AUTOMATIC ROLE RECOGNITION

HUGUES ERIC SALAMIN

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

SCHOOL OF COMPUTING SCIENCE

COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

APRIL 2013

Abstract

The computing community is making significant efforts towards the development of automatic approaches for the analysis of social interactions. The way people interact depends on the context, but there is one aspect that all social interactions seem to have in common: humans behave according to roles. Therefore, recognizing the roles of participants is an essential step towards understanding social interactions and the construction of socially aware computer.

This thesis addresses the problem of automatically recognizing roles of participants in multiparty recordings. The objective is to assign to each participant a role. All the proposed approaches use a similar strategy. They all start by segmenting the audio into turns. Those turns are used as basic analysis units. The next step is to extract features accounting for the organization of turns. The more sophisticated approaches extend the features extracted with features from either the prosody or the semantic. Finally, the mapping of people or turns to roles is done using statistical models. The goal of this thesis is to gain a better understanding of role recognition and we will investigate three aspects that can influence the performance of the system:

- We investigate the impact of modelling the dependency between the roles.
- We investigate the contribution of different modalities for the effectiveness of role recognition approach.
- We investigate the effectiveness of the approach for different scenarios.

Three models are proposed and tested on three different corpora totalizing more than 90 hours of audio. The first contribution of this thesis is to investigate the combination of turn-taking features and semantic information for role recognition, improving the accuracy of

role recognition from a baseline of 46.4% to 67.9% on the AMI meeting corpus. The second contribution is to use features extracted from the *prosody* to assign roles. The performance of this model is 89.7% on broadcast news and 87.0% on talk-shows. Finally, the third contribution is the development of a model robust to change in the social setting. This model achieved an accuracy of 86.7% on a database composed of a mixture of broadcast news and talk-shows.

Table of Contents

List of Tables

List of Figures

Acknowledgements

The work in this thesis would not have been possible without the help and encouragement of several people. First and foremost, I am especially grateful to my supervisor Alessandro Vinciarelli. His input and insight has been invaluable in my research. I would like to thank him for helping me in this journey (both in the figurative and literal sense).

I would also like to thank Louis-Philippe Morency and everybody I had the chance to meet during my internship at the Institute for Creative Technology.

The first years of my PhD would have been much more difficult without the support and encouragement of my friends and colleagues Sarah Favre and Hari Parthasarathi. They have always been there, either for a chat or for advice on my research.

Une telle entreprise n'aurais pas été possible sans le soutient de ma famille et en particulier de mes parents, Anne-Marie et Jean-Michel, qui m'ont toujours encouragé, donné de bon conseils et m'ont fait profiter de leur experience. Un grand merci aussi à mon frère Thomas pour les nombreuses discussions où sa franchise légendaire m'a éclairé. Finalement, merci à ma sœur Camille et sa famille pour tous les bons moments passés ensemble qui ont étés une source d'énergie et de joie.

Je veux aussi remercier les nombreuses personnes qui m'ont aidé durant cette période (souvent sans qu'elles ne s'en rendent compte), que ce soit par des discussions, des encouragements ou des bon moments partagés. En particulier, un grand merci à Cynthia, Fabian, Brice, Michel et Bastien.

Finally, my time in Glasgow would not have been the same without my friends and colleagues Donny and Craig. I'm in their debts for many discussions and idea (not all of them sound) as well as hours of fine entertainment.

Author's declaration

I hereby declare that all of the work presented in this thesis was performed personally unless otherwise stated. No part of this work has been submitted for consideration as part of any other degree or award.

Chapter 1

Introduction

1.1 Introduction

Human beings are social animals [1]. Most of our activities, both during work and leisure, revolve around interactions with other human beings. As humans, we are always communicating and sending signals. Even during our sleep, we communicate through our position and movements [2]. Given this constant immersion, most humans beings are very effective at perceiving and interpreting the signals sent by other humans.

As the computing community moves toward a human-centred approach to computing [3] where computers interact with human using natural human interaction techniques (such as speech synthesis or virtual agents) and where automated systems support human-human interaction, the interest in socially-aware systems is increasing. Therefore, the computing community is making significant efforts to develop automatic approaches for the analysis of social interactions (see [4, 5] for extensive surveys of the domain).

However, the way people interact socially is extremely complex. Social psychologists have been studying human interaction for almost a century and new findings are still being discovered and explored. Therefore, in order to keep this thesis tractable, we decided to focus on a single concept, namely *roles*. The way people interact depends on the context, but there is one aspect that all social interactions seem to have in common:

> *People do not interact with one another as anonymous beings. They come together in the context of specific environments and with specific purposes. Their*

> *interactions involve behaviors associated with defined statuses and particular roles. These statuses and roles help to pattern our social interactions and provide predictability* [6].

As the above suggests that roles as a universal key to understand social interactions and because interactions are commonly captured in multimedia data, this thesis revolves around approaches for the automatic recognition of roles in multi-party recordings.

1.2 Motivation

Our aim, in this thesis, is to develop an automatic approach for the recognition of roles. There are three main motivations for this work that are detailed below: the ubiquitous presence of social interaction, the influence of roles on people's behaviours, and the possible applications for roles recognition.

The first motivation is the ubiquitous presence of social interaction in everyday life. It is not restricted to the recognition of roles and a large community is working on the development of automatic approaches for the analysis of social interactions [4, 5]. This is not surprising as social interactions are not only one of the most important aspects of our everyday life, but also an ubiquitous theme in multimedia data: radio and television programs (debates, news, talk-shows, movies, etc.) rarely show something other than social interactions and even less common kinds of data (meeting recordings, surveillance material, call center conversations, etc.) revolve in general around interactions between individuals. It is crucial that if computers want to interact in a natural way with humans, computers need to be able to respond appropriately to human social interactions. In this respect, role recognition can help to achieve the long-term goal of bringing social intelligence into machines.

The second motivation is specific to role recognition and follows from the two main aspects of the relation between *roles* and *behaviours* [7]. The first aspects is that roles are associated with *shared* expectations that people hold about their own behaviour as well as about the behaviour of others [8]. Thus, roles contribute to the overall predictability of social interaction, a key condition for making reasonable guesses about others and participating effectively in social exchanges. The second is that roles typically result into "*characteristic behaviour*

patterns" [8] that can be identified and recognized as such by interaction participants (see the seminal work in [9, 10]). It follows from those two aspects that being able to identify roles can help a computer understand the behaviours of humans and also provide a set of appropriate behaviours for the current interaction.

The third motivation is found in the possible applications of role recognition. From an applied point of view, roles can be useful in several applications and the list given here is not exhaustive. In media browsers, the information about the role of the person speaking at a given moment can help users to quickly identify segments of interest. In summarization, the role of an individual can be used as a criterion to select more or less representative segments of the data [11][12]. In Information Retrieval, roles can be used as an index to enrich the content description of the data. Furthermore, roles can be used to segment the data into semantically coherent segments [13][14].

1.3 Thesis Statement

The main goal of this thesis is to investigate automated systems for role recognition. By automated system, we mean a system that does not need the intervention of a human expert once it has been trained. Human experts still need to be involved in the creation of the training set, in particular in the labelling. We propose to use machine learning to map features extracted from the recording of interaction to roles. As mentioned earlier, roles can be tied to behaviours and we propose to use features extracted from the non-verbal behaviours of participants.

Non-verbal communication is a phenomenon that psychologists have been studying for more than a century, and consists of the wide spectrum of non-verbal behaviours through which humans communicate what cannot be said with words, including feelings and attitudes toward others [15, 2]. Non-verbal communication can be considered as one of the physical, detectable, and measurable evidences of our inner life, the other being the content of our verbal messages. But unlike the latter, non-verbal communication is typically *honest* [16, 17] and reliable because it is mostly out of the reach of conscious control, thus it leaks information about our actual state, and not about what we want to show as such.

Non-verbal communication is composed of a multitude of non-verbal cues that can be grouped

into five codes according to the communication modality [4]: physical appearance, gestures and postures, face and eye behaviour, vocal behaviour, and space and environment behaviour. This thesis is based on the Social Signal Processing paradigm and focuses on vocal nonverbal behaviours, i.e. everything that can be observed in speech except words, as evidence to identify roles. Nonverbal vocal behaviour includes the way people speak (pitch, rhythm, energy and modulation of those aspects) and turn-taking (who speaks when and how long).

1.4 Research Objectives

Overall, the goal of the approach proposed in this thesis is to automatically recognise roles in multiparty recordings. In particular, the focus is on the use of vocal social signals (i.e., nonverbal behaviours that carry information about social interactions between people in speech) for detecting roles. The approach developed in the experiments presented in the thesis is composed of three steps:

1. Detecting participants involved in the social interaction.
2. Extracting audio behavioural cues.
3. Assigning roles to participants using the behavioural cues.

The input of my approach is an audio recording. During the first step, the participants are identified and the recordings are segmented into speaker turns. In the second step, behavioural cues such as pitch and turn-taking pattern are extracted and associated with the participants. In the last step, the cues are automatically interpreted in terms of roles.

The experiments presented in this thesis are performed over three different corpora for a total of around 90 hours of material. This allows for an easy comparison of the results in the different chapters since performance measures, data and the experimental setup are similar. The main research objective is to better understand the performance of a fully automated system for role recognition. In particular, we will investigate three aspects that can influence the performance of the system.

- We investigate the use of model for the dependency between the roles.

- We investigate the contribution of different modalities for the effectiveness of role recognition approach.
- We investigate the effectiveness of the approach for different scenarios.

1.5 Main Contributions

The main novelties and distinctive aspects of this thesis with respect to the *state-of-the-art* are, to the best of our knowledge, as follows:

- This is the first work that combines turn-taking features and semantic information for role recognition (chapter 5).
- This is the first work that uses features extracted from *prosody* to assign roles (chapter 6).
- This is the first work that can deal with roles from two different setups and trains one classifier to identify roles in the two settings (chapter 6).

1.6 Organisation of the Thesis

The rest of the thesis is organised into 6 chapters.

Chapter 2 contains a description of the state of the art of the field. This chapter briefly discusses the main work in the area of social signal processing. It also describes the groundwork in psychology that motivated the approach presented in this thesis and the methodology used in this thesis.

Chapter 3 contains a description of the statistical models used in this work, namely graphical models. It presents the mathematical background as well as the techniques used for the estimation of model parameters during the experiments.

Chapter 4 is the first chapter presenting results from experiments. The focus of the chapter is on the effect of modelling the dependence between the roles.

Chapter 5 contains a description of the experiment carried out using semantic information. The objective of the experiment was to investigate the respective importance of two channels in the voice (spoken word and intonation) for the role recognition problem.

Chapter 6 presents the results of the experiment using prosody. Furthermore, the approach presented in this chapter is independent of the format of the interaction.

Finally, *chapter 7* describes the contributions of this thesis. This chapter also contains suggestions for future work and a summary of the results.

1.7 Publications list

During this thesis, the following publications were produced :

- A. Vinciarelli, H. Salamin, G. Mohammadi, and K. Truong. More than words: inference of socially relevant information from nonverbal vocal cues in speech. *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, pages 23–33, 2011.

- H. Salamin and A. Vinciarelli. Introduction to sequence analysis for human behavior understanding. *Computer Analysis of Human Behavior*, pages 21–40, 2011.

- H. Salamin and A. Vinciarelli. Automatic role recognition in multiparty conversations: an approach based on turn organization, prosody and conditional random fields. *IEEE Transactions on Multimedia*, PP(99):1, 2011.

- H. Salamin, A. Vinciarelli, K. Truong, and G. Mohammadi. Automatic role recognition based on conversational and prosodic behaviour. In *Proceedings of the International Conference on Multimedia*, pages 847–850. ACM, 2010.

- H. Salamin, S. Favre, and A. Vinciarelli. Automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. *IEEE Transactions on Multimedia*, 11(7):1373–1380, 2009.

- A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent*

Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on, pages 1–4. IEEE, 2009.

- A. Vinciarelli, H. Salamin, and M. Pantic. Social signal processing: Understanding social interactions through nonverbal behavior analysis. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 42–49. IEEE, 2009.

- S. Favre, H. Salamin, J. Dines, and A. Vinciarelli. Role recognition in multiparty recordings using social affiliation networks and discrete distributions. In *Proceedings of the 10th International Conference on Multimodal interfaces*, pages 29–36. ACM, 2008.

- N.P. Garg, S. Favre, H. Salamin, D. Hakkani Tür, and A. Vinciarelli. Role recognition for meeting participants: an approach based on lexical information and social network analysis. In *Proceeding of the 16th ACM International Conference on Multimedia*, pages 693–696. ACM, 2008.

Chapter 2

State of the art

2.1 Introduction

Automatic analysis of social interactions has attracted significant attention in the last few years (see [4] for an extensive survey). In this context, role recognition is one of the problems most commonly addressed and the resulting *state-of-the-art*, while being at a relatively early stage, includes an increasingly wider spectrum of scenarios and approaches.

This chapter will be divided in five parts. The first part will introduce the theory from social psychology that has been used in this thesis. The main approach used for role recognition will be presented. The third part contains the detail on the evaluation methodology used in this thesis. We then present the main databases available in the field. Finally, the fifth part will present results of automatic approaches applied to roles.

2.2 Psychological Theory

This section will present the main results from psychology that justify and motivate our approach. First, we give a definition for roles. We will then introduce the thin slice theory that justifies the use of relatively short time windows to infer roles. Finally, we will present the links between roles and non-verbal behaviours.

When people interact in a social context, they behave in some ways predictably depending on their social identities and the situation. These characteristic behaviour patterns are called

roles [8]. Those roles consist of a set of shared expectations about the behaviour of the participants. The main functions of roles are to avoid surprises [8], to organize social interaction and to facilitate a smoother interaction.

From the definition or roles, we can derive two important characteristics of roles. The first is that roles only determine part of the behaviour of a person, the rest of the behaviour will be determined by other aspects of the person. Therefore, machine learning and probabilistic approaches seem adequate way to map the behaviours with the roles. Those approaches can handle the variability that will be introduced by the different people having the same roles.

The second characteristic is that roles depend both on the social identity of the person as well as the situation. Therefore, the same person in different situation will have a different role. For example, when the university professor crosses the door step of his home, his role changes and the way she behave also changes. From those characteristics, it follows that the data collections used in the work must present a certain uniformity in the settings used in the different recordings.

The definition given here for roles does not restrict role to explicit functions (i.e. teachers, father, anchorman). Roles can also represent more implicit functions from the theories of human interactions [9] such as the attacker, the defender or the gate-keeper. The main challenge with roles that represent an implicit function is the labelling of the data collections. Those implicit roles can change during the interaction and their boundary are not always well defined both in the duration (when does somebody start to be the attacker) as well as between roles. In this thesis, we focus on explicit roles with an unambiguous labelling. There is however no reason for the approach presented here not to work on implicit roles, given the availability of suitable training data.

Given the task, automatic approaches will have to attribute roles to people based only on the content of one interaction, with no prior knowledge of the participants involved in the interaction. For the approach to be feasible, there needs to be enough information contained in the interaction. Research in psychology suggests that this is the case. Humans can make surprisingly accurate judgements about a social situation even when presented with very limited information or when being exposed to another person for a very short amount of time. This phenomena has been referred in the psychology literature as *thin slices theory* [18] and several experiments have been conducted to investigate it. For example, teachers' non verbal

behaviours in silent short video clips (under 30 s) predicted end-of semester evaluations of those teachers [19]. Similar results have been reproduced for a variety of traits including sexual orientation [20], personality and intelligence tests [21], and personality disorders [22].

Thin slices theory indicates that there is enough information in short window of interaction for humans to infer higher level social information. This theory also indicates what information people use to make those judgements. Humans rely heavily on non-verbal communication to assess and understand their surroundings [15, 2]. As this thesis focuses on audio data, we will focus on the results concerning the nonverbal communication in the voice. The vocal nonverbal behaviour can be decomposed in five parts parts: prosody, linguistic and non-linguistic vocalizations, silences, and turn-taking patterns [2]. Prosody corresponds to how things are said (i.e loudness, rhythm and pitch). Linguistic vocalizations are speech productions used as words but that are not words (such as "ehm", "ah-ah"). Non-linguistic vocalizations are speech productions not used as words such as laughter and cries. Turn-taking patterns account for who talks when in a conversion [23].

There has been very little research in psychology on the relation between specific roles and vocal non-verbal behaviours. The only results we are aware of are that turn-taking patterns in news interviews have been shown to differ significantly from normal conversations [24]. More importantly for this thesis, this article also puts forward difference in turn-taking behaviours between the interviewees and the interviewers. This is evidence that people that play different roles have different turn-taking patterns.

Turn-taking behaviours can also indicate how people interact. In a news broadcast, for example, knowing who talks when also indicates who interacts with whom. Interaction between people in small groups has been studied in psychology under the name *Social Network Analysis* [25]. This concepts predate social networks as most people known (i.e. Facebook, MySpace, etc.) them and is not related. Social network analysis has been successfully applied to the understanding of roles [26]. In particular, people that play the same role will relate to other people in the same way. This finding is the main motivation for the use of Social Network analysis to extract features in the experiments presented in Chapter 4 and Chapter 5.

The last non-verbal vocal behaviour we consider is *prosody*. There is a rich tradition (starting with Darwin and the link between emotion and prosody [27]) of research between prosody

and several social attributes. Most of the studies decompose human prosody in three main dimensions: the *pitch*, the *duration* and the *energy* [28]. The pitch is the related to the frequency at which the vocal fold vibrates and accounts for the tone of the voice is perceived [29]. For analysis, the formants defined as the peak in the spectrum of the voice are commonly used. Usually, vowels will have four distinguishable formants, with the lowest formant usually in the range 80 Hz – 300 Hz. Human speech is composed of segments where the vocal folds vibrate called voiced segments and segments where the vocal folds are not vibrating, called unvoiced segments. The duration corresponds to the length of the voiced and unvoiced segments in the speech signal [29]. The duration can capture the speed at which somebody is talking. Finally, the power is simply the energy of the speech signal and account for how loud a person is speaking. These three dimensions of prosody have been extensively used for the recognition of personality and emotion in speech [30]. Prosody is also related to other social constructs such as status and dominance [31]. Those findings indicate that the three dimensions of prosody capture important information about the social interaction between people. Therefore we decided to use features derived from prosody for the experiment in Chapter 6.

In this section, we have seen that roles play an essential part in social interaction by helping structure the interaction. Furthermore, humans use non-verbal behaviours to communicate and understand the social environment. By using cues extracted from those non-verbal behaviours, we plan to recognize the roles played by people. This concludes our overview of the findings in psychology.

2.3 General Approach

Role recognition is the task of automatically recognizing roles of participants in an interaction recording. The goal is to assign to every participant in the recording of an interaction (usually and audio recording or video recording) a role. That is, we want to know at every time in the recording what is the role of the participants. For example, given the recording of a meeting, the goal could be to identify the chairman. The set of possible roles is typically predefined and the interaction is usually given in the form of an audio or video file.

Current approaches typically include three main steps: *Feature Extraction*, *Model Training*

and *Role Assignment*. The Feature Extraction step takes as input the signal captured during the interaction (audio, video, etc.) and extracts, for each unit of analysis (e.g., a turn or speaker), a feature vector. The second step, Model Training, takes as input feature vectors with associated labels (i.e. the roles to be recognized) and produces a statistical model linking the feature vectors and the labels. The labels are usually obtained by manually annotating the corpus. The labels also provide a way to measure the performance of the approach. The third step, Role Assignment, maps the feature vectors into roles. This assignment is usually done using a statistical model. This overall approach works on many different types of data. However, the technical details of each part will change depending on the method used for the recording and depending on the type of roles we want to identify.

The first step, *Feature Extraction*, transforms audio and video signals into feature vectors. Those vectors represent behavioural evidences that we are interested in detecting. This step uses techniques developed in signal processing. In this step, participants need to be identified, or at least detected. This step can be either fully automated or make use of manual annotation. The use of manual annotation in the current state of the art allows the modelling of more complex features (e.g, overlapping speech, word transcript). In terms of detectable behavioural patterns, most of the works presented in the literature make use of features related to turn-organization (see below). These can be accompanied by other sources of evidence such as lexical choices (e.g. the word distribution in what people say) or movement (e.g., fidgeting). It is also important to note that for feature extraction, the type of features that can be extracted automatically is dependent on the devices used for the capture. For example, overlapping speech and interruptions can easily be captured using close-talk microphone, but are very difficult to extract automatically from ambient microphones.

The second step, *Model Training*, takes labelled data and builds a statistical model. Usually, the models aim at capturing the relation between the features extracted in step one and the social aspects (roles in this case) under consideration. Currently, most of the approaches to take into account the temporal dependency use one form or another of a graphical model. For turn-based approaches (i.e. approaches that classify one turn at a time), Support Vector Machines produce state of the art results.

The last step, *Role Assignment*, uses the model from the second step to identify the roles of participants. Role attribution can be done at the participant level (each participant is given

a roles, independently of the roles attributed to the other participants) or at the interaction levels, where the roles are attributed respecting a certain set of constraints (number of occurrences of a given role, etc.).

2.4 Evaluation Methodology

As we have seen, the role recognition problem can been seen either as a problem of classification or as a problem of labelling. In the case of classification, the goal is to assign to every participant a role. In the case of labelling, the goal is to segment the interaction and assign to each segment a role. In both cases, a model has to be trained and evaluated. In this section, we will first present the protocol for training a model. We will then introduce the most common used performance metrics. We will conclude this section by presenting the statistical test used for analyzing the results.

As we use machine learning, it is important to keep a clean separation between the *training set* and the *test set* [32]. The training set is the subset of the data used to train the model. The test set is used to compute the performance measure. There are two possibilities for achieving this separation.

The first possibility is to simply partition the data in two parts. The first subset is used to learn the parameters of the model. If the model has meta-parameters (for example the number of Gaussians in a Gaussian mixture Model), some part of the training set can be used as a *validation set*. Several models with different values for the meta-parameters are trained on the training set, and the model with the best performance on the validation set is selected. The second subset of the data is used for the evaluation of the model. The model developed on the first part of the data is used to predict the labels on the test sets. The results on the test set are then used to evaluate the performance of the model.

The second possibility for separating the training set and the test set is to use *cross-validation* [33]. For each round, the dataset is partitioned in a training set and test set. Several rounds are performed and the results are aggregated over the rounds. For the role recognition problem, the most common forms of cross-validation are the *K-fold cross validation* [34] and the *Leave-one-out cross-validation* [35]. In K-fold cross-validation, the data collection is divided in K disjoint subsets and each subset is used once as the test sets. The remaining

$K-1$ subsets are used for the training, and K rounds of cross-validation are effectuated. Common values for K are 5 or 10. In the Leave-one-out cross-validation, each element in the data collection is used as a test set and all the remaining elements are used for the training set. One round is done per element in the data collection.

Both methods allow for a proper separation of the training set and the test set. We will briefly discuss the advantages and disadvantages of each methods. Setting aside a part of the data for the test sets is commonly done for challenges [36]. In the challenge case, the participants receive only the training set. Each submission is then evaluated against the test set. The main advantage of this method is that is that it guarantees that the participants do not try several approaches and only report on successes. However, this approach needs larger data collections. Therefore, given the size of the data collections used, all of the experiments done in this thesis were done using cross-validation.

The two approaches for cross-validation (K-Fold and Leave-One-Out) produces very similar results. Using between 5 and 20 folds has been show to be in practice a very reliable method [37]. The main draw-back for leave-one-out cross validation is that it is computationally expensive. One round of training needs to be done for every element in the data collections. This is not the case for the K-Fold cross-validation as only k rounds of training need to be conducted. The advantages of the Leave-One-Out approach is that it maximizes the size of the training sets. This is particularly important if the data collection is small or the model is complicated. In this thesis, the experiments were conducted using Leave-One-Out (Chapter 5) and 5-fold cross-validation (Chapter 4 and Chapter 6)

The last part of the evaluation is the compare the performance of two models. Usually, a simple model is selected as a base-line and the performance of the different models are compared against this base-line. To compute most performance measures, the labels obtained from human annotation are compared to the labels given by the model. The most common metric used for role recognition is the *accuracy*. Accuracy is defined as the percentage of time correctly labelled (in the case of segmentation) or the percentage of people assigned to the correct role (in the case of classification). Accuracy is also helpful in giving a good approximation of the *Generalization Error* of the model: the expected error on a never seen before interaction. A good estimator of the generalization error is the expected error on the training set which also corresponds to $1 -$ Accuracy.

To compare the performance of two models, it is important to use *statistical hypothesis testing* [32]. The performance of both models depends on the data collection and the variation in performance could be explained by randomness. By testing the null hypothesis that the performances of the two models are similar, we can verify if the difference in performance is meaningful. The hypothesis is rejected if its probability is below a threshold, called the p-value. If the hypothesis is rejected, the difference is said to be statistically significant. In this thesis, we have used a non-parametric test: The Kolmogorov-Smirnov test [38]. The main advantage of this test is that it does not make assumptions about the distribution of the performance (unlike the t-test that assumes the performances follow a Gaussian distribution) and it is adapted to continuous distributions

The use of statistical hypothesis testing has one important practical consequence. A statistical test can make two types of errors: rejecting a true hypothesis (also called a False Negative) and accepting a false hypothesis (called a False Positive). The probability of a test rejecting a true hypothesis is the p-value and can be controlled. Controlling for the probability of a test accepting a false hypothesis is extremely difficult in practice. However, increasing the size of the data collection will reduce the probability of a false positive. In particular, for the problem of classification, a larger data collection permit to estimate the generalization error (or any error measure) with a smaller confidence interval. This allows, for example, better comparison between models that are close in term of performance.

This concludes the section on evaluation methodology. The next section will be presenting the data collections that have been used for the role recognition problems.

2.5 Data Collections

This section gives more details about the different data collections used in the results presented in Table 2.1. As the field is still relatively young and no extensive campaign of evaluation has been conducted, there is currently no standard data collection in the domain of Social Signal Processing. Most of the largest collections presented below, such as the AMI corpus or the NIST TREC SDR Corpus, have been adapted from other fields. For each collection, we will give an overview of its main characteristics (duration, type of data and content, role annotation, other annotations).

The collections can be divided into two groups and the rest of the section is accordingly split. The first group consists of recordings of broadcast news and the roles are based on the profession of the participants (journalist versus non-journalist) and the function in the broadcast (for example, the Anchor). The second group of data collections consists of meetings. In that case, the roles are either related to the function of the person in the meeting (the project manager or the presenter) or to the relation between the meeting participants. We will conclude this section by motivating the choice of data collections used in this thesis.

2.5.1 Broadcast News Corpus

Several data collections used for the role recognition problems consist of recordings of broadcast news or radio programs. There are two main reasons for that. The first reason is the availability of large and well annotated corpora from other fields such as the document retrieval fields. The second reason is that roles are usually easy to define for the participants in that type of data as they can be directly derived from the function of the person.

The data set used in [39] is a subset of the TREC-7 SDR track corpus [40], a data set developed for the evaluation of spoken document retrieval. The TREC-7 SDR data set contains 100 hours of news data, including both radio and television sources. The data set is segmented by stories and speakers. A full manual transcription is available. However, this data set does not contain labelling for the roles of the participants, only speaker identities are included. In [39], the authors utilized a subset of 35 recordings lasting half an hour each from the program "All Things Considered". The speaker roles were manually labelled by matching the speaker IDs in the database with the list of anchors and journalists available from the website of the program. Every speaker was attributed to one of three classes: *Anchor*, *Journalist* and *Program Guest*. This data collection provide a good illustration of the fact that although large amounts of multimedia data with rich annotation are available, social information and in particular roles information is usually not available and must be derived from other sources.

