# Social Network Analysis for Automatic Role Recognition

THÈSE N$^{o}$4931 (2010)

PRESENTÉE LE 8 DECEMBRE 2010

IDIAP RESEARCH INSTITUTE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

**Sarah Favre**

acceptée sur proposition du jury:

Prof. P. Vandergheynst, Président du jury

Prof. H. Bourlard, Directeur de thèse

Dr. A. Vinciarelli, Co-directeur de thèse

Dr. F. Pianesi, rapporteur

Prof. S. Renals, rapporteur

Prof. J.-Ph. Thiran, rapporteur

# ABSTRACT

The computing community has shown a significant interest for the analysis of social interactions in the last decade. Different aspects of social interactions have been studied such as dominance, emotions, conflicts, etc. However, the recognition of roles has been neglected whereas these are a key aspect of social interactions. In fact, sociologists have shown not only that people play roles each time they interact, but also that roles shape behavior and expectations of interacting participants. The aim of this thesis is to fill this gap by investigating the problem of automatic role recognition in a wide range of interaction settings, including production environments, e.g. news and talk-shows, and spontaneous exchanges, e.g. meetings.

The proposed role recognition approach includes two main steps. The first step aims at representing the individuals involved in an interaction with feature vectors accounting for their relationships with others. This step includes three main stages, namely segmentation of audio into turns (i.e. time intervals during which only one person talks), conversion of the sequence of turns into a social network, and use of the social network as a tool to extract features for each person. The second step uses machine learning methods to map the feature vectors into roles. The experiments have been carried out over roughly 90 hours of material. This is not only one of the largest databases ever used in literature on role recognition, but also the only one, to the best of our knowledge, including different interaction settings. In the experiments, the accuracy of the percentage of data correctly labeled in terms of roles is roughly 80% in production environments and 70% in spontaneous exchanges (lexical features have been added in the latter case). The importance of roles has been assessed in an application scenario as well. In particular, the thesis shows that roles help to segment talk-shows into stories, i.e. time intervals during which a single topic is discussed, with satisfactory performance.

The main contributions of this thesis are as follows: To the best of our knowledge, this is the first work where social network analysis is applied to automatic analysis of conversation recordings. This thesis provides the first quantitative measure of how much roles constrain conversations, and a large corpus of recordings annotated in terms of roles. The results of this work have been published in one journal paper, and in five conference articles.

**Keywords: Social Network Analysis, Role Recognition, Semantic Segmentation, Broadcast Data, Meeting Recordings, Turn-Taking Analysis, Bayes Classifiers, Hidden Markov Models, Statistical Language Models**.

# RÉSUMÉ

Ces dix dernières années, la communauté scientifique a montré un certain intérêt pour l'analyse des interactions sociales. Différents aspects des interactions sociales ont déjá été étudiés tels que la reconnaissance des personnes dominant les conversations, des émotions, et des conflits prśents lors d'interactions. Néanmoins, bien que ce soit un aspect clé des interactions sociales, la reconnaissance des rôles a été négligée. Les sociologues ont démontré non seulement que les gens jouaient un rôle à chaque fois qu'ils entraient en interaction, mais également que les rôles modifiaient les comportements et les attentes des protagonistes. Le but de cette thèse est de combler le vide existant en analysant le problème de la reconnaissance automatique des rôles dans un large choix de types d'interactions, comprenant des informations et des débats, ainsi que des types d'interactions plus spontanées comme par exemple les réunions.

L'approche proposée pour la reconnaissance des rôles comprend deux étapes principales. La première vise à représenter les individus interagissant par des caractéristiques définissant leurs relations avec les autres. Cette étape se décompose elle-même en trois sous-étapes principales: le repérage des différentes interventions des participants, l'identification du réseau social, et l'utilisation de ce dernier en tant qu'outil permettant d'extraire des particularités pour chacun des intervenants. La deuxième étape utilise des méthodes de classification afin d'attribuer un rôle à chaque intervenant.

L'expérimentation a porté sur près de 90 heures de matériel audio. Il s'agit là non seulement de l'une des bases de données les plus importantes utilisée dans la reconnaissance des rôles, mais encore la seule, à notre connaissance, comprenant différentes formes dinteractions. Lors de nos expérimentations, le pourcentage des données correctement étiquetées est d'environ 80% dans les informations et les débats et de 70% dans le cas de conversations plus spontanées comme les réunions (dans ce dernier cas, des caractéristiques lexicales ont été ajoutées).

L'importance des rôles a également été utilisée dans le développement d'une application. La thèse montre en particulier que les rôles aident à segmenter, de manière tout à fait satisfaisante, les débats en intervalles de temps pendant lesquels un sujet unique est abordé.

Les contributions essentielles de cette thèse sont les suivantes: à notre connaissance, c'est la première fois qu'un travail d'analyse de réseau sociaux porte sur l'étude de conversations. Cette thèse fournit pour la première fois la preuve quantifiée de la façon dont les rôles modèlent les conversations. De plus, les travaux de recherche ont porté sur un très grand nombre d'enregistrements annotés en terme de rôle.

Les résultats de ces travaux ont été publiés dans une revue ainsi que dans cinq articles de conférence.

**Mots clés: Social Network Analysis, Reconnaissance des rôles, Segmentation sémantique, Informations radiophoniques, Débats, Réunions, Analyse des interventions, Classificateurs bayésiens, Modèles de Markov cachés, Modèles de languages statistiques**.

# ACKNOWLEDGMENTS

Now that the endpoint of my thesis work has been finally laid down, it's a real pleasure for me to thank my thesis director, Professor Hervé Bourlard, who gave me the chance to study in his institute, and for his confidence.

I would like also to warmly thank Dr. Alessandro Vinciarelli, who was my supervisor during these four years. Without him I won't have learnt so many things and won't have enjoyed so much the subject of my thesis. Alessandro was a fair, grateful, always available and generous supervisor. But he also was a demanding supervisor, who teach me to verify my statements, to support my knowledge and to improve my way of expressing me. He had a demanding side which made me progressing but knew also how to always motivate me and congratulate me once it was done. He helped me to be a complete PhD student, by combining knowledge and capacity to share them. Alessandro shared lots of his knowledge with me, I will always be grateful to him for this. Merci Alessandro!

I also wish to thank my colleague and friend Hugues Salamin for his help, for our collaborative work, for the numerous hours of discussions we shared, and of course for the great time we had together remodelling the world!

I am grateful to the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM2) through the Multimodal Content Abstraction IM2.MCA, and to the IDIAP Research Institute, who have funded my research.

I am quite honored that Prof. Steve Renals, Prof. J.-P. Thiran, and Dr. Fabio Pianesi have accepted to be the members of my jury thesis, as well as Prof. P. Vandergheynst as President.

My gratitude also goes to all my IDIAP colleagues and specially to my officemates Gelareh, Nicolae and Alfred and to Nadine and Sylvie for their administrative support, their availability and their smiles. A number of colleagues deserve a special mention for their scientific and technical support: Dr. Fabio Valente, Dr. John Dines, Dr. Phil Garner, Dr. Petr Motlicek, the system guys (always available and nice), and the developer guys for their help and kindness. I am also thankful to Guillaume, Tristan,

Bastien, Yann, Pierre, Hari, Ganga, and others for the great time and fun we had together.

Special thanks to my closer friends: Loriane, Sophie, and Stéphanie who often listened to me about the ongoing of my research work.

Finally, I wish to warmly thank my family, my parents, my dear sister, my grandmother and my husband. Their unconditional support, encouragement and love provided me with that precious peace of mind which is an essential ingredient to the completion of such an intellectual task. Merci Greg pour ta patience et de m'avoir permis d'aller jusqu'au bout de ce travail.

# Contents

**1 Introduction** **7**

1.1 Motivation . . . 7

1.2 Contributions . . . 9

1.3 Organization of the Thesis . . . 10

**2 State-of-the-art** **11**

2.1 Social Network Analysis . . . 11

2.2 Role Recognition . . . 13

2.2.1 Recognition of Formal Roles . . . 14

2.2.2 Recognition of Informal Roles . . . 15

2.3 Analysis of Social Interactions in Small Groups . . . 17

2.4 Conclusions . . . 18

**3 Role Recognition** **19**

3.1 Feature Extraction . . . 21

3.1.1 Speaker Diarization for Broadcast Data . . . 22

3.1.2 Speaker Diarization for Meeting Data . . . 24

3.1.3 Affiliation Network Extraction . . . 25

3.1.4 Duration Distribution Extraction . . . 26

3.1.5 Summary . . . 27

3.2 Role Recognition Approach based on Bayesian Classifiers . . . 27

3.2.1 Modeling Interaction Patterns . . . 28

3.2.2 Modeling Durations . . . 30

3.2.3 Estimating Role Probabilities . . . 30

3.3 Databases, Experimental Protocols and Performance Measures . . . 32

3.3.1 Data and Roles . . . . . 32
3.3.2 Experimental Setup . . . . . 35
3.3.3 Performance Measures . . . . . 37
3.4 Role Recognition Results for Bayesian Classifiers based Approach . . . . . 39
3.4.1 Results . . . . . 39
3.4.2 Influence of interaction patterns . . . . . 43
3.5 Role Recognition Approach based on Probabilistic Sequential Models . . . . . 44
3.5.1 Modeling Sequential Information . . . . . 45
3.5.2 Experiments and Results . . . . . 46
3.6 Combination of Interaction and Lexical Patterns . . . . . 49
3.6.1 Social Affiliation Networks Based Role Recognition . . . . . 50
3.6.2 Lexicon Based Role Recognition . . . . . 51
3.6.3 Combination Approach . . . . . 52
3.6.4 Experiments and Results . . . . . 53
3.7 Role Recognition Discussion . . . . . 54

**4 Semantic Segmentation 57**

4.1 Structure Detection . . . . . 58
4.1.1 Social Network Analysis for Structure Detection . . . . . 59
4.1.2 Duration Distribution Modeling for Structure Detection . . . . . 61
4.1.3 Experiments and Results . . . . . 63
4.2 Story Segmentation . . . . . 64
4.2.1 Story Segmentation Approach . . . . . 65
4.2.2 Experiments and Results . . . . . 66
4.3 Semantic Segmentation Discussion . . . . . 68

**5 Conclusion 69**

5.1 Conclusions . . . . . 69
5.2 Future Research Directions . . . . . 69

**Bibliography 71**

**A Curriculum Vitae** **79**

# List of Figures

3.1 Role recognition approach. The picture shows the two main stages of the approach: the features extraction and the actual role recognition. . . . . . . . . . . . . . . . . . . . . . . . 20

3.2 Social Affiliation Network extraction. The events of the network correspond to the segments $e_j$ and the actors are linked to the events when they talk during the corresponding segment. The actors are represented using n-tuples $\mathbf{x}_a$ where the components account for the links between actors and events. . . . . . . . . . . . . . . . . . . . . . . . . . . 26

3.3 Distribution of the recording lengths. The histograms show the distribution of the recording lengths for corpora C1 and C3 (in corpus C2 each recording lasts for exactly 1 hour). 33

3.4 Distribution of recording participants. The histograms show the distribution of the number of persons participating in each recording for corpora C1 and C2 (in corpus C3 each recording involves exactly 4 persons). . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

3.5 Influence of hyperparameter $D$ over C1 database. The curve shows the role recognition performance in terms of accuracy for different $D$ values. . . . . . . . . . . . . . . . . . . . 36

3.6 Influence of hyperparameter $D$ over C2 database. The curve shows the role recognition performance in terms of accuracy for different $D$ values. . . . . . . . . . . . . . . . . . . . 37

3.7 Influence of hyperparameter $D$ over C3 database. The curve shows the role recognition performance in terms of accuracy for different $D$ values. . . . . . . . . . . . . . . . . . . . 38

3.8 Overview of the role recognition approach combining two types of behavioral cues: interaction and lexical patterns. The two parallel paths produce separate decisions that are combined at the end of the process. . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50

4.1 Social Network: this figure shows the Social Network extracted from one of the recordings in our collection. We can see that the two speakers labeled as spk19 and spk8 are the two main central speakers. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 59

4.2 Average number of stories vs time. This plot shows the average number of stories (estimated at two minutes long time steps) as a function of the time. . . . . . . . . . . . . . . 61

4.3 Story Segmentation approach. This figure shows the three stages of the story segmentation approach: the first splits the audio into a sequence of single speaker segments (i.e. turn-taking), the second converts the turn-taking into features using SAN, the third maps the features into stories using HMMs. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 64

# LIST OF TABLES

2.1 Synopsis of role recognition results. The table provides a brief description of the data used in literature, as well as the performance achieved in the different works. . . . . . . . . . . 12

3.1 Summary of parameters in the training and inference steps in the automatic speech segmentation system for meeting data. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 24

3.2 Corpora. The table reports the main characteristics of the corpora used in the experiments. From left to right: number of recordings, interaction setting, total time, average recording length, average number of participants. Note that the length is the same (one hour) for all recordings in C2, and the number of participants is constant (four) in C3. In all other cases, the figures change from one recording to the other. . . . . . . . . . . . . . . . . . . . 33

3.3 Role distribution in broadcast data. The table reports the percentage of time each role accounts for in C1 and C2. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34

3.4 Role distribution in meetings. The table reports the percentage of time each role accounts for in the AMI meeting corpus (C3). . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 35

3.5 Role recognition performance based on Bayes classifiers 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 $\sigma$ of the performance achieved over the single recordings. The upper part of the table reports the results obtained over the output of the speaker segmentation (diarization), the lower part reports the results obtained over the manual speaker segmentation. . . . . . . . . . . . . 40

3.6 Role recognition performance based on Bayes classifiers 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 $\sigma$ of the performance achieved over the single recordings. The upper part of the table reports the results obtained over the output of the speaker segmentation (diarization), the lower part reports the results obtained over the manual speaker segmentation. . . . . . . . . . . . . . . . . . . . . . . . 41

