Syst. Video Technol. **18**(8), 1006–1015 (2008)
50. Zancanaro, M., Lepri, B., Pianesi, F.: Automatic detection of group functional roles in face to face interactions. In: 8th International Conference on Multimodal Interfaces (ICMI'06), pp. 28–34 (2006)