The data collections used in [41] is a subset of the TDT4 broadcast news data [42]. TDT stands for Topic Detection and Tracking, and this data collection aims at helping develop technologies for understanding news broadcasts. The whole data include audio as well as manual transcription and segmentation in stories (defined as two or more declarative inde-

pendent clauses about a single event). For the role recognition, only a subset consisting of a collection of 363 shows in Mandarin, totalling 170 hours, was used. As no speaker ID information is available in the original data, the audio was manually labelled in speaker turns and roles. The three roles considered are: *Anchor*, *Reporter* and *Other*. However, there is not enough data publicly available on the roles annotation to assess the reliability of the annotation. The number of annotators is missing and there is no measure of the agreement between the annotators.

The EPAC corpus [43] is composed of recordings from three French broadcasters (France Inter, RFI, France Culture). The data has been selected to be of a conversational nature and the total duration is 100 hours. The data is composed of radio programs and thus is only audio. It has been manually annotated for speaker segmentations and speaker roles. The roles are: the *Anchor*, the *Journalist* and *Others*. Both the anchor and the journalists are professional speakers. Others include all the non-professional speakers appearing in the broadcast. They can be interviewees, people from the street or guests present in the studio.

[13] and [44] have used a corpus composed of 96 news broadcast, totalling 19 hours, from the "Radio Suisse Romande" (French speaking Swiss National Broadcast). Each broadcast corresponds to a hourly news bulletin. The data consists of audio only and has been manually segmented in speaker turns. Each speaker in each broadcast was assigned a role. There are six possible roles: the *Anchorman*, the *Second Anchorman*, the *Guest*, the *Interview Participant*, the *Headline Reader*, and the *Weather Man*. The role assignation was given by the broadcaster. Some participants appear in several broadcasts but do not always play the same roles. The set of roles used by this data collection is more detailed than the sets used in the two previous data collections. In particular, there are two anchormen with slightly different roles. There is also a distinction between a guest (a person invited to report about a single and specific issue) and an interview participant (includes both interviewees and interviewers) not presents in the previous data sets. This larger set of roles is the main specificity of those data collections.

The Radio Talk Shows data collection used in [44], is composed of 27 talk-show lasting one hour each broadcasted on the "Radio Suisse Romande". The total duration is 27 hours. The data has been manually segmented in speaker turns and each speaker was assigned a role for each, by the broadcaster. The five roles present in the corpus are: the *Anchorman*, the

Second Anchorman, the *Guest*, *Headline Person*, and the *Weatherman*. The main difference between this data set and the previous one is the type of program. This data sets contains talk-shows, where the aim is to entertain, with a much larger participation of the guests, and a less constrained structure.

In conclusion, all the data collections derived from broadcast news have a similar set of roles: the Anchor, the Journalist and the Guest. The datasets from the "Radio Suisse Romande" have a slightly larger set of roles, however those roles are a refinement of the basic set.

2.5.2 Meeting Corpus

The second group of data collections is composed of recordings of meetings. Meetings have several interesting properties for the role recognition problem. First, meetings offer a natural setting for small group interaction and roles. In meetings, people tend to play different roles and those roles have less constraints than in the broadcast news. This makes role recognition in meetings more challenging and interesting. Second, from a technical point of view, meetings can be recorded easily as they happen in one room and people do not tend to move around the room. Finally, meetings play an important part in most organization and therefore are important from an application perspective. We will present three corpora composed of meetings.

The AMI meeting corpus [45] is a large collection of meetings recorded in three smart rooms. This corpus was developed to support human interaction in meetings and to help in developing meeting browsers. The whole data set is publicly available. The rooms were equipped with several microphones and video cameras. Each participant was recorded using a close-up camera and a lapel microphone. This corpus has been annotated with a lot of details including speech transcription, dialogue act and group activities. For the role recognition research, only a subset of the collection was suitable. This subset was selected because the participants have a clear role in those meetings. This subset consists of 138 meeting recordings for a total of 45 hours and 38 minutes of material. The meetings are a *simulation* of a design meeting and the role set contains the *Project Manager*, the *Marketing Expert*, the *User Interface Expert*, and the *Industrial Designer*. It is important to note that the meetings are a simulation and that each participant appears in 4 meetings, playing each of the roles in turn.

Another meeting corpus [46] was recorded at Carnegie Mellon University. The corpus includes both audio and video feeds. The audio was captured using a close-talking head-mounted microphone and the video was captured using a single camera. The meetings are real meetings involving various faculty, staff and students of the university. Two meetings, for a total duration of 45 minutes, were manually annotated for the participant roles. The five roles considered were the *presenter*, the *discussion participator*, the *information provider*,the *information consumer* and *undefined*) . The interesting aspect of those roles is that the same person may play a different role at different times in the meeting. However, this also implies that the annotation process is extremely costly. This explains the relatively small size of this data collection.

The Mission Survival Corpus is a data collection of 11 meetings involving 4 participants [47, 48]. The meetings are based on the survival task [49], a commonly used task in psychology to elicit small group discussion. The participants have to rank a list of 12 items according to their importance for survival. The meeting are recorded using 5 cameras and 4 microphones (one close-talk microphone per participant). The meetings are manually annotated for two types of roles. The first set of roles for the facilitation and coordination of the tasks the group is involved in and consists of: the *Orienteer* (responsible for the conduct of the meeting), the *Giver* (providing factual information), the *Seeker* (requesting information), the *Recorder* (keeping track of the evolution of the meeting) and the *follower* (listening). The second set of roles accounts for the relationships between the participants and includes: the *Attacker* (expresses disapproval), the *Gate-keeper* (moderating the discussion), the *Protagonist* (actively participating), the *Supporter* (cooperative attitude) and the *Neutral*. Each participant is at each instant playing two roles (one from each set) and those roles change over the meeting. The first set of roles (related to the task) is similar to roles used in the previous data collection (the data recorded at Carnegie Mellon University). The second set of roles accounting for the relationships is unique to this corpus.

2.5.3 Data collection used in this Thesis

In this thesis, we have used three data collections: the news broadcast from “Radio Suisse Romande”, the talk-shows from “Radio Suisse Romande” and the AMI meeting corpus. Those three data sets are described in more details in Section 4.1. Those datasets account

for the most common setting in roles recognition (news broadcast, talk-shows and meetings) and thus allow for the comparison of our approach on different setups. At the beginning of this thesis, the AMI meeting corpus was to only corpus whose role annotation was publicly available. The other two corpora from the "Radio Suisse Romande" were developed and annotated at the Idiap Research Institute and were also available for this research. The rest of the data collection presented in this section did not have available annotation and there were limited resources during the thesis that make the annotations of new data impracticable. This concludes the presentation of the data collections used in this state of the art. The next section will be presenting results for the role recognition problems.

2.6 Role Recognition Results

From a role point of view, the approaches proposed in the literature can be split into two broad groups. The first includes works aimed at the recognition of roles related to *norms* (explicit prescriptions about behaviours to be associated to a role), the second includes approaches aimed at the recognition of roles related to *preferences* (personal attitudes influencing the behaviours associated to a role) and *beliefs* (subjective assessments of what behaviours should be associated to a role) .

Table 2.1 gives an overview of the results for role recognition found in the literature. For each result, the datasets used in the experiments as well as it size are given. The datasets are described in more detail in Section 2.5. We also indicate the type of features used as well the statistical model used for the recognition. The column expectation indicates which type of roles is recognized. N stands for *norms* and those results are detailed in the first subsection. BP stands for *preferences* and *beliefs* and those results will be detailed in the second subsection.

The rest of the section is divided into three parts. The first part will present the roles related to *norms*. All the results from this section were obtained on data collection of news broadcasts and talk-shows. The roles set is very similar between the different experiment reported. The second part gives results obtained for roles related to *preferences* and *beliefs*. Those results were derived from the data sets of meetings. The direct comparison of performance is more difficult in this part as the role sets are not comparable. Finally, the third part reports results

Ref.	Data	Time	Exp.	Evidence	Approach	Performance
[39]	NIST TREC SDR Corpus (35 recordings, 3 roles)	17h	N	Term distribution, speaking time	BoosTexter, Maximum Entropy Classifier	80.0% of the news stories correctly labeled in terms of role
[41]	TDT4 broadcast news (336 shows, 3 roles)	170h	N	Distribution of bigrams and trigrams	Hidden Markov Models	77.0% of the news stories correctly labeled in terms of role
[13]	Radio news bulletins (96 recordings, 6 roles)	25h	N	Turn organization, social networks (centrality, nodes degree, etc.)	Bayesian Classifiers	85% of the data time correctly labeled in terms of role
[44]	Radio news, Talk shows, meetings	90h	NBP	Turn organization, social networks (centrality, nodes degree, etc.)	Bayesian Classifiers	Up to 85% of the data correctly labeled in terms of role (45% for the meetings)
[50]	Movies and TV shows (13 recordings , 2 roles)	21h	N	Co-occurrence of faces, social networks	Bayiesian c;assifiers	85% to 95% of recognition rate depending on the role
[51]	EPAC Corpus (Broadcast data, 3 roles)	100h	N	Turn organization, prosody	Gaussian Mixture Models, SVM, k Nearest Neighbours	92% of role recognition rate
[46]	Meetings (2 recordings, 5 roles)	45m	BP	turn organization	Decision tree	53.0% of segments correctly classified
[47]	Mission Survival Corpus (11 recordings, 5 roles)	4h 30m	BP	speaking activity, fidgeting	Support Vector Machines	70% of analysis windows (10 seconds) correctly classified
[48]	Mission Survival Corpus (11 recordings, 5 roles)	4h 30m	BP	speaking activity, fidgeting	Support Vector Machine	90% of analysis windows (around 10 seconds long) correctly classified
[52]	Mission Survival Corpus (11 recordings, 5 roles)	4h 30m	BP	speaking activity, fidgeting	Influence Model, Support Vector Machines	75% of roles correctly assigned
[53]	AMI Meeting Corpus (138 recordings, 4 roles)	45h	BP	speaking activity, talkspurts	Conditional Random Fields	53% of the time correctly labeled
[54]	AMI Meeting Corpus (138 recordings, 4 roles)	45h	BP	speaking activity, term distribution	Bayesian Classifiers, BoosTexter	75% of the time correctly labeled

Table 2.1: Synopsis of role recognition results. The table reports the main results on role recognition presented in the literature. The time is expressed in hours (h) and minutes (m), the expectations in terms of *norms* (N), *beliefs* (B) and *preferences* (P).

on analysis of social interaction that are not directly roles. Those results were included as they share tools, techniques and data sets that are very similar to those used in fro role recognition.

2.6.1 Roles Driven by Norms

The upper part of Table 2.1 reports the main aspects of the works dedicated to the recognition of roles for which the expectations are expressed as norms.

In 2000, Barzilay et al. [39] used features extracted from automatic speech transcriptions to identify roles in radio broadcast data (the DAPPER Broadcast News Corpus). The three predefined roles were *journalist*, *guest* and *anchorman*. The speaker labels and turn boundaries were extracted from the manual turn segmentation of the audio. The features extracted were:

- n-grams (from uni-grams to 5-grams)
- features from the surrounding context (labels of the n previous turns or all the features of the n previous turns)
- Duration of turn, extracted from the boundary of the manual segmentation
- Explicit speaker introduction, computed from n-grams, word frequency and position in the segment.

The roles were assigned for each turn using Maximum Entropy Model and BoosTexter. Contextual information was helpful and improved the classification accuracy with respect to using only the text of the current turn. The main limitation of this approach is that it cannot be easily automated. In the test on the automatic data, some information from the manual annotation were used (duration of the turn, speaker id).

The work in [41] addresses a similar problem (three roles in broadcast news). A Hidden Markov Model is used to align the sequence of the turns with a sequence of roles, and each turn is represented with the distribution of bi-grams and tri-grams in the transcription of what is said. The sequence of roles is modelled with a 6-gram language model. A Maximum Entropy Model (MEM) was also used and yielded similar performance to the HMM model. In the case of the MEM, there was no language model to take into account the information

contained in the sequence roles, the bi-grams and tri-grams of the previous turn and the next turn were added to the features. For the experiment, a corpus of 170 hours in mandarin were used. The experiments were run only on the manually annotated data.

In both previous works [39, 41], the words at the beginning of each turn appear to be more discriminant than the others. The reason is probably that the beginning of the turn contains a self-introduction of the speaker that often mentions explicitly her role.

The work in [51] adopts features that account for turn organization and prosody and maps each person, as detected with a speaker diarization approach, into one of three roles accounting for the general aspects of broadcast data, namely *anchorman*, *journalist* and *others*. There are two experiments presented in this paper. In the first experiment, an unsupervised clustering algorithm (K-mean clustering) is used to check that the features selected and the role definition are meaningful. The approach used in the paper to check the meaningfulness of the role definition suffers two problems. First the number of clusters is a meta-parameter that is not validated against a validation set and thus the purity reported is probably overestimated. Second, as the number of clusters is not limited, it is always possible to increase the number of clusters to increase the purity (to the degenerated case of one cluster per observation). In the second experiment, the classification is performed using Gaussian Mixture Models or Support Vector Machines. Another interesting technique presented in the article is to use a hierarchy of models (first classify anchorman versus non-anchorman then separate journalist from others).

The works in [44, 13, 50] extract automatically social networks from the data in order to assign each person involved in a broadcast recording a different role. The approach in [13] segments the data (audio recordings of news) into turns and then uses the adjacency in the speaker sequence to build a social network. Social Network based features (e.g. the centrality) are then used to represent each person and map her into one of six predefined roles. In a similar way, the approach in [50] uses the co-presence of two faces (automatically detected and extracted from Hollywood movies) in the same scene as an evidence of direct interaction to build a social network. Features like those applied in [13] are then used to detect the main characters of the movie (the "*leading*" roles) as well as the members of the communities possibly associated to each of the main characters. Finally, the approach proposed in [44] uses the turn-taking to build a Social Affiliation Network based on the proximity in time of

different speakers. The structure of the network edges is then represented with patterns that are fed to Bayesian Classifiers and mapped into roles.

The work in [55] investigated the detection of "soundbites" in news broadcasts. Soundbites are segments of interview that can be clearly attributed to one speaker. They usually correspond to turns that can be attributed to a person being interviewed. As such, they correspond to turns rich in content. The data-set used in the experiments is composed of 24 half-hour news-broadcast. The data was automatically transcribed and manually segmented in speaker turns. The following prosodic features were extracted:

- speaking rate,
- pitch minimum, maximum, mean, range and slope,
- energy minimum, maximum, mean and slope
- turn duration.

Structural features extracted were:

- Normalised position of turn,
- Speaker change,
- turn position,
- speaker distribution
- previous and next speaker,
- top-ranking speaker.

Lexical features finally were also extracted:

- number of words in the turn,
- cue phrase,
- distribution of cue words.

Those features were used to train Conditional Random Fields (CRF) and MEM. The results obtained are significantly higher than chance. However, the annotation seems to be very noisy (high variability in the different folds in term of performance). Also the use of the cue phrase, based on manual inspection, weakens the generality of the approach. This work was preliminary, but no further publications are available.

2.6.2 Roles Driven by Beliefs and Preferences

The work in [46] applies decision trees to assign meeting participants roles corresponding to different ways of participating in a discussion (*presenter*, *discussion participator*, *information provider*, *information consumer* or *undefined*). The data used in the experiments was a collection of meetings recorded at Carnegie Mellon University. The corpus includes both audio and video feeds. The behavioural evidences are extracted from short temporal windows and include:

- number of speaker changes,
- number of participants in the window,
- number of overlapping speech segments,
- average length of the overlaps.

The proposed approach can detect the state of the meeting at the end of the windows. For the role recognition, the following behavioural evidences were extracted for each participant:

- total participant speaking time,
- total overlapping speech involving the participant
  - initiated by the participant
  - initiated by another participant.

Decision trees were used for the classification. The trees were learnt using the C4.5 learning algorithm [56]. The main achievement is the introduction of a taxonomy for meeting states and meeting participants. The performance of the classifier was below 60% percent but well

above the chance level. The main drawback of the approach is the selection of the windows length meta-parameter. Furthermore, the approach can not be used to do segmentation for the meeting state, but only do labelling of the meeting state at the end of the window.

Zancanaro et al. [47, 48] studied functional roles in face to face interaction. The roles considered were of two types:

- Task area roles consist of roles relating to the management of the task the group is involved in. They were the orienter, the giver, the seeker, the recorder and the follower.
- Socio-Emotional area roles concern the relationship between members of the groups. They were the attacker, the gate-keeper, the protagonist, the supporter and the neutral

In these works, the behavioural evidence is given by speaking activity (e.g., silence versus speaking) and movement (e.g., total amount of fidgeting). The role recognition is performed over short time intervals (2-40 seconds) that are aligned with a sequence of roles using probabilistic sequential models (e.g., Factorial Hidden Markov Models) and Support Vector Machines.

Several participants can play the same role and the role of a participant can evolve over the interaction. The approach was tested on eleven meetings involving four people. The goal was to identify the role of the current speaker. The features were:

- manual annotation of fidgeting (i.e. localised repetitive motions such as tapping the fingers on the table)
- speaking activity derived from close talk microphone
- overlapping speech.

Features were extracted on time windows of varying length. For long enough windows, SVM classification is significantly statistically better than the baseline. The limitation of the approach is the limited number of features. The model is also not taking into account temporal dependencies.

This work was extended [52] to use an influence model to replace the SVM. The main goal was to reduce the number of parameters to learn and to take into account the time dynamic

of the interaction. The influence model is based on coupled Hidden Markov Models and allows to model jointly and efficiently the evolution of a group of hidden variables (one per participant in this case). The parameters that are learnt are called influence parameters. The model performed better than linear SVM and close to inter-agreement between annotators. The experiments were conducted on the same corpus.

Finally two works aim at the recognition of roles corresponding to a position in a company in the AMI Meeting Corpus (*Project Manager*, *Marketing Expert*, *Industrial Designer*, *User Interface Expert*) [53, 54]. The approach in [53] uses Conditional Random Fields to align behavioural evidences extracted from short time intervals (e.g., number of times a person talks, total number of speaking attempts of all meeting participants, etc.) with a sequence of roles. The approach in [54] combines lexical choices (distributions of uttered words) and Social Network features like those applied in [44]. The former features are recognized using the BoosTexter, while the latter using Bayesian classifiers based on discrete distributions.

Laskowski et al. [53] investigated the detection of roles, gender and seniority in the AMI meeting corpus and the ICSI meeting corpus. The features extracted are related to the vocal interaction pattern of the participants. More exactly, the probability of a participant talking at time t given to speech activity at time $t-1$. Those features were extracted from the manual segmentation of the meetings. The experiments are extremely well documented and conducted. They achieve a role classification accuracy of 53 % on the AMI meetings (chance level is 25 %) and 75 % accuracy for the project manager versus the rest. There was not conclusive results on the gender detection. For seniority, performance of 61 % (versus 45 %) were attained. One limitation of the approach used, is the need to have accurate detection and annotation of overlapping speech (including who is participating in the overlapping speech). This is currently only possible with a close-talk microphone or a microphone array.

2.6.3 Analysis of Social Interactions

This section proposes a short survey of works that consider other aspects of social interactions than roles (see [4, 5] for extensive reviews). The techniques and statistical models used for those other aspects are very close to the tools used for role recognition. Most of the literature in this domain is dedicated to meeting analysis not only for the availability of large annotated corpora (e.g., see [48, 57]), but also because most social phenomena taking

place in small groups (meetings rarely involve more than 10 people) are equivalent to those happening at any social scale, while being easier to model and analyse [58].

The literature has tackled three major problems: group action recognition, dominance detection, and interest level measurement. The recognition of collective actions (discussions, presentations, etc.) has been addressed, e.g., in [59, 60, 61, 62, 63]. The common aspect of these works is that they model jointly streams of features extracted from multiple modalities. In [60], hand movements and speaking activity are modelled with hidden Markov models and then fused with different strategies (concatenation of feature vectors extracted from different streams or multiplication of likelihoods estimated using HMMs applied to different streams). The same approach is applied in [59], where different streams are fused with coupled and asynchronous HMMs, and [61], where a hierarchic layered HMM models individual participant actions and, at an upper level, collective actions. A similar approach has been proposed in [62], where actions are modelled with Dynamic Bayesian Networks, and [63], where Hidden Conditional Random Fields are shown to improve the action recognition performance with respect to the other approaches.

The problem of detecting the most dominant person in a group has been investigated in [64, 65, 66, 67]. The approaches in [64, 65] are based on vocal behaviour (speaking time, number of turns, interruptions, etc.) and apply a Support Vector Machine to map people into three dominance classes (low, normal and high). The other works include similar audio features and combine them with information about gaze behaviour [66] (Dynamic Bayesian Networks are used to model the effect of one person on another one), and kinesics [67] (people are classified using Support Vector Machines into dominance categories).

The last topic significantly investigated in this domain is interest, the degree of engagement of people in interactions [68, 69, 70]. The approach in [68] is closely related to the one described in [59] (same features and same combination approach for multiple modalities). The approaches in [69, 70] combine through early fusion a wide spectrum of visual and audio features, including facial expression, eyes behaviour, non-linguistic vocalisations (e.g., laughter) and lexical information, then use Support Vector Machines to measure the interest level.

2.7 Conclusion

In this chapter, we have provided a large background on the role recognition problems. The psychological foundations of the problems were given in Section 2.2 and show the importance of roles and non-verbal behaviours in the human social life. Sections 2.3 and 2.4 describe the technical tools used to map the low-level non-verbal behaviours to the roles and the methods to evaluate the results. Finally, Section 2.5 and 2.6 presented the main data collections used in the field and the results already obtained. This chapter has highlighted the position of the role recognition problem at the boundary between psychology and machine learning. The rest of the thesis and particularly the next chapter will be more focused on machine learning and the technical aspects of role recognition.

Chapter 3

Graphical Models

In this chapter, we will present the mathematical tool we used to recognize roles: *graphical models*. Those are the probabilistic models we use to associate the feature vectors to the roles. Given a set of observation $\boldsymbol{X} = (x_1, \ldots, x_N)$, where x_t is generally a D-dimensional vector, there are two problems we can solve using machine learning. The first is called *classification* and it consists of assigning to the observations $\boldsymbol{X}$ a class c belonging to a predefined set $C = \{c_1, \ldots, c_K\}$. The second problem is called *labeling* and it corresponds to mapping $\boldsymbol{X}$ to a set of labels $\boldsymbol{Z} = (z_1, \ldots, z_M)$, with $M \leq N$, where each z_t belongs to a discrete set $\boldsymbol{S} = \{s_1, \ldots, s_T\}$. In the case of labelling, we may label only a subset of the observations or have several observations associated with one label. Classification can been seen as an extreme case of labelling, where all the observations are mapped to one label. In the case of role recognition, both approaches can be used. In the case of labelling, one role is associated to each turn (and each turn can be represented by several observations) and one speaker can be attributed several different roles over the whole interaction. In the case of classification, one role is associated to each speaker. Both approaches have been used for the work presented in this thesis. In Chapters 4 and 5, we used the classification approach. In Chapters 6 we used the labelling approach.

In both cases, the problem can be thought of as finding the value $\hat{\boldsymbol{Y}}$ satisfying the following equation:

$$\hat{\boldsymbol{Y}} = \arg\max_{\boldsymbol{Y}} \mathrm{P}(\boldsymbol{X}, \boldsymbol{Y}) \tag{3.1}$$

where $\hat{\boldsymbol{Y}}$ can be one of the classes belonging to C, or a set $\boldsymbol{Z}$. In this respect, the main chal-

lenge is to find a model $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})$ suitable for the problem at hand, i.e. an actual expression of the probability to be used in the equation above. We present the unifying framework of *graphical models* [71] to introduce the two probabilistic models used to estimate $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})$ in this work, namely Bayesian Networks (in particular the Naive Bayes Model) and Conditional Random Fields [72, 73].

This chapter focuses in particular on two major aspects of the modelling of a set of random variables: On one hand, the role that conditional independence assumptions have in making the problem tractable and, on the other hand, the relationship between independence assumptions and the particular factorization that the models mentioned above show. This chapter provides some details about inference and training as well, including pointers to the relevant literature.

The rest of the chapter is organized as follows: Section 3.1 describes the graphical model framework, Section 3.2 and 3.3 introduce Bayesian Networks and Conditional Random Fields respectively, Section 3.4 proposes training and inference methods and finally Section 3.5 draws some conclusions.

3.1 Graphical Models

The main problem in estimating $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})$ is that the state spaces of the random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ increase exponentially with the length of $\boldsymbol{X}$. The resulting challenge is to find a suitable trade-off between two conflicting needs: to use a compact and tractable representation of $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})$ on one side and to take into account dependencies between the labels on the other side. Probability theory offers two main means to tackle the above, the first is to *factorize* the probability distribution, i.e. to express it as a product of factors that involve only part of the random variables in $\boldsymbol{X}$ and $\boldsymbol{Y}$ (e.g., only a subsequence of $\boldsymbol{X}$). In this way, the global problem is broken into small, possibly simpler, problems. The second is to make *independence assumptions* about the random variables, i.e. to make hypotheses about what are the variables that actually influence one another in the problem.

As an example of how factorization and independence assumptions can be effective, consider the simple case where $\boldsymbol{Y}$ is a sequence of binary variables. By applying the chain rule, it is

possible to write the following:

$$\mathrm{P}(Y_1, \ldots, Y_N) = \mathrm{P}(Y_1) \prod_{i=2}^{N} \mathrm{P}(Y_i \,|\, Y_1, \ldots, Y_{i-1}). \tag{3.2}$$

As the number of possible sequences is 2^N, a probability distribution expressed as a table of experimental frequencies (the percentage of times each sequence is observed) requires $2^N - 1$ parameters.

In this respect, the factorization helps to concentrate on a subset of the variables at a time and maybe to better understand the problem (if there is a good way of selecting the order of the variables), but still it does not help in making the representation more compact, the number of the parameters is the same as before the factorization. In order to decrease the number of parameters, it is necessary to make independence assumptions like, e.g., the following (known as *Markov property*):

$$\mathrm{P}(Y_i \,|\, Y_1, \ldots, Y_{i-1}) = \mathrm{P}(Y_i \,|\, Y_{i-1}). \tag{3.3}$$

The above transforms Equation (3.3) as follows:

$$\mathrm{P}(Y_1, \ldots, Y_N) = \mathrm{P}(Y_1) \prod_{i=2}^{N} \mathrm{P}(Y_i \,|\, Y_{i-1}), \tag{3.4}$$

where the number of parameters is only $2(N-1)+1$, much lower than the original $2^N - 1$. The number of parameters can be reduced to just 3 if we consider that $\mathrm{P}(Y_i \,|\, Y_{i-1})$ is independent of i, thus it does not change depending on the particular point of the sequence. The combination of factorization and independence assumptions has thus made it possible to reduce the number of parameters and model long sequences with a compact and tractable representation.

Probabilistic graphical models offer a theoretic framework where factorization and independence assumptions are equivalent. Distributions $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})$ are represented with graphs where the nodes correspond to the random variables and the missing edges account for the independence assumptions. More in particular, the graph acts as a filter that, out of all possible $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})$, selects only the set DF of those that *factorize over the graph* (see below what this means depending on the type of graph). In parallel the graph acts as a filter that selects the

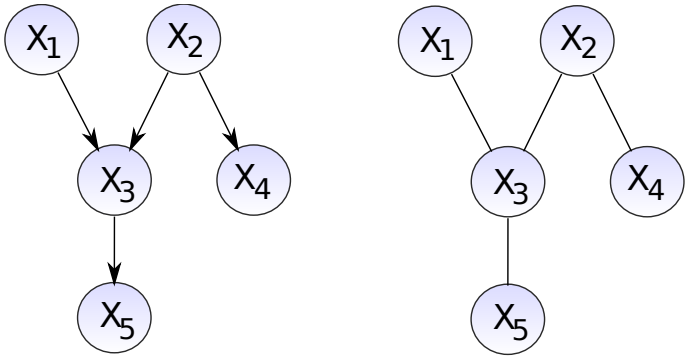

Figure 3.1: Probabilistic graphical models: each node corresponds to a random variable and the graph represents the joint probability distribution over all of the variables. The edges can be directed (left graph) or undirected (right graph).

set DI of those distributions $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})$ that respect the independence assumptions encoded by the graph (see below how to identify such independence assumptions). The main advantage of graphical models is that $DF = DI$, i.e. factorization and independence assumptions are equivalent (see [32] for an extensive description of this point). Furthermore, inference and training techniques developed for a certain type of graph can be extended to all of the distributions encompassed by the same type of graph (see [74] for an extensive account of training techniques in graphical models).