3.7 Features comparison. Role recognition performance based on Bayes classifiers for C1 and C2. The table reports both the overall accuracy and the accuracy for each role. “SAN” stands for interaction features extracted through SAN, “T” stands for fraction of speaking time as features, and SNA + T stands for the combination of both types of features. . . . 44

3.8 Features comparison. Role recognition performance based on Bayes classifiers for C3. The table reports both the overall accuracy and the accuracy for each role. “SAN” stands for interaction features extracted through SAN, “T” stands for fraction of speaking time as features, and SNA + T stands for the combination of both kinds of features. . . . . . . . 44

3.9 Role recognition performance based on probabilistic sequential models over C1 and C2. The table reports both the overall accuracy and the accuracy for each role. The overall accuracy is accompanied by the standard deviation $\sigma$ of the performance achieved over the single recordings. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 47

3.10 Role recognition based on probabilistic sequential models over C3. The table reports both the overall accuracy and the accuracy for each role. The overall accuracy is accompanied by the standard deviation $\sigma$ of the performance achieved over the single recordings. . . . . 47

3.11 $PP$ stands for the perplexity measure of the different $n$-gram and $PP/|\mathcal{R}|$ is the proportion of the dictionary that has a probability higher than 0 to produce the $n$-gram sequence. . . 48

3.12 Diversity assessment. The table reports the accuracy of the percentage of data where the two approaches are both correct (C), both wrong (W), or one wrong and the other correct. 49

3.13 Role recognition results when combining interaction features (SAN) and lexical features over the meetings C3. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 53

4.1 Structure detection results. The table reports the accuracy (percentage of time correctly labeled in terms of semantic class) obtained using SNA and DDM approaches. . . . . . . 63

4.2 Story segmentation performance. The table reports the purity as a function of the number of segments $D$ used to split the news bulletins and capture the interaction patterns, and of the amount of variance retained after PCA. . . . . . . . . . . . . . . . . . . . . . . . . . . 67

4.3 Effect of the speaker diarization errors. The reported results have been obtained using D=14, for both manual and automatic speaker segmentations. . . . . . . . . . . . . . . . . . 68

# ACRONYMS

| | |
|---|---|
| AM | Anchorman |
| AMI | Augmented Multi-party Interaction |
| DDM | Duration Distribution Modeling |
| ECA | Embodied Conversational Agent |
| GMM | Gaussian Mixture Models |
| GT | Guest |
| HMM | Hidden Markov Model |
| HR | Headline Reader |
| ID | Industrial Designer |
| IP | Interview Participant |
| ME | Marketing Expert |
| MFCC | Mel Frequency Cepstral Coefficients |
| MLP | Multilayer Perceptron |
| PCA | Principal Component Analysis |
| PM | Project Manager |
| PP | Perplexity |
| PSP | Poisson Stochastic Process |
| SA | Second Anchorman |
| SAN | Social Affiliation Networks |
| SLM | Statistical Language Models |
| SNA | Social Networks Analysis |
| UI | User Interface Expert |
| WM | Weather Man |

# NOTATIONS

| Notation | Definition | Section |
|---|---|---|
| $M$ | initial number of states in the HMM for speaker diarization for broadcast data | 3.1.1 |
| $q$ | sequence of states in the HMM | 3.1.1 |
| $\mathcal{Q}$ | the set of all possible sequences of states in the HMM | 3.1.1 |
| $O$ | the sequence of the observation vectors in the HMM | 3.1.1 |
| $\Theta$ | the parameter set of the HMM | 3.1.1 |
| $S = \{(s_k, \Delta t_k)\}$ | sequence of turn-taking | 3.1.3 |
| $N$ | the number of turns detected at the speaker diarization process | 3.1.3 |
| $s_k$ | the label corresponding to the voice detected in the $k^{th}$ turn | 3.1.3 |
| $\Delta t_k$ | the duration of the $k^{th}$ turn | 3.1.3 |
| $G$ | total number of speakers/actors | 3.1.3 |
| $A = \{a_1, \ldots, a_G\}$ | set of $G$ unique speaker labels as provided by the speaker diarization process | 3.1.3 |
| $D$ | number of segments used as events to capture the interaction features | 3.1.3 |
| $\mathbf{x}_a = (x_{a1}, \ldots, x_{aD})$ | tuple representing the interaction features of each actor $a$ | 3.1.3 |
| $\tau_a$ | fraction of the total time of a recording attributed to each actor $a$ | 3.1.3 |
| $\mathbf{y}_a = (\tau_a, \mathbf{x}_a)$ | tuple representing each actor $a$ containing the interaction features as well as the fraction of the total time of the recording actor $a$ talks. Dimension is $D$+1 | 3.1.3 |
| $\mathcal{R}$ | set of roles | 3.2 |
| $\varphi : A \rightarrow \mathcal{R}$ | function mapping the actors to their actual role | 3.2 |
| $\varphi(a)$ | the role of actor $a$ | 3.2 |
| $Y = \{\mathbf{y}_a\}_{a \in A}$ | set of observations | 3.2 |
| $\mathcal{R}^A$ | the set of all possible functions mapping actors into roles | 3.2 |
| $\overrightarrow{\mu} = (\mu_1, \ldots, \mu_D)$ | the parameter vector of the Bernoulli distribution used to model the interaction patterns | 3.2.1 |
| $\overrightarrow{\mu}_r$ | parameter of the distribution for a given role $r$ | 3.2.1 |
| $\|A_r\|$ | the number of actors in the training set playing the role $r$ | 3.2.1 |

| Notation | Definition | Section |
|---|---|---|
| $\overrightarrow{z}_j = (z_{j1}, \ldots, z_{jT})$ | the parameter vector of the Multinomial distribution used to model the interaction patterns | 3.2.1 |
| $T$ | the maximum number of times that an actor can talk during a given event | 3.2.1 |
| $\mathcal{N}(\tau \mid \mu_r, \sigma_r)$ | Gaussian distribution, where $\mu_r$ and $\sigma_r$ are the sample mean and variance respectively | 3.2.2 |
| $C_g$ | set of functions or classes where each role is assigned the same number of times to actor $a$ | 3.2.3 |
| $\mathbf{w}_a$ | output of the application of PCA to the tuples $\mathbf{y}_a$ which results into L-dimensional projections $\mathbf{w}_a$, where $L \leq D + 1$ | 3.5 |
| $W = (\mathbf{w}_a 1, \ldots, \mathbf{w}_a N)$ | sequence of tuples representing the actor interactions for each recording | 3.5 |
| $R = (r_1, \ldots, r_N)$ | sequence of roles of length $N$ | 3.5 |
| $\mathcal{R}^N$ | set of all possible role sequences of length $N$ | 3.5 |
| $C1$ | corpus containing 96 news bulletins and accounting for 18 hours and 56 minutes of material | 3.3.1 |
| $C2$ | corpus containing 27 one hour long talk-shows | 3.3.1 |
| $C3$ | corpus containing 137 meeting recordings and accounting for 45 hours and 38 minutes of material | 3.3.1 |
| $\pi$ | the purity a measure for the effectiveness of the diarization process and the story segmentation approach | 3.3.3 4.2.2 |
| $\alpha$ | the accuracy role recognition performance measure | 3.3.3 |
| $\sigma$ | the standard deviation of the accuracies achieved over the different recordings of each corpus | 3.4 |
| $B$ | stands for Bernoulli distribution used to model interaction patterns | 3.4 |
| $M$ | stands for Multinomial distribution used to model interaction patterns | 3.4 |
| $I$ | stands for the estimatation of the a-priori role probabilities considering the roles as independent | 3.4 |
| $D$ | stands for the estimatation of the a-priori role probabilities considering the roles as dependent | 3.4 |
| $C$ | stands for correct role classification | 3.5.2 |
| $W$ | stands for wrong role classification | 3.5.2 |

| Notation | Definition | Section |
|---|---|---|
| $\mathbf{d}_i$ | tuple representing the transcription of the interventions of meeting participant $i$ | 3.6.2 |
| $\beta$ | factor for the combination of Social Affiliation Networks based role recognition and lexicon based role recognition | 3.6.3 |
| $X$ | sociomatrix | 4.1.1 |
| $C(a_i)$ | centrality of speaker $a_i$ | 4.1.1 |
| $d(a_i, a_j)$ | geodesic distance between two speakers $a_i$ and $a_j$ | 4.1.1 |
| $\tau(a_k)$ | time at which the last intervention of speaker $a_i$ ends | 4.1.1 |
| $t^*$ | the transition time between news and talk-show | 4.1.1 |
| $n(t)$ | the average number of transitions from news to talk-show in the data set at a given time $t$ | 4.1.2 |
| $P$ | total number of recordings in a data set | 4.1.2 |
| $\mathbf{u}_a$ | output of the application of PCA to the tuples $\mathbf{x}_a$ which results into L-dimensional projections $\mathbf{u}_a$, where $L \leq D$ | 4.2.1 |
| $U = (\mathbf{u}_a 1, \dots, \mathbf{u}_a N)$ | sequence of tuples representing the actor interactions for each recording | 4.2.1 |
| $H = (h_1, \dots, h_N)$ | sequence of stories of length $N$ | 4.2.1 |
| $\mathcal{H}$ | set of all possible story sequences $H$ | 4.2.1 |
| $S$ | maximum number of stories | 4.2.1 |

# Chapter 1

# INTRODUCTION

## 1.1 Motivation

Following one of the most famous statements of Western philosophy (Aristotle, Politika ca. 328 BC)[1]:

> Man is by nature a social animal; an individual who is unsocial naturally and not accidentally is either beneath our notice or more than human.

Almost twenty-five centuries after these words have been written for the first time, several disciplines confirm the intuition of Aristotle by grounding the social nature of humans into measurable and observable aspects of human biology, psychology and behavior. Neuroscientists have identified brain structures, called *mirror neurons* [Rizzolatti 04], that seem to have no other goal than improving our awareness of others, whether this means to share their feelings [Iacoboni 09] or to learn through imitation [Frith 07]. Biologists and physiologists have shown that our ears are tuned to human voices more than to any other sound [Pickles 82], that the only facial muscles present in every human being (the others can be absent) are those we use to communicate the six basic emotions [Waller 08] and, more in general, that evolution has shaped our body and senses around social contacts. Furthermore, human sciences (psychology, anthropology, sociology, etc.) have shown how social interactions dominate our perception of the world [Kunda 99] and shape our daily behavior by attaching social meaning to acts as simple and spontaneous as gestures, facial expressions, intonations, etc. [Knapp 72, Richmond 95].

The computing community could not remain immune from this wave of interest for the "*social animal*". Nowadays, computers are leaving their original role of improved versions of old tools [Vinciarelli 09a]

[1] At the time this thesis is being written, the sentence "*Man is by nature a social animal*" returns 1.6 millions of documents when submitted to Google as a query (only documents including the whole statement are counted).

and move towards a new, human-centered vision of computing [Pantic 08] where intelligent machines seamlessly integrate and support human-human interactions [Crowley 06], embody natural modes of human communication for interacting with their users [Bickmore 05], and are the platform through which large scale social activities take place on-line [Wang 07]. In such a new context, the gap between social animal and unsocial machine is no longer acceptable and social adept computers become a crucial need and challenge for the future of computing [Pentland 05, Lazer 09].

This thesis is part of the above effort as it aims at making machines capable of understanding one of the most important aspects of social interactions: the roles. In fact, people play roles each time they interact:

> *"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"* [Tischler 90].

Despite its importance, the role recognition problem was still largely neglected in literature at the beginning of this thesis. In this respect, the work presented in the next chapters has been one of the first attempts to tackle the problem. Nowadays, the topic attracts more interest and is addressed by a larger number of groups (see e.g. [Pianesi 07][Laskowski 08]). However, efforts so far have focused on specific scenarios (e.g. meetings) and the problem is still open.

As roles shape behavior and allow someone to reasonably predict what others do during an interaction, role recognition can be useful in any social context involving both humans and machines, like in the following examples (the list is not exhaustive): Alex Pentland and his group at MIT have shown that the analysis of social interactions could help for negotiating salary, hiring interviews or conducting speed-dating conversations [Pentland 07]. In such contexts, role recognition can help to assigning specific roles to the group participants, and thus, for example, decide who is going to lead the group for a maximal effectiveness and collaboration in a company. In human-computer interactions, roles have a real importance either in the context of interactions with Embodied Conversational Agents (ECAs) or in role-playing games. ECAs demonstrate many of the same properties as humans in face-to-face conversations [Bickmore 05], but further developments are still to be done and the role recognition can help to this. In fact, the recognition of the roles played by the participants in conversations provides information about the nature of the social interactions and could help the agents to adopting an expected behavior. In education, Justine Cassell and her group have shown how agents can encourage children's active exploration of narrative, linguistic creativity and verbal play with their *Story Listening Systems* [Cassell 04]. In health care, a study shows that embodied agents could help in improving the pupil's way of learning in the case of child suffering from autism [Robins 05]. In both previous examples, the recognition of the role of the agent during the interaction is important in order to adapt it to the interlocutor and obtain a positive collaboration. In role-playing games, it is an evidence that the recognition of the roles played by the different participants help these lasts to perform actions respecting the system of rules and guidelines underlying such games [Tychsen 06]. Finally, roles can be useful in several multimedia content analysis applications: in media browsers, the role of the person speaking at a given time can help users to quickly identify segments of interest; in summarization, the role can be used as a criterion to select representative segments of the data; and in Information Retrieval, the role can be used as an index to enrich the content description of the data [Laskowski 08]. In this thesis, we have addressed such an application scenario by showing how roles can help in performing semantic segmentation. In particular, we show how roles can be used to structure a radio program and perform story segmentation.

## 1.2 Contributions

The main contributions of this work, to the best of our knowledge, are as follow:

- This thesis is the first work where social network analysis, i.e. Social Affiliation Networks (SAN) [Wasserman 94], is applied to automatic analysis of conversation recordings, more particularly to automatic role recognition.