The rest of this section introduces notions and terminology that will be used throughout the rest of this chapter.

3.1.1 Graph Theory

The basic data-structure used in the chapter is the graph.

Definition 1. A *graph* is a data structure composed of a set of nodes and a set of edges. Two nodes can be connected by a directed or undirected edge.

We will denote by $G = (\boldsymbol{N}, \boldsymbol{E})$ a graph, where $\boldsymbol{N}$ is the set of nodes and $\boldsymbol{E}$ is the set of the edges. We write $n_i \to n_j$ when two nodes are connected by a directed edge and n_i — n_j when they are connected by an undirected one. If there is an edge between n_i and n_j, we say that these are *connected* and we write that $n_i \rightleftharpoons n_j$. An element of $\boldsymbol{E}$ is denoted with (i, j) meaning that nodes n_i and n_j are connected.

Definition 2. If $n \rightleftharpoons m$, then m is said to be a *neighbour* of n (and vice-versa). The set of all neighbours of n is called the *neighbourhood* and it is denoted by $\mathrm{Nb}(n)$. The set of the *parents* of a node n contains all nodes m such that $m \to n$. This set is denoted by $\mathrm{Pa}(n)$. Similarly, the set of the *children* of a node n contains all nodes m such that $n \to m$. This is set is denoted by $\mathrm{Ch}(n)$.

Definition 3. A *path* is a list of nodes $(p_1, \ldots, p_n)$ such that $p_i \to p_{i+1}$ or p_i — p_{i+1} holds for all i. A *trail* is a list of nodes $(p_1, \ldots, p_n)$ such that $p_i \rightleftharpoons p_{i+1}$ holds for all i.

The difference between a trail and a path is that a trail can contain $p_i \leftarrow p_{i+1}$ edges. In other words, in a trail it is possible to follow a directed edge in the wrong direction. In undirected graphs, there is no difference between paths and trails

Definition 4. A *cycle* is a path $(p_1, \ldots, p_n)$ such that $p_1 = p_n$. A graph is *acyclic* if there are no cycles in it.

3.1.2 Conditional Independence

Consider two random variables X and Y that can take values in $Val(X)$ and $Val(Y)$, respectively.

Definition 5. Two random variables X and Y are *independent*, if and only if $\mathrm{P}(Y|X) = \mathrm{P}(Y)\ \forall x \in \mathrm{Val}(X), \forall y \in \mathrm{Val}(Y)$. When X and Y are independent, we write that $\mathrm{P} \models (X \perp Y)$.

The definition can be easily extended to sets of variables $\boldsymbol{X}$ and $\boldsymbol{Y}$:

Definition 6. Two sets of random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *independent*, if and only if $\mathrm{P}(Y|X) = \mathrm{P}(Y)\ \forall X \in \mathrm{Val}(\boldsymbol{X}), \forall Y \in \mathrm{Val}(\boldsymbol{Y})$. When $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, we write that $\mathrm{P} \models (\boldsymbol{X} \perp \boldsymbol{Y})$.

Definition 7. Let $\boldsymbol{X}, \boldsymbol{Y}$, and $\boldsymbol{Z}$ be sets of random variables. We say that $\boldsymbol{X}$ is *conditionally independent* of $\boldsymbol{Y}$ given $\boldsymbol{Z}$ if and only if:

$$\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = \mathrm{P}(\boldsymbol{X} \mid \boldsymbol{Z})\, \mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{Z})$$

We write that $\mathrm{P} \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$.

The rest of the chapter shows how the notion of conditional independence is more useful, in practice, than the simple independence. For example, the Markov property (see above) can be seen as a conditional independence assumption where the future X_{t+1} is conditionally independent of the past $(X_1, \ldots, X_{t-1})$ given the present X_t. Such an assumption might not be true in reality (X_t is likely to be dependent on $X_1, \ldots, X_{t-1}$), but it introduces a simplification that makes the simple model of Equation (3.2) tractable.

3.2 Bayesian Networks

Bayesian Networks [75, 76, 77] are probabilistic graphical models encompassed by Directed Acyclic Graphs (DAGs), i.e. those graphs where the edges are directed and no cycles are allowed. The rest of the section shows how a probability distribution factorizes over a DAG and how the structure of the edges encodes conditional independence assumptions. As factorization and independence assumptions are equivalent for graphical models, it is possible to say that all of the distributions that factorize over a DAG respect the conditional independence assumptions that the DAG encodes. Inference and training approaches will not be presented for directed models because each directed graph can be transformed into an equivalent undirected one and related inference and training approaches can be applied. The interested reader can refer to [78, 74] for extensive surveys of these aspects.

3.2.1 Factorization

Definition 8. Let $\boldsymbol{X} = (X_i, \ldots, X_N)$ be a set of random variables and G be a DAG whose node set is $\boldsymbol{X}$. The probability distribution P over $\boldsymbol{X}$ is said to *factorize* over G if

$$P(\boldsymbol{X}) = \prod_{i=1}^{n} \mathrm{P}(X_i \,|\, \mathrm{Pa}(X_i)). \tag{3.5}$$

A pair (G, P) where P factorizes over G is called *Bayesian Network*.

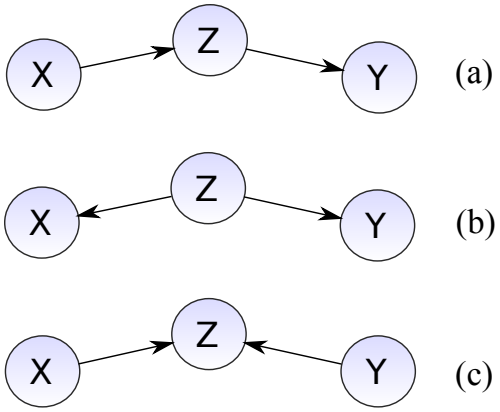

Figure 3.2: The picture shows the three ways it is possible to pass through a node along a path: head-to-tail, tail-to-tail and head-to-head.

3.2.2 The d-Separation Criterion

A DAG allows one to read conditional independence assumptions through the concept of *d-separation* for directed graphs.

Definition 9. Let (G, P) be a Bayesian Network and $X_1 \rightleftharpoons \ldots \rightleftharpoons X_N$ a path in G. Let $\boldsymbol{Z}$ be a subset of variables. The path is blocked by $\boldsymbol{Z}$ if there is a node W such that either:

- W has converging arrows along the path ($\rightarrow W \leftarrow$) and neither W nor its descendants are in $\boldsymbol{Z}$
- W does not have converging arrows ($\rightarrow W \rightarrow$ or $\leftarrow W \rightarrow$), and $W \in \boldsymbol{Z}$

Definition 10. The set $\boldsymbol{Z}$ d-separates $\boldsymbol{X}$ and $\boldsymbol{Y}$ if every undirected path between any $X \in \boldsymbol{X}$ and any $Y \in \boldsymbol{Y}$ is blocked by $\boldsymbol{Z}$

The definition is more clear if we consider the three structures depicted in Figure 3.2. In the case of Figure 3.2 (a), $\boldsymbol{Z}$, d-separates X and Y and we can write the following:

$$\mathrm{P}(X, Y, Z) = \mathrm{P}(X)\,\mathrm{P}(Z \mid X)\,\mathrm{P}(Y \mid Z) = \mathrm{P}(Z)\,\mathrm{P}(X \mid Z)\,\mathrm{P}(Y \mid Z). \tag{3.6}$$

As $\mathrm{P}(X, Y, Z) = \mathrm{P}(X, Y \mid Z)\,\mathrm{P}(Z)$, the above means that $\mathrm{P} \models (X \perp Y \mid Z)$. The case of Figure 3.2 (b) leads to the same result (the demonstration is left to the reader), while the structure of Figure 3.2 (c) has a different outcome:

$$\mathrm{P}(X, Y \mid Z) = \mathrm{P}(X \mid Z)\,\mathrm{P}(Y \mid Z)\,\mathrm{P}(Z). \tag{3.7}$$

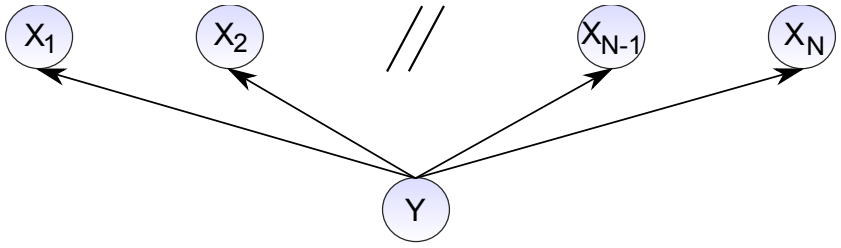

Figure 3.3: The figure depicts the Bayesian Networks representing a Naive Bayes Model. X_1 to X_N represent the observations (the set $\boldsymbol{X}$) and Y is the class.

In this case, Z does not d-separate X and Y and it is not true that $\mathrm{P} \models (X \perp Y \mid Z)$, even if $\mathrm{P} \models (X \perp Y)$. This phenomenon is called *explaining away* and it is the reason for the condition on the nodes with converging arrows in the definition of d-separation.

In more general terms, the equivalence between d-separation and conditional independence is stated as follows:

Theorem 1. *Let* (G, P) *be a Bayesian Network. Then if* $\boldsymbol{Z}$ *d-separates* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$, $\mathrm{P} \models (\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z})$ *holds.*

Thus, the conditional independence assumptions underlying a Bayesian Network can be obtained by simply applying the d-separation criterion to the corresponding directed graph.

3.2.3 Naive Bayes Models

One of the simplest models for classification is the *Naive Bayes Model* (NB). This model has been used for more than 50 years in Information Retrieval [79] as well as automated medical diagnosis [80]. This model can be used to estimate $\mathrm{P}(\boldsymbol{X}, Y)$ where $\boldsymbol{X}$ are observations and Y is a class. The observations is a set of random variables. The main assumption made by the model is that the observations are independent given the class. This assumption can easily be represented graphically (see figure 3.3). It is easy to check that Y d-separates any pair of observations.

From the graph of the naive Bayes, we can read the factorization for the model:

$$\mathrm{P}(\boldsymbol{X}, Y) = \mathrm{P}(Y) \prod_i \mathrm{P}(X_i \mid Y) \tag{3.8}$$

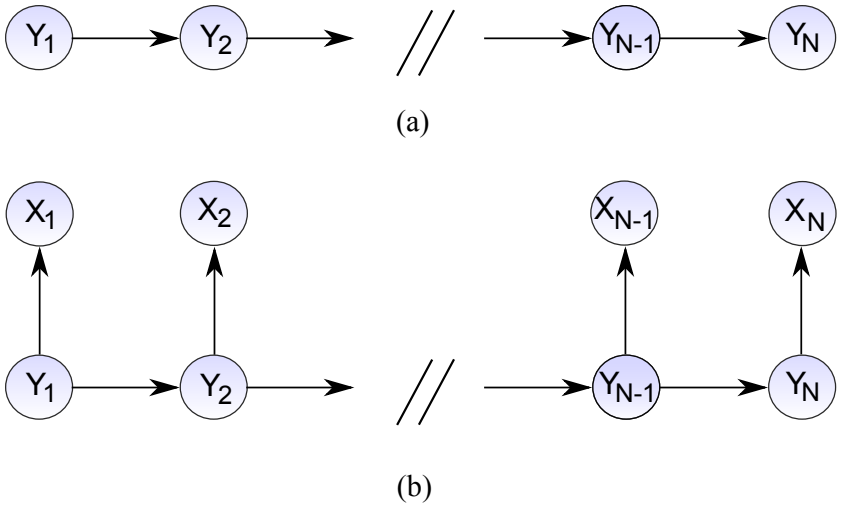

Figure 3.4: The figure depicts the Bayesian Networks representing a Markov Model (a) and a Hidden Markov Model (b).

$\mathrm{P}(Y)$ is usually called the *a-priori* probability of Y and, as Y is discrete, can be modelled by a multinomial distribution. In order to fully determine this model, we need to determine the conditional distribution $\mathrm{P}(X_i \,|\, Y)$ of the observations. In our work, we used either a multinomial distribution if X_i was discrete or a normal distribution in the case of continuous observations.

3.2.4 Hidden Markov Models

The example presented in Section 3.1, known as *Markov Model*, can be thought of as a Bayesian Network where $Pa(Y_t) = \{Y_{t-1}\}$:

$$\mathrm{P}(Y_1, \ldots, Y_N) = \mathrm{P}(Y_1) \prod_{i=2}^{N} \mathrm{P}(Y_i \,|\, Y_{i-1}) = \prod_{i=1}^{N} \mathrm{P}(Y_i \,|\, \mathrm{Pa}(Y_i)), \tag{3.9}$$

The DAG corresponding to this distribution is a linear chain of random variables.

An important related model is the Hidden Markov Model (HMM) [81, 82], where the variables can be split into two sets, the states $\boldsymbol{Y}$ and the observations $\boldsymbol{X}$:

$$\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{P}(Y_1)\, \mathrm{P}(X_1 \,|\, Y_1) \prod_{t=2}^{N} \mathrm{P}(X_t \,|\, X_{t-1})\, \mathrm{P}(X_t \,|\, Y_t) \tag{3.10}$$

where the terms $\mathrm{P}(X_t \,|\, X_{t-1})$ are called *transition probabilities*, the terms $\mathrm{P}(X_t \,|\, Y_t)$ are

called *emission probability functions*, and the term $\mathrm{P}(Y_1)$ is called *initial state probability*. The underlying assumptions are the Markov Property for the states and, for what concerns the observations, the conditional independence of one observation with respect to all of the others given the state at the same time.

HMMs have been used extensively for both classification and labeling problems. In the first case, different sequences of states $\boldsymbol{Y}_i$ are used to estimate the probability $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y}_i)$ and the one leading to the highest value is retained as the winning one:

$$k = \arg \max_{i \in [1,N]} = \mathrm{P}(\boldsymbol{X}, \boldsymbol{Y}_i), \tag{3.11}$$

where k is assigned to $\boldsymbol{X}$ as class. In the labeling case, the sequence of states $\hat{\boldsymbol{Y}}$ that satisfies the following equation:

$$\hat{\boldsymbol{Y}} = \arg \max_{\boldsymbol{Y} \in \mathcal{Y}} = \mathrm{P}(\boldsymbol{X}, \boldsymbol{Y}), \tag{3.12}$$

is used to label the observations of $\boldsymbol{X}$ ($\mathcal{Y}$ is the set of the state sequences of the same length as X). Each element X_t is labeled with the value y_t of variable $\hat{Y}_t$ in $\hat{\boldsymbol{Y}}$.

3.3 Conditional Random Fields

Conditional Random Fields [71, 72, 73] differ from Bayesian Networks under two main respects: The first is that they are encompassed by undirected graphical models, the second is that they are *discriminative*, i.e. they model $\mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{X})$ and not $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})$. The former aspect influences the factorization as well as the way the graph encodes conditional independence assumptions. The latter aspect brings the important advantage that no assumptions about $\boldsymbol{X}$ need to be made (see below for more detail).

3.3.1 Factorization and Conditional Independence

Definition 11. Let $G = (\boldsymbol{N}, \boldsymbol{E})$ be a graph such that the random variables in $\boldsymbol{Y}$ correspond to the nodes of G and let P be a joint probability distribution defined over $\boldsymbol{Y}$. A pair (G, P)

is a Markov Random Field if:

$$\mathrm{P}(Y \mid \boldsymbol{Y} \setminus \{Y\}) = \mathrm{P}(Y \mid \mathrm{Nb}(Y)) \forall Y \in \boldsymbol{Y}. \tag{3.13}$$

The factorization of P is given by the following theorem:

Theorem 2. *Le (G, P) be a Markov Random Field, then there exists a set of functions $\{\varphi_c \mid c$ is a clique of $G\}$ such that*

$$\mathrm{P}(\boldsymbol{Y}) = \frac{1}{Z} \prod_c \varphi_c(\boldsymbol{Y}|c), \tag{3.14}$$

where $\boldsymbol{Y}|c$ is the subset of $\boldsymbol{Y}$ that includes only variables associated to the nodes in c, and Z is a normalization constant:

$$Z = \sum_{\boldsymbol{y}} \prod_c \varphi_c(\boldsymbol{y}|c), \tag{3.15}$$

where $\boldsymbol{y}$ iterates over all possible assignments on $\boldsymbol{Y}$.

The functions φ_c are often called potentials. They need to be positive functions but they do not necessarily need to be probabilities, i.e. they are not bound to range between 0 and 1. The conditional independence assumptions underlying the factorization above can be inferred by considering the definition of the Markov Network. Each variable is conditionally independent of all of the others given the variables that correspond to the nodes in its neighborhood: $\mathrm{P} \models (Y \perp \boldsymbol{Y} \setminus \{Y, Nb(Y)\} \mid Nb(Y))$.

Conditional Random Fields are based on Markov Networks and are defined as follows:

Definition 12. Let $G = (\boldsymbol{N}, \boldsymbol{E})$ be a graph such that the random variables in $\boldsymbol{Y}$ correspond to the nodes of G. The pair $(\boldsymbol{X}, \boldsymbol{Y})$ is a *Conditional Random Field* (CRF) if the random variables in $\boldsymbol{Y}$ obey the Markov property with respect to the graph G when conditioned on X:

$$\mathrm{P}(Y \mid \boldsymbol{X}, \boldsymbol{Y} \setminus Y) = \mathrm{P}(Y \mid \boldsymbol{X}, Nb(Y)). \tag{3.16}$$

the variables in $\boldsymbol{X}$ are called *observations* and those in $\boldsymbol{Y}$ *labels*.

The definition above does not require any assumption about $\boldsymbol{X}$ and this is an important

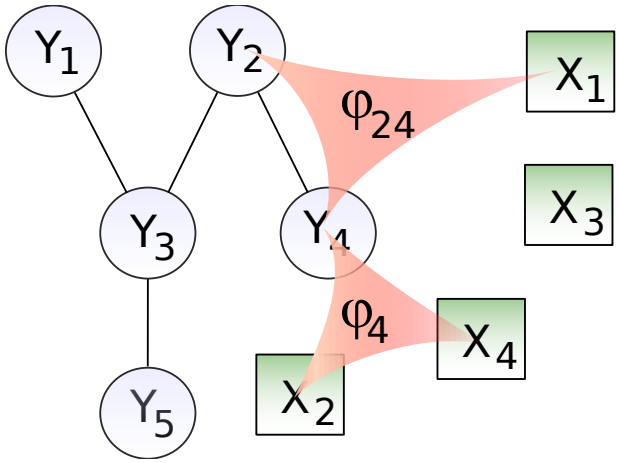

Figure 3.5: Conditional Random Fields. The potentials are defined over cliques and have as argument the variables corresponding to the nodes of the clique and an arbitrary subset of the observation sequence X.

advantage. In both labelling and classification problems, $\boldsymbol{X}$ is a constant and the value of $\mathrm{P}(\boldsymbol{X}, \boldsymbol{Y})$ must be maximized with respect to $\boldsymbol{Y}$:

$$\hat{\boldsymbol{Y}} = \arg\max_{\boldsymbol{Y}} \mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{X})\, \mathrm{P}(\boldsymbol{X}) = \arg\max_{\boldsymbol{Y}} \mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{X}) \tag{3.17}$$

Thus, modelling explicitly $\boldsymbol{X}$ (as it is done, e.g., in Hidden Markov Models) is not really necessary. The model does not require conditional independence assumptions for the observations that might make the models too restrictive for the data and negatively affect the performance. In this respect, modelling $\mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{X})$ makes the model more fit to the actual needs of labelling and classification (see equation above) and limits the need of conditional independence assumptions to the only $\boldsymbol{Y}$.

The factorization of Conditional Random Fields is as follows:

Theorem 3. *Let (G, P) be a Markov Network, then there exists a set of functions $\{\varphi_c \mid c$ is a clique of $G\}$ such that*

$$\mathrm{P}(\boldsymbol{y} \mid \boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x})} \prod_c \varphi_c(\boldsymbol{y}|c, \boldsymbol{x}). \tag{3.18}$$

Z is is a normalization constant called the partition function*:*

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \prod_c \varphi_c(\boldsymbol{y}|c, \boldsymbol{x}), \tag{3.19}$$

where $\boldsymbol{y}$ iterates over all possible assignments on $\boldsymbol{Y}$.

The problem left open so far is the definition of the potentials. As this chapter focuses on sequence analysis, the rest of this section will consider the particular case of *Linear Chain Conditional Random Fields*, one of the models most commonly applied for the sequence labelling problem.

3.3.2 Linear Chain Conditional Random Fields

In linear chain CRFs, the cliques are pairs of nodes corresponding to adjacent elements in the sequence of the labels or individual nodes (see Figure 3.6):

Definition 13. A graph is a *chain* if and only if $E = \{(y_i, y_{i+1}), 1 \leq i < |Y|\}$.

Where E is the set of the edges and (y_i, y_{i+1}) represents the edge between the nodes corresponding to elements Y_i and Y_{i+1} in $\boldsymbol{Y}$.

The following assumptions must be made about the potentials to make the model tractable:

1. The potential over $\{y_t, y_{t+1}\}$ depends only on y_t and y_{t+1}.
2. The potential over $\{y_t\}$ depends only on y_t and x_t.
3. The potentials are the same for all t.
4. The potentials are never zero.

This first three assumptions mean that the marginal distribution for y_t is fully determined by y_{t-1}, y_{t+1} and x_t. The fourth assumption means that every sequence of labels $\boldsymbol{Y}$ has a probability strictly greater than zero. This last assumption is important in practice because it allows the product of potentials to be replaced by the exponential of a sum as follows [71] :

$$\mathrm{P}(Y|X) = \frac{\exp\left(\sum_{t=1}^{N} f_1(y_t, \vec{x}_t) + \sum_{t=1}^{N-1} f_2(y_t, y_{t+1})\right)}{Z(X)}$$

$$Z(X) = \sum_{Y \in \mathcal{Y}^N} \exp\left(\sum_{t=1}^{N} f_1(y_t, \vec{x}_t) + \sum_{t=1}^{N-1} f_2(y_t, y_{t+1})\right)$$

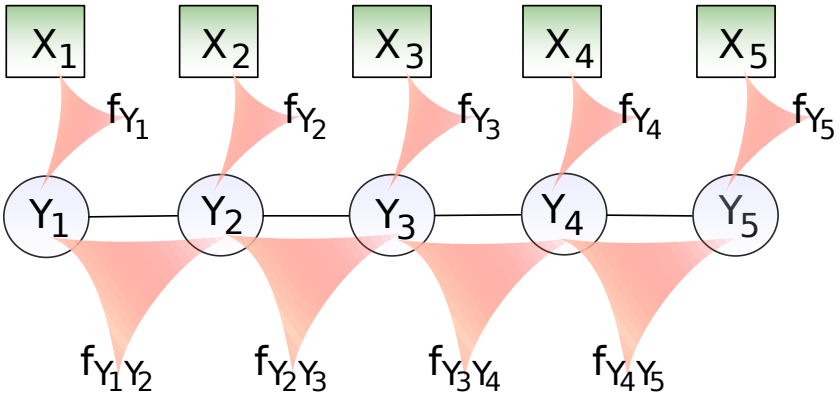

Figure 3.6: Linear Chain Conditional Random Fields. The cliques in a chain are pair of adjacent labels or individual labels. The potentials are function of adjacent nodes or of a node and the corresponding observation.

where f_1 and f_2 represent potentials having as arguments only one label y_t or a pair of adjacent labels $\{y_t, y_{t+1}\}$. Thus, the potentials have been represented as a linear combination of simpler terms called *feature functions*.

In general, the feature functions used for f_1 are as follows:

$$f_{y,t}(y_t, \boldsymbol{x}) = \begin{cases} x_t & \text{if } y_t = y \\ 0 & \text{otherwise} \end{cases} \tag{3.20}$$

where x_t is the observation at time t. This family of feature functions can capture linear relations between a label and an observation x_t. For f_2, the feature functions are typically as follows:

$$f_{y,y'}(y_t, y_{t+1}) = \begin{cases} 1 & \text{if } y_t = y \text{ and } y_{t+1} = y' \\ 0 & \text{otherwise} \end{cases} \tag{3.21}$$

In summary, Linear Chain CRFs estimate $p(\boldsymbol{Y}|\boldsymbol{X})$ as follows:

$$p(\boldsymbol{Y}|\boldsymbol{X}, \alpha) = \frac{1}{Z(\boldsymbol{X})} \exp \left(\begin{array}{l} \sum_{t=1}^{N} \sum_{y \in \mathcal{Y}} \alpha_y f_{y,t}(y_t, x_t) + \\ \sum_{t=1}^{N-1} \sum_{(y,y') \in \mathcal{Y}^2} \alpha_{y,y'} f_{y,y'}(y_t, y_{t+1}) \end{array} \right) \tag{3.22}$$

The weights α_y of the feature functions of form $f_{y,t}(\boldsymbol{X}, \boldsymbol{Y})$ account for how much the value of a given observation is related to a particular label. The weights of the feature functions of

form $f_{y,y'}(\boldsymbol{X}, \boldsymbol{Y})$ account for how frequent it is to find label y followed by role y'.

3.4 Training and Inference

The models presented so far cannot be used without appropriate *training* and *inference* techniques. The training consists in finding the parameters of a model (e.g., the transition probabilities in a Hidden Markov Model or the α coefficients in a Conditional Random Field) that *better fit* the data of a training set, i.e. a collection of pairs $\mathcal{T} = \{(\boldsymbol{X}_i, \boldsymbol{Y}_i)\}$ $(i = 1, \ldots, |\mathcal{T}|)$ where each observation is accompanied by a label supposed to be true. By "better fit" it is meant the optimization of some criterion such as the maximization of the likelihood or the maximization of the entropy (see below for more details).

The inference consists in finding the value of $\boldsymbol{Y}$ that better fits an observation sequence $\boldsymbol{X}$, whether this means to find the individual value of each Y_j that better matches each $\boldsymbol{X}$:

$$\mathrm{P}(Y_j = y \mid \boldsymbol{X}) = \sum_{\boldsymbol{Y} \in \{\boldsymbol{Y}, Y_j = y\}} \mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{X}) \tag{3.23}$$

or finding the sequence $\hat{\boldsymbol{Y}}$ that globally better matches $\boldsymbol{X}$:

$$\hat{\boldsymbol{Y}} = \arg\max_{\boldsymbol{Y}} \mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{X}). \tag{3.24}$$

We will start by presenting the inference and training techniques for the Naive Bayes Model in Section 3.4.1. This model presents no major difficulties.

In the case of the other models, the main difficulty is that the number of possible sequences increases exponentially with the size of $\boldsymbol{Y}$. In the case of classification, this is not a problem as the size of$\boldsymbol{Y}$ is one and $\hat{Y}$ can be estimated directly from equation (3.24). However, in the case of labelling, direct enumeration is not possible and a more sophisticated approach is needed. The most commonly used approach is message passing that will be presented in Section 3.4.2.

The problem for the training arises from the fact that there is no known closed form expression for maximum likelihood parameters in the case of HMM and linear chain CRF. The

parameters have to be iteratively estimated in this case and we present in Section 3.4.2 the method used for linear chain CRF.

3.4.1 Naive Bayes Inference and Training

In the case of the Naive Bayes Model, the probability distribution defined by the model is given by:

$$\mathrm{P}(\boldsymbol{X}, Y) = \mathrm{P}(Y) \prod_i \mathrm{P}(X_i \,|\, Y) \tag{3.25}$$

and there is a finite number of discrete classes. In the case of inference, the observations X are known and thus finding the optimal value (the value that maximize Equation (3.24)) is usually trivial. It is possible to each class and select the best one:

$$\hat{Y} = \arg\max_Y \mathrm{P}(\boldsymbol{X} \,|\, Y) \tag{3.26}$$

For the training phase, we are given a training set $\{(\boldsymbol{X^{(i)}}, Y^{(i)})\}$ and we are interested in finding the model that maximizes the likelihood on this training set:

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \prod_i \left(\mathrm{P}(\boldsymbol{X^i}, Y^i \,|\, \boldsymbol{\alpha})\right) \tag{3.27}$$

$$= \arg\max_{\alpha} \prod_i \left(\mathrm{P}(Y^i \,|\, \alpha) \prod_j (X_i^j \,|\, Y_i, \alpha) \right) \tag{3.28}$$

where $\boldsymbol{\alpha}$ are the parameters of the model. In the case of the NB model, the parameters for the different conditional probability are independent and we can write $\boldsymbol{\alpha} = \{\alpha_0, \alpha_1, \ldots, \alpha_N\}$ where α_0 denote the parameters for the a-priori distribution and the α_j are the parameters for the conditional distribution of observation j. Those parameters can be optimized separately. The problem can be simplified in the following sub-problems:

$$\hat{\alpha}_0 = \arg\max_{\alpha} \prod_i \mathrm{P}(Y^i \,|\, \alpha_0) \tag{3.29}$$

$$\hat{\alpha}_k = \arg\max_{\alpha} \prod_i \mathrm{P}(X_j^i \,|\, Y^i \alpha_j) \tag{3.30}$$

Therefore, any probability distribution that can be learned by maximum likelihood can easily

be used as conditional distribution for a NB model. For example, if a multinomial distribution is used for $\mathrm{P}(Y)$, then the a-priori probability of a class is simply the frequency of that class in the training set.

3.4.2 Message Passing

We can now consider the more complicated case of the CRF. In that case, it is usually not possible to use a brute force approach and try all the possible values for the labels. We will see in this section that one of the main issues in both training and inference is to estimate the probability $\mathrm{P}(Y_j = y)$ that a given label Y_j takes the value y. The *Message Passing* algorithm allows one to perform such a task in an efficient way by exploiting the local structure of the graph around the node corresponding to Y_j (see [83] for an extensive survey of the subject). In particular, the key idea is that the marginal distribution of a node Y_j can be determined if the value of the variables corresponding to its neighbouring nodes are known. In practice, those values are unknown, but it is possible to estimate the *belief* that measures the relative probability of the different values. For this reason, the message passing algorithm is sometimes referred to as *belief propagation*.

This section will focus in particular on the message passing algorithm for Pairwise Markov Networks, namely Markov Networks where the cliques include no more than two nodes. While being an important constraint, still it includes cases of major practical importance such as chains, trees and grids (the Linear Chain Conditional Random Fields fall in this class).

The beliefs are defined as follows:

$$b_j(y_j) = \rho\varphi_j(y_j) \prod_{k \in Nb(Y_j)} m_{kj}(y_j) \tag{3.31}$$

where $\varphi_j(y_j)$ is the potential for node Y_j, m_{kj} is the message from node Y_k to node Y_j (see below for the definition of the messages), and ρ is a normalization constant (the beliefs must sum to 1). Formally, a belief is a function that maps each possible value of Y_j into a real number.

A message is another function that maps the value of one node into a real number and it

represents the influence that the sending node has on the receiving one:

$$m_{kj}(y_j) = \sum_{y_k} \left(\varphi_k(y_k)\varphi_{jk}(y_j, y_k) \prod_{n \in Nb(Y_k)\setminus\{Y_j\}} m_{nk}(y_k) \right) \tag{3.32}$$

where φ_{jk} is the potential of the clique including Y_j and Y_k (this equation motivates the name *sum-product* algorithm that it is used sometimes for this algorithm).

The belief propagation requires the variables to be ordered and this might create problems when a graph contain cycles. When cycles are absent (which is the case for the models considered in this chapter), the following procedures allow one to find a suitable ordering:

1. Choose a root node.
2. Compute messages starting at the leaf moving to the root.
3. Compute messages starting at the root, going to the leaves.

It is important to note that the value of the message is independent of the order in which the messages are passed.

At the end of the procedure, each node is associated with a belief that can be used to compute the marginal probabilities as shown by the following:

Theorem 4. *Let G be a pairwise random field on $\mathbf{Y}$ and b_j the beliefs computed using the message passing algorithm, then the following holds:*

$$\mathrm{P}(Y_j = y_j) = \frac{b_j(y_j)}{\sum_{y_i} b_j(y_i)}. \tag{3.33}$$

Furthermore, the normalization constant ρ (see above) can be computed as follows:

$$\rho = \sum_{y_j} b_j(y_j) \tag{3.34}$$

In the case of Conditional Random Fields, the observations in $\boldsymbol{X}$ have to be taken into account. The message and the beliefs are now dependent on $\boldsymbol{X}$:

$$b_j(y_j, \boldsymbol{X}) = \varphi_j(y_j, \boldsymbol{X}) \prod_{Y_k \in Nb(Y_j)} m_{kj}(y_j, \boldsymbol{X}) \tag{3.35}$$

$$m_{kj}(y_j, \boldsymbol{X}) = \sum_{y_k, \boldsymbol{X}} \left(\varphi_k(y_k, \boldsymbol{X}) \varphi_{jk}(y_j, y_k, \boldsymbol{X}) \prod_{Y_n \in Nb(Y_k) \setminus \{Y_j\}} m_{nk}(y_k, \boldsymbol{X}) \right) \tag{3.36}$$

$$\tag{3.37}$$

As $\boldsymbol{X}$ is a constant and it is known a-priori, it is possible to apply exactly the same equations as those used for the Markov Networks.

Inference

There are two possible inference scenarios (see beginning of this section): The first consists of finding, for each label, the assignment that maximizes the marginal probability. The second consists of finding the assignment that maximizes the joint probability distribution over the entire labels sequence $\boldsymbol{Y}$.

The first case is a straightforward application of the message passing algorithm. For a given label Y_j, it is sufficient to use the beliefs to find the particular value $\hat{y}$ that maximizes the following probability:

$$\hat{y} = \arg\max_y \mathrm{P}(Y_j = y) = \arg\max_y b_j(y). \tag{3.38}$$

It can be demonstrated that this particular way of assigning the values to the labels minimizes the misclassification rate.

In the second case, the expression of the messages in Equation (3.32) must be modified as follows:

$$m_{kj}(y_j) = \max_{y_k} \left(\varphi_k(y_k) \varphi_{jk}(y_j, y_k) \prod_{n \in Nb(Y_k) \setminus \{Y_j\}} m_{nk}(y_k) \right) \tag{3.39}$$

where the initial sum has been changed to a maximization. This ensures that the message received by the node corresponding to label Y_j brings information about the sequence $(Y_1, \ldots, Y_{j-1})$ with the highest possible probability rather than about the sum of the probabilities over all possible sequences.

It is again possible to assign to each Y_j, the value $\hat{y}_j$ that maximize the beliefs obtained using the modified messages:

$$\hat{y}_j = \arg\max_y b_j(y). \tag{3.40}$$

It can be shown that the resulting assignment $\hat{\boldsymbol{Y}} = \{\hat{y}_1, \ldots, \hat{y}_n\}$ is the sequence with the maximum probability:

$$\hat{\boldsymbol{Y}} = \arg\max_Y \mathrm{P}(Y) \tag{3.41}$$

Training

The last important aspect of probabilistic sequential models is the training. The topic is way too extensive to be covered in detail and the section will focus in particular on Markov Networks as this can be a good starting point towards training Conditional Random Fields. If the assumption is made that the potentials are strictly greater than zero, then Markov Networks can be factorized as follows:

$$\mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{\alpha}) = \frac{1}{Z} \exp\left(\sum_c \sum_{i=1}^{n_c} \alpha_c^i f_c^i(Y|c)\right) \tag{3.42}$$

$$Z = \sum_Y \exp\left(\sum_c \sum_{i=1}^{n_c} \alpha_c^i f_c^i(Y|c)\right) \tag{3.43}$$

where the $f_c^i(Y|c)$ are feature functions defined over a clique c. The same expression as the one used for Conditional Random Fields, but without the observations $\boldsymbol{X}$.

Training such a model means to find the values of the coefficients α that optimize some criteria over a training set. This section considers in particular the maximization of the likelihood:

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \sum_j \log \mathrm{P}(\boldsymbol{Y^j} \mid \boldsymbol{\alpha}) \tag{3.44}$$

where the $\boldsymbol{Y}_j$ are the sequences of the training set.

The main problem is that solving the above equation leads to an expression for the $\boldsymbol{\alpha}$ coefficients which is not in closed form, thus it is necessary to apply gradient descent techniques. On the other hand, these are effective because of the following:

Theorem 5. *The log-likelihood function is convex with respect to the weights.*

In practice, the LBFGS algorithm [84] works well and this has two main motivations: The first is that the algorithm approximates the second derivative and thus converges faster, the second is that it has a low memory usage and works well on large scale problems. One of the main steps of the LFBGS is the estimation of the derivative of the log-likelihood with respect to $\boldsymbol{\alpha}$.

$$\frac{\partial}{\partial \alpha_i^c} \sum_j \log \mathrm{P}(Y^j) = \frac{\partial}{\partial \alpha_i^c} \sum_j \log \left(\frac{1}{Z} \exp \left(\sum_c \sum_{i=1}^{n_c} \alpha_c^i f_c^i(\boldsymbol{Y^j}|c) \right) \right) \tag{3.45}$$

$$= \frac{\partial}{\partial \alpha_i^c} \sum_j \left(\sum_c \sum_{i=1}^{n_c} \alpha_c^i f_c^i(\boldsymbol{Y^j}|c) \right) - \frac{\partial}{\partial \alpha_i^c} \sum_j \log Z \tag{3.46}$$

$$= \sum_j \left(f_c^i(\boldsymbol{Y_j}|c) - E[f_c^i] \right) \tag{3.47}$$

The equation above shows that the optimal solution is the one where the theoretical expected value of the feature functions is equal to their empirical expected value. This corresponds to the application of the Maximum Entropy Principle and it further explains the close relationship between Conditional Random Fields and Maximum Entropy Principle introduced in this section.

3.5 Conclusions

This chapter has introduced the graphical model used for role recognition and more generally for sequence analysis in machine learning. The problem has been formulated in terms of two major issues, namely classification (assigning a label to an entire sequence of observation) and labelling (assigning a label to each observation in a sequence). This chapter has introduced some of the most important statistical models for sequence analysis, Hidden Markov Models and Conditional Random Fields as well as the Naive Bayes Models used in some of our experiments. The unifying framework of Probabilistic Graphical Models has been used in both cases and the accent has been put on factorization and conditional independence assumptions. Some details about training and inference issues have been provided for Conditional Random Fields and, more generally, for undirected graphical models.

The models introduced in this chapter are not aimed in particular at human behaviour understanding, but they have been used successfully in the domain (see [4] for an extensive survey

of the domain). Sequences arise naturally in many behaviour analysis problems, especially in the case of social interactions where two or more individuals react to one another and produce sequences of social actions [85].

While trying to provide an extensive description of the sequence analysis problem in machine learning, this chapter cannot be considered exhaustive. However, the chapter, and the references therein, can be considered a good starting point towards a deeper understanding of the problem. In particular, graphical models have been the subject of both tutorials (see, e.g., [86] and Chapter 8 of [32]) and dedicated monographies [71], the same applies to Hidden Markov Models (see, e.g., [82] for a tutorial and [81] for a monography) and Conditional Random Fields (see, e.g., [87] for a tutorial and [71] for a monography).

Last, but not least, so far Human Sciences and Computing Science (in particular machine learning) have looked at the sequence analysis problem in an independent way. As the cross-pollination between the two domains improves, models more explicitly aimed at the human behaviour understanding problem are expected to be developed.

Chapter 4

Modelling Role Dependency in Automatic Role Recognition

In this chapter, we will present our first work on automatic analysis of social interaction. The techniques and approach presented in this chapter were published in :

- S. Favre, H. Salamin, J. Dines, and A. Vinciarelli. Role Recognition in Multiparty Recordings using Social Affiliation Networks and Discrete Distributions. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, pages 29–36, 2008.

- A. Vinciarelli. Speakers role recognition in multiparty audio recordings using Social Network Analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(6):1215–1226, 2007.

My main contribution was on the introduction of models for the dependency between the roles. In previous works, the role assigned to each speaker were considered independent in order to make the model used tractable.

This chapter will also introduce the methodology used in the rest of the thesis as well as a presenting the data-set used. The approach includes three main stages (see Figure 4.1). The first stage is the *Speaker Diarization*. The second stage is the *feature extraction* which, in this experiment, involves the automatic construction of a Social Affiliation Network [25] as well as its conversion into feature vectors that represent each person in terms of her relationships

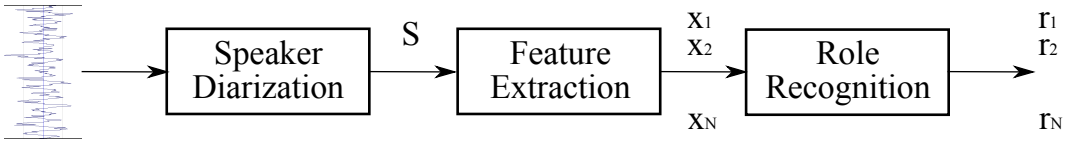

Figure 4.1: Role recognition approach. The picture shows the three main stages of the approach: the speaker diarization, the features extraction and the role recognition.

with the others. The third stage is the *role recognition*, i.e. the mapping of the feature vectors extracted in the first stage into roles belonging to a predefined set. In this case, the model used is the Naive Bayes Model, where the conditional probability for the observations are Bernoulli or Multinomial distributions [32] for the Affiliation Network features and Gaussian distributions for the intervention lengths associated to each role.

The experiments were performed over three different corpora (see Section 4.1 for more details):

- a collection of radio news bulletins (around 20 hours),
- a dataset of radio talk-shows (around 25 hours),
- the AMI meeting corpus (around 45 hours) [45]

For the first two datasets, the accuracy (i.e. the percentage of recording time correctly labelled in terms of role) ranged from 60 to 85%, for the third dataset the accuracy was approximately 45%. A possible explanation of the difference is that roles are easier to model when they are *norms*, i.e. correspond to functions that impose more or less rigorous constraints on the way people behave and interact with the others (like in the case of broadcast data). In contrast, roles are harder to model when they are *beliefs and preferences*, i.e. when they correspond to a position in a given social system (e.g. manager in a company) and do not necessarily impose tight constraints on the way people behave and interact (like in the case of the meetings). However, the accuracy significantly outperforms chance for both broadcast and meeting recordings.

The rest of the chapter is organized as follows: Section 4.1 presents the datasets used in the experiments, Section 4.2 describes the feature extraction stage, Section 4.3 describes the role recognition stage and Section 4.4 presents experiments and results.

Corpus	AM	SA	GT	IP	HP	WM
C1	41.2%	5.5%	34.8%	4.0%	7.1%	6.3%

Table 4.1: Role distribution in broadcast data. The table reports the percentage of data time each role accounts for in C1.

4.1 Corpora

This section presents the datasets used in the experiments:a collection of radio news bulletins (Section 4.1.1), a dataset of radio talk-shows (Section 4.1.2), and the AMI meeting corpus (Section 4.1.3).

4.1.1 Radio news bulletins

The first corpus, hereafter referred to as C1, contains 96 news bulletins with an average length of 11 minutes and 50 seconds. The corpus contains all the news bulletins broadcasted by *Radio Suisse Romande* (the French speaking Swiss National broadcasting service) during February 2005 and can thus be considered a representative sample of this kind of programs.

The roles in this corpus are: the *Anchorman* (AM), i.e. the person managing the program, the *Second Anchorman* (SA), i.e. the person supporting the AM, the *Guest* (GT), i.e. the person invited to report about a single and specific issue, the *Interview Participant* (IP), i.e. interviewees and interviewers, the *Headline Person* (HP), i.e. the speaker reading a short abstract at the beginning of the program, and the *Weatherman* (WM), i.e. the person reading the weather forecasts. The time distribution across roles is given in Table 4.1. Figure 4.2 shows the distribution of the number of persons across different recordings for the news bulletins and the distribution of the length of each recording is given in Figure 4.3.

4.1.2 Talk-shows

The second corpus, hereafter referred to as C2, contains 27 one hour long talk-shows broadcasted by *Radio Suisse Romande* (see above) during February 2005. This corpus can also be considered a representative sample of this specific kind of program.

The roles are: the *Anchorman* (AM), i.e. the person managing the program, the *Second*

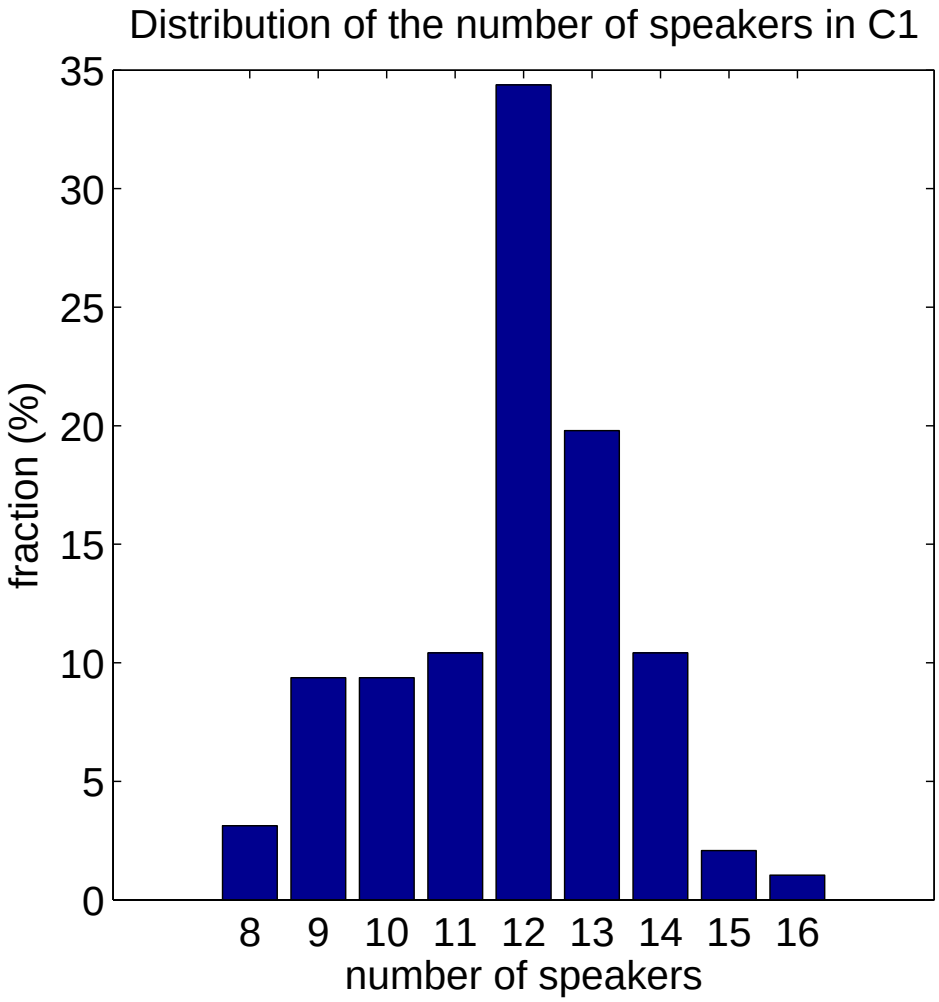

Figure 4.2: Distribution of recording participants. The histograms show the distribution of the number of people participating in each recording for corpora C1.

Anchorman (SA), i.e. the person supporting the AM, the *Guest* (GT), i.e. the person invited to report about a single and specific issue, the *Interview Participant* (IP), i.e. interviewees and interviewers, the *Headline Person* (HP), i.e. the speaker reading a short abstract at the beginning of the program, and the *Weatherman* (WM), i.e. the person reading the weather forecasts. However, even if the roles have the same name and correspond to roughly the same functions, they are played in a different way in C1 and C2 (e.g., consider how different is the behaviour of an anchorman in news supposed to inform and in talk-shows supposed to entertain). Table 4.2 shows the percentage of time that each role accounts for in talk-shows dataset. Figure 4.4 gives the distribution of the number of participants for the talk-shows.

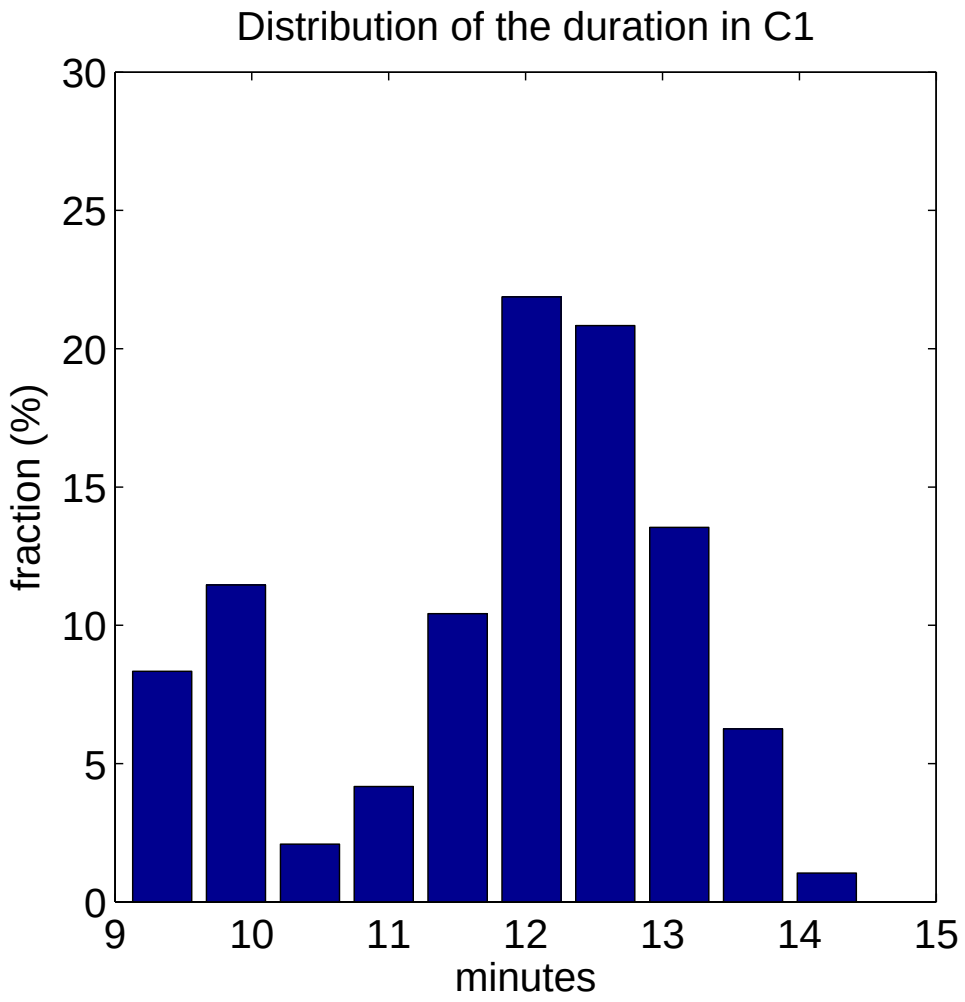

Figure 4.3: Distribution of the recording lengths. The histograms show the distribution of the recording lengths for news broadcasts.

4.1.3 AMI corpus

The third corpus, hereafter referred to as C3, is the AMI corpus [45]. This is a collection of 138 recorded meetings which contain a total of 45 hours and 38 minutes of material. The AMI meetings are based on a scenario where the participants are playing the roles of members of a team working on the development of a new remote control. The meetings are a *simulation*, the participants act roles they do not play in their real life. In C3, the role set contains the *Project Manager* (PM), the *Marketing Expert* (ME), the *User Interface Expert* (UI), and the *Industrial Designer* (ID). The role distribution for this corpus is given in Table 4.3. In this database, the number of participants is always 4. However, the length of every meeting is variable. Figure 4.5 give the distribution of the length of the meeting.

Corpus	AM	SA	GT	HP	WM
C2	17.3%	10.3%	64.9%	4.0%	1.7%

Table 4.2: Role distribution in broadcast data. The table reports the percentage of data time each role accounts for in C2.

Corpus	PM	ME	UI	ID
C3	36.6%	22.1%	19.8%	21.5%

Table 4.3: Role distribution in meetings. The table reports the percentage of data time each role accounts for in the AMI meeting corpus (C3).

The different corpora are compared in Table 4.4. For each corpus, we gave the total length, the type of roles, the number of roles present in the corpus and the number of participants in each interaction. C1 and C2 have very similar characteristics as they both contains data from radio broadcast. The main difference is the C1 contains news broadcasts whereas C2 is composed of talk-shows. For both corpora, the roles are related to the function of the speaker. C3 is the corpus of meeting. It has less speaker in each interaction and a smaller role set. In all the corpora, the roles are static. Each speaker play one and only one role for the whole duration of the interaction.

4.2 Features Extraction

This section presents the feature extraction stage aimed at extracting and representing the interaction pattern of each person. The stage includes two steps: the first is the segmentation of the recordings into single speaker segments (speaker diarization), the second is the extraction of a Social Affiliation Network from the resulting speaker sequence (see left dotted box in Figure 4.1).

The experiments involve two kinds of data: radio programs, where there is a single audio

Corpus	total length	type of roles	# of roles	# of participants
C1	$\sim$ 20 h.	norms	6	8 – 16
C2	27 h.	norms	5	22 – 44
C3	$\sim$ 45 h.	beliefs, preferences	4	4

Table 4.4: Overview of the corpora used in this thesis

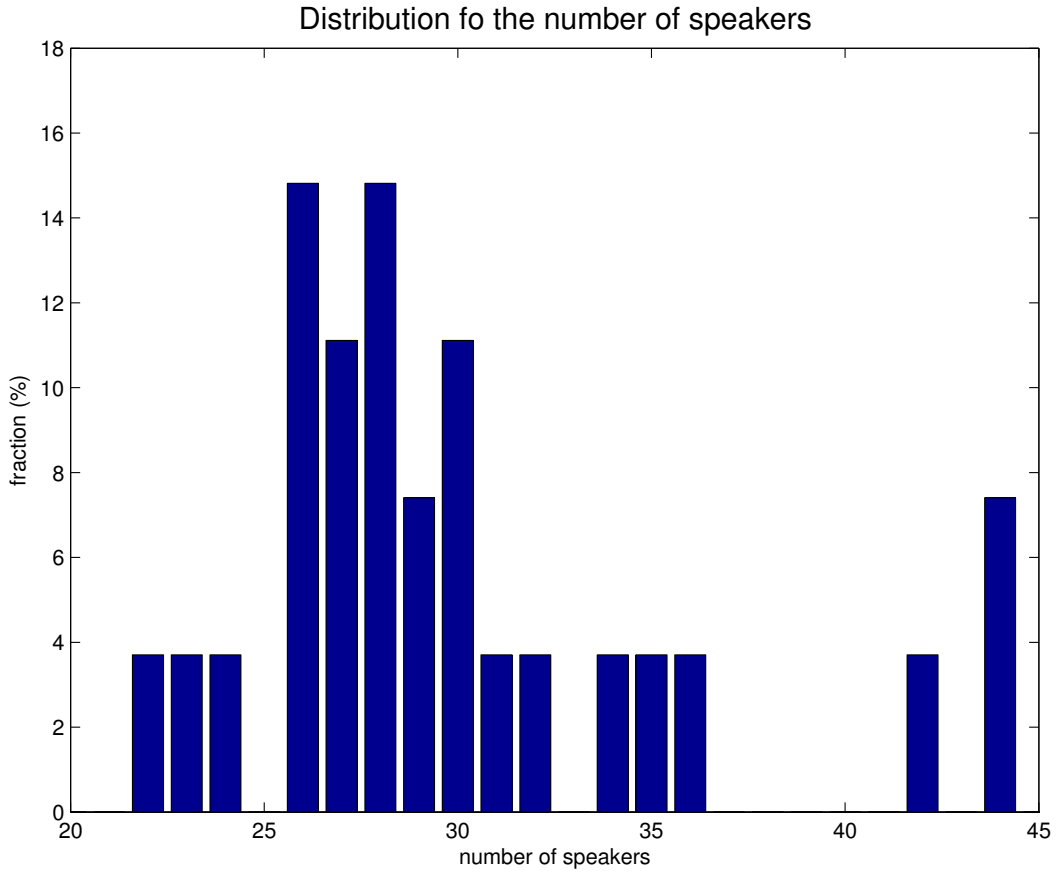

Figure 4.4: Distribution of recording participants. The histograms show the distribution of the number of people participating in each recording for corpora C2.

channel, and meeting recordings, where each participant wears a headset microphone. This requires the application of different speaker diarization techniques described in Section 4.2.2 and Section 4.2.3 for the radio and meeting data, respectively.