- During this thesis, extensive role recognition experiments have been performed over one of the larger data sets ever used in literature for this task, and including for the first time, different human-human interaction settings, i.e. production environments (news and talk-shows) and spontaneous exchanges (meetings).

- This thesis is one of the first role recognition works based on interaction features, i.e. turn-taking (who talks when and how much) and is the first combining interaction and lexical features in meetings.

- This thesis provides for the first time a measure of how much roles constrain conversations. This is useful for defining how likely a role recognition approach is to be effective in a given interaction setting.

- This work proposes a new approach for modeling the dependence between roles played by different individuals in the same interaction.

- This work is the first attempt of using role recognition in an application scenario, i.e. semantic segmentation.

## 1.3 Organization of the Thesis

The thesis is organized as follows:

- **Chapter 1** introduces the problem of recognizing automatically the roles of persons interacting and its importance in the analysis of social interactions. It also states the motivation and contributions of this work.

- **Chapter 2** starts by introducing Social Network Analysis. It then provides a review of the state-of-the-art in the role recognition task. This chapter also includes a survey of the major works dedicated to other aspects of social interactions than just role recognition.

- **Chapter 3** presents approaches for the automatic detection of the roles of the persons interacting in different situations, such as production environment contexts (e.g., news and talk-shows) and spontaneous exchanges (e.g., meetings).

- **Chapter 4** shows how roles can be used to perform semantic segmentation.

- **Chapter 5** finally concludes this thesis by providing a summary of the performed work and obtained results, as well as the main achievements of this thesis. Possible directions for future research are also suggested.

Chapter 2

# STATE-OF-THE-ART

This section first introduces Social Network Analysis (SNA) (Section 2.1) and then reviews the existing literature related to the role recognition task (Section 2.2). For the sake of completeness, this section also includes a survey of the major works dedicated to other aspects of social interactions than just role recognition, with special attention to the analysis of meetings (Section 2.3).

## 2.1 Social Network Analysis

Social Network Analysis (SNA) [Wasserman 94] is a corpus of mathematical techniques used by sociologists to analyze social interactions, i.e. the analysis of relationships among social entities. The concept of a network emphasizes the fact that individuals has ties to other individuals, each of whom in turn is tied to a few, some, or many others, and so on. The phrase *social network* refers to the set of actors and the relation defined on them. Such networks allow to extract patterns and implications of the relationships shared by the actors.

In this thesis, we consider a particular kind of social networks, namely Social Affiliation Networks (SAN), which is a two-mode network which represents the affililation of a set of actors with a set of social occasions, i.e. events. The importance of studying affiliation networks is grounded in the theoretical importance of individuals' memberships in collectivities. In fact, multiple group affiliations (e.g. with family, voluntary organizations, etc) are fundamental in defining social identity of individuals.

The interesting information extracted by Affiliation Networks and used in this work, is the fact that the affiliation of actors with events constitute a direct linkage, either between the actors through memberships in events, or between the events through common members. We are mainly interested by the former, the

**Tab. 2.1:** Synopsis of role recognition results. The table provides a brief description of the data used in literature, as well as the performance achieved in the different works.

| Ref. | Data | Amount | Roles | Features | Approach | Performance |
|---|---|---|---|---|---|---|
| [Barzilay 00] | NIST TREC SDR Corpus (35 recordings, publicly available, 3 roles) | 17h.00m | formal | Lexical features | BoosTexter and Maximum Entropy Method | 80.0% of the news stories correctly labeled in terms of role |
| [Liu 06] | TDT4 Mandarin broadcast news (336 shows, 3 roles) | 170h.00m | formal | Lexical and contextual features | HMM and Maximum Entropy method | 77.0% of the news stories correctly labeled in terms of role |
| [Vinciarelli 07] | Radio news bulletins (96 recordings, 6 roles) | 25h.00m | formal | Turn-taking and speaking time duration | Social Networks Analysis and Bayes classifier | 85% of the data correctly labeled in terms of role |
| [Weng 09] | Movies and TV shows (10 movies and 3 TV shows , 9-20 roles) | 21h.00m | formal | Visual features (co-occurrence of face's individuals in the scenes) | Social Networks Analysis | 95% of leading roles correctly assigned and 84.3% of community roles correctly assigned |
| [Raducanu 09] | "The Apprentice" TV-reality-show (14 meetings, 1 role) | 1h.30m | formal | Nonverbal speech features | Rank-based classification method | 85.0% of the meeting chairman correctly detected |
| [Banerjee 04] | Meetings (2 recordings, 5 roles) | 0h.45m | informal | simple speech features | Decision tree | 53.0% of segments (up to 60 seconds long) correctly classified |
| [Zancanaro 06] | The Mission Survival Corpus (11 recordings, publicly available, 5 roles) | 4h.30m | informal | speech and fidgeting features | Support Vector Machines | Up to 65% of analysis windows (around 10 seconds long) correctly classified in terms of task area roles and 70% in terms of socio area roles |
| [Pianesi 07] | The Mission Survival Corpus (11 recordings, publicly available, 5 roles) | 4h.30m | informal | speech and fidgeting features of each participant and of all the other participants | Support Vector Machine | 90% of analysis windows (around 10 seconds long) correctly classified in terms of task area roles and 95% in terms of socio area roles |
| [Dong 07] | The Mission Survival Corpus (11 recordings, publicly available, 5 roles) | 4h.30m | informal | speech and fidgeting features of each participant and of all the other participants | Influence Model | 75% of task area roles and socio area roles correctly assigned |
| [Lepri 09] | The Mission Survival Corpus (11 recordings, publicly available, 5 roles) | 4h.30m | informal | speech honest signals | Influence Model | Up to 77% of analysis windows (around 1-minute long) correctly classified in terms of task area roles and 74% in terms of socio area roles |
| [Laskowski 08] | AMI Meeting Corpus (138 recordings, publicly available, 4 roles) | 45h.00m | informal | speech features, talkspurts | Maximum Likelihood criteria | 53% of the data correctly labeled in terms of role |
| [Jayagopi 08a] | AMI Meeting Corpus (subset, 1 role, publicly available) | 5h.00m | informal | multimodal nonverbal features | Social Networks Analysis | 68.0% of the Project Manager correctly detected |

subsets of actors who participate in the same social activities (events).

In this thesis, the events correspond to time intervals and the actors (i.e. the individuals involved in the conversations) are linked to events when they talk during the events. In this way, the events capture the proximity in time of the actors interventions, and the direct linkage between the actors is that they are likely to interact with one another if they talk during the same event (i.e. the same interval of time).

## 2.2 Role Recognition

This section reviews the existing literature related to the role recognition task after shortly introducing the types of roles represented in this thesis.

Even if the concept of *role* is one of the most popular ideas in the social sciences, a formal definition is hard to find. Some role theorists focus on the person as an individual and think of roles as the evolving, coping strategies that are adopted by the person. Others more focus on the person as representative of a social position and thus conceive roles as patterns of behavior that are typical of persons whose structural positions are similar. We care about the latter in this work, and consider roles as *characteristic behavior patterns*, as defined by the role theorist Biddle [Biddle 86]. We thus presume that persons are members of social positions, and that expectations are the major generators of roles.

Most of the existing works focusing on the recognition of roles in the computing community do similar assumptions and such works can be divided into two groups studying two types of roles: those who are associated to a task in a specific social context, e.g. the accomplishment of specific functions such as the moderator in a debate, and those who are associated to a position in a social system, e.g. the manager in a company. In the rest of this thesis, we define the former as *formal* roles and the latter as *informal* roles. In both cases, the characteristic behaviors of persons occupy social positions within a stable social system and shared norms governing their behaviors. However, in the case of formal roles, the norms govern the general behaviors of the persons interacting, whereas the norms govern only the relationships between the persons in the case of informal roles, i.e. norms may vary among individuals. We find formal roles in *production environment data* (movies, news, talk-shows, etc.) where people have to accomplish specific tasks and have more or less rigorous constraints on their interactions. Informal roles are typical of *spontaneous exchanges* (meetings, call center conversations, etc.) where people do not necessarily respect predefined constraints on their interactions, but still follow the norms imposed by their social position.

The approaches discussed in this survey are presented in two sections corresponding to the two types of roles (Sections 2.2.1 and 2.2.2).

### 2.2.1 Recognition of Formal Roles

The upper part of Table 2.1 contains experimental setup and role recognition performance for each work discussed in this section.

The work in [Barzilay 00] describes the recognition of three roles in news (*anchor*, *journalist*, and *guest*) with the goal of finding the structure of English broadcast news. This work exploits the lexical information found in the speech transcriptions and aims at recognizing speaker role without any apriori information about the identity of the speaker. The features used as role evidence are : distribution of terms (i.e. what is said, key words), speaking time length, and participant introductions at the beginning of their interventions (where the speaker names were manually labeled). Contextual features are also taken into account: the labels of the $n$ previous segments, and all the features of $n$ previous segments. The ratio between the intervention length and the length of the previous intervention is shown to be a good role predictor, as well as the presence or absence of speaker introductions. Lexical features are selected using the BoosTexter categorization approach. Role recognition is performed with two classifiers: Boostexter and Maximum Entropy Model.

A similar task is addressed in [Liu 06] where three roles (*anchor*, *reporter*, and *other*) are recognized in Mandarin broadcast news. A new method is proposed for the role recognition: the application of a Hidden Markov Model (HMM). The states correspond to the roles and the observations are the words at the beginning and end of each person intervention labeled manually. This HMM method is further combined with a Maximum Entropy classifier. Contextual information is also taken into account by considering the roles of the persons talking before and after an individual under exam. The beginning and the end sentences in a speaker's intervention are shown to be good cues for role identification and the contextual information helps in improving the role recognition performance.

The work in [Vinciarelli 07] addresses the recognition of six different roles in broadcast news: *anchorman*, *second anchorman*, *guest*, *headline reader*, *weather man*, and *interview participant.* The novelty of this work consists in extracting automatically a social network from turn-taking (i.e. who talks when and how much) and then uses it to extract features for each person. Each individual is then assigned the role corresponding to the highest a-posteriori probability estimated with Bayesian classifier. The advantage of using SNA [Wasserman 94] for assigning roles to individuals is that it takes into account only relational data and is thus independent of speakers identity and recording length. The main limitation of the approach used in [Vinciarelli 07] is that the number of individuals interacting must be high enough (more than 8-10 persons) to build meaningful social networks. In fact, if the number of persons involved in the conversation is small, all the nodes in the network will be connected and it will be difficult to extract characteristic patterns. The use of Social Affiliation Networks (SNA) [Wasserman 94] in this thesis overcome this limitation as such networks represent the evidence of interactions in terms of proximity in time and thus makes possible the analysis of small groups. Moreover, the dependence among the roles is not modeled in [Vinciarelli 07] and each person is assigned the most probable role independently of the role of the others.

Another approach is proposed in [Weng 09] to analyze Hollywood movies (e.g. *You've Got Mail*, *Catch Me If You Can*, etc.) and TV shows (e.g. *Sex and the City* and *Friends*) from the perspective of social relationships. Social Networks are applied to extract the leading roles (*hero*, *heroine*) and their respective communities (*hero's friends* and *colleagues*). The approach uses the co-occurrence of the faces of the individuals in the same scene as an evidence of the interaction between people and between roles. A graph is constructed with two types of nodes, i.e. the scenes and the roles, and the edges between the two types of nodes represent which role appears in which scene. The leading roles can be detected using the *Centrality* measure [Wasserman 94]. The community roles are groups of nodes within which the connections are dense but between which the connections are sparse. The results show that almost perfect performance is achieved for the leading roles determination and is very promising for community roles identification.

Finally, the work in [Raducanu 09] addresses the novel problem of role analysis in competitive meetings. The work recognizes the talk-show host (i.e. meeting chairman) of a popular US reality TV show: "The Apprentice", where participants aim at getting a real job in a firm. Manually extracted speech features such as the participants speaking time and turns, and interruptions are employed to this end. The centrality (i.e. person's position in a group [Wasserman 94]) is also considered. The meeting chairman role is recognized using a rank-based classification.

### 2.2.2 Recognition of Informal Roles

The lower part of Table 2.1 contains experimental setup and role recognition performance for each work discussed in this section. The approach in [Banerjee 04] uses simple speech features to first classify the meetings into two defined meeting states: "discussion and "information flow. Typically a discussion between a group of persons is characterized by each person raising issues, asking questions, making comments, etc. On the other hand, an information flow is a meeting state where essentially one person is giving information to one or more meeting participants. Once the meeting has been segmented into meeting states, the roles of the participants are detected for any given segment of the meeting. This work is one of the first attempts aimed at assigning meeting roles (*presenter*, *discussion participator*, *information provider*, *information consumer*, and *undefined*). The features are manually extracted and aim at estimating the participant activities in short sliding windows: number of speaker changes, number of meeting participants that have spoken, number of overlapping speech segments, etc. A decision tree is then used to classify the resulting features into the different roles.

The group of researchers in [Zancanaro 06] use a similar window-based technique for the detection of participant roles in multiparty recordings using Support Vector Machines rather than decision trees. They develop an approach for the recognition of two types of roles appearing in the Mission Survival Corpus (see [Pianesi 07]): *task* roles (*follower*, *orienteer*, *giver*, *seeker*, and *recorder*) and *socio-emotional* roles (*neutral*, *gate-keeper*, *supporter*, *protagonist*, and *attacker*). Furthermore, they extract features accounting for participant speech (i.e. presence/absence of speech ) and body activities (i.e. fidgeting for hands and body), as well as the number of simultaneous speakers during each window. The work proposed in [Zancanaro 06] is further extended in [Pianesi 07] to predict the role of each individual participant by using also features corresponding to all meeting participants. The performance improves, but the approach suffers from curse of dimensionality and overfitting. These issues are addressed in [Dong 07] with an influence model that reduces significantly the number of model parameters and can thus take into account the features of the other participants in a more robust and generalizable way. In [Lepri 09] the same relational roles are addressed using the same influence model proposed in [Dong 07], but exploiting a much larger set of features, i.e. 16 speech honest signals grouped into five classes (*Consistency*, *Spectral Center*, *Activity*, *Mimicry*, and *Influence*) and two body gestures (hand and body fidgeting). The extended set of features are part of honest signals defined by Pentland [Pentland 08] as following: " honest signals are behaviors that are sufficiently expensive to fake that they can form the basis for a reliable channel of communication". Lepri et al. also compare the independent vs. joint classification of tasks and socio-emotional roles, and observed that it is advantageous to model the relationship between tasks and social roles.