The next step is to use the segmentation of the audio to extract the features. The segmentation is also used to attribute roles to each audio segment.

4.2.1 Diarization

The techniques used to segment the data in speakers are only briefly described, as they are not the main element of interest in this work. The interested reader can refer to [88, 89, 90] for a full description. Section 4.2.4 shows how the output of the speaker diarization is used to build a Social Affiliation Network and represent people with tuples accounting for their interaction pattern.

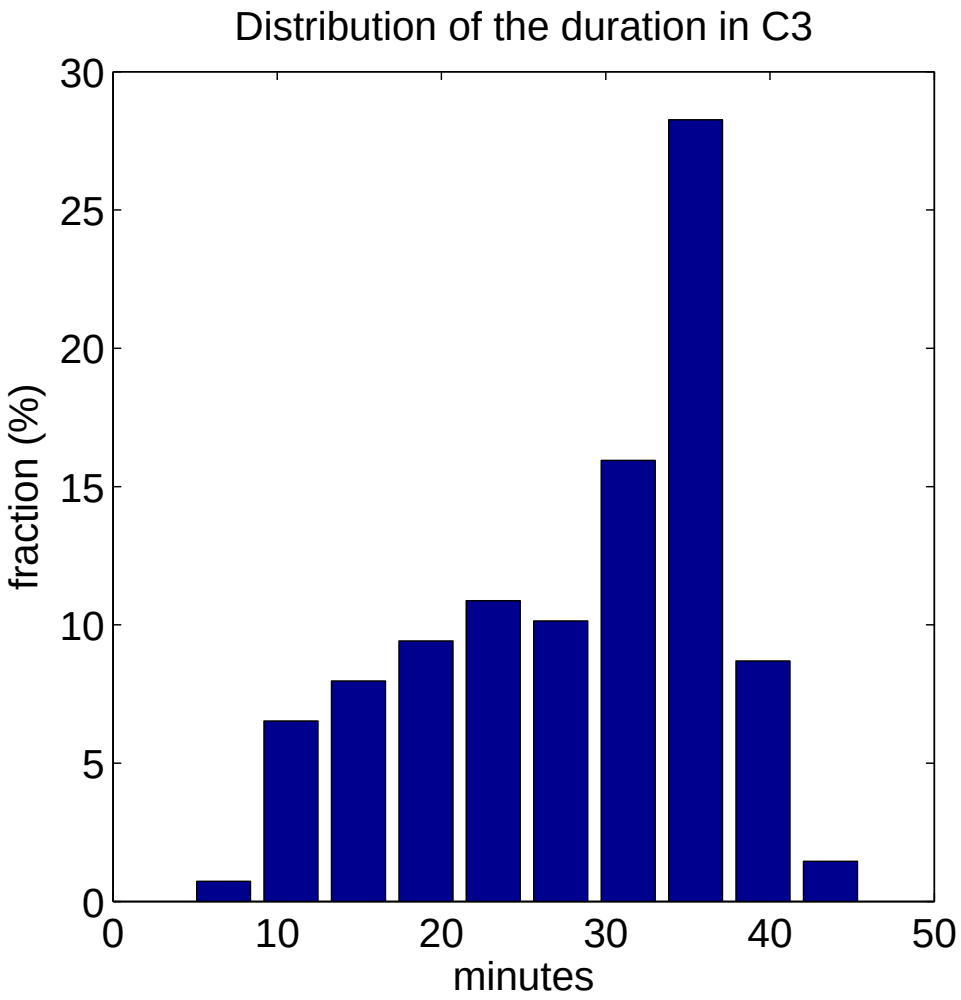

Figure 4.5: Distribution of the recording lengths. The histograms show the distribution of the recording lengths for the AMI meetings.

The first step is to detect the speakers. This step is essential in a fully automated approach, as we need on one hand to associate the features extracted to the corresponding person and also identify the segment of the interaction that belongs to each person. In this thesis, we focus only on audio and we will present the technique used to segment audio recordings. Depending on the devices used for the capture, there are two main cases to consider. In the first case, only one audio file is available for all the participants, either due to the use of far-field microphone or of production. In this case, the audio must be segmented in voices using clustering technique presented in Section 4.2.2. In the second case, one audio file is available per speaker due to the use of close talk microphones. In this case, one need only to remove the cross-talk and segment the audio of each participants in either speaking or not speaking. This technique is presented in Section 4.2.3. Another difference is the fact that in the second case, the number of participants in known a-priori.

4.2.2 Speaker Diarization for Broadcast Data

In the case of the radio programs (single audio channel), the diarization is performed with an unsupervised speaker clustering technique based on an ergodic HMM where each state corresponds to a cluster and, in principle, to a single voice. A full description of the algorithm is given in [88, 89].

The audio signal is first converted into a sequence of 12-dimensional observation vectors corresponding to the *Mel Frequency Cepstral Coefficients* (MFCC) extracted every $10\ ms$ from a $30\ ms$ long window [91]. MFCC features are used because they have, on average, higher performance in speaker recognition tasks (they are thus effective in capturing speaker voice characteristics). Furthermore, extensive experiments have shown that they lead to better results in speaker clustering experiments [91]. The observation sequence is then iteratively aligned with the ergodic HMM (see above) where the emission probabilities are modeled with Gaussian Mixture Models (GMM) [32]. At each iteration, the two most similar states (in terms of GMM parameters) are merged because they are supposed to correspond to the same voice. The number of parameters is kept constant from one iteration to the other so that the likelihood improves as long as the merged states actually correspond to a single voice, while it starts to decrease when states corresponding to different voices are merged. Reaching the likelihood peak is the stopping criterion. The approach does not use any *a-priori*

information about the number of speakers (thus about the number of necessary states in the HMM). The initial number of states is thus set arbitrarily to a value significantly higher than the expected number of speakers. In this way, after a sufficient number of iterations where similar states are merged, the number of states is expected to correspond to the actual number of speakers.

4.2.3 Speaker Diarization for Meeting Data

In the meeting recordings, the diarization can be performed by simply segmenting into speech and non-speech the output of the headset microphones that each of the meeting participants wears. A full description of the approach used for this task is given in [90]. In summary, the approach employs a *Multilayer Perceptron* (MLP) for estimating the posterior probability of audio frames of belonging to speech or non-speech classes. The frames are represented with feature vectors including 12 MF-PLP features, and features specifically designed for the detection of cross-talk in headset microphone recordings, as this has been found to be a major source of segmentation errors in meeting data [92].

The segmentation is carried out using HMMs where the states correspond to speech and non-speech. Minimum duration and insertion penalty constraints are applied to ensure that the segmentation is consistent with that observed for the groundtruth. Emission probabilities for the HMM states are estimated as scaled likelihoods in which MLP posterior probabilities are divided by their respective prior class probability.

4.2.4 Affiliation Network Extraction

The result of the speaker diarization process is that each recording is split into a sequence of turns. Each turn is encoded as a start time, an end time and a set of labels, one for each speaker speaking during the turn. Depending on the type of social interaction, the amount of over-lapping speech varies. This representation is not unique but the goal is to be able to determine for every time t who is speaking. In our work, we only consider turns that last at least 2 seconds, so that any short pause in the speech are not encoded. For every recording, there is also a manual segmentation available.

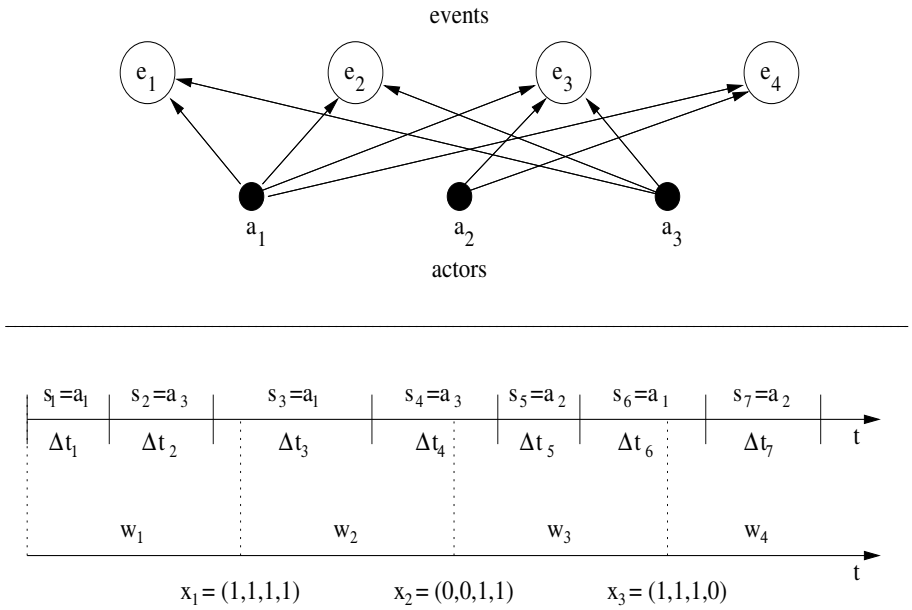

Figure 4.6: Interaction pattern extraction. The picture shows the Social Affiliation Network extracted from a speaker segmentation. The events of the network correspond to the segments w_j and the actors are linked to the events when they talk during the corresponding segment. The actors are represented using tuples $\vec{x}_a$ where the components account for the links between actors and events.

The result of the speaker diarization process is that each recording is split into a sequence $S = \{(s_i, \Delta t_i)\}$, where $i = 1, \ldots, |S|$, s_i is the label assigned to the speaker voice detected in the i^{th} segment of audio, and Δt_i is the duration of the i^{th} segment. The label s_i belongs to the set A of unique speaker labels, output by the speaker diarization process (see lower part of Figure 4.6). The sequences extracted from the speaker diarization are used to create a Social Affiliation Network (SAN) representing the relationships between the roles. A SAN is a graph with two kinds of nodes: the *actors* and the *events* [25]. Actors can be linked to events, but no links are allowed between nodes of the same kind (see upper part of Figure 4.6). In the experiments, the actors correspond to the people involved in the recordings, and the events correspond to uniform non-overlapping segments spanning the whole length of the recordings. In this work, the events do not have any meanings. They are introduced in order to discretize the time in the interaction. The length of the events was selected using cross-validation.

The rationale behind this choice of features is that actors speaking in the same interval of time are more likely to talk with one another (i.e. of interacting with one another) than actors speaking in different intervals of time. Thus, the SAN encodes information about *who*

interacts with whom and when. Research in psychology [25] has shown that the way people interact is related to the roles the play.

One of the main advantages of this representation is that each actor a can be represented by a tuple $\mathbf{x}_a = (x_{a1}, \ldots, x_{aD})$, where D is the number of segments used as events and the component x_{aj} accounts for the participation of the actor a in the j^{th} event. The experiments used two kinds of representation. In the first, component x_{aj} is 1 if the actor a talks during the j^{th} segment and 0 otherwise (the corresponding tuples are shown at the bottom of Figure 4.6). In the second, x_{aj} is the number of times that actor a talks during the j^{th} segment. In the first case the tuples are binary, in the second case they have integer components higher or equal to 0. In both cases, people that interact more with each other tend to talk during the same segments and are represented by similar tuples. If the roles influence the structure of the relationships between people, similar tuples should correspond to the same role.

4.3 Role Recognition

The problem of role recognition can be formalized as follows: given a set of actors A and a set of roles $\mathcal{R}$, find the function $\varphi : A \to \mathcal{R}$ mapping the actors into their actual role. In other words, the problem corresponds to finding the function φ such that $\varphi(a)$ is the role of actor a. In our problem, the roles are statics and therefor the function φ map each speaker in each interaction to only one role.

Section 4.2 has shown that the interaction pattern of each actor a is represented with a tuple $\mathbf{x}_a = (x_{a1}, \ldots, x_{aD})$, where D is the number of segments, that can have either binary or positive integer components. Furthermore, every actor a talks for a fraction τ_a of the total time of the recording. Thus, each actor corresponds to a couple $\mathbf{y}_a = (\tau_a, \mathbf{x}_a)$.

Given a function $\varphi : A \to \mathcal{R}$ and the set of observations $Y = \{\mathbf{y}_a\}_{a \in A}$, the problem of assigning a role to each actor can be thought of as the maximization of the *a-posteriori* probability $\mathrm{P}(\varphi \,|\, Y)$. By applying the Bayes Theorem and by taking into account that $\mathrm{P}(Y)$ is constant during the recognition, this problem is equivalent to finding $\hat{\varphi}$ such that:

$$\hat{\varphi} = \arg \max_{\varphi \in \mathcal{R}^A} \mathrm{P}(Y \,|\, \varphi)\, \mathrm{P}(\varphi). \tag{4.1}$$

where $\mathcal{R}^A$ is the set of all possible functions mapping actors into roles.

In order to simplify the problem, two assumptions are made: the first is that the observations are mutually conditionally independent given the roles. The second is that the observation $\mathbf{y}_a$ of actor a only depends on its role $\varphi(a)$ and not on the role of the other actors. Equation (4.1) can thus be rewritten as:

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} \mathrm{P}(\varphi) \prod_{a \in A} \mathrm{P}(\mathbf{y}_a \,|\, \varphi(a)). \tag{4.2}$$

The above expression is further simplified by assuming that the speaking time τ_a and the interaction tuples $\mathbf{x}_a$ of actors a are statistically independent given the role $\varphi(a)$, thus the last equation becomes:

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} \mathrm{P}(\varphi) \prod_{a \in A} \mathrm{P}\left(\mathbf{x}_a \,|\, \varphi(a)\right) \mathrm{P}\, \tau_a \,|\, \varphi(a)). \tag{4.3}$$

The probabilities appearing in the last equation have been estimated using different models to take into account the two representations of $\mathbf{x}_a$ described above, and to model the constraints in the distribution of roles (e.g. there must be only one *anchorman* in a given talk-show), i.e. to explicitly take into account the dependence between the roles.

The next sections show how $\mathrm{P}(\mathbf{x}_a \,|\, \varphi(a))$, $\mathrm{P}(\tau_a \,|\, \varphi(a))$, and $\mathrm{P}(\varphi)$ are estimated in the experiments.

4.3.1 Modeling Interaction Patterns

This section shows how the probability $\mathrm{P}(\mathbf{x}_a \,|\, \varphi(a))$ is estimated for both binary and multinomial tuples $\mathbf{x}_a$ (see Section 4.6).

When the components of the tuple $\mathbf{x}_a$ are binary, i.e. $x_{aj} = 1$ when actor a talks during segment j and 0 otherwise, The most natural way of modelling $\mathbf{x}_a$ is to use independent Bernoulli discrete distributions:

$$\mathrm{P}(\mathbf{x} \,|\, \mu) = \prod_{j=1}^{D} \mu_j^{x_j} (1 - \mu_j)^{1 - x_j}, \tag{4.4}$$

where D is the number of events in the network (see Section 4.2), and $\mu = (\mu_1, \ldots, \mu_D)$

is the parameter vector of the distribution. A different Bernoulli distribution is trained for each role. The maximum likelihood estimates of the parameters μ_r for a given role r are as follows [32]:

$$\mu_{rj} = \frac{1}{|A_r|} \sum_{a \in A_r} x_{aj}, \tag{4.5}$$

where A_r is the set of actors playing the role r in the training set, and $\mathbf{x}_a$ is the tuple representing the actor a.

When the components $\mathbf{x}_j$ correspond to the number of times that actor a talks during event j, i.e. when the components are integers greater or equal to 0, they can be represented with a vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{iT})$ where T is the maximum number of times that an actor can talk during a given event, $z_{ij} \in \{0, 1\}$, and $\sum_{j=1}^{T} z_{ij} = 1$. In other words, x_i is represented with a T-dimensional vector where all the components are 0 except one, i.e. the component $z_{in} = 1$, where n is the number of times that the actor represented by $\mathbf{x}$ talks during event i. As a result, $\mathbf{x}$ is represented as a tuple of vectors $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_D)$ and can be modeled as a product of independent Multinomial distributions:

$$\mathrm{P}(\mathbf{z} \mid \mu) = \prod_{i=1}^{D} \prod_{j=1}^{T} \mu_{ij}^{z_{ij}}. \tag{4.6}$$

The parameters μ can be estimated by maximizing the likelihood of $\mathrm{P}(\mathbf{z} \mid \mu)$ over a training set $\mathcal{X}$. This leads to a closed form expression for the parameters:

$$\mu_{ij} = \frac{1}{|A_r|} \sum_{a \in A_r} z_{aij}, \tag{4.7}$$

where A_r is the set of actors playing role r.

4.3.2 Modeling Durations

Given a labeled training set, there is a set A_r of actors playing role r, $\mathrm{P}(\tau \mid r)$ is estimated using a Gaussian Distribution $\mathcal{N}(\tau \mid \mu_r, \sigma_r)$, where μ_r and σ_r are the sample mean and variance respectively:

$$\mu_r = \frac{1}{|A_r|} \sum_{a \in A_r} \tau_a, \tag{4.8}$$

$$\sigma_r = \frac{1}{|A_r|} \sum_{a \in A_r} (\tau_a - \mu_r)^2. \tag{4.9}$$

This corresponds to a Maximum Likelihood estimate, a different Gaussian distribution is obtained for each role.

4.3.3 Estimating Role Probabilities

This subsection describes how the *a-priori* probability $\mathrm{P}(\varphi(a))$ of actor a playing role $\varphi(a)$ is estimated. Two approaches are proposed: the first is based on the assumption that roles are independent and does not take into account the constraints that the role distribution across different participants in a given recording must respect, e.g. there is only one *Anchorman* in a talk-show, there is only one *Project Manager* in a meeting, etc. The second approach considers the roles dependent and takes into account the above constraints.

Modeling Independent Roles

The first approach assumes that the roles are independent and thus that $\mathrm{P}(\varphi)$ is simply the product of the a-priori probabilities of the roles assigned through φ to the different actors:

$$\mathrm{P}(\varphi) = \prod_{a \in A} \mathrm{P}(\varphi(a)) \tag{4.10}$$

The a-priori probability of observing the role r can be estimated as follows:

$$\mathrm{P}(\varphi(a)) = \frac{N_{\varphi(a)}}{N}, \tag{4.11}$$

where N and $N_{\varphi(a)}$ are the total number of actors and the total number of actors playing role $\varphi(a)$ in the training set.

Using the above approach, Equation (4.2) boils down to

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} \prod_{a \in A} \mathrm{P}(\mathbf{x}_a \,|\, \varphi(a))\, \mathrm{P}(\tau_a \,|\, \varphi(a))\, \mathrm{P}(\varphi(a)). \tag{4.12}$$

and the role recognition process simply consists in assigning each actor the role $\varphi(a)$ that maximizes the probability $\mathrm{P}(\mathbf{x}_a \,|\, \varphi(a))\, \mathrm{P}(\tau_a \,|\, \varphi(a))\, \mathrm{P}(\varphi(a))$.

Modeling Dependent Roles

The second approach for the estimation of $\mathrm{P}(\varphi)$ takes into account the constraints that the role distribution in a given recording must respect, e.g. there must be only one *Anchorman* in a talk show while the number of *Guests* changes at each edition of the talk show. In this case, the roles played by the different recording participants cannot be considered independent, and $\mathrm{P}(\varphi)$ can not be written as the product of the a-priori probabilities of the roles (like in Equation 4.10).

A given mapping $\varphi \in \mathcal{R}^A$ corresponds to a distribution of roles across the different recording participants where each role is played by a certain number of actors. The constraints to be respected are expressed in terms of the number of actors that can play a given role (e.g., only one actor can be the *Anchorman*). Thus, $\mathrm{P}(\varphi)$ must be different from 0 only for those distributions of roles that respect the constraints. For some roles, the number of possible actors playing it is actually predetermined (i.e. exactly n_r actors must play role r), while for others the only available a-priori information is that at least one person must play the role (i.e. $n_r > 0$).

According to the above, $\mathrm{P}(\varphi)$ is modeled with a product of Multinomial distributions [32]:

$$\mathrm{P}(\varphi) = \prod_{r \in \mathcal{R}} \mathrm{P}(\mathbf{z}_r \,|\, \mu_r) \tag{4.13}$$

where $\mathbf{z}_r$ is a *one-out-of-K* (see Section 4.3.1) representation of the number of times a role can be played in a given recording, and μ_r is the parameter vector.

We can divide the set $\mathcal{R}^A$ in classes $\{C_g\}$ where all mappings lead to a role distribution where the same role is played always the same number of times. We assume that all mappings φ in the same class have the same probability. Thus, the probability of observing a given assignment is:

$$\mathrm{P}(\varphi) = \frac{\prod_{r \in \mathcal{R}} \mathrm{P}(\mathbf{z}_r \,|\, \mu_r)}{|C_g|}. \tag{4.14}$$

Then in the second model, Equation (4.2) can be rewritten as:

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} \mathrm{P}(\varphi) \prod_{a \in A} \mathrm{P}(\mathbf{x}_a \,|\, \varphi(a))\, \mathrm{P}(\tau_a \,|\, \varphi(a)). \tag{4.15}$$

where $\mathrm{P}(\varphi)$ is the expression of Equation 4.14. Maximizing this product using a brute-force approach is not tractable if the number of actors is high. Therefore, we used simulated annealing [93] to approximate the best mapping for each recording.

4.4 Experiments and Results

The next three sub-sections describe performance measures, experimental setup and role recognition results.

4.4.1 Speaker Diarization Results

The interaction patterns used at the role recognition step are extracted from the speaker segmentation obtained with the diarization process. Errors in the diarization (e.g. people detected as speaking when they are silent, or multiple voices attributed to a single speaker) lead to spurious interactions that can mislead the role recognition process.

The effectiveness of the diarization is measured with the *Purity* π, a metric showing on one hand to what extent all feature vectors corresponding to a given speaker are detected as belonging to the same voice, and on the other hand to what extent all vectors detected as a single voice actually correspond to a single speaker. The Purity ranges between 0 and 1 (the higher the better) and it is the geometric mean of two terms: the *average cluster purity* π_c and the *average speaker purity* π_s. The definition of π_c is as follows:

$$\pi_c = \sum_{k=1}^{N_c} \sum_{l=1}^{N_s} \frac{n_k}{N} \frac{n_{lk}^2}{n_k^2}, \tag{4.16}$$

where N is the total number of feature vectors, N_s is the number of speakers, N_c is the number of voices detected in the diarization process, n_{lk} is the number of vectors belonging to speaker l that have been attributed to voice k, and n_k is the number of feature vectors in voice k. The definition of π_s is as follows:

$$\pi_s = \sum_{l=1}^{N_s} \sum_{k=1}^{N_c} \frac{n_l}{N} \frac{n_{lk}^2}{n_l^2} \tag{4.17}$$

(see above for the meaning of the symbols).

The application of the speaker diarization process in the case of radio programs requires the setting of the initial number of states M in the fully connected Hidden Markov Model (see Section 4.2). The value of M must be significantly higher than the number of expected speakers for the diarization process to work correctly. In our experiments, we set *a-priori* $M = 30$ for C1 and $M = 90$ for C2. No other values have been tested. The average purity is 0.81 for C1 and 0.79 for C2. The average purity for C3 is 0.99. The difference in purity is explained by the different experimental conditions and methods used to obtain the speaker segmentation (see Sections 4.2.2 and 4.2.3 for more details).

4.4.2 Experimental Setup

The experiments are based on a K-fold crossvalidation approach [32]. The corpora are split into K equally sized parts of which $K-1$ are used as training set, while the remaining one is used as test set. Each of the K parts is used iteratively as test set so that the experiments can be performed over the whole dataset at disposition while still preserving a rigorous separation between training and test set. In the case of our experiments, $K = 5$ and each subset contains 20% of the data. The only hyperparameter to be set is the number D of segments used as events in the Social Affiliation Network. At each iteration of the K-fold crossvalidation, the value of D giving the highest role recognition results *over the training set* has been retained for test. *In this way, a rigorous separation between training and test set has been observed for the setting of the hyperparameter as well.*

Statistical significance of performance differences is assessed with the Kolmogorov-Smirnov test [38]. The advantage of this test is that it does not make assumptions about the distribution of the performance (unlike the t-test that assumes the performances following a Gaussian distribution) and it is adapted to continuous distributions (unlike the χ^2-test that requires the distributions to be made discrete through histogramming).

4.4.3 Role Recognition Results

Table 4.5 reports the results achieved for C1 and C2, Table 4.6 reports the results those obtained for C3. The performance is measured in terms of *accuracy*, i.e. the percentage of data time correctly labeled in terms of role in the test set. Each accuracy value is accompanied by

the standard deviation of the accuracies achieved over the different recordings of each corpus. The distribution used to model the interaction patterns is indicated with *B* (Bernoulli) and *M* (Multinomial). The approach used to estimate the *a-priori* role probabilities is indicated with *I* (independence) and *D* (dependence).

Modeling the dependence between roles leads to statistically significant improvements for C2 and C3, while it decreases the performance for C1. One probable explanation is that C1 presents more variability in the number of people playing a given role, thus $\mathrm{P}(\varphi)$ (see Section 4.3.3) cannot be estimated as reliably as for the other corpora. However, these results suggest that taking into account the dependence across roles is beneficial as long as $\mathrm{P}(\varphi)$ can be estimated reliably. To the best of our knowledge, this is the first attempt to model explicitly the dependence between roles and the results provide a first assessment of what can be expected, at least for the approach proposed here, in terms of performance improvement.

For the three corpora, *the differences between the performances achieved using Bernoulli and Multinomial distributions are not statistically significant*. This suggests that the information about the number of times a speaker talks during an event (conveyed by the Multinomial) does not add information with respect to the simple absence or presence (conveyed by the Bernoulli distribution). This is not surprising because the most important aspect encoded by Social Affiliation Networks (at least for the approach proposed in this work) is who interacts with whom and not how long someone interacts with someone else.

Overall, roles in meeting data appear to be harder to model for several reasons. On one hand, roles in meeting are based on *preferences and beliefs*, i.e. they correspond to a position in a given social system and do not correspond to stable behavioral patterns like in the case of the roles based on *norms* in broadcast data. On the other hand, the meetings in C3 are not real-world, i.e. the participants *act* in a scenario that does not correspond to their real lives. Not surprisingly, the meeting role recognized with highest accuracy is the *Project Manager* (PM). In fact, the PM plays also the role of *chairman*, i.e. a role based on *norms* that influences the actual interaction pattern of the people that play it.

The performance difference when passing from manual to automatic speaker diarizations is statistically significant for C1 and C2 (see Tables 4.5 and 4.6). The difference is not significant for C3 because the purity of the speaker segmentation for a such a corpus is 0.99, i.e. it corresponds almost perfectly to the groundtruth speaker segmentation. In contrast, the

Table 4.5: Role recognition performance for C1 and C2. The table reports both the overall accuracy and the accuracy for each role. "B" stands for *Bernoulli*, "M" stands for *Multinomial*, "I" stands for roles *Independence*, and "D" stands for roles *dependence*. The overall accuracy is accompanied by the standard deviation σ of the performances achieved over the single recordings. The upper part of the table reports the results obtained over the output of the speaker segmentation, the lower part reports the results obtained over the manual speaker segmentation.

	all (σ)	AM	SA	GT	IP	HP	WM
Automatic Speaker Segmentation							
C1 (B,I)	81.7 (6.9)	98.0	4.0	92.0	5.6	55.9	76.8
C1 (B,D)	62.7 (16.5)	89.9	4.2	68.9	9.0	11.0	10.1
C1 (M,I)	**82.4** (7.1)	97.8	4.8	92.2	4.2	64.3	78.2
C1 (M,D)	62.3 (16.7)	88.7	3.4	70.2	4.5	7.0	15.4
C2 (B,I)	83.2 (6.7)	75.0	88.3	91.5	N/A	29.1	9.0
C2 (B,D)	87.5 (4.4)	77.1	92.1	93.2	N/A	91.0	17.7
C2 (M,I)	84.0 (6.5)	68.7	92.2	89.7	N/A	83.7	15.4
C2 (M,D)	**87.8** (4.3)	77.1	92.1	93.2	N/A	98.4	16.3
Manual Speaker Segmentation							
C1 (B,I)	95.1 (4.6)	100	88.5	98.3	13.9	100	97.9
C1 (B,D)	66.7 (12.5)	96.9	5.2	66.9	11.8	21.9	12.5
C1 (M,I)	97.0 (4.2)	100	86.5	98.7	61.5	100	97.9
C1 (M,D)	67.5 (9.6)	99.0	6.2	72.0	3.3	6.2	10.4
C2 (B,I)	96.2 (2.6)	96.3	100	96.6	N/A	100	70.4
C2 (B,D)	96.1 (5.8)	96.3	96.3	97.7	N/A	100	33.3
C2 (M,I)	95.8 (7.7)	96.3	96.3	95.7	N/A	100	81.5
C2 (M,D)	98.1 (2.1)	100	100	98.6	N/A	100	48.1

difference is significant for C1 and C2 because in this case the speaker diarization process produces more errors and the purity is around 0.8, i.e. the output of the speaker diarization is significantly different from the groundtruth speaker segmentation. The difference in accuracy is around 10 percent (statistically significant) and this is mostly due to the small differences (2 seconds on average) between the actual speaker changes and the changes as detected by the diarization process. The sum of all the displacements amounts, on average, to roughly 10 percent of the recording length and this is the probable explanation of the performance difference when passing from manual to automatic speaker segmentations.

The rest of the error is due to limits of the role recognition approach that cannot distinguish between different roles when the associated interaction patterns are too similar. This is the case, e.g., of the low performance on IP in corpus C1. The interaction pattern of the IP role

Table 4.6: Role recognition performance for C3. The table reports both the overall accuracy and the accuracy for each role. "B" stands for *Bernoulli*, "M" stands for *Multinomial*, "I" stands for roles *Independence*, and "D" stands for roles *dependence*. The overall accuracy is accompanied by the standard deviation σ of the performances achieved over the single recordings. The upper part of the table reports the results obtained over the output of the speaker segmentation, the lower part reports the results obtained over the manual speaker segmentation.

	all (σ)	PM	ME	UI	ID
Automatic Speaker Segmentation					
C3 (B,I)	46.0 (24.7)	79.6	13.1	41.4	20.3
C3 (B,D)	**46.4** (30.0)	68.7	26.0	32.9	25.7
C3 (M,I)	39.3 (24.9)	67.4	18.0	19.3	25.6
C3 (M,D)	43.7 (31.3)	67.4	28.7	22.0	24.3
Manual Speaker Segmentation					
C3 (B,I)	51.2 (24.2)	83.3	15.9	42.0	29.0
C3 (B,D)	56.0 (33.0)	76.1	37.7	40.6	41.3
C3 (M,I)	43.7 (27.3)	67.4	17.4	39.1	21.7
C3 (M,D)	52.6 (27.6)	76.8	29.0	34.1	33.3

is similar to the one of the GT, but this last has higher *a-priori* probability, so it is usually favoured as output of the recognizer.

A qualitative comparison with other approaches is possible only for some works using the same data, at least in part, as this work. Both [94][67] perform experiments over a subset of the AMI meeting corpus (around 5 hours of material). The performance in [94] is around 80%, almost twice as our approach over the same data (see Section 4.4). However, as the goal is to detect the two most dominant persons, the probability of assigning each person the correct role is 50%, while it is only 25% in our case. The work in [67] reports a 65% recognition rate of the Project Manager, while our work achieves, over the same role, an accuracy of 79%. Considering that our experiments are performed over the whole AMI meeting corpus, while the experiments of [94][67] take into account only a subset of 5 hours, our approach seems to be more effective in both cases, though the task is not the same. The work in [53] uses the whole AMI corpus, but it applies a different experimental setup. However it performs exactly the same task as this work and the role recognition rate is around 60%.

4.5 Conclusion

This chapter has presented an approach for the automatic recognition of roles in multi-party recordings. The problem of role recognition has been addressed only recently in the literature, but it attracts an increasingly growing interest because it is a key point in the automatic analysis of social interactions [4, 95]. The proposed approach has been tested over roughly 90 hours of material, one of the biggest datasets ever used in the literature for this task. To the best of our knowledge, this is the first work that compares the performance of an approach over both roles based on *beliefs and preferences* and roles based on *norms* showing how the role typology influences the effectiveness of the recognition.

The results show that the recognition accuracy is higher than 85% in the case of broadcast data, and it is around 45% in the case of meeting recordings. There are several possible reasons for such a difference. The first, and probably most important, is that broadcast data include roles based on norms, while meetings include roles based on preferences. Roles based on norms are easier to model because they impose constraints on the behaviour of people that can be detected, represented and modelled with probabilistic approaches (like in the case of this work). In contrast, roles based on beliefs and preferences do not necessarily constrain behaviour and this makes difficult the automatic recognition through approaches like the one presented in this work, at least for the aspect of behaviour used as role evidence in this work, i.e. *who talks with whom and when*.

The second is that the broadcast data is real, while the meeting data is acted. The meetings do not involve people playing the role they actually have in their life, but volunteers that simulate an artificially assigned role they have never played before. This is likely to reduce significantly the performance of any role recognition method.

In the case of the broadcast data, the performance is sufficient to effectively browse the data (users can quickly find segments corresponding to a given role and the mismatch between the groundtruth and the automatic output rarely exceeds few seconds). In the case of meeting recordings, the approach is effective only to identify the Project Manager. This allows one to effectively follow the progress of the meeting because the PM plays the chairman role as well and, as such, is responsible for following the agenda through her/his interventions.

The main limitation of the approach presented in this chapter is that it does not take into

account any sequential information. The role of the person speaking at turn n is likely to have a statistical influence on the role of the person speaking at turn $n + 1$. Furthermore, the approach proposed in this chapter uses only the turn-taking patterns as a role evidence, while other behavioral cues can be extracted from audio (e.g., prosodic features). We will see in the next chapter how we can integrate information from the words used by the speaker to improve the model. In chapter 6, we will see how Conditional Random Field allows one to easily integrate new features in the model and to take into account the dependency between successive turns.

Chapter 5

Role Recognition Based on Lexical Information and Social Network Analysis

This chapter presents experiments on the automatic recognition of roles in meetings. The proposed approach combines two sources of information: the lexical choices made by people playing different roles on one hand, and the Social Networks describing the interactions between the meeting participants on the other. This chapter can been seen as an extension of the previous chapter. We had a new modality (the words spoken) with respect to the features used in the previous chapter. Those results were published in:

- N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür, and A. Vinciarelli. Role recognition for meeting participants: an approach based on lexical information and Social Network Analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 693–696, 2008.

Both sources lead to role recognition results significantly higher than chance when used separately, but the best results are obtained with their combination. The experiments were conducted over the AMI meeting corpus (denoted C3, see Section 4.1) and they show that around 70% of the time is labelled correctly in terms of role.

5.1 Introduction

The overall scheme of the approach is depicted in Figure 5.1: the first step is the application of a speaker diarization approach that identifies the time intervals where each speaker talks. The subsequent steps follow two parallel paths corresponding to the two behavioural cues mentioned above. The right path describes the modelling of the lexical choice and it includes two stages: extraction of the lexical features from the automatic speech transcriptions, and mapping of the lexical features into roles using the BoosTexter text categorization approach [96]. The left path corresponds to the interaction pattern modelling and it also includes two stages: extraction of a Social Affiliation Network [25] representing social interactions, and assignment of roles to people using a Bernoulli distribution [32]. The main advantage of the behavioural cues is that they are, to a large extent, identity-independent. This enables one to address the general case where an individual plays different roles in different circumstances (as is actually the case in the data used in this work).

The main novelty of this chapter is the combination of approaches based on both lexical features and social networks that so far have been applied only separately (see above). This is expected to make the recognition approach more robust with respect to the two major sources of *noise* in the experiments, i.e. the errors of the Automatic Speech Recognition (ASR) system used to transcribe the recordings, and the errors of the speaker diarization approach used to segment the data into single speaker intervals. The experiments of this work are performed over the AMI corpus [45], a collection of 138 meetings with a total duration of 45 hours and 38 minutes. Each meeting involves four participants playing different predefined roles (see Section 4.1). The two datasets from the "Radio Suisse Romande" were not used because those corpora have not been manually transcribed. Furthermore, both data sets are in French and no reliable automatic transcription system is available.

The results show that, on average, roughly 70% of the meeting time is labelled correctly in terms of role. The accuracy is higher for the roles associated with well defined and stable behavioural patterns, while it is lower for the roles that do not exhibit predictable behaviours. However, the performance of the system is significantly higher than a random guess for all roles. The combination of the two approaches described above slightly improves the performance of the best role recognizer (based on the lexical choice). However, the improvement appears to be significant for the roles most represented in terms of time. The overall approach

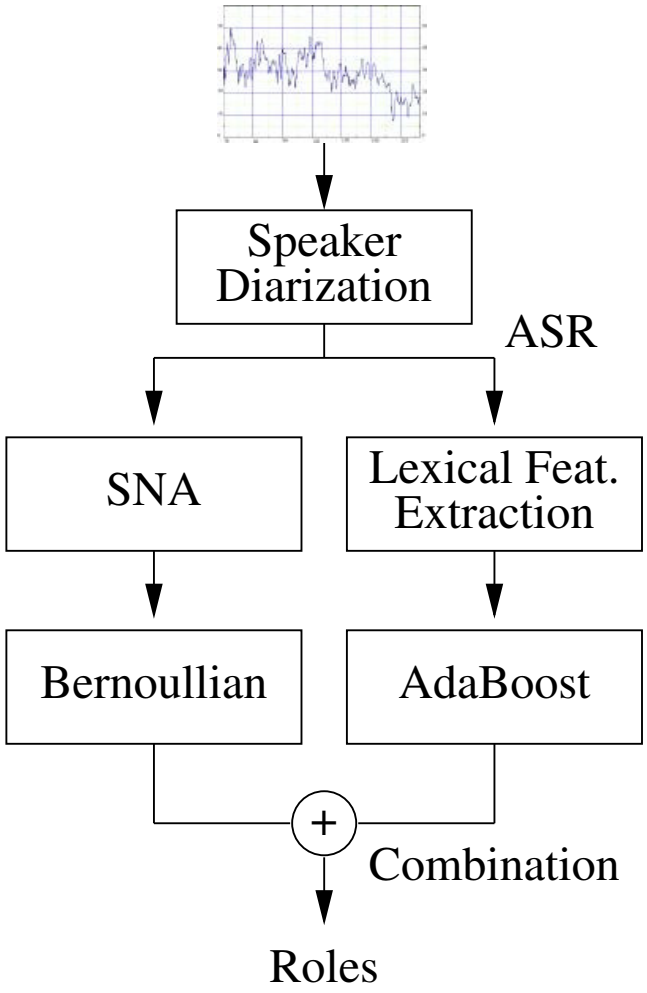

Figure 5.1: Overview of the approach. The two parallel paths produce separate decisions that are combined at the end of the process.

seems to be more robust to the errors of the speaker diarization step than to the speech recognition errors.

The rest of the chapter is organized as follows: Section 5.2 describes the approach proposed in this work, Section 5.3 presents experiments and results, and Section 5.4 draws some conclusions.

5.2 The approach

This section describes the recognition approach based on the lexical features (right path of Figure 5.1), the one based on Social Network Analysis (left path of Figure 5.1), and the combination approach. Due to space limitations, no details are given about speaker diarization

and Automatic Speech Recognition approaches applied in this work (see Section 4.2.1 for a full description). The diarization accuracy (percentage of data time correctly labelled in terms of speaker) is 97.0%, while the Word Error Rate is between 35 and 40% depending on the specific recording of the corpus used for the experiments.

5.2.1 Lexicon Based Role Recognition

The role recognition approach based on lexical features recognizes the roles of speakers using the lexical content of their utterances. The intuition here is that the meeting structure and content are correlated with the roles of its participants, and lexical cues related to structure and topics can be useful for determining speaker roles. For example, the person leading the discussion can use phrases to return to aimed discussion, when a topic shift to an unrelated topic occurs. Also, due to his/her functional role, a speaker may only talk about certain related topics.

We model speaker role detection as a multi-class classification task, where there is one class for each speaker role, and the goal is to assign a role to a speaker in every meeting. Note that, sometimes, a speaker can play different roles in different meetings, but the role is constant in a single meeting. For classification, we use BoosTexter, a multi-class classification tool. Boosting aims to combine *weak* base classifiers to come up with a *strong* classifier [96]. This is an iterative algorithm, where at each iteration, a weak classifier is learned so as to minimize the training classification error. The algorithm begins by initializing a uniform distribution, $D_1(i, r)$, over training examples, i, and labels (i.e., speaker roles), r. After each round this distribution is updated so that the example-class combinations which are easier to classify (e.g. the examples that are classified correctly with the weak learners learned so far) get lower weights and vice versa. The intended effect is to force the algorithm to concentrate on examples and labels that will most improve the classification rule. To represent every example i (i.e. every meeting participant in the training corpus), we use word n-grams ($n = 1, 2,$ and 3) from all the turns of a speaker in a meeting as features.

The weak classifiers check the presence or absence of word n-grams in the speaker's turns, and can therefore be used for analysis purposes. The final strong classifier is a linear combination of the individual weak classifiers. We use a held-out data set to compute the optimum

number of iterations for the classifier. The classifier outputs a probability for each role and for each speaker.

If $\vec{d_i}$ is the vector representing the transcription of the interventions of meeting participant i, then the BoosTexter approach estimates the probability $p(\vec{d_i} \mid r)$ of the participant playing role r by combining the weak classifiers described above. The participant i is assigned the role r^* that satisfies the following expression:

$$r^* = \arg\max_{r \in \mathcal{R}} p(\vec{d_i} \mid r), \tag{5.1}$$

where $\mathcal{R}$ is the set of the predefined roles.

5.2.2 Social Networks Based Role Recognition

This role recognition approach is based on the Affiliation Networks (see upper part of Figure 4.6) [25], i.e. Social Networks where there are two kinds of nodes, the *actors* and the *events*, and only links between different kinds of nodes are allowed. The rationale behind this representation is that people participating in *similar* sets of events are more likely to interact with one another. Thus, actor nodes with similar sets of connections are expected to represent individuals with high mutual interaction likelihood.

The set of the connections of an actor node a_i is represented with a binary vector $\vec{x}_i = (x_{i1}, \ldots, x_{iD})$, where D is the number of events, and $x_{ij} = 1$ if actor a_i participates in event e_j and 0 otherwise. The more two vectors $\vec{x}_i$ and $\vec{x}_l$ are similar, the more actors a_i and a_l are likely to interact because they participate together in many events. In the case of the meeting recordings, the actors are the participants, and the events are segments of uniform length that span the whole duration of a meeting (see lower part of Figure 4.6). If D is the total number of segments for a meeting, then the event e_n corresponds to the time interval $[(n-1)T/D, nT/D]$, where T is the total duration of the meeting. Actors are said to participate in an event when they talk during the corresponding meeting segment. Thus, the actors are supposed to have a higher probability of interaction when they talk during the same intervals of time (i.e., when they participate in the same events) than when they talk in different intervals of time.

The most natural way of modelling binary vectors is to use Bernoulli discrete distributions:

$$p(\vec{x}_i \,|\, \vec{\mu}_r) = \prod_{j=1}^{D} \mu_{rj}^{x_{ij}} (1 - \mu_{rj})^{1-x_{ij}}, \tag{5.2}$$

where $\vec{\mu}_r = (\mu_{r1}, \ldots, \mu_{rD})$ is the parameter vector of the distribution related to role r. The maximum likelihood estimates of the μ_{ri} parameters are as follows [32]:

$$\mu_{ri} = \frac{1}{N_r} \sum_{n=1}^{N_r} x_{ni}, \tag{5.3}$$

where N_r is the number of people playing the role r in the training set, and x_{nj} is the j^{th} component of the vector representing the n^{th} person playing the role r. A different Bernoulli distribution can be trained for each role, and an actor represented with a vector $\vec{x}$ will be assigned the role $\hat{r}$ satisfying the following equation:

$$\hat{r} = \arg\max_{r \in \boldsymbol{R}} p(\vec{x} \,|\, \vec{\mu}_r), \tag{5.4}$$

where $\boldsymbol{R}$ is the set of the predefined roles.

5.2.3 Combination Approach

Both role recognition approaches described above estimate the probability of a meeting participant playing a role r. We use the weighted product rule [97] for the combination of the probabilities from the two models. The joint probability can be written as

$$p(\vec{d}, \vec{x} \,|\, r, \vec{\mu}_r) = p(\vec{d} \,|\, r)^{\beta} \cdot p(\vec{x} \,|\, \vec{\mu}_r)^{(1-\beta)}. \tag{5.5}$$

We can derive an estimate for the roles based on this combination:

$$\begin{aligned} \hat{r} &= \arg\max_{r \in \boldsymbol{R}} p(\vec{x}, \vec{d} \,|\, r, \vec{\mu}_r) \\ &= \arg\max_{r \in \boldsymbol{R}} \beta \log p(\vec{d} \,|\, r) + (1 - \beta) \log p(\vec{x} \,|\, \vec{\mu}_r), \end{aligned} \tag{5.6}$$

where the factor β ensures that both terms are of the same order of magnitude and contribute

Role	PM	ME	UI	ID
Fraction	36.6%	22.1%	19.8%	21.5%

Table 5.1: Role distribution. The table reports the average fraction of time each role accounts for in a meeting.

to the final decision. The value of β is constrained between 0 and 1. In the extreme case where β is 1, the model using the SNA is ignored. In the other extreme case, where β is 0, the lexical model is ignored. For our experiment, the β value is selected through cross validation (see next section) in order to maximize the accuracy of the combined model. The techniques to estimate $p(\vec{d} \mid r)$ and $p(\vec{x} \mid \vec{\mu}_r)$ are explained in the previous subsections.

5.3 Experiments and Results

This section presents the data, the experiments and the results obtained in this work.

5.3.1 Data and Roles

The experiments of this work are performed over the AMI corpus [57], a collection of 138 meeting recordings for a total of 45 hours and 38 minutes of material. The meetings are simulated and are based on a scenario where the participants are the members of a team working on the development of a new remote control. Each meeting involves four participants playing one of the following roles: the *Project Manager* (PM), the *Marketing Expert* (ME), the *User Interface Expert* (UI), and the *Industrial Designer* (ID). Each participant plays a different role, and all roles are represented in each meeting. The same person can play different roles in different meetings, and the fraction of meeting time that each role accounts for, on average, is reported in Table 5.1.

5.3.2 Experiments

The training of the role recognition system is performed using a *leave-one-out* approach: all the meetings in the corpus are used for training the models with the exception of one that is used as test set. Training and test are repeated as many times as there are meetings in the

approach	all	PM	ME	UI	ID
SNA (aut.)	43.1	75.7	16.4	41.2	13.4
lex. (aut.)	67.1	78.3	71.9	38.1	53.0
SNA+lex. (aut.)	67.9	84.0	69.8	38.1	50.1
SNA (man.)	49.5	79.0	20.3	44.9	24.6
lexical (man.)	76.7	92.0	70.3	60.1	60.9
SNA+lex. (man.)	78.0	95.7	68.8	60.1	61.6

Table 5.2: Role recognition results. The upper part of the table shows the accuracies obtained over automatic (aut.) speaker diarization and speech recognition. The lower part reports the accuracies obtained over manual (man.) speaker segmentation and speech transcriptions.

corpus (138 in the case of the AMI corpus), and each time a different meeting is *left out* as test set. In this way, the whole corpus can be used as test set while still keeping rigorously separated training and test set, as required to assess correctly the system performance. The hyperparameters of the system (number of AdaBoost iterations for the lexicon based approach, and β factor for the combination) are tuned over a subset of 20 meetings randomly selected in the training set.

The performance is measured in terms of the *accuracy* α, i.e. the percentage of data time correctly labeled in terms of role. Table 5.2 reports the accuracies obtained by using only Social Network Analysis, only lexical choices, and the combination of the two. The lower part of the table shows the results obtained using groundtruth speaker segmentation and speech transcripts, while the upper part of the table shows the results obtained using the output of automatic speaker diarization and speech recognition systems. The results are reported for the overall meetings, as well as for the single roles separately.

The lexical choice appears to be, at least for the AMI corpus, a more reliable cue for the recognition of the role. The overall accuracy of the lexicon based system is significantly higher for both groundtruth (76.7% against 49.5%) and automatic data (67.1% against 43.1%). A possible explanation is that the AMI corpus is particularly suitable for lexical analysis, while it is rather unfavourable to the application of SNA. On one hand, the content of the interventions is constrained by the role and this helps the former approach, on the other hand, the small number of participants significantly limits the latter approach because the social networks tend to be more meaningful when the number of people increases [25].

The SNA based system appears to be more robust when passing from the groundtruth data

to the output of the automatic systems for speaker segmentation and speech recognition. A possible explanation is that the SNA based approach uses only the speaker segmentation that is performed with high accuracy (around 97%), while the lexical based approach uses the speech transcriptions that are affected by a much higher error rate (around 40%). As a result, while the overall performance remains significantly different, the accuracy for PM and UI is comparable for both systems (see upper part of Table 5.2). Thus, the systems have similar performance over more than 50% of the data time because PM and UI account together for roughly 57% of the total AMI corpus time (see Table 5.1).

The combination of the two systems improves only slightly the performance of the best system (see table 5.2). The main reason is probably that the performance of the SNA approach is too close to chance (around 25%) for at least two roles (ME and ID). Thus, the SNA does not bring useful information in the combination, but simply some random noise. This seems to be confirmed by the case of the PM role, where the combination improves by almost 6% the performance of the best classifier. Not surprisingly, the performance of the SNA system over the PM is significantly better than chance.

5.4 Conclusions

This chapter has presented a role recognition approach based on the combination of two systems relying on lexical choices and interaction patterns, respectively. The results show that roughly 70% of the data time is labelled correctly in terms of role, and that the combination improves the best classifier, in particular for the PM role.

The main limit of the approach presented in this chapter is that getting an automated transcript is costly and language dependant. Since the integration of different modalities seems to be the most effective technique to analyze social interactions [47], we will replace the automatic speech transcription with a new modality in the next chapter: the prosody. We will also use a Conditional Random Field, which allows for a better fusion of different modalities.

Chapter 6

Turn-based Approach to Role Recognition

This chapter proposes an approach for the automatic recognition of roles in settings like news and talk-shows, where roles correspond to specific functions like Anchorman, Guest or Interview Participant. The approach is based on purely nonverbal vocal behavioural cues, including who talks when and how much (turn-taking behaviour), and statistical properties of pitch, formants, energy and speaking rate (prosodic behaviour). The experiments have been performed over the corpora C1 and C2 (see Section 4.1), totalizing around 50 hours of broadcast material. The main difference with the approach presented in Chapter 4 is that the features are extracted for every turn, and take into account the local behaviour of the participant. The accuracy, percentage of time correctly labelled in terms of role, is up to 89%. These results were published in

- H. Salamin and A. Vinciarelli. Automatic role recognition in multiparty conversations: an approach based on turn organization, prosody and conditional random fields. *IEEE Transactions on Multimedia*, PP(99):1, 2011.

Both turn-taking and prosodic behaviour lead to satisfactory results. Furthermore, on one database, their combination leads to a statistically significant improvement.

6.1 Introduction

In this chapter, the goal is to present the results obtained on the role recognition problem using Conditional Random Fields. The main contributions are:

- Turn-based approach allows the use of features extracted on a turn by turn basis, allows for the same person to play different roles in the same interaction and allows for the role assignment to be performed even if only part of the interaction is actually available.
- Use of prosody (pitch, energy and rhythm),
- No Meta-parameter (in particular, the windows-length parameter used in Chapter 4 is not needed),
- Use of larger amounts of observation (via regularization),
- Work across different broadcast type with the same model.

This chapter proposes an approach for the recognition of roles based on norms (particularly in news and talk-shows). The approach is based on turn-taking and prosodic behaviour. Turn-taking was presented in Chapter 4 and accounts for who talks when and how much and provides a description of how each person participates in a conversation. Prosodic behaviour accounts for the way people talk, i.e. their pitch, loudness and speaking rate.

The approach includes three main steps (see Figure 6.1). The first is the segmentation of the data into turns, time intervals during which only one person is talking. The second is the extraction of turn-taking and prosodic features from each turn. The third is the mapping of the feature vectors extracted from each turn into a sequence of roles with Conditional Random Fields.

Furthermore, this is the first work, to the best of our knowledge, where prosodic and turn-taking behaviour are combined to provide a full description of non-verbal vocal behaviour in conversations. With respect to previous approaches, the main novelty is not only the use of prosodic behaviour, but also that the role assignment is performed for each turn rather than for each person. This is a major improvement because it ensures that the approach can be

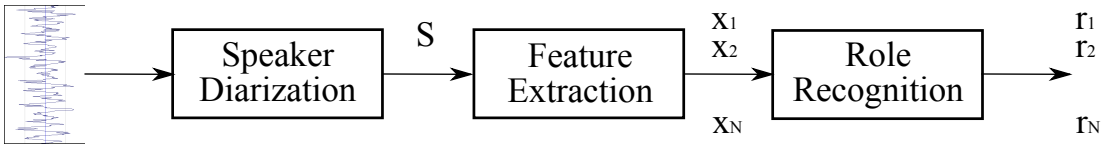

Figure 6.1: The figure depicts the role recognition approach presented in this work: The audio data is first segmented into turns (single speaker intervals), then converted into a sequence of feature vectors and mapped into a sequence of roles.

used in situations where the same person is having a different role, at different times in the interaction.

The results show that both prosodic and turn-taking behavior, when used individually, achieve satisfactory performances (up to 89% accuracy). Moreover, the combination of the two leads to a statistically significant improvement with respect to the best individual performance on one of the databases. The performance achieved seems to confirm that people playing different roles display different prosodic behaviours: that is, they exhibit specific ways of speaking.

The rest of the chapter is organized as follows: Section 6.2 describes the proposed approach, Section 6.3 describes experiments and results, and Section 6.4 draws some conclusions.