The works in [Laskowski 08][Jayagopi 08a] use the AMI meeting corpus [McCowan 05a] (see Section 3.3.1) and try to recognize different sets of predefined roles. The work in [Laskowski 08] tries to assign each meeting participant to one of the four predefined roles (the *Project Manager*, the *Marketing Expert*, the *User Interface Expert*, and the *Industrial Designer*). The extracted features are low-level speech activity features, namely talkspurts defined as contiguous intervals of speech, with interval pauses no longer than 0.3 seconds. Probabilities are estimated from different talkspurts scenarios such as the number of talkspurts initialized during a silence, the number of talkspurts initialized when someone else is speaking or the number of talkspurts initialized when a participant in a specific other role is speaking. The classification of the four participants into one of the four possible roles is performed using a maximization of the a-posteriori probabilities obtained by the different talkspurt scenarios. Compared to [Vinciarelli 07], this work explores also behavior during vocalization overlap and considers the features from all participants rather than characterizing each participant independently from the roles of the other group participants.

The work in [Jayagopi 08a] uses nonverbal features extracted both from audio (speaking activity, interventions length, etc.) and video (movement energy, total amount of movement, visual focus of attention, etc.) to investigate the relationship between dominance and a specific informal role: the project manager of the team, which is supposed to correspond to higher status. The role assignment is performed using the *Centrality* measure [Wasserman 94]. The study shows that 65% of the time a project manager was also perceived as the most dominant. Overall, features extracted from audio seem to be more reliable than those extracted from video. This is probably because the latter depend more on the experimental setup (lighting conditions, arrangement of cameras, etc.) and, in general, are thus less generalizable to different data sets.

## 2.3 Analysis of Social Interactions in Small Groups

This section proposes a short survey of works that consider other aspects of analyzing social interactions than performing roles recognition (see [Vinciarelli 09b][Gatica-Perez 09] for extensive reviews). Most of the literature in this domain is dedicated to meeting analysis not only for the availability of large annotated corpora such as the Mission Survival Corpus [Pianesi 08] and AMI corpus [Carletta 07], 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 analyze [Levine 98].

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 [McCowan 05b][McCowan 03][Zhang 06][Dielmann 07][Reiter 07]. The common aspect of these works is that they model jointly streams of features extracted from multiple modalities. In [McCowan 03], hand movements and speaking activity are modeled with HMM 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 [McCowan 05b], where different streams are fused with coupled and asynchronous HMMs. In [Zhang 06], the same features are classified using a hierarchic layered HMM into individual participant actions and, at a higher level, into collective actions. A similar approach has been proposed in [Dielmann 07], where actions are modeled with Dynamic Bayesian Networks, and in [Reiter 07], 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 [Rienks 06b], and in [Rienks 06a] [Otsuka 05] [Jayagopi 09]. The approaches in [Rienks 06a][Rienks 06b] are based on vocal behavior (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 behavior [Otsuka 05] (Dynamic Bayesian Networks are used to model the effect of one person on another one), or kinesics [Jayagopi 09] (people are classified using Support Vector Machines into dominance categories).

The last topic significantly investigated in this domain is the level of interest, i.e. the degree of engagement of people in interactions [Wrede 03][Gatica-Perez 05][Schuller 07][Schuller 09]. The approach in [Gatica-Perez 05] is closely related to the one described in [McCowan 05b] (same features and same combination approach for multiple modalities). The approaches in [Schuller 07][Schuller 09] combine through early fusion a wide spectrum of visual and audio features, including facial expression, eyes behavior, non-linguistic vocalizations (e.g., laughter) and lexical information, and then use support vector machines to measure the interest level.

## 2.4 Conclusions

The main novelty of our work, according to this review, is the extensive use of social networks to perform automatic role recognition, and more generally, automatic analysis of social interactions. An advantage of using SAN is its independence to language and identity, and thus the fact of being suitable for different kinds of interaction contexts. Moreover, SAN can be applied to any groups of persons interacting, independently of the size of the group.

Chapter 3

# ROLE RECOGNITION

This chapter includes the works presented in the following papers:

- "Automatic Role Recognition in Multiparty Recordings: Using Social Affiliation Networks for Feature Extraction", H. Salamin, S. Favre and A. Vinciarelli, in IEEE Transactions on Multimedia, 2009, Volume 11, Number 7, pages 1373-1380
- "Automatic Role Recognition in Multiparty Recordings Using Social Networks and Probabilistic Sequential Models", S. Favre, A. Dielmann and A. Vinciarelli, in the 2009 Proceedings of International Conference on Multimedia (ACM), pages 585-588
- "Role Recognition in Multiparty Recordings using Social Affiliation Networks and Discrete Distributions", S. Favre, H. Salamin, J. Dines and A. Vinciarelli, in the 2008 Proceedings of ACM International Conference on Multimodal Interfaces (ICMI), pages 29-36
- "Role Recognition for Meeting Participants: an Approach Based on Lexical Information and Social Network Analysis", N. Garg, S. Favre, H. Salamin, D. Hakkani-Tür and A. Vinciarelli, in the 2008 Proceedings of the ACM International Conference on Multimedia (ACM), pages 693-696

As mentioned in the previous chapter (see Section 2.2), we have considered the following definition for roles provided by Biddle [Biddle 79]:

> "Role theory concerns one of the most important features of social life, characteristic behavior patterns or *roles*. It explains roles by presuming that persons are members of *social positions* and hold *expectations* for their own behaviors and those of other persons" [Biddle 86].

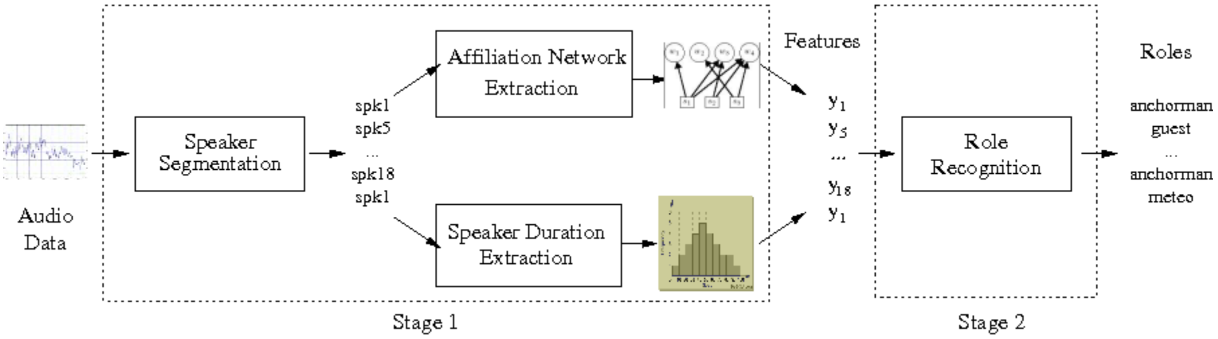

**Fig. 3.1:** Role recognition approach. The picture shows the two main stages of the approach: the features extraction and the actual role recognition.

According to this, we have developed approaches for automatic role recognition based on physical, machine detectable characteristic behavior patterns.

The presented role recognition approach includes two main stages (see Figure 3.1): the first is the *feature extraction* and it involves the automatic construction of a Social Affiliation Network (SAN) [Wasserman 94] as well as its conversion into features that represent each person in terms of their interactions with the others. The second stage is the *role recognition*, i.e. the mapping of the features extracted in the first stage into roles belonging to a predefined set.

The main contributions of this thesis with respect to the state-of-the-art in role recognition are as follows:

- We believe that this work presents the first extensive exploration of social networks as a tool for automatic role recognition and, more generally, automatic analysis of social interactions.

- This work proposes a new approach for modeling the dependence between roles played by different individuals in the same interaction. This is important because it takes into account the constraints that the role distribution across different interacting individuals must respect (see Section 3.2.3 for details).

- This thesis presents for the first time an approach for the role recognition in meetings combining interaction features with lexical features. This allows to increase the role recognition performance over the meetings.

- This is the first work, to the best of our knowledge, that identifies a quantitative measure of how formal a role set is (see Section 2.2 for a definition of formal roles). This is important because it assesses how much the roles under consideration constrain behavior patterns, thus how likely an approach is to be effective in a given interaction setting.

- This work is probably the first one to report experiments performed over different interaction contexts, i.e. production environment data involving formal roles and spontaneous exchanges involving informal roles, (see Section 2.2 for the difference between the two types of role).

The rest of this chapter is organized as follows: Sections 3.1 and 3.2 detail the two stages of the role recognition approach, Section 3.3 describes the data and the experiments performed, Section 3.4 presents the role recognition results, Section 3.5 details a second role recognition approach, Section 3.6 presents a specific role recognition approach for the meetings, and Section 3.7 draws some conclusions on the presented role recognition approaches.

## 3.1 Feature Extraction

This section presents the feature extraction stage aimed at extracting and representing the interaction patterns of each person (see first Stage in Figure 3.1).

The feature extraction stage includes three steps: the first is the segmentation of the conversations into single speaker segments. This detects the persons involved in the conversations and the sequence of their interventions, i.e. the turn-taking informing on who talks when and how much (see left side of Stage 1 in Figure 3.1). The second stage is the extraction of a Social Affiliation Network (SAN) [Wasserman 94] from the resulting turn-taking. The SAN represents each person in terms of their interactions with the others (see upper part of right side of Stage 1 in Figure 3.1). The third step is the extraction of the fraction of time a person is talking, computed from the resulting turn-taking obtained at the first step (see lower part of right side of Stage 1 in Figure 3.1).

In our experiments, we considered two kinds of data: broadcast material where there is a single audio channel, and meeting recordings [McCowan 05a], where each participant wears a headset microphone. This requires the application of different speaker diarization techniques: in the first case (single audio channel), an unsupervised speaker diarization technique identifies the voices of the different persons involved in the conversations (see Section 3.1.1). In the second case (headset microphones), the diarization splits the channel of each microphone into speech and non-speech segments (see Section 3.1.2). Section 3.1.3 shows how the output of the speaker diarization is used to build a Social Affiliation Network and represent the persons with n-tuples accounting for their interaction patterns. Section 3.1.4 shows how we extract the fraction of the total time a person is talking from the output of the speaker diarization process, and finally Section 3.1.5 summarizes the feature extraction stage.

### 3.1.1 Speaker Diarization for Broadcast Data

This section provides a description of the speaker diarization approach used for broadcast data, where there is one audio channel. For a full description, see [Ajmera 04][Ajmera 03].

The diarization is performed with an unsupervised speaker clustering technique based on an ergodic Hidden Markov Model (HMM), where each state corresponds to a cluster and, in principle, to a single speaker voice.

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 [Huang 01]. 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 show that they lead to good results in speaker clustering experiments. The observation sequence is then iteratively aligned with the ergodic HMM where the emission probabilities are modeled with Gaussian Mixture Models (GMM) [Bishop 06].

The method needs an oversegmentation of the number of states (to be sure not having an undersegmentation and miss some speakers) as there is no information *a-priori* about the number of speakers (thus about the number of necessary states in the HMM). To this, the initial number of states is set arbitrarily to a value significantly higher than the expected number of speakers. The process thus starts by segmenting the audio into $M$ uniform non-overlapping segments, where $M$ is the initial number of states in the HMM. The initial GMM set of parameters is defined using the Viterbi algorithm to find the best sequence of states (i.e. speakers) given the uniform segmentation of the data:

$$q^{(0)} = \arg\max_{q \in \mathcal{Q}} \mathrm{p}(q \mid O, \Theta^{(0)}) \tag{3.1}$$

where $q$ is a specific state sequence, $\mathcal{Q}$ is the set of all possible state sequences, and $O = \{\vec{o}_1, \ldots, \vec{o}_K\}$ is the sequence of the observation vectors. The alignment results into a segmentation different from the uniform one used for the initialization. The HMM can thus be retrained and a new parameter set $\Theta^{(1)}$ is obtained:

$$\Theta^{(1)} = \arg\max_{\Theta} \mathrm{p}(q^{(0)} \,|\, O, \Theta) \tag{3.2}$$

where $\Theta = \{\theta_1, \ldots, \theta_M\}$, i.e. the parameter set of the HMM, can be thought of as a set of GMM parameters, if the transition probabilities and the initial state probabilities are kept uniform.

Since the number $M$ is higher than the actual number of speakers, the data is oversegmented and there are clusters that should be merged since they contain data belonging to the same voice. For this reason, the two most similar states are merged at each iteration when the following condition is met:

$$\log \mathrm{p}(O_{m+n} \,|\, \theta_{m+n}) \geq \log \mathrm{p}(O_m \,|\, \theta_m) + \log \mathrm{p}(O_n \,|\, \theta_n) \tag{3.3}$$

where $O_m$, $O_n$ and $O_{m+n}$ are the observation vectors attributed to cluster $m$, $n$ and their union respectively, $\theta_m$ and $\theta_n$ are the parameters of GMMs in states $m$ and $n$ and $\theta_{m+n}$ are the parameters of a GMM trained with Expectation-Maximization on $O_{m+n}$.