6.2 The Approach

The overall approach is depicted in Figure 6.1. The input data is the audio recording of a multi-party conversation and the first step is the segmentation into turns via a speaker clustering approach. The rest of the process includes the feature extraction applied to each turn and the mapping of the resulting observations into roles.

6.2.1 Speaker Diarization

The extraction of turns is performed with a speaker diarization approach that does not require to know in advance the number and identity of speakers. The diarization process is fully described in [88] and it will not be further presented here as it is not an original part of this work. The technique used is similar to the technique presented in Section 4.2.1. As a

reminder, the output of the diarization is a list of triples:

$$S = \{(s_1, t_i, \Delta t_i), \ldots, (s_N, t_N, \Delta t_N)\} \tag{6.1}$$

where N is the number of turns extracted by the diarization approach, $s_i \in A = \{a_1, \ldots, a_G\}$ is a speaker label, G is the total number of speakers detected during the diarization, t_i is the starting time of turn i, and Δt_i is its length. The label s_i is not the name of the speaker, but an arbitrary label automatically assigned by the diarization approach. The set A is not defined a-priori, but it is a result of the diarization process. In general, $G \ll N$ and several turns share the same speaker label. This means that the speaker is the same for different turns.

6.2.2 Feature extraction

The turn sequence S provides information about *who talks when and how much*. This makes it possible to extract features accounting for the overall organization of turns as well as for the prosodic behavior of each speaker. The turn organization is important because it conveys information about the social actions carried out by different interaction participants [98], typically through "*systematically ordered features*" [23] or appropriate sequences called *preference structures* [99]. The prosody is important because it influences the perception of a large number of socially relevant aspects including competence and expressivity [100], personality [101], and emotional state [102].

Since the earliest works on role theory, both turn organization and prosody have been recognized as one of the main evidences of the role people play. However, while the turn organization has been extensively used in the role recognition literature, the prosody has been, to the best of our knowledge, largely neglected. This work tries to fill this gap and it proposes to use two sets of features for each turn, the first relates to turn organization while the second relates to prosody.

From each turn, two types of features are extracted, turn-taking and prosody related, respectively. The former are expected to account for who talks when and how much, the latter for how people talk during their interventions.

The first set includes features that account for the way an individual interaction participant contributes to the turn organization. The following features were extracted:

- turn duration (in seconds),
- total number of turns for the current speaker,
- time from the beginning of recording to first turn of the current speaker (in seconds),
- average time between two turns of the current speaker (in seconds),
- time after last turn of the current speaker (in seconds),
- time from previous to current turn of the current speaker (in seconds),
- number of unique speakers in the T neighboring turns ($T = 3, 5, 7$).

Those features can be divided into two types. The first type corresponds to how a particular turn contributes to the overall turn organization (turn duration, time after last turn of the current speaker, number of unique speakers in the neighboring turns). The second type of feature captures how one speaker contributes to the turn organization (total number of turns for current speaker, time from the beginning of recording to first turn of current speaker, average time between two turns of current speaker). Those features have the same value for all of the turns where the speaker is the same. All of these features have already been applied in the role recognition literature and they have been shown to be effective. The features are clearly non-independent, but this is not a problem because Conditional Random Fields (see below) do not make any assumption about the independence of the observations.

The second set includes the prosodic features, namely the pitch, the first two formants, the energy and the length of each voiced and unvoiced segment. These measurements are made with Praat [103] over short analysis windows ($30\ ms$) at regular time steps ($10\ ms$) and account for short-term speech aspects. Longer term aspects can be obtained by estimating statistical properties of each feature over the entire turn. In this work, for each feature f we use the relative entropy.

The extraction of prosody related features includes two steps. The first is the extraction of the low-level features, and the second is the extraction of the turn-level features. Low-level features include:

- pitch,

- 2nd, 3rd and 4th and formants,
- energy and
- segmentation into voiced and unvoiced intervals, i.e. segments during which there is emission of voice or not.

The extraction of the low-level features is performed with Praat [103], one of the most commonly applied tools in speech analysis. Low-level features are extracted from $30\ ms$ long segments at regular time steps of $10\ ms$. Thus, low-level features account only for short-term phenomena and are not suitable in their raw form to represent turns that can last from several seconds up to minutes.

The approach applied to address the above problem is to extract turn-level features, i.e. statistics accounting for the distribution of the low-level features on the scale of a turn. In this work, the statistics correspond to the entropy of the low-level features. If f is a low-level feature, the entropy is estimated as follows:

$$H(f) = \frac{\sum_{i=1}^{|F|} p(f_i) \log p(f_i)}{\log |F|} \tag{6.2}$$

where $F = \{f_1, \ldots, f_{|F|}\}$ is the set of f values observed in a turn, $|F|$ is the cardinality of F, and f corresponds to one of the low-level features mentioned above. The turn-level features are expected to capture the variability of each low-level feature: the higher the entropy, the higher the number of f values represented a large number of times during the turn, and vice-versa. Entropy was selected as a characteristic that captures the speaking style of a person and influences the social perception that others develop about her [104][105].

The turn-level features are not extracted from the whole turn, but from a fraction of the turn centred in its middle and with length corresponding to 90% of the total turn length. The reason is that the speaker clustering process is affected by errors and the turn boundaries are not detected correctly. Thus, initial and final part of the turn might include noise.

The data used in this experiment contained very little overlapping speech and no special treatment for it has been done. In more spontaneous data, the amount of overlapping speech can be important. The turn-level features extracted assume that only one voice is present in the turn. In the case of overlapping speech and data mixed into on audio channel, this is not

case and reliable features can not be extracted. If individual channel are available for every speaker involved in the overlapping speech, then the features can be extracted without any problems.

6.2.3 Role Recognition

After the feature extraction step, the sequence S of turns is converted into a sequence $\boldsymbol{X} = \{x_1, \ldots, x_N\}$ of observations, where the components of vectors x_i correspond to the features described in the previous section. The role recognition step is performed by labelling the sequence of observations $\boldsymbol{X} = \{x_1, \ldots, x_N\}$ with a Conditional Random Field (CRF) (see Section 3.3). This corresponds to finding the sequence of roles $\hat{\boldsymbol{Y}}$ satisfying the following expression:

$$\hat{\boldsymbol{Y}} = \arg\max_{\boldsymbol{Y} \in \mathcal{Y}} \mathrm{P}(\boldsymbol{Y} \mid \boldsymbol{X}, \alpha) \tag{6.3}$$

where the α_i are the model parameters , $\mathcal{Y}$ is the set of all possible sequences $\boldsymbol{Y}$, and $\boldsymbol{Y} = \{y_1, \ldots, y_N\}$ is the sequence of roles (y_t is the role assigned to the person talking at turn t). The experiments of this work use a linear chain CRF. This model corresponds to the assumption that two labels are conditionally independent given the observations and one label between them. More formally, the core assumption of this model is that the label y_{t-i} is conditionally independent of the label y_{t+j} given the label y_t and all the observations $\boldsymbol{X}$, for any t and for any i and j greater than 0.

CRF are used in this work as they have been shown to perform very well for sequence labelling [87] and they out perform Hidden Markov Model and Maximum Entropy Model, the two other most common models used in sequence labelling. The main advantage of CRFs with respect to other probabilistic sequential models is that they do not require any conditional independence assumption about the observations of $\boldsymbol{X}$. This is particularly important in this work because some of the features depend on long term dependencies (e.g., the distance with respect to the last turn of the current speaker) and others have the same value for all of the turns of a certain speaker (e.g., the number of turns for the current speaker). In both cases, models based on the assumption that the observations are conditionally independent given an underlying variable (e.g., Hidden Markov Models) would not be appropriate.

Another possible approach would have been to use a Support Vector Machine (SVM) to clas-

sify each turn. Such an approach has several drawbacks with respect to sequential models. The main drawback is that SVM cannot use the label of the previous turn and the next turn to help the classification. However, human interaction is naturally sequential, and the labels of the neighbouring turn contain useful information. Furthermore, SVM are not probabilistic models as they only output the class of an observation instead of the probabilities of belonging to different classes. This makes it difficult to combine or reuse their output with another model. Finally, most of the uses of SVM presented in the state of the art has been on problems with a small number of classes (2 or 3). In the experiment of this section, the number of classes is higher and the use of multi-class SVM would have added unwarranted complexity.

As presented in Chapter 3, for any CRF, the probability distribution can be represented as a product of potentials. Each potential is associated with a clique in the graph and can in theory depend on the whole set of observations. In our case, the maximal cliques are pairs of adjacent role assignments $\{y_t, y_{t+1}\}$. We will consider potential not only on the maximal cliques but also on single role assignment y_t. We will make the following assumptions with respect to the potentials to make the model tractable:

1. The potential over two consecutive labels $\{y_t, y_{t+1}\}$ depends only on y_t and y_{t+1} and not on the observations $\boldsymbol{X}$.

2. The potential over a label $\{y_t\}$ depends only on y_t and the observation at that time x_t. Feature taking into account global information (such as the number of turns for a speaker) can be introduced by copying it in the features of every turn.

3. The potentials are the same for all t. We assume that the behaviours associated with a role are not changing over the duration of the interaction.

4. The potentials are never zero.

This first three assumptions mean that the marginal distribution for a label y_t is fully determined by the previous label y_{t-1}, the next label y_{t+1} and the observation x_t. The fourth assumption means that every role assignment has a probability strictly greater than zero and is important in practice as it allows the product of potential to be replaced by the exponential

of a sum. We can now write a first expression for the probability distribution

$$\mathrm{P}(\boldsymbol{Y}|\boldsymbol{X}) = \frac{\exp\left(\sum_{t=1}^{N} f_1(y_t, x_t) + \sum_{t=1}^{N-1} f_2(y_t, y_{t+1})\right)}{Z(X)}$$
$$Z(\boldsymbol{X}) = \sum_{Y \in \mathcal{Y}^N} \mathrm{P}(\boldsymbol{Y}|\boldsymbol{X})$$

$Z(\boldsymbol{X})$ is called the partition and is simply a normalization constant. f_1 and f_2 represent the potentials on one turn and two adjacent turns respectively. They will be represented as a linear combination of simpler functions called feature functions. We denote by $\boldsymbol{R}$ the set of roles. Then, for every $y \in \boldsymbol{R}$ and every index in the observations vector, we have a feature function f_1 given by:

$$f_{y,i}(y_t, \vec{x}_t) = \begin{cases} x_t^{(i)} & \text{if } y_t = y \\ 0 & \text{otherwise} \end{cases} \tag{6.4}$$

where $x_t^{(i)}$ is the i^{th} component of $\vec{x}_t$. This family of feature functions can capture linear relations between a role and an observation $x_t^{(i)}$. For f_2, the following feature function was used (for every possible pair of roles $(y, y') \in \boldsymbol{R}^2$:

$$f_{y,y'}(y_t, y_{t+1}) = \begin{cases} 1 & \text{if } y_t = y \text{ and } y_{t+1} = y' \\ 0 & \text{otherwise} \end{cases} \tag{6.5}$$

This family of feature functions captures the dependency of the roles between adjacent turns. Putting it all together, Linear Chain CRFs estimate the a posteriori probability of a role sequence as follows:

$$p(\boldsymbol{Y}|\boldsymbol{X}, \alpha) = \frac{1}{Z(\boldsymbol{X})} \exp\left(\begin{array}{l} \sum_{t=1}^{N} \sum_{y \in R} \sum_{i} \alpha_{y,i} f_{y,i}(y_t, x_t) + \\ \sum_{t=1}^{N-1} \sum_{(y,y') \in R^2} \alpha_{y,y'} f_j(y_t, y_{t+1}) \end{array}\right) \tag{6.6}$$

The weights $\alpha_{y,i}$ of the feature functions of form $f_{y,i}(Y, X)$ account for how much the value of a given feature is related to a particular role. The weights of the feature functions of form $f_{y,y'}(Y, Y)$ account for how frequent it is to find role y followed by role y'.

Given a training set $\{(\boldsymbol{X}^{(j)}, \boldsymbol{Y}^{(j)})\}$ of labelled interaction, the weights α are learnt using a

maximum likelihood approach

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \sum_{j} \log \mathrm{P}(\boldsymbol{Y}^{(j)} \mid \boldsymbol{X}^{(j)}, \boldsymbol{\alpha}). \tag{6.7}$$

In the case of CRFs, this maximization can be accomplished using gradient ascent techniques and a regularization term is added to avoid over-fitting. Training a CRF boils down to finding the vector α satisfying the following equation:

$$\hat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} \sum_{j} \log \mathrm{P}(\boldsymbol{Y}^{(j)} \mid \boldsymbol{X}^{(j)}, \boldsymbol{\alpha}) - \frac{\|\boldsymbol{\alpha}\|_2}{\sigma^2} \tag{6.8}$$

where $X^{(j)}$ and $Y^{(j)}$ are training sequences, and the second element of the difference is a regularization term (σ is a hyper-parameter to be set via cross-validation) aimed at avoiding overfitting (its expression is based on the assumption that the α_i follow a normal distribution). The maximization of the right hand side of the above equation is performed using gradient ascent.

6.3 Experiments and Results

The next sections describe in detail the data and the recognition results obtained during the experiments of this work.

The experiments have been performed over two of the corpora presented in Section 4.1, referred to as C1 and C2, containing 96 news bulletins (19 hours in total) and 27 talk-shows (27 hours in total), respectively. The set of roles is the same for both corpora and it includes the Anchorman (AM), the Second Anchorman (SA), the guest (GT), the Interview Participant (IP), the Weather Man (WM), and the Headline Reader (HR). However, the distribution of the roles is different in the two corpora (see Table 6.2) and, even if the roles have the same name, they do not correspond exactly to the same function (e.g., the anchorman is expected to inform in the news and to entertain in the talk shows). The experiments are performed using a k-fold approach ($k = 5$), each corpus has been split into k subsets of equal size and $k - 1$ of them have been used for training while the k^{th} one has been left out for testing. The experiment has been repeated leaving out for test each of the k partitions. In this way,

Corpus	P	T	PT
C1 (A)	83.0%	89.7%	89.3%
C2 (A)	69.5%	84.2%	**87.0**%
C1+C2 (A)	68.1%	86.4%	86.7%
C1 (M)	87.1%	99.1%	99.1%
C2 (M)	76.2%	96.9%	96.2%
C1+C2 (M)	75.8%	96.6%	96.5%

Table 6.1: Results. This table reports the recognition results, A stands for "*automatic*" (results obtained over the output of the speaker clustering, M for "*manual*" (results obtained over the groundtruth speaker segmentation), P for prosody, T for turn-taking, $P + T$ for the combination of prosody and turn-taking. The value typed in bold corresponds to a statistically significant improvement of P+T with respect to P and T.

it is possible to test the approach over the whole corpus while keeping a rigorous separation between training and test sets.

The experiments have been performed not only on C1 and C2 separately, but also on their union. In this last case, the role IP has been converted into GT because C2 does not include people playing the IP role (see Table 6.2).

The accuracy, percentage of time in the test set correctly labeled in terms of role, is reported in Table 6.1 for the different experiments. The results are shown for both automatic and manual speaker segmentation. In the first case, the system works over othe output of the speaker clustering system described in Section 6.2, in the second case, the system works over the groundtruth speaker segmentation. This allows one to assess the effect of the speaker clustering errors that corresponds, on average, to roughly 10% decrease of the performance. The reason is that, each time there is a speaker change, the speaker clustering approach takes $1 - 2$ seconds to switch speaker. The accumulation of this error over all turns amounts to roughly 10% of the time in the different corpora.

The two types of features work to a satisfactory extent when they are applied separately. On the manual segmentation, their combination does not lead to statistically significant changes. The main reason is probably that the performance of the turn-taking features (close to 100%) is too high to leave an actual margin for improvement. On the automatic segmentation, the combination leads to a statistically significant (p-value < 0.05 measured using the Kolmogorov-Smirnov test) improvement of the performance on C2. On C1, the performance of the turn taking features is already very high and the remaining error is mainly due

Role	AM	SA	GT	HR	WM	IP
C1	41.2%	5.5%	34.8%	7.1%	6.3%	4.0%
C2	17.3%	10.3%	64.9%	4.0%	1.7%	0.0%

Table 6.2: Role distribution. The table reports the percentage of time each role accounts for in the two corpora.

to the small delays between actual and detected speaker changes. This source of error can be eliminated only by improving the speaker clustering approach and not by working on the features or the role modeling.

In several cases, it has not been possible to extract all the features for a turn. This applies, e.g., to turns that are too short ($2-3$ seconds) to extract a meaningful distribution of prosodic features, or to turns that are too close to the boundaries to count the number of speakers in the N-neighboring turns (see Section 6.2). The missing values have been set to the mean of the corresponding feature over the training set. This seems not to affect the performance of the model and represents a good approach to deal with missing data, at least in the case of these experiments.

The approach has been tested on the union of C1 and C2 to assess its robustness with respect to the presence of multiple settings in the data. The results show that the performance is comparable to that obtained over the two corpora separately. Thus, the approach actually seems to deal effectively with different settings at the same time.

6.3.1 Recognition Results

The recognition experiments have been performed using a k-fold approach ($k = 5$): each corpus has been split into k disjoint subsets and, iteratively, each one of these has been used as a test set while the others have been used as training set. The k-fold approach allows one to use the entire dataset at disposition for testing purposes while still keeping a rigorous separation between training and test data [106].

Table 6.3 reports the overall recognition results obtained over C1 and C2 separately, as well as on their union. The performance is reported in terms of accuracy, i.e. the percentage of time correctly labeled in terms of role in the test set. The upper part of the table shows the recognition results when using an automatic speaker diarization, while the lower part

Corpus	P	T	PT
C1 (A)	83.0%	89.7%	89.3%
C2 (A)	69.5%	84.2%	87.0%
C1+C2 (A)	68.1%	86.4%	86.7%
C1 (M)	87.1%	99.1%	99.1%
C2 (M)	76.2%	96.9%	96.2%
C1+C2 (M)	75.8%	96.6%	96.5%

Table 6.3: Accuracy. The table reports accuracy values when using only prosodic features (*P*), only turn-organization features (*T*), or the combination of the two (*PT*). The upper part of the table reports the results achieved over the turns extracted automatically (A), while the lower parts reports those achieved over the manual speaker segmentation (M).

reports the results when segmenting the audio data into turns manually. In the former case the segmentation is affected by errors, while in the latter case it corresponds to the actual turns in the data. The performance over the manual segmentation is higher than 95% for all of the corpora and this seems to suggest that the features adopted in this work capture, at least in part, the behavior patterns associated to the roles. The performance loss when moving to the automatic speaker segmentation is typically higher than 10%. The main reason is that speaker changes are detected with a certain delay ($1 - 2$ seconds) and the accumulation of these misalignments sums up, on average, to roughly $10 - 12\%$ of the recording's length.

The performance is reported using only prosodic features (column *P*), only turn organization features (column *T*), and the combination of the two (column *P*+*T*). Prosodic features lead to performances significantly higher than chance, but still significantly lower than the results obtained with turn-organization features. These results seem to suggest that the prosodic features are not effective, but the high performance of turn organization features (see accuracies higher than 95% on the manual speaker segmentation) might actually hide the contribution of prosody. Not surprisingly, the combination of prosody and turn-organization leads to statistically significant improvements only for C2, where the turn-organization features show the lowest accuracy.

The recognition experiments have been performed not only over C1 and C2 separately, but also over their union. As the results are comparable to those obtained over C1 and C2 individually, the role recognition approach seems to be robust with respect to a higher variability in the behavioural patterns through which roles are played.

Table 6.5 reports the results with the details for every role in each corpus. The performance

varies significantly across the roles for the same settings. The roles that are more represented (the Anchorman and Guest in C1, the Guest in C2) tend to be recognized better. This ensures that the overall accuracy remains high and that the roles that account for the highest fraction of time in the data are recognized correctly. Another interesting observation is that only the prosody is very bad at capturing the second anchorman. This is probably due to the fact that the second anchorman and the anchorman tend to speak in a very similar way.

Previous works have presented results obtained over C1 and C2 [44]. The approaches proposed here and in [44] are different in several respects: This work uses a probabilistic sequential model taking into account the sequence of the roles in a conversation, while the previous one uses a social network to represent the overall structure of the turns. This work assigns the roles turn by turn, while the previous one assigns the roles person by person. Furthermore, this work uses prosodic features and turn organization, while the previous one is based only on turn organization. The performances obtained in this work over the same data are higher and the difference is statistically significant. This seems to suggest that taking into account sequential aspects and prosody leads to significant improvements with respect to the approach proposed in [44].

The approach proposed here achieves very high performance on the broadcast news and the talk-shows. This high performance depends on two factors: the availability of a good speaker segmentation and the fact that the features capture the non-verbal behaviours associated with the roles. The speaker segmentation is a prerequisite for the extraction of features and therefore the approach proposed here can not be applied if the automatic speaker segmentation fails. This can be the case if the data was collected in a noisy environment or contains a substantial part of overlapping speech. This first limitation of the approach can be partially avoided if the experimenter controls the recording of the data by ensuring that the data is recorded using a lapel microphone. The second limitation of the approach depends on the type of roles that need to be recognized. If the features extracted from the prosody and the turn taking do not capture behaviours associated with the roles, the accuracy will be very low. This limitation is more difficult to avoid and can only be addressed on a case by case basis depending on the roles. The development of new features or the use of new modalities to capture those behaviours will need to be tested on new data collections. It is not possible to predict the performance of the approach on a new set of roles and only careful

Corpus	P	T	PT
C1 (A)	0.84	0.94	0.94
C2 (A)	0.84	0.93	0.93
C1 (M)	0.93	0.99	0.99
C2 (M)	0.84	0.98	0.98

Table 6.4: Purity. The table reports the purity of the role assignment, i.e. the coherence between speaker label and role.

experimentation will show if the approach presented here is adequate.

There is one last particularity of the model we need to investigate. As a role is assigned to each turn, the same person can be assigned multiple roles as the conversation evolves. The model does not enforce the constraint of one role per speaker. This is a desirable characteristic of the approach because in many scenarios individuals can play different roles in the same conversation. In the case of the data collection used in this experiment, however, this is a disadvantage as the roles considered are static. Each participant plays only one role in each interaction.

To assess if adding the constraints of one role per participant was useful, we investigate the coherence between speaker labels and roles. The coherence of the role assignment is measured with the *Average speaker purity* π_s, a metric showing to what extent all the speaking time of one speaker has been attributed the same role:

$$\pi_s = \sum_{l=1}^{N_s} \sum_{k=1}^{N_r} \frac{d_l}{D} \frac{d_{lk}^2}{d_l^2} \tag{6.9}$$

where D is the total duration, N_s is the number of speakers, N_r is the number of roles, d_{lk} is the duration where role k has been attributed to speaker l, and d_k is the total duration of role k and d_l is the total speaking time of speaker l. It is important to note that this definition does not refer to the actual role of the person. A very high purity does not mean that the correct role has been assigned to the person, but rather that the same role has been assigned to the person over the interaction.

Table 6.4 reports the purity of the role assignment, i.e. the coherence between speaker labels and roles. For T and PT features, the purities are always higher than 0.9 and this clearly suggests that the same person tends to be assigned always the same role. Therefore, it is not

Role	AM	SA	GT	HR	WM	IP
C1 (A, P)	66.1%	0.0%	60.5%	88.2%	90.2%	0.0 %
C2 (A, P)	43.1%	9.7%	92.1%	94.7%	0.9%	N/A
C1+C2 (A, P)	37.8%	0.6%	72.0%	83.6%	67.8%	N/A
C1 (M, P)	94.7%	77.8%	93.3%	100%	93.9%	31.2%
C2 (M, P)	70.1%	15.4%	94.6%	96.3%	0.0%	N/A
C1+C2 (M, P)	71.1%	34.2%	92.7%	98.4%	69.1%	N/A
C1 (A, T)	96.5%	11.7%	94.1%	97.8%	96.0%	13.7%
C2 (A, T)	72.6%	85.8%	92.6%	95.0%	13.3%	N/A
C1 + C2 (A, T)	92.6%	16.2%	93.1%	94.8%	72.6%	N/A
C1 (M, T)	99.9%	96.6%	99.0%	100%	99.1%	93.0%
C2 (M, T)	99.4%	95.4%	98.8%	96.3%	81.5%	N/A
C1 + C2 (M, T)	99.7%	92.4%	99.3%	100%	84.6%	N/A
C1 (A, PT)	96.5%	11.6%	94.1%	97.4%	93.5%	12.3%
C2 (A, PT)	76.4%	85.8%	92.9%	95.0%	41.5%	N/A
C1 + C2 (A, PT)	91.2%	18.2%	92.3%	92.5%	74.3%	N/A
C1 (M, PT)	99.7%	96.6%	98.8%	100%	97.9%	88.7%
C2 (M, PT)	98.2%	85.3%	97.5%	100%	74.1%	N/A
C1 + C2 (M, T)	99.0%	90.8%	99.0%	100%	87.6%	N/A

Table 6.5: Role accuracy. The table reports, for each feature set and for each corpus, the performance for the different roles. Results are reported for only prosodic features (*P*), only turn-organization features (*T*), or the combination of the two (*PT*), over both the turns extracted automatically (A) and the manual speaker segmentation (M). Each column corresponds to a role.

necessary to enforce one role per person as the model is already performing as expected.

6.4 Conclusion

This chapter has proposed an approach for automatic role recognition based on turn-taking and prosodic behavior. To the best of our knowledge, this is the first work showing that roles, at least in the settings considered, are associated to specific ways of speaking corresponding to different regions of the prosodic features space. Furthermore, the experiments show that, in some cases, the combination of turn-taking and prosodic features improves the performance to a statistically significant extent the performance. The recognition step is performed with linear chain CRFs where the feature functions allow one to capture relationships between roles and observation values or between roles following one another in the turn sequences.

The main source of error in the automatic case is the speaker clustering. The delay between the actual and detected speaker changes results into an accuracy loss of more than 10% that can be eliminated only by obtaining a better speaker segmentation. This means that further progress on role modelling can be obtained only by working on other, possibly more spontaneous, data and roles that are less constrained than those considered in this work, and possibly relevant to more general human-human interaction scenarii, like, e.g., those described in general theories of social interaction [58]. This might help to identify better directions for the improvement of the models such as the use of kernels exploiting the correlations between features.

Chapter 7

Conclusion

This thesis has addressed the problem of automatic role recognition in conversational broadcast data and meetings. Roles are present in every human interactions and consist of a set of shared expectations about the behaviour of the participants. The main functions of roles are to avoid surprises [8], to organize social interaction and to facilitate a smoother interaction. In order to automatically detect the roles, we have applied machine learning techniques to map behavioural cues to a predefined set of roles. The approach used in this thesis is composed of three steps:

1. Detecting participants involved in the social interaction.
2. Extracting audio behavioural cues.
3. Assigning roles to participants using the behavioural cues.

This approach has been tested using different types of features on three data sets: a corpus of meetings, a corpus of broadcast news and a corpus of talk shows.

The three motivations for addressing this problem are: the ubiquitous presence of social interaction in everyday life, the relation between roles and expected behaviours, and the multiple application of roles. The first motivation is not only a direct consequences of the fact that social interactions are one of the most important aspect of our everyday life but also that social interactions are present in most multimedia data such as radio and television programs. The second motivation comes from the field of psychology (see Section 2.2). The work in this field demonstrates that roles play an important part in shaping the behaviour of

participants in a social interaction. The understanding of roles is therefore one of the keys to understand social interactions. Finally, roles can be used in at least two ways. First, they can be used to enrich the annotation of multimedia data for summarization and information retrieval. Second, the can be used by an interactive system to make a sense of the surroundings and to select appropriate behaviours.

The rest of this chapter is organized in two sections. In Section 7.1, the main results of this thesis are summarized. Finally, in Section 7.2, some directions for possible future work are detailed.