After the merging, the HMM has fewer states and it can be realigned with the data in order to obtain a new segmentation which can be used to train again the HMM. The new states satisfying the above condition will be thus merged again and the whole procedure will be iterated. The merging between states is performed by keeping constant the number of parameters of the Gaussians from one iteration to the other:

$$|\theta_{m+n}| = |\theta_m| + |\theta_n| \tag{3.4}$$

the above condition is achieved by setting the number of Gaussians in the state resulting from the merging to the sum of the numbers of Gaussians in the merged states. In this way, the likelihood should improve until the merged states actually correspond to a single voice, while the likelihood should start decreasing when the states corresponding to different voices are merged. Reaching this likelihood peak is the stopping criterion for the iteration process. In this way, after a sufficient number of iterations where states corresponding to similar voices are merged, the number of states is expected to correspond to the actual number of speakers.

| Step | Parameter | Setting | Step | Parameter | Setting |
|---|---|---|---|---|---|
| Training | Training examples | > 22M | Inference | Minimum duration | 20 states |
| | Feature sampling rate | 100 Hz | | Insertion penalty | -40 |
| | Feature dimensionality | 54 | | Silence/speech prior | 0.8/0.2 |
| | Input layer | 810 (54 × 15) units | | Silence collar | 100 ms |
| | Hidden layer | 25 arctan units | | Silence merge | 250 ms |

**Tab. 3.1:** Summary of parameters in the training and inference steps in the automatic speech segmentation system for meeting data.

### 3.1.2 Speaker Diarization for Meeting Data

In the meeting recordings, the diarization can be performed by simply segmenting the output of the headset microphones that each of the meeting participants wears into speech and non-speech. A full description of the approach used for this task is given in [Dines 06].

The audio frames are represented with feature vectors including 12 MF-PLP features [Hermansky 90] and HTKBook [1], augmented by 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 [Wrigley 05]. The input features are summarized as follows:

- 12 *Mel filterbank perceptual linear predictive coefficients* (MF-PLP) [Hermansky 90] and HTKBook [1] including $C0$, plus normalized log-energy,

- *Log cross-channel normalized energy* [Dines 06] which is estimated as the logarithm of the energy of the current headset microphone minus the logarithm of the sum of energies across all headset microphones for the current meeting,

- *Signal kurtosis* [Wrigley 05] (i.e. the normalized fourth-order cumulant of the signal), which should be higher during single speaker activity (since speech signals tend to be super-Gaussian) than during cross-talk (since, in accordance with the central limit theorem, mixtures of speech signals will tend towards a Gaussian distribution),

- *Mean cross-channel correlation* and *Maximum cross-channel correlation* [Wrigley 05], where, for a given time frame, we take the peak cross-correlation between the current headset microphone channel and each of the other headset microphones channels and obtain the *mean* measure as

[1] http://htk.eng.cam.ac.uk/

[1] http://htk.eng.cam.ac.uk/

the arithmetic mean of these cross-correlation values and the *maximum* measure as the maximum cross-correlation value across all channels.

In practice, we concatenate the first and second order differences of these features, thus giving a feature dimensionality of 54. Finally, we take several consecutive frames and provide these as input to a *Multi-Layer Perceptron* (MLP) [Bishop 95] for estimating the posterior probability of audio frames belonging to speech or non-speech classes.

The segmentation of the output of the headset microphones 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 ground truth. Emission probabilities for the HMM states are estimated as scaled likelihoods in which MLP posterior probabilities are divided by their respective prior class probability [Bourlard 93]. Table 3.1 summarizes the main parameters in the training and inference steps.

### 3.1.3 Affiliation Network Extraction

The result of the speaker diarization process is that each recording is split into a sequence of turns, i.e. into a sequence $S = \{(s_k, t_k, \Delta t_k)\}$, where $k \in \{1, \ldots, N\}$, $s_k$ is the label corresponding to the voice detected in the $k^{th}$ turn, $t_k$ is the beginning of speaker $s_k$ intervention, and $\Delta t_k$ is the duration of the $k^{th}$ turn. The label $s_k$ belongs to the set $A = \{a_1, \ldots, a_G\}$ of $G$ unique speaker labels as provided by the speaker diarization process (see lower part of Figure 3.2). $G$ is the total number of speakers in the conversation. The sequence of turns $S$ extracted from the speaker diarization can be used to extract a Social Affiliation Network (SAN), capturing the interaction patterns between the speakers. A SAN is a bipartite graph with two types of nodes: the *actors* and the *events* [Wasserman 94]. Actors can be linked to events, but no links are allowed between nodes of the same type, following the definition of bipartite graphs (see upper part of Figure 3.2). In our experiments, the actors correspond to the persons involved in the conversations, detected during the diarization process. The events correspond to uniform non-overlapping segments spanning the whole length of the recordings (see lower part of Figure 3.2), thus capturing the proximity in time of the persons interventions. Actors participating in the same events (i.e. participants talking during the same interval of time) are likely to interact with one another. Each recording is thus split into a number of $D$ uniform, non-overlapping events. Actors are said to participate in an event if they talk during it, and then the corresponding nodes are linked. One of the main advantages

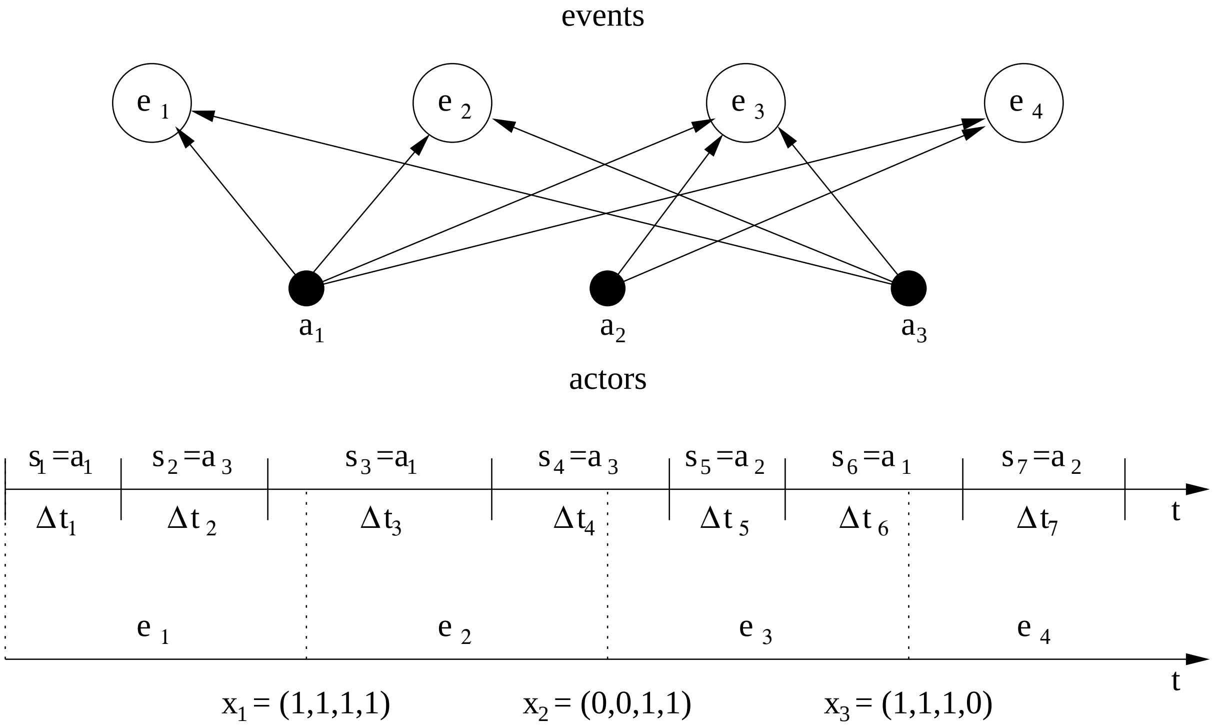

**Fig. 3.2:** Social Affiliation Network extraction. The events of the network correspond to the segments $e_j$ and the actors are linked to the events when they talk during the corresponding segment. The actors are represented using n-tuples $\mathbf{x}_a$ where the components account for the links between actors and events.

of this representation is that each actor $a$ can be represented by a n-tuple $\mathbf{x}_a = (x_{a1}, \ldots, x_{aD})$, where $D$ is the number of events and the component $x_{aj}$ accounts for the participation of the actor $a$ in the $j^{th}$ event. The experiments make use of two kinds of representation. In the first one, component $x_{aj}$ is 1 if the actor $a$ talks during the $j^{th}$ event and 0 otherwise (the corresponding n-tuples are shown at the bottom of Figure 3.2). In the second one, $x_{aj}$ is the number of times that actor $a$ talks during the $j^{th}$ event. In the first case the n-tuples are binary, in the second case they have integer components higher or equal to 0. In both cases, the persons that interact more with each other tend to talk during the same events and are represented by similar n-tuples.

### 3.1.4 Duration Distribution Extraction

As explained previously, the result of the speaker diarization process is that each recording is split into a sequence of turns $S = \{(s_k, t_k, \Delta t_k)\}$, where $k \in \{1, \ldots, N\}$, $s_k$ is the label corresponding to the voice detected in the $k^{th}$ turn, $t_k$ is the beginning of speaker $s_k$ intervention, and $\Delta t_k$ is the duration of the $k^{th}$ turn.

From the turn-taking sequence $S$, we can easily obtain the fraction $\tau$ of the total time of a recording attributed to each voice. In fact, $\tau$ is obtained by summing the durations $\Delta t_k$ of the $k^{th}$ turns during which the same voice is speaking, and dividing the sum by the total length of the recording under process ($\sum_{k=1}^{N} \Delta t_k$).

In Section 3.1.3, we have seen that each actor $a$ could be represented by a n-tuple $\mathbf{x}_a = (x_{a1}, \ldots, x_{aD})$. Furthermore, as every actor talks for a fraction $\tau_a$ of the total time of the recording, each actor corresponds thus to a pair $\mathbf{y}_a = (\mathbf{x}_a, \tau_a)$ with a dimension $D + 1$.

### 3.1.5 Summary

We have defined the features representing each actor $a$ as a n-tuple $\mathbf{x}_a$ accounting for the interaction patterns (see Section 3.1.3), that can have either binary or positive integer components, and a fraction $\tau_a$ of the total time of a recording (see Section 3.1.4). In this way, each actor is represented by a pair $\mathbf{y}_a = (\mathbf{x}_a, \tau_a)$. This last expression will be used in the rest of this chapter.

The only hyperparameter introduced in the feature extraction stage is the number of events $D$ used to capture the interaction patterns in the Affiliation Networks. This hyperparameter must be defined via crossvalidation. Its influence over the role recognition performance will be studied in details later in this chapter (see Section 3.3.2).

## 3.2 Role Recognition Approach based on Bayesian Classifiers

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 $\varphi : A \rightarrow \mathcal{R}$ mapping the actors into their actual role. In other words, the problem corresponds to finding the function $\varphi$ such that $\varphi(a)$ is the role of actor $a$.

Section 3.1 has shown that each actor corresponds to a pair $\mathbf{y}_a = (\mathbf{x}_a, \tau_a)$. Thus, given the set of observations $Y = \{\mathbf{y}_a\}_{a \in A}$ and the function $\varphi : A \rightarrow \mathcal{R}$, the problem of assigning a role to each actor can be formulated as the maximization of the *a-posteriori* probability $\mathrm{p}(\varphi \,|\, Y)$. By applying Bayes Theorem, and by taking into account that $\mathrm{p}(Y)$ is constant during recognition, this problem is equivalent to finding

$\hat{\varphi}$ such that:

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} \mathrm{p}(Y \,|\, \varphi)\, \mathrm{p}(\varphi) \tag{3.5}$$

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 (3.5) can thus be rewritten as:

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} \mathrm{p}(\varphi) \prod_{a \in A} \mathrm{p}(\mathbf{y}_a \,|\, \varphi(a)) \tag{3.6}$$

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

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} \mathrm{p}(\varphi) \prod_{a \in A} \mathrm{p}\left(\mathbf{x}_a \,|\, \varphi(a)\right) \mathrm{p}(\tau_a \,|\, \varphi(a)) \tag{3.7}$$

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 in Section 3.1.3 (i.e. $\mathbf{x}_a$ can have either binary or positive integer components). We have also considered different models accounting for 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 $\mathrm{p}(\varphi)$ have been estimated in the experiments.

### 3.2.1 Modeling Interaction Patterns

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

$$\mathrm{p}(\mathbf{x} \,|\, \overrightarrow{\mu}) = \prod_{j=1}^{D} \mu_j^{x_j} (1 - \mu_j)^{1 - x_j} \tag{3.8}$$

where $D$ is the number of events used to capture the interaction patterns in the SAN, and $\overrightarrow{\mu} = (\mu_1, \dots, \mu_D)$ is the parameter vector of the distribution. A different Bernoulli distribution like the one in equation 3.8 is trained for each role. The maximum likelihood estimates of the parameters $\overrightarrow{\mu}_r$ for a given role $r$ are as follows [Bishop 06]:

$$\mu_{rj} = \frac{1}{|A_r|} \sum_{a \in A_r} x_{aj} \tag{3.9}$$

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

When we considered the number of times that actor $a$ talks during event $j$, the components of the n-tuples $\mathbf{x}_a$ are multinomial, i.e $x_{aj}$ are integers greater to 0 if actor $a$ talks during event $j$ and 0 otherwise. In this case, each component $x_{aj}$ can be represented with a vector $\overrightarrow{z}_j = (z_{j1}, \dots, z_{jT})$ where $T$ is the maximum number of times that an actor can talk during a given event $j$, $z_{ji} \in \{0, 1\}$, and $\sum_{i=1}^{T} z_{ji} = 1$ (*one-out-of-K*). In other words, $x_{aj}$ is represented with a $T$-dimensional vector where all the components are 0 except one, i.e. the component $z_{jn} = 1$, where $n$ is the number of times that the actor represented by $\mathbf{x}_a$ talks during event $j$. As a result, $\mathbf{x}_a$ is represented as a n-tuple of vectors $\overrightarrow{z} = (\overrightarrow{z}_1, \dots, \overrightarrow{z}_D)$ and can be modeled as a product of independent Multinomial distributions:

$$\mathrm{p}(\overrightarrow{z} \mid \overrightarrow{\mu}) = \prod_{j=1}^{D} \prod_{i=1}^{T} \mu_{ji}^{z_{ji}} \tag{3.10}$$

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

$$\mu_{rji} = \frac{1}{|A_r|} \sum_{a \in A_r} z_{aji} \tag{3.11}$$

where $|A_r|$ is the number of actors in the training set playing the role $r$, and $\overrightarrow{z}_j$ is the vector representing the n-tuple $\mathbf{x}_a$ composed of the interaction patterns of actor $a$.

### 3.2.2 Modeling Durations

$\mathrm{p}(\tau \mid r)$ is estimated using a Gaussian Distribution $\mathcal{N}(\tau \mid \mu_r, \sigma_r^2)$, where $\mu_r$ and $\sigma_r$ are the sample mean and variance respectively, and $A_r$ is a set of actors playing role $r$ given a labeled training set:

$$\mu_r = \frac{1}{|A_r|} \sum_{a \in A_r} \tau_a \tag{3.12}$$

$$\sigma_r^2 = \frac{1}{|A_r|} \sum_{a \in A_r} (\tau_a - \mu_r)^2 \tag{3.13}$$

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

### 3.2.3 Estimating Role Probabilities

This subsection shows how the *a-priori* probability $\mathrm{p}(\varphi)$, for each actor $\mathrm{p}(\varphi(a))$, 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 to be dependent and takes into account the above constraints.

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 $\varphi$ to the different actors:

$$\mathrm{p}(\varphi) = \prod_{a \in A} \mathrm{p}(\varphi(a)) \tag{3.14}$$

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

$$\mathrm{p}(\varphi(a)) = \frac{|A_r|}{G} \tag{3.15}$$

where $G$ is the total number of actors and $|A_r|$ the total number of actors playing role $\varphi(a)$ in the training set.

Using the above approach, (3.6) boils down to

$$\hat{\varphi} = \arg \max_{\varphi \in \mathcal{R}^A} \prod_{a \in A} \mathrm{p}(\mathbf{x}_a \,|\, \varphi(a))\, \mathrm{p}(\tau_a \,|\, \varphi(a))\, \mathrm{p}(\varphi(a)) \tag{3.16}$$

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

The second approach aims at modeling the constraints that the role distribution of a given recording must respect. For example, in a talk show (i.e one kind of data considered in this thesis), some roles must be assign only once (e.g. the *Anchorman* role) whereas other roles (e.g. the *Guest* role) can be assigned a different number of times at each edition of the talk show (see Section 3.3.1 for a description of the different set of roles). In this case, the roles played by the different recording participants cannot be considered independent, and $\mathrm{p}(\varphi)$ cannot be written as the product of the a-priori probabilities of the roles (like in (3.14)).

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. For some roles, the number of actors playing them is actually predetermined (i.e. exactly $n_r$ actors must play role $r$). This is the case for example in the talk show data set where only one actor can play the *Anchorman* role. For other roles, the only available a-priori information is that at least one person must play the role (i.e. $n_r > 0$). Thus, $\mathrm{p}(\varphi)$ must be different from 0 only for those distributions of roles that respect the constraints.

According to the above, $\mathrm{p}(\varphi)$ is modeled with a product of Multinomial distributions [Bishop 06]:

$$\mathrm{p}(\varphi \epsilon C_g) = \prod_{r \in \mathcal{R}} \mathrm{p}(\overrightarrow{z}_r \,|\, \overrightarrow{\mu}_r) \tag{3.17}$$

which represents the probability of observing a certain class of functions, where $\overrightarrow{z}_r$ is a *one-out-of-K* representation of the number of times a role can be played in a given recording, $\overrightarrow{\mu}_r$ is the parameter vector, and $C_g$ is a set of functions where each role is assigned the same number of times to actor $a$.

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 $\varphi$ in the same class have the same probability. Thus, the probability of observing a given assignment is:

$$\mathrm{p}(\varphi) = \frac{\prod_{r \in \mathcal{R}} \mathrm{p}(\overrightarrow{z}_r \mid \overrightarrow{\mu}_r)}{|C_g|} \tag{3.18}$$

Then in the second model, Equation (3.6) can be rewritten as:

$$\hat{\varphi} = \arg\max_{\varphi \in \mathcal{R}^A} \mathrm{p}(\varphi) \prod_{a \in A} \mathrm{p}(\mathbf{x}_a \mid \varphi(a))\, \mathrm{p}(\tau_a \mid \varphi(a)) \tag{3.19}$$

s where $\mathrm{p}(\varphi)$ is the expression of (3.18). Maximizing this product using a brute-force approach is not tractable if the number of actors is high. Therefore, we used simulated annealing [Kirkpatrick 83] to approximate the best mapping for each recording.

## 3.3 Databases, Experimental Protocols and Performance Measures

The three next sections describe data and roles, experimental setup, and performance measures.

### 3.3.1 Data and Roles

The experiments of this work have been performed over three different corpora for a total amount of roughly 90 hours of material (one of the largest databases used for role recognition the literature). The first, referred to as C1 in the following, contains 96 news bulletins with an average length of 11 minutes and 50 seconds (the shortest recording is 9 minutes and 4 seconds long, while the longuest one lasts 14 minutes and 28 seconds). The total duration of C1 accounts for 18 hours and 56 minutes of material. The corpus contains all 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 these kind of programs. The second corpus, referred to as C2 in the following, contains 27 one hour long talk-shows broadcasted by *Radio Suisse Romande* (see above) during February 2005 and thus accounts for a total of 27 hours of material. Also in this case, the corpus can be considered a representative sample of this specific kind of program. The third corpus, referred to as C3 in the following, is the AMI

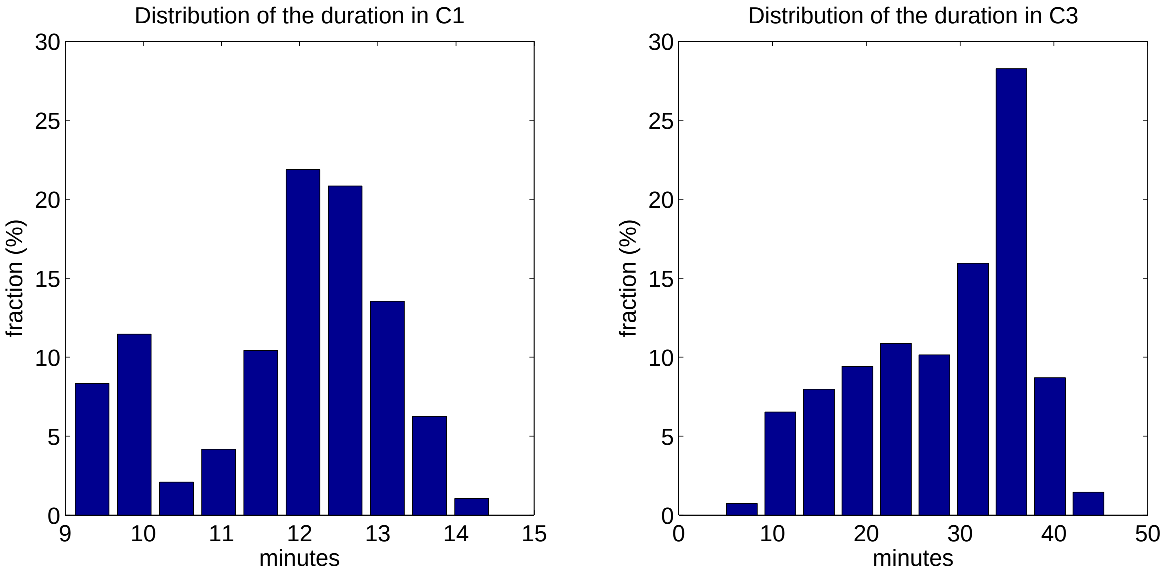

**Fig. 3.3:** Distribution of the recording lengths. The histograms show the distribution of the recording lengths for corpora C1 and C3 (in corpus C2 each recording lasts for exactly 1 hour).

| DB | recs. | setting | tot. $t$ | avg. $t$ | avg. $G$ |
|---|---|---|---|---|---|
| C1 | 96 | news | 18h 56m | 11m 50s | 12 |
| C2 | 27 | talk-show | 27h 00m | 1h 00m | 30 |
| C3 | 137 | meeting | 45h 38m | 19m 50s | 4 |

**Tab. 3.2:** Corpora. The table reports the main characteristics of the corpora used in the experiments. From left to right: number of recordings, interaction setting, total time, average recording length, average number of participants. Note that the length is the same (one hour) for all recordings in C2, and the number of participants is constant (four) in C3. In all other cases, the figures change from one recording to the other.

meeting corpus [McCowan 05a][1], a collection of 137 meeting recordings for a total of 45 hours and 38 minutes of material. The average length of the meetings is 19 minutes and 50 seconds (the shortest recording is 9 minutes long, while the longest one lasts 43 minutes). 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. The distribution of the recording lengths for C1 and C3 is shown in Figure 3.3 (in C2 all recordings last for exactly one hour).

Figure 3.4 shows the distribution of the number of persons across different recordings for corpora C1

[1] The corpus is publicly available at the following URL: `http://corpus.amiproject.org/`

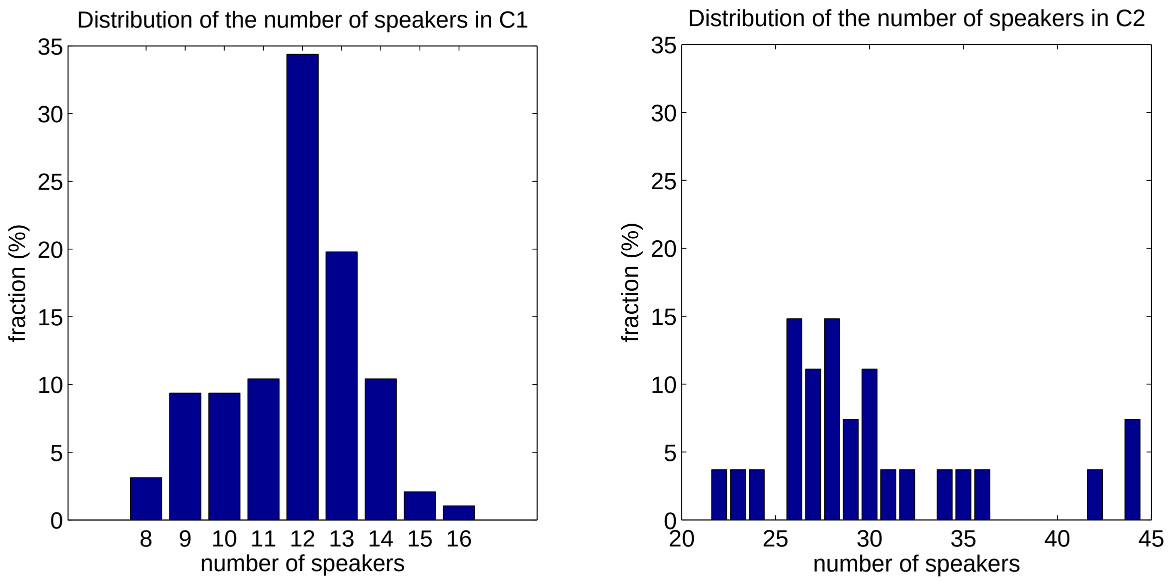

**Fig. 3.4:** Distribution of recording participants. The histograms show the distribution of the number of persons participating in each recording for corpora C1 and C2 (in corpus C3 each recording involves exactly 4 persons).

**Tab. 3.3:** Role distribution in broadcast data. The table reports the percentage of time each role accounts for in C1 and C2.

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

and C2 (in C3 the number of meeting participants is always 4). The number of persons varies from 8 to 16 with an average number of 12 for C1 and varies from 22 to 44 with an average number of 30 persons for C2.

Table 3.2 summarizes the main characteristics of C1, C2, and C3. The roles of C1 and C2 share the same names and correspond to similar functions: 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 Reader* (HR), i.e. the speaker reading a short abstract at the beginning of the program, and the *Weather Man* (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

**Tab. 3.4:** Role distribution in meetings. The table reports the percentage of time each role accounts for in the AMI meeting corpus (C3).

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

different way in C1 and C2 (e.g., consider how different is the behavior of an anchorman in news supposed to inform and in talk-shows supposed to entertain). In C3, the role set is different and contains the *Project Manager* (PM), the *Marketing Expert* (ME), the *User Interface Expert* (UI), and the *Industrial Designer* (ID).

Table 3.3 shows the distribution of speaking time across the roles in C1 and C2. Even though C1 and C2 include roles (i.e. functions) with the same name, the fraction of data each one of these accounts for changes significantly between the two corpora. This enables one to test the robustness of the approach with respect to changes of this aspect of the data. Table 3.4 reports the same information for corpus C3.

### 3.3.2 Experimental Setup

The experiments are performed using a leave-one-out approach [Bishop 06]. We have selected all the recordings of the corpus in the training set (i.e. for training the role's models) with the exception of one that is used as test set. Training and test are repeated as many times as there are recordings in the corpus, and each time a different recording 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. We have chosen to left out only one recording for model selection and classification because it implies a large training set and thus minimize the variance between the different role's models [Kohavi 95].

The hyperparameter of the system, i.e. the number $D$ of events in the Social Affiliation Network, is tuned at each iteration of the leave-one-out process. At each iteration, the hyperparameter giving the highest role recognition results over the training set has been retained for testing. In this way, a rigorous separation between the training and test set has been observed for the setting of the hyperparameter as well.

Figures 3.5, 3.6 and 3.7 show the influence of the hyperparameter $D$. Figure 3.5 illustrates the overall accuracy performance (see Section 3.3.3) for the role recognition process over the database C1.

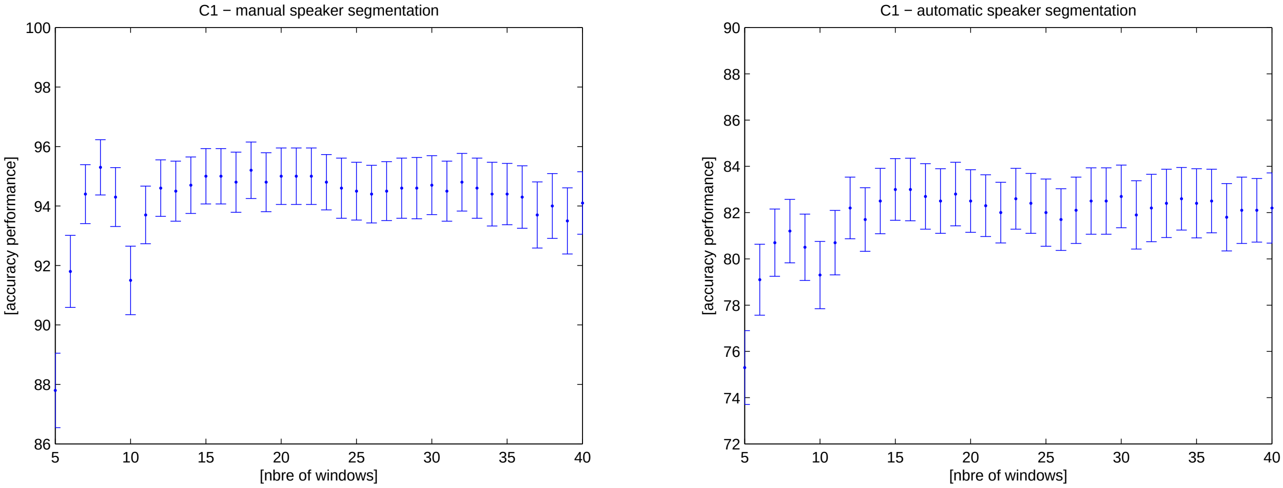

**Fig. 3.5:** Influence of hyperparameter $D$ over C1 database. The curve shows the role recognition performance in terms of accuracy for different $D$ values.

Figure 3.6 shows the performance over C2, and Figure 3.7 shows the accuracy performance over C3. The role recognition performance is represented for both groundtruth speaker diarization and automatic speaker diarization. The accuracy performance estimation is accompanied with an error for a trust interval of 95%.

The $D$ value has been varied from 5 to 40 in our experiments, covering a large scale for the duration of the events used to capture the interaction patterns in terms of proximity in time (see Section 3.1.3). The maximum $D$ value has been set to 40 corresponding to a minimum event's duration of 20 seconds for C1, 30 seconds for C2 and 1.5 minute for C3. These duration are long enough to contain an entire person's intervention or to capture a short conversation. The minimum $D$ value has been set to 5 corresponding to maximum event's duration of 2.4 minutes for C1, 12 minutes for C2 and 4 minutes for C3. These duration are short enough to capture characteristical interaction patterns.

The curves representing the role recognition performance in Figures 3.5, 3.6, and 3.7 do not show an unique maximum, but rather flat shapes. These results show that $D$ does not influence significantly the role recognition performance and that $D$ can be determined by maximizing the role recognition performance over the training set without using a validation set. In corpora C1 and C2 the selected $D$ is almost the same for each training set showing that the system is stable. The values are around 16 and 28 for C1 and C2 respectively, corresponding to high values of the performance in figures 3.5 and

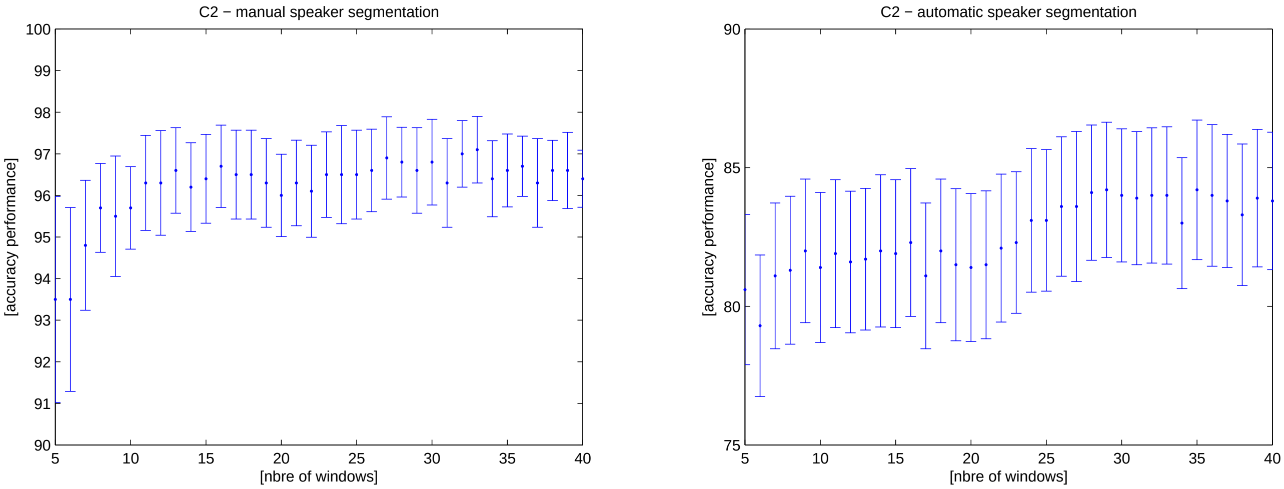

**Fig. 3.6:** Influence of hyperparameter $D$ over C2 database. The curve shows the role recognition performance in terms of accuracy for different $D$ values.

3.6. $D = 16$ corresponds to events of 45 seconds duration in C1 (the average length duration in C1 is 12 minutes) and $D = 28$ corresponds to events of 2 minutes long in C2 (the duration of recordings in C2 is one hour long). In C3, the selected $D$ varies from 20 to 40 and changes from one training set to another one. According to the figure 3.7, this is not a problem as the role recognition performance is not significantly influenced by the hyperparameter $D$. The selected $D$ correspond to events of 30 seconds to 1 minute long (the average length duration in C3 is 20 minutes). The overall $D$ selected for the different corpus correspond to events of roughly 1 minute. As suggested by Pentland [Pentland 08], one minute is large enough to compute speech features in a reliable way, while being small enough to capture transient nature of social behavior.

### 3.3.3 Performance Measures

The role recognition performance is measured with the *accuracy* $\alpha$, i.e. with the percentage of data time correctly labeled in terms of role. The statistical significance of performance differences presented in this thesis is assessed with the Kolmogorov-Smirnov test [Massey Jr. 51]. 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 performance following a Gaussian distribution) and it is adapted to continuous distributions (unlike the $\chi^2$-test that requires discrete distributions).

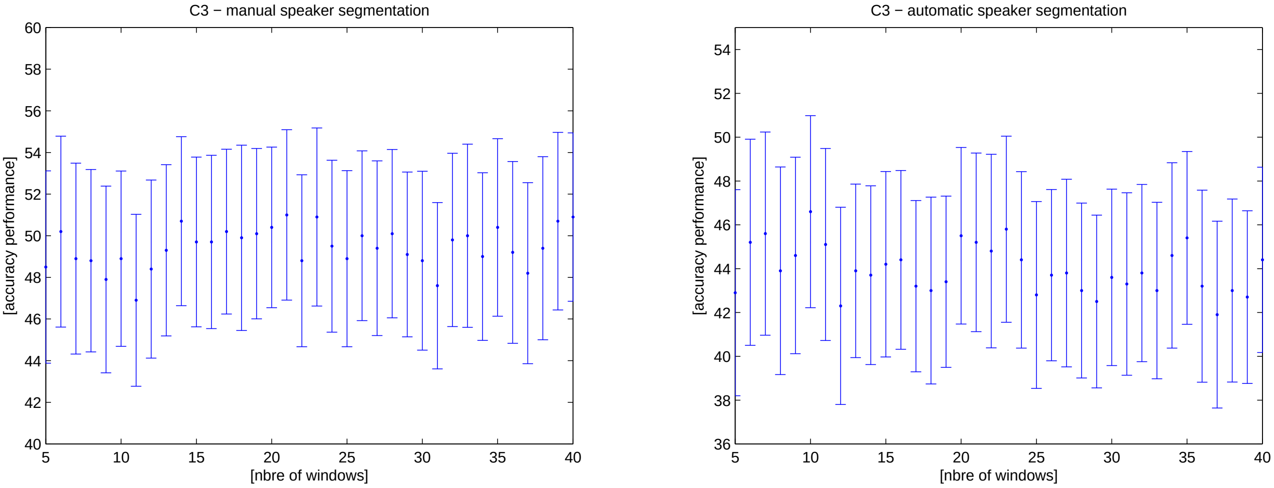

**Fig. 3.7:** Influence of hyperparameter $D$ over C3 database. The curve shows the role recognition performance in terms of accuracy for different $D$ values.

The role recognition performance also depends on the effectiveness of the diarization process as it is one of the steps of the whole role recognition process. Its effects are discussed in the next paragraph.

### Speaker Diarization Results

The interaction patterns used at the role recognition step are extracted from the speaker segmentation obtained with the two different diarization processes (see Sections 3.1.1 and 3.1.2). Errors in the diarization (e.g. persons 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* $\pi$, 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. We choose to use the Purity measure as it is the common metric for speaker diarization evaluation [Ajmera 02]. The Purity ranges between 0 and 1 (the higher the better) and it is the geometric mean of two terms: the *average cluster purity* $\pi_c$ and the *average speaker purity* $\pi_s$. The definition of $\pi_c$

is as follows:

$$\pi_c = \sum_{k=1}^{N_c} \sum_{l=1}^{N_s} \frac{n_k}{N} \frac{n_{lk}^2}{n_k^2} \tag{3.20}$$

where $N$ is the total number of feature frames, $N_s$ 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 $\pi_s$ is as follows:

$$\pi_s = \sum_{l=1}^{N_s} \sum_{k=1}^{N_c} \frac{n_l}{N} \frac{n_{lk}^2}{n_l^2} \tag{3.21}$$

(see above for the meaning of the symbols).

The application of the speaker diarization process in the case of broadcast news requires the setting of the initial number of states $M$ in the fully connected Hidden Markov Model (see Section 3.1.1). 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.

## 3.4 Role Recognition Results for Bayesian Classifiers based Approach

### 3.4.1 Results

Table 3.5 reports the results achieved over C1 and C2, Table 3.6 those obtained for C3. Each overall 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 (see Section 3.2.1) is indicated with $B$ (Bernoulli) and $M$ (Multinomial). The approach used to estimate the *a-priori* role probabilities (see Section 3.2.3) is indicated with $I$ (Independence) and $D$ (Dependence).

For the three corpora, *the differences between the performance achieved using Bernoulli and Multinomial distributions are not statistically significant* (according to the Kolmogorov-Smirnov test [Massey Jr. 51]).

**Tab. 3.5:** Role recognition performance based on Bayes classifiers 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 $\sigma$ of the performance achieved over the single recordings. The upper part of the table reports the results obtained over the output of the speaker segmentation (diarization), the lower part reports the results obtained over the manual speaker segmentation.

| | all ($\sigma$) | AM | SA | GT | IP | HR | WM |
|---|---|---|---|---|---|---|---|
| **Automatic** Speaker Segmentation | | | | | | | |
| C1 (B,I) | **82.5** (6.9) | 98.0 | 3.6 | 97.8 | 8.0 | 64.6 | 79.9 |
| C1 (B,D) | 53.3 (17.8) | 83.6 | 6.4 | 48.6 | 4.6 | 12.8 | 12.7 |
| C1 (M,I) | 81.5 (7.1) | 97.8 | 3.4 | 92.0 | 3.4 | 56.0 | 78.4 |
| C1 (M,D) | 55.2 (15.3) | 87.7 | 8.5 | 48.4 | 2.8 | 13.9 | 13.1 |
| C2 (B,I) | 82.6 (6.8) | 75.0 | 88.3 | 91.6 | N/A | 18.3 | 6.7 |
| C2 (B,D) | **86.6** (6.5) | 75.3 | 88.5 | 92.8 | N/A | 91.3 | 14.8 |
| C2 (M,I) | 84.3 (6.8) | 68.5 | 92.1 | 89.8 | N/A | 83.7 | 18.3 |
| C2 (M,D) | 86.5 (7.7) | 73.9 | 92.1 | 91.9 | N/A | 98.4 | 18.6 |
| **Manual** Speaker Segmentation | | | | | | | |
| C1 (B,I) | 95.2 (4.7) | 100 | 88.5 | 98.0 | 17.1 | 100 | 97.9 |
| C1 (B,D) | 61.2 (14.3) | 94.8 | 8.3 | 52.6 | 10.7 | 22.9 | 17.7 |
| C1 (M,I) | 97.0 (4.2) | 100 | 84.4 | 98.4 | 72.5 | 98.4 | 96.9 |
| C1 (M,D) | 62.5 (11.2) | 96.9 | 11.5 | 56.3 | 7.9 | 14.6 | 15.6 |
| C2 (B,I) | 96.1 (2.7) | 96.3 | 100 | 96.2 | N/A | 100 | 70.4 |
| C2 (B,D) | 97.4 (2.0) | 100 | 100 | 98.0 | N/A | 100 | 33.3 |
| C2 (M,I) | 95.7 (7.7) | 96.3 | 96.3 | 95.7 | N/A | 100 | 85.2 |
| C2 (M,D) | 98.6 (2.1) | 100 | 100 | 98.9 | N/A | 100 | 63.0 |

This suggests that the important information is presence/absence (conveyed by the Bernoulli distribution) and not the number of times a speaker talks during an event (conveyed by the Multinomial). 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 much someone interacts with someone else. According to this result, we will consider binary interaction patterns in the rest of this report.

*Modeling the dependence between roles leads to improvement for C3, leads to statistically significant improvements for C2, while it decreases the performance for C1.* One probable explanation for the improved performance when considering the C2 database is that roles are very constrained in this case. In fact, each emission contains exactly one AM, one SA, one WM, and one HR. In contrast, C1 presents more variability in the number of persons playing a given role from one news bulletin to another one and some roles even do not appear in every recording. It implies that a lot of role's combinations must

**Tab. 3.6:** Role recognition performance based on Bayes classifiers 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 $\sigma$ of the performance achieved over the single recordings. The upper part of the table reports the results obtained over the output of the speaker segmentation (diarization), the lower part reports the results obtained over the manual speaker segmentation.

| | all ($\sigma$) | PM | ME | UI | ID |
|---|---|---|---|---|---|
| **Automatic** Speaker Segmentation | | | | | |
| C3 (B,I) | 43.5 (23.9) | 75.3 | 15.1 | 40.0 | 15.1 |
| C3 (B,D) | 44.9 (29.2) | 68.0 | 21.7 | 36.3 | 23.0 |
| C3 (M,I) | 42.9 (27.2) | 64.4 | 26.6 | 30.8 | 28.4 |
| C3 (M,D) | **46.7** (30.9) | 64.7 | 30.3 | 31.4 | 32.2 |
| **Manual** Speaker Segmentation | | | | | |
| C3 (B,I) | 49.0 (24.5) | 79.0 | 21.7 | 42.8 | 19.6 |
| C3 (B,D) | 51.5 (31.8) | 72.5 | 31.2 | 39.1 | 32.6 |
| C3 (M,I) | 45.1 (27.8) | 71.0 | 21.7 | 37.0 | 21.7 |
| C3 (M,D) | 51.8 (28.2) | 76.1 | 30.4 | 29.7 | 36.2 |

be tested and complexity is added to the model. Similarly as in C2, the model for C3 is simplified as the exactly four roles are represented during each meeting. Moreover, the errors due to the speaker diarization process is almost negligible in the case of meetings as the process performs a purity of 0.99. In the case of broadcast news, i.e. C1 and C2, the errors in the speaker diarization process can influence the role recognition performance. In fact, as the role's distributions are really narrow, an error during the voice's detection implies an error in the role's assignment.

However, these results suggest that taking into account the dependence across roles is beneficial as long as p($\varphi$) (see Section 3.2.3) 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 and the different databases used, in terms of performance improvement.

For binary interaction patterns (corresponding to B in Tables 3.5 and 3.6), and considering independence between the roles (I in Tables 3.5 and 3.6), the overall $\alpha$ is above 80% for both C1 and C2, and around 43% for C3. *The roles in meeting data (C3) are harder to model.* A probable explanation is that the roles in meetings (C3) are *informal*, 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 *formal* roles in broadcast data (C1 and C2). Moreover, the meetings in C3 are not real-world data, i.e. the participants are asked to *act* in a scenario. It can thus happen that the participants have to play roles they are not used to and this might result into non ecologically valid data. For example, it happens during some meetings that the participants do not remember which role they are asked to play. This implies that their social positions in the group are not distinct anymore and that similar interaction patterns are produced for different acted roles.

Not surprisingly, the only meeting role recognized with a high accuracy is the *Project Manager* (PM). The reason is that the PM acts as a *chairman*, having a specific task to achieve, and thus having distinct behavioral turn-taking patterns, comparing to the less formal roles such as the domain experts ID, ME, and UI.

The overall accuracy $\alpha$ over 80% achieved for both C1 and C2 means that the role recognition approach is robust with respect to changes in the time distribution across the roles (see Table 3.3). This is important because it shows that the presented approach is capable of adapting automatically to the different role scenarios.

*The performance difference when passing from manual (ground truth) to automatic speaker diarization is statistically significant for C1 and C2* (see Table 3.5). The difference is not significant for C3 because the purity of the speaker segmentation for this corpus is 0.99 (see Section 3.3.3), i.e. it corresponds almost perfectly to the groundtruth speaker segmentation.

In contrast, the 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 (see Section 3.3.3), i.e. the output of the speaker diarization is significantly different from the groundtruth speaker segmentation. The difference in accuracy is around 10% (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 misalignments, on average, corresponds to roughly 10% 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 role recognition errors are due to limits of the role recognition approach that cannot distinguish between different roles when the associated interaction patterns are too similar. This is true for example, in the case of the low performance of the IP in corpus C1. The interaction pattern of the IP role is similar to that of the Guest, but the latter has higher *a-priori* probability, so it is usually favored as the output of the recognizer. This is also true in the meetings C3 where the domain experts ID, ME, and UI have similar interaction patterns, and thus our approach is not able to distinguish between these three roles.

*A qualitative comparison with other approaches is possible only for some works which use parts of the same data as ours.* Both [Jayagopi 08a][Jayagopi 08b] perform experiments over a subset of the AMI meeting corpus (around 5 hours of material). The performance in [Jayagopi 08b] is around 80%, almost twice as much as our approach over the same data. 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 [Jayagopi 08a] reports a 65% recognition rate of the Project Manager, while our work achieves, over the same role, an accuracy of 75%. Considering that our experiments are performed over the whole AMI meeting corpus, while the experiments of [Jayagopi 08b][Jayagopi 08a] 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 [Laskowski 08] 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%.

### 3.4.2 Influence of interaction patterns

This section shows how the different features (interaction patterns extracted through SAN and fraction of speaking time, see Sections 3.1.3 and 3.1.4) influence the role recognition performance. We want to compare the role recognition performance achieved when considering the different types of features separately or as a combination. The first row of Tables 3.7 and 3.8 reports the role recognition performance achieved when considering only the interaction patterns as features extracted through SAN, i.e. $\mathrm{p}(\mathbf{x}_a \,|\, \varphi(a))\,\mathrm{p}(\varphi)$. The second row reports the role recognition performance achieved with only the fraction of speaking time as features, i.e. $\mathrm{p}(\tau_a \,|\, \varphi(a))\,\mathrm{p}(\varphi)$. The last row reports the role recognition performance obtained when combining both the SAN and speaking time features, i.e. $\mathrm{p}(\mathbf{x}_a \,|\, \varphi(a))\,\mathrm{p}(\tau_a \,|\, \varphi(a))\,\mathrm{p}(\varphi)$ (see (3.16)).

For both C1 and C2, the SAN features are more effective than the speaking time features, particularly for less frequent roles, i.e. roles accounting for a smaller percentage of time (cf Table 3.3). The combination of the two types of features improve statistically significantly the role recognition performance achieved when considering speaking time features only. These results highlight the relevance of using interaction patterns as features when performing the role recognition task.

In C3, the SAN and the speaking time features have similar influence over the role recognition per-

**Tab. 3.7:** Features comparison. Role recognition performance based on Bayes classifiers for C1 and C2. The table reports both the overall accuracy and the accuracy for each role. "SAN" stands for interaction features extracted through SAN, "T" stands for fraction of speaking time as features, and SNA + T stands for the combination of both types of features.

| | all ($\sigma$) | AM | SA | GT | IP | HR | WM |
|---|---|---|---|---|---|---|---|
| Results over C1 | | | | | | | |
| C1 SAN | 83.4 (6.2) | 97.8 | 2.3 | 95.7 | 0.0 | 61.7 | 84.8 |
| C1 T | 73.0 (5.8) | 98.0 | 0.0 | 93.5 | 2.4 | 0.0 | 0.0 |
| C1 SAN + T | 82.5 (6.9) | 98.0 | 3.6 | 91.8 | 8.0 | 64.6 | 79.9 |
| Results over C2 | | | | | | | |
| C2 SAN | 81.3 (8.1) | 72.0 | 88.4 | 90.3 | N/A | 22.2 | 0.0 |
| C2 T | 73.6 (8.6) | 55.2 | 64.6 | 86.9 | N/A | 0.0 | 0.0 |
| C2 SAN + T | 82.6 (6.8) | 75.0 | 88.3 | 91.6 | N/A | 18.3 | 6.7 |

**Tab. 3.8:** Features comparison. Role recognition performance based on Bayes classifiers for C3. The table reports both the overall accuracy and the accuracy for each role. "SAN" stands for interaction features extracted through SAN, "T" stands for fraction of speaking time as features, and SNA + T stands for the combination of both kinds of features.

| | all ($\sigma$) | PM | ME | UI | ID |
|---|---|---|---|---|---|
| Results over C3 | | | | | |
| C3 SAN | 43.1 (23.5) | 81.9 | 11.5 | 33.7 | 12.5 |
| C3 T | 42.7 (27.5) | 46.7 | 41.7 | 57.3 | 3.6 |
| C3 SAN + T | 43.5 (23.9) | 75.3 | 15.1 | 40.0 | 15.1 |

formance and their combination does not improve statistically significantly the performance. The SAN features are not relevant for the ME, UI and ID roles because they have similar interaction patterns, and thus our system is not able to distinguish between them. The speaking time features does not improve either the role recognition performance. It is due to the fact that the distribution of the speaking time is similar for the PM, ME and UI roles. It seems that both types of features are not relevant enough for the role recognition task over the meetings C3. However, it is hard to determine whether the low performance achieved over C3 is due to the limitation of our approach, or whether it is due to the data set itself, which is composed of acted data and not real data (see Section 3.3.1 for a description of the metting corpus C3).

## 3.5 Role Recognition Approach based on Probabilistic Sequential Models

The main limitation of the automatic role recognition approach presented in Section 3.2 is that it does not take into account any sequential information, whereas it should be important as we consider conversations. In fact, 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 $n+1$. This is the reason why we have considered a second role recognition approach modeling sequential information using probabilistic sequence models (i.e. Hidden Markov Models (HMM) and statistical language models (SLM)).

### 3.5.1 Modeling Sequential Information

The core idea of the approach we propose is that the sequence of actors talking during a conversation is the observable, machine detectable, evidence of an underlying, hidden, sequence of roles $R$. The role recognition problem can thus be thought of as finding the best role sequence $R^*$ given the sequence of observation features.

Section 3.1 has shown that each actor corresponds to a pair $\mathbf{y}_a = (\mathbf{x}_a, \tau_a)$ of dimension $D+1$. We have reduced the dimensionality of the tuples representing the interaction patterns through Principal Component Analysis (PCA) [Bishop 06]. The application of PCA to the $\mathbf{y}_a$ tuples results into L-dimensional projections $\mathbf{w}_a$, where $L \leq D+1$. Therefore, each recording can be represented through a sequence of tuples $W = (\mathbf{w}_a 1, \ldots, \mathbf{w}_a N)$, where $N$ is the number of turns detected at the speaker diarization step, and $\mathbf{w}_a k$ is the tuple representing the actor $a$ talking at turn $k$.

Thus, given the sequence of observations $W$, the role recognition problem can be formulated as finding the role sequence $R^*$, satisfying the following expression:

$$R^* = \arg\max_{R \in \mathcal{R}^N} p(W, R) p(R) \tag{3.22}$$

where $R = (r_1, \ldots, r_N)$ is a sequence of roles of length $N$, $r_i \in \mathcal{R}$ ($\mathcal{R}$ is a predefined set of roles), and $\mathcal{R}^N$ is the set of all possible role sequences of length $N$. In intuitive terms, the above equation says that $R^*$ is the sequence of roles that better explains the sequence of turns actually observed during a conversation.

In our experiments, the joint probability $p(W, R)$ was estimated with a fully connected, ergodic, HMM [Rabiner 89] where each state corresponds to a role $r \in \mathcal{R}$. Each state can be reached from any other state, meaning that transitions between any pair of roles are allowed. The emission probability function associated to each state are Gaussians.

The *a-priori* probability $p(R)$ is estimated using a $n$-gram ($n \geq 1$) statistical language model [Rosenfeld 00]:

$$p(R) = \prod_{k=1}^{N} p(r_k | r_{k-1}, r_{k-2}, \ldots, r_{k-n+1}) \tag{3.23}$$

HMMs and SLMs have been implemented with two publicly available packages, the Hidden Markov Model Toolkit (HTK) [1], and the SRI Language Model Toolkit [2].

### 3.5.2 Experiments and Results

The same experimental setup as the one used for the role recognition approach based on Bayes classifiers (see Section 3.2) has been applied in order to compare the obtained role recognition results (see Section 3.3.2).

Table 3.9 reports the results achieved over C1 and C2, Table 3.10 those obtained for C3. Each overall accuracy value is accompanied by the standard deviation of the accuracies achieved over the different recordings of each corpus. The first row (HMM) shows the results when using only HMMs, the others show the accuracy achieved with language models of increasing order (HMM+$n$-gram). For each corpus, the last row reports, for comparison purposes, the performance achieved with the Bayes approach (see Tables 3.5 and 3.6). These results are obtained when considering independence between the roles, in order to have similar setup and comparable results.

To reduce the dimensionality of interaction features, PCA has been applied. The amount of variance to be retained after PCA has been selected through cross-validation over the training set. The minimum amount of variance to be retained after PCA has been set to 70%.

We have modeled the emission probability function in the HMM using a single Gaussian, as it was sufficient to capture the necessary information for the role recognition. Mixtures of Gaussians do not improve the role recognition performance, but simply increase the number of parameters.

Even if the training material available is sufficient to train language models of order up to 6, no performance improvements are observed for $n > 3$. This seems to suggest that higher order models do

[1] http://htk.eng.cam.ac.uk/

[2] http://www.speech.sri.com/projects/srilm/

| | all ($\sigma$) | AM | SA | GT | IP | HP | WM |
|---|---|---|---|---|---|---|---|
| Results over C1 | | | | | | | |
| HMM | 75.3 (9.2) | 97.8 | 9.8 | 69.7 | 27.8 | 58.1 | 73.1 |
| HMM + 1-gram | 79.6 (8.1) | 97.8 | 10