7.1 Results

In this section, we will give an overview of all the results presented in this thesis. Those results were presented in Chapter 4, Chapter 5 and Chapter 6. Overall, three models were proposed and tested on three different corpora composed of more than 90 hours of audio. The use of the same corpora for all the experiments allows an easy comparison between the different models.

Over the course of the thesis, the model was improved to take into account more modalities (spoken words and prosody), to allow for a better modelling of the dependency between the roles, and to take into account sequential aspects of the role assignment. For each model presented in the thesis, we will highlight their main contributions and how well they perform on the different data collections. We have also highlighted how each result informs us with respect to the three aspects we investigated during this thesis:

- the use of models for the dependency between roles,
- the contribution of different modalities to the effectiveness of roles recognition approach,
- and the effectiveness of our approach for different settings.

7.1.1 Modelling dependency between roles

The first aspect investigated in this thesis is the exploitation of the dependency between the roles. When people interact, the role they play is dependent on the role played by the other participants in the interaction. For example, in a broadcast news, there can be only on anchor.

The first model (Chapter 4) is a Naive Bayes model that uses only features extracted from social network analysis. Those features account for who interacts with whom and when. The model also takes into account role dependency explicitly. The experiment of this chapter allow us to answer the first research question on the influence of modelling the role dependency.

- Modelling the dependence between roles leads to statistically significant improvements for radio talk shows and meetings, while it decreases the performance for Radio news broadcasts

The main practical consequence is that the dependency between the roles should be modelled only if results on a validation set demonstrate an improvement of the performance.

The model from this chapter also serves as the baseline system for the rest of the results and was tested on all three corpora. The detailed performance is reported in Table 4.5 and Table 4.6. The performance for the fully automatic approach of the model was measured using accuracy (the percentage of time correctly labelled) and is 82.4% on C1 (radio news broadcast), 87.8% on C2 (radio talk shows) and 46.4% on C3 (AMI meetings). Our approach perform much better using the manual segmentation as input, with an accuracy of 97.0% on C1, 98.1% on C2 and 56% on C3. This difference in performance between the manual segmentation and the automatic segmentation is only statistically significant for C1 and C2 and is due to errors in the automatic segmentation. Those errors are either inaccurate turn boundaries causing a decrease in accuracy or spurious speakers (either one speaker recognized as two different speakers by the system or two speakers recognized as one) causing errors in the computation of the SNA features. Those results also indicate that this model is well suited for roles based on norms, as those roles impose constrains on the behaviour of people, but may not be indicated for roles based on beliefs and preferences, as those roles lead to a more free interaction.

7.1.2 Verbal and non-verbal features

The experiment in Chapter 5 investigates the use of two modalities for the role recognition problem. The first type of features are the features extracted from the Social Network Analysis and are similar to the features from the previous model. The second type of features are extracted from the automatic speech transcription. Those features, accounting for the lexical choices, are n-grams with $n = 1, 2,$ and 3 and the model used is the BoosTexter model. This model was tested only on the AMI meeting corpus because we could not get access to a transcript of the data in French.

For the combination, two models were trained (one for each modality) and their output was fused by combining the probabilities given by each model. The results of this experiment (detailed in Table 5.2) give us an insight on the contribution of those two modalities when applied to the role recognition problem.

- Both modalities lead to role recognition results significantly higher than chance when used separately, but the best results are obtained with their combination.

From a practical point of view, adding uncorrelated modalities to a model for role recognition improves the performance. This improvement occurs even when one of the two models perform significantly worse than the other.

In term of accuracy, the performance of the lexical model is 67.1% using a fully automatic approaches and 76.7% using the manual segmentation and transcription. The performance of the combination of the two modalities is 67.9% on the automatic segmentation and 78.0% on the manual annotation. Those numbers represent a significant improvement with respect to the model based on SNA only. However, the need of a transcript of the meeting made this model more costly to run and language dependent. This need is the most important limitation of this model.

7.1.3 Prosody features

The experiment in Chapter 6 investigate the use of prosody and turn-taking features for recognizing roles. Prosody features account for the way people speak. The turn taking

features account for how the turns are organized between the different participants. The model used in this chapter is based on the Conditional Random Field (CRF) model that has several advantages (see Chapter 3 for a more detailed discussion). The first advantage is the ability to take into account the dependence between the turns and to exploit some of the temporal dependency of roles (which roles follow or precede the current role). CRF can also learn the relative weights of different modalities.

The experiments were conducted on the data from the radio news broadcasts and the talk-shows. The data from the meetings was not used as previous experiments (Chapter 4) have shown that this type of approach is not particularly well suited on this data.

The results with this model confirm, on radio news broadcast and talk-shows, the conclusion made previously on the meeting data with respect to the combination of different modalities:

- Prosody and Turn-taking both lead to role recognition results significantly higher than chance when used separately, but the best results are obtained with their combination.

In term of accuracy, the performance of this model was 89.7% on the news broadcasts and 87.0% on the talk-shows when using the fully automatic approaches (see Table 6.3 for the detailed results). When using the manual segmentation, the performance increase to 99.1% on the news broadcasts and 96.2% on the talk-shows. From those performances, we can make two conclusions concerning the performance of this role recognition system:

- The model based on CRF and using features from prosody and turn-taking have the best performance on news broadcasts and talk-shows.
- The performance obtained using manual segmentation suggest that improving the speaker segmentation is the surest way to improve the performance of a fully automated system.

The main limitation is that the results presented in this section have been obtained on conversational data where the presence of professional speaker (the anchor and the journalists) impose a structure on the discussion. The performances presented here may not be confirmed on more spontaneous data.

7.1.4 Combined Data Collection

One more experiment was conducted in Chapter 6 by training one model on both talk-show and news broadcast. Both data collection have a similar set of roles (Anchorman, Second Anchorman, Guest, Weather Man and Headline Reader) but the different settings lead to different behaviours associated with the roles. The setup was similar to the previous experiment. The model was based on Conditional Random Field and the features were extracted from the prosody and the turn-taking patterns.

The aim of that experiment was to assess to robustness of the model to change in the roles. The accuracy for this model was 86.7% when using the automatic segmentation and 96.5% when applied to the manual segmentation. Those performances are very close to the best results of the models trained only on the talk-shows or the news broadcast presented in the previous section.

We can conclude from this experiment that:

- Model based on CRF and trained on broadcast news and talk shows performs similarly than models specialized on each data set

In other words, we were able to train a single model for all the broadcast data that was able to detect roles that were similar but not identical with close to state of the art performance. Thus, this approach seems to deal effectively with different settings at the same time.

From a practical point of view, the model can be applied indiscriminately to different settings, without the need to detect or provide the social setting beforehand. This model could, for example, be applied to a radio broadcast that has not been segmented into shows.

7.2 Future Work

We will now focus on possible area for future researches. Given the relatively early stage of the research on role recognition, there are several possible directions worth investigating. The first direction is the improvement of the data collections currently available. The second direction is the identification and evaluation of possible application for the role recognition.

The third direction worth exploring is to extend to work to new modalities and more challenging settings. Finally, the fourth direction is the use of unsupervised approaches for role detection. In the rest of this section, we will present each direction with their main challenges.

7.2.1 Data Collections

One of the challenge for developing better models for roles is the lack of large annotated corpora for roles. As mentioned in the state of the art (Section 2.5), there is currently no standard data collection in the domain of Social Signal Processing. Another limitation of the currently available data is that they cover only a very limited number of settings: meetings and broadcast news. However, human interactions and roles occur in a much wider set of settings. Collecting more varied data is an essential steps for improving the state of the art of roles recognition.

Collecting the data is only half of the battle, annotating the collected data in term of roles is the other challenge. Producing high quality annotations poses many challenges. The first is that this requires a good understanding of the underlying psychological theory. The annotation process would benefit from collaboration with researchers in social sciences in order to insure that the roles are well defined and adequate for the data at hand. The second challenge is that annotating data is time consuming. An interesting way to solve this problem is to use crowdsourcing for the annotation. The idea behind crowdsourcing is to ask a large number of person, not necessarily experts, to annotate a part of the data in exchange for remuneration. This approach has been shown to be effective [107] and should be investigated to obtain role annotation over a large dataset.

The collection and annotation of data is certainly not a very exciting task, but efforts along this direction will definitely help the research on role recognition. For example, once such a data set has been collected, an evaluation campaign or a challenge could be organized to assess and compare the effectiveness of different approach for roles recognition. This direction of research, by itself, may not bring many new and exiting results, but is at the start of all the future work proposed below. Only the collection of new data will allow to evaluate the progress made.

7.2.2 Applications

An interesting research question is the identification and evaluation of applications for the automatic detection of roles. Given that the role is low level information, simply displaying them to the user may not be useful. However, role information could be used as input for other automated systems. The first challenge is to identify applications that would benefit the most from role annotations. As roles are tightly related to the behaviours of participants, tasks aimed at understanding human interaction are possibly the best candidates. For example, group activity detection could benefit from knowing the roles of the different participants. Another possible application is the detection of the moderator in a debate for the analysis of conflict. Once a suitable application has been selected, it is important to evaluate the contribution of the role detection. The most obvious way to measure it would be by investigating the influence of role on the performance of the approach.

Another field with applications that could benefit from role recognition is the human computer interfaces field. In that context, roles information could be applied in two ways. First, role recognition could be used for sensing the social context of the user and used to improve the interaction. In that situation, role recognition is used to enhance the information available to the system and help provide a more meaning-full interaction. The second possibility is to take advantage of the relation between roles and behaviour. Roles could be used to select appropriate behaviours for the system in order to make it behave more humanly. In both situations, extensive user studies need to be conducted to evaluate the effectiveness of this approach. The evaluation is made more difficult by the fact that any change to the system leads to a change in the experience by the user and new data has to be collected.

7.2.3 Settings and Modalities

As already mentioned, the approaches proposed in this thesis have been tested on a small set of social settings. A natural extension of the current work is to investigate different settings and roles. Obviously, this extension of the work is dependant on the data and annotation available. One question that has arisen often during this work is the influence of the culture and language on the proposed approach. In that context, it is worth investigating the use of *transfer learning* [108] for role recognition. The idea behind transfer learning is to use

annotation from a different data collection to help train a model on a new data collection. This technique can be used to reduce the need for labelling of new datasets. As the role recognition task offers a wide variety of settings, it is well adapted to the development of transfer strategy. For example, one could investigate how to adapt models based on prosody in French to a corpus of English data. Another interesting approach would be to use data collected in one setting (meetings) to improve the model used in another settings (informal family reunion). Finally, by using transfer learning, the researcher can access to larger annotated data sets and more complex models can be trained.

The approach in this thesis was focused on audio data and relatively short time scales (each interaction lasted between 5 minutes and 1 hour). By using new sensors giving access to new modalities, roles could be studied on a much larger time scale. For example, most smart-phones are now equipped with a GPS and this allow to track people over an extended period of time (see [109] for an example of such a data collection). Detecting the roles played by people on those larger scales is still an open problem. Similarly, social networks and online forums are another place of constant human interaction, involving large communities. That settings with text as a modality offer a rich environment for the detection of roles. Those are only two possible examples of new modalities. As role are present in most human interaction, their automatic detection can be investigated on a large variety of settings.

7.2.4 Unsupervised Roles Detection

All the approaches presented in this thesis use supervised machine learning approaches for the role recognition. The models learn a mapping from the features to the roles using the labelled data. The final direction of research we propose is to use *unsupervised* machine learning techniques to automatically discover roles in the data. Unsupervised machine learning does not use the labels to learn a mapping, but tries to discover hidden structures in unlabelled data. For example, common algorithm used for unsupervised learning are k-means and mixture models. From a role recognition perspective, this approach has the advantage that the set of role is not needed a-priori but can be discovered by the algorithm.

The main challenge facing this approach is that human interaction, in general, has a lot of structures and not all those structures are related to roles. Therefore, the structure discovered

by the algorithm may not be related to roles. From a technical point of view, finding the correct representation for the features and the correct algorithm is the main difficulty.

Another challenge faced by unsupervised approaches is the interpretation of the hidden structure discovered by the algorithm. This aspect offers a great opportunity for collaboration between social scientists and computer scientists. On the one hand, social scientists can explain and put into perspective the structures discovered by the algorithms. On the other hand, the computer scientists can explore large data collections automatically and bring to light patterns that would have been extremely difficult to find by manual exploration of the data. Of all the work proposed in this section, the unsupervised approach looks the most ambitious and promising.

7.3 Final words

This work has addressed the role recognition problem in machine intelligence terms, i.e. by trying to maximize the accuracy of the approach. During this thesis, the following novelties and aspects with respect to the state of the art were developed:

- This is the first work that combines turn-taking features and semantic information for role recognition (chapter 5).
- This is the first work that uses features extracted from the *prosody* to assign roles (chapter 6).
- This is the first work that works on roles from two different setups and trains one classifier to identify roles in two settings (chapter 6).

No attempt has been made to explain what are the behavioural patterns the roles corresponds to. There is still a lot to explore in role recognition problems, in particular for roles in less constrained settings.

Bibliography

[1] Aristotle, *Politics*, 350 BC, vol. 1.

[2] V. Richmond and J. McCroskey, *Nonverbal Behaviors in interpersonal relations*. Allyn and Bacon, 1995.

[3] M. Pantic, A. Nijholt, A. Pentland, and T. Huanag, "Human-centred intelligent human? computer interaction (hci^2): how far are we from attaining it?" *International Journal of Autonomous and Adaptive Communications Systems*, vol. 1, no. 2, pp. 168–187, 2008.

[4] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[5] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: a review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[6] H. Tischler, *Introduction to Sociology*. Harcourt Brace College Publishers, 1990.

[7] J. Scott and G. Marshall, Eds., *Dictionary of Sociology*. Oxford University Press, 2005.

[8] B. Biddle, "Recent developments in role theory," *Annual Review of Sociology*, vol. 12, pp. 67–92, 1986.

[9] R. Bales, "A set of categories for the analysis of small group interaction," *American Sociological Review*, vol. 15, no. 2, pp. 257–263, 1950.

[10] P. Slater, "Role differentiation in small groups," *American Sociological Review*, vol. 20, no. 3, pp. 300–310, 1955.

[11] S. Maskey and J. Hirschberg, "Automatic summarization of broadcast news using structural features," in *Eighth European Conference on Speech Communication and Technology*. Citeseer, 2003.

[12] A. Vinciarelli, "Sociometry based multiparty audio recordings summarization," in *18th International Conference on Pattern Recognition*, vol. 2. IEEE, 2006, pp. 1154–1157.

[13] ——, "Speakers role recognition in multiparty audio recordings using Social Network Analysis and duration distribution modeling," *IEEE Transactions on Multimedia*, vol. 9, no. 6, pp. 1215–1226, 2007.

[14] C. Weng, W. Chu, and J. Wu, "Movie analysis based on roles' social network," in *IEEE International Conference on Multimedia and Expo*, 2007, pp. 1403–1406.

[15] M. Knapp and J. Hall, *Nonverbal Communication in Human Interaction*. Harcourt Brace College Publishers, 1972.

[16] P. Ekman, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003.

[17] A. Pentland, *Honest signals: how they shape our world*. MIT Press, 2008.

[18] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis." *Psychological Bulletin*, vol. 111, no. 2, pp. 256–274, 1992.

[19] ——, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness." *Journal of Personality and Social Psychology*, vol. 64, no. 3, pp. 431–441, 1993.

[20] N. Ambady, M. Hallahan, and B. Conner, "Accuracy of judgments of sexual orientation from thin slices of behavior." *Journal of Personality and Social Psychology*, vol. 77, no. 3, pp. 538–547, 1999.

[21] P. Borkenau, N. Mauer, R. Riemann, F. Spinath, and A. Angleitner, "Thin slices of behavior as cues of personality and intelligence." *Journal of personality and social psychology*, vol. 86, no. 4, pp. 599–614, 2004.

[22] T. Oltmanns, J. Friedman, E. Fiedler, and E. Turkheimer, "Perceptions of people with personality disorders based on thin slices of behavior," *Journal of Research in Personality*, vol. 38, no. 3, pp. 216–229, 2004.

[23] H. Sacks, E. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[24] D. Greatbatch, "A turn-taking system for british news interviews," *Language in society*, vol. 17, no. 03, pp. 401–430, 1988.

[25] S. Wasserman and K. Faust, *Social network analysis*, 1994.

[26] C. Winship and M. Mandel, "Roles and positions: A critique and extension of the blockmodeling approach," *Sociological methodology*, vol. 1984, pp. 314–344, 1983.

[27] C. Darwin, *The descent of man*, 1871, vol. 1.

[28] R. Cowie and E. Douglas-Cowie, "Automatic statistical analysis of the signal and prosodic signs of emotion in speech," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 3. IEEE, 1996, pp. 1989–1992.

[29] G. Fant, *Acoustic theory of speech production*. Walter de Gruyter, 1970, no. 2.

[30] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[31] M. Mast, "Dominance as expressed and inferred through speaking time," *Human Communication Research*, vol. 28, no. 3, pp. 420–450, 2002.

[32] C. Bishop, *Pattern recognition and machine learning*. Springer New York, 2006.

[33] S. Geisser, *Predictive inference*. Chapman & Hall/CRC, 1993, vol. 55.

[34] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *The Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2004.

[35] M. Kearns and D. Ron, "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," *Neural Computation*, vol. 11, no. 6, pp. 1427–1453, 1999.

[36] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Weninger, F. Eyben, T. Bocklet *et al.*, "The interspeech 2012 speaker trait challenge," *Interspeech, Portland, Oregon*, 2012.

[37] I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[38] F. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[39] R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker, "The rules behind the roles: identifying speaker roles in radio broadcasts," in *Proceedings of the 17th National Conference on Artificial Intelligence*, 2000, pp. 679–684.

[40] J. Garofolo, E. Voorhees, C. Auzanne, V. Stanford, and B. Lund, "1998 trec-7 spoken document retrieval track overview and results," in *Broadcast News Workshop'99 Proceedings*. Morgan Kaufmann Pub, 1999, p. 215.

[41] Y. Liu, "Initial study on automatic identification of speaker role in broadcast news speech," in *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, June 2006, pp. 81–84.

[42] J. Kong and D. Graff, "Tdt4 multilingual broadcast news speech corpus," *Linguistic Data Consortium, http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp*, 2005.

[43] Y. Esteve, T. Bazillon, J. Antoine, F. Béchet, and J. Farinas, "The epac corpus: manual and automatic annotations of conversational speech in french broadcast news," *LREC, Malta*, 2010.

[44] H. Salamin, S. Favre, and A. Vinciarelli, "Automatic Role Recognition in Multiparty Recordings: Using Social Affiliation Networks for Feature Extraction," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1373–1380, 2009.

[45] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal *et al.*, "The AMI meeting corpus: A pre-announcement," in *Lecture notes in computer science*, vol. 3869, 2005, pp. 28–39.

[46] S. Banerjee and A. Rudnicky, "Using simple speech based features to detect the state of a meeting and the roles of the meeting participants," in *proceedings of International Conference on Spoken Language Processing*, 2004, pp. 2189–2192.

[47] M. Zancanaro, B. Lepri, and F. Pianesi, "Automatic detection of group functional roles in face to face interactions," in *proceedings of International Conference on Multimodal Interfaces*, 2006, pp. 47–54.

[48] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon, and B. Lepri, "Multimodal support to group dynamics," *Personal Ubiquitous Computing*, vol. 12, no. 3, pp. 181–195, 2008.

[49] J. Hall and W. Watson, "The effects of a normative intervention on group decision-making performance." *Human Relations*, 1970.

[50] C. Weng, W. Chu, and J. Wu, "Rolenet: Movie analysis from the perspective of social networks," *IEEE Transactions on Multimedia*, vol. 11, no. 2, pp. 256–271, 2009.

[51] B. Bigot, I. Ferrané, J. Pinquier, and R. André-Obrecht, "Speaker role recognition to help spontaneous conversational speech detection," in *Proceedings of International Workshop on Searching Spontaneous Conversational Speech*, 2010, pp. 5–10.

[52] W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi, and M. Zancanaro, "Using the influence model to recognize functional roles in meetings," in *Proceedings of the 9th International Conference on Multimodal Interfaces*, 2007, pp. 271–278.

[53] K. Laskowski, M. Ostendorf, and T. Schultz, "Modeling vocal interaction for text-independent participant characterization in multi-party conversation," in *In proceedings of the 9th ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, June 2008, pp. 148–155.

[54] N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür, and A. Vinciarelli, "Role recognition for meeting participants: an approach based on lexical information and Social Network Analysis," in *Proceedings of the ACM International Conference on Multimedia*, 2008, pp. 693–696.

[55] S. Maskey and J. Hirschberg, "Soundbite Detection in Broadcast News Domain," in *Proceedings of Interspeech 2006*, 2006, pp. 1542–1546.

[56] J. Quinlan, *C4. 5: programs for machine learning*. Morgan Kaufmann, 1993.

[57] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[58] J. Levine, *Small groups*. Psychology Press, 2006.

[59] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 305–317, 2005.

[60] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner, and H. Bourlard, "Modeling human interaction in meetings," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 4. IEEE, 2003.

[61] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Modeling individual and group actions in meetings with layered HMMs," *IEEE Transactions on Multimedia*, vol. 8, no. 3, pp. 509–520, 2006.

[62] A. Dielmann and S. Renals, "Recognition of dialogue acts in multiparty meetings using a switching DBN," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 7, pp. 1303–1314, 2008.

[63] S. Reiter, B. Schuller, and G. Rigoll, "Hidden conditional random fields for meeting segmentation," in *IEEE International Conference on Multimedia and Expo*, 2007, pp. 639–642.

[64] R. Rienks and D. Heylen, "Dominance detection in meetings using easily obtainable features," *Machine Learning for Multimodal Interaction*, pp. 76–86, 2006.

[65] R. Rienks, D. Zhang, D. Gatica-Perez, and W. Post, "Detection and application of influence rankings in small group meetings," in *Proceedings of the 8th International Conference on Multimodal interfaces*. ACM, 2006, pp. 257–264.

[66] K. Otsuka, J. Yamato, Y. Takemae, and H. Murase, "Quantifying interpersonal influence in face-to-face conversations based on visual attention patterns," in *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 2006, pp. 1175–1180.

[67] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations from non-verbal activity cues," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, no. 3, pp. 501–513, 2009.

[68] D. Gatica-Perez, I. McCowan, D. Zhang, and S. Bengio, "Detecting Group Interest-Level in Meetings," in *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 489–492.

[69] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous *et al.*, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," *Proc. INTERSPEECH*, pp. 2253–2256, 2007.

[70] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker, and H. Konosu, "Being bored? Recognising natural interest by extensive audiovisual integration for real-life application," *Image and Vision Computing*, vol. 27, no. 12, pp. 1760–1774, 2009.

[71] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[72] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of International Conference on Machine Learning*, 2001, pp. 282–289.

[73] C. Sutton and A. McCallum, "An introduction to conditional random fields for relational learning," in *Introduction to statistical relational learning*, L. Getoor and B. Taskar, Eds. The MIT Press, 2007.

[74] M. Jordan, *Learning in graphical models*. Kluwer Academic Publishers, 1998.

[75] F. Jensen, *An introduction to Bayesian Networks*. UCL press London, 1996.

[76] F. Jensen and T. Nielsen, *Bayesian Networks and decision graphs*. Springer Verlag, 2007.

[77] J. Pearl, *Bayesian networks: A model of self-activated memory for evidential reasoning*. Computer Science Department, University of California, 1985.

[78] D. Heckerman, "A tutorial on learning with bayesian networks," in *Innovations in Bayesian Networks*, D. Holmes and L. Jain, Eds. Springer Berlin / Heidelberg, 2008, pp. 33–82.

[79] M. Maron, "Automatic indexing: an experimental inquiry," *Journal of the ACM*, vol. 8, no. 3, pp. 404–417, 1961.

[80] H. Warner, A. Toronto, L. Veasey, and R. Stephenson, "A mathematical approach to medical diagnosis," *The Journal of the American Medical Association*, vol. 177, no. 3, p. 177, 1961.

[81] F. Jelinek, *Statistical methods for speech recognition*. the MIT Press, 1997.

[82] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[83] J. Yedidia, W. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," in *Exploring artificial intelligence in the new millennium*, G. Lakemeyer and B. Nebel, Eds. Morgan Kaufman, 2003, pp. 239–270.

[84] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, vol. 45, pp. 503–528, 1989.

[85] I. Poggi and F. D'Errico, "Cognitive modelling of human social signals," in *Proceedings of the 2nd International Workshop on Social Signal Processing*, 2010, pp. 21–26.

[86] K. Murphy, "An introduction to graphical models," University of British Columbia, Tech. Rep., 2001.

[87] H. Wallach, "Conditional Random Fields: An introduction," Department of Computer and Information Science, University of Pennsylvania, Tech. Rep. MS-CIS-04-21, 2004.

[88] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *Signal Processing Letters, IEEE*, vol. 11, no. 8, pp. 649–651, 2004.

[89] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," in *Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding*, 2003, pp. 411–416.

[90] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," in *Ninth International Conference on Spoken Language Processing*. Citeseer, 2006.

[91] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2001.

[92] S. Wrigley, G. Brown, V. Wan, and S. Renals, "Speech and crosstalk detection in multichannel audio," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 84–91, 2004.

[93] S. Kirkpatrick, C. Gelatt Jr, M. Vecchi, and A. McCoy, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, pp. 671–679, 1983.

[94] D. Jayagopi, B. Raducanu, and D. Gatica-Perez, "Characterizing conversational group dynamics using nonverbal behaviour," in *Proceedings of IEEE International Conference on Multimedia and Expo*, 2009, pp. 370–373.

[95] A. Vinciarelli, M. Pantic, H. Bourlard, and A. Pentland, "Social signal processing: state-of-the-art and future perspectives of an emerging domain," in *Proceeding of the 16th ACM international conference on Multimedia*. ACM, 2008, pp. 1061–1070.

[96] R. E. Schapire and Y. Singer, "BoosTexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000. [Online]. Available: citeseer.ist.psu.edu/schapire00boostexter.html

[97] D. Tax, R. Duin, and M. Van Breukelen, "Comparison between product and mean classifier combination rules," in *Proc. Workshop on Statistical Pattern Recognition, Prague, Czech*, 1997.

[98] G. Psathas, *Conversation analysis: The study of talk-in-interaction*. Sage Publications, 1995.

[99] J. Bilmes, "The concept of preference in conversation analysis," *Language in Society*, vol. 17, no. 2, pp. 161–181, 1988.

[100] P. Ekman, W. Friesen, M. O'Sullivan, and K. Scherer, "Relative importance of face, body, and speech in judgments of personality and affect," *Journal of Personality and Social Psychology*, vol. 38, no. 2, pp. 270–277, 1980.

[101] C. Nass and S. Brave, *Wired for speech*. MIT press, 2005.

[102] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[103] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, no. 9/10, pp. 341–345, 2002.

[104] D. Addington, "The relationship of selected vocal characteristics to personality perception," *Communication Monographs*, vol. 35, no. 4, pp. 492–503, 1968.

[105] G. Ray, "Vocally cued personality prototypes: An implicit personality theory approach," *Communication Monographs*, vol. 53, no. 3, pp. 266–276, 1986.

[106] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the International Joint Conference on Artificial Intelligence*, vol. 14, 1995, pp. 1137–1145.

[107] P. Hsueh, P. Melville, and V. Sindhwani, "Data quality from crowdsourcing: a study of annotation selection criteria," in *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Association for Computational Linguistics, 2009, pp. 27–35.

[108] S. Pan and Q. Yang, "A survey on transfer learning," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 22, no. 10, pp. 1345–1359, 2010.

[109] J. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T. Do, O. Dousse, J. Eberle, and M. Miettinen, "The mobile data challenge: Big data for mobile computing research," in *Mobile Data Challenge by Nokia Workshop, in conjunction with Int. Conf. on Pervasive Computing, Newcastle, UK*, 2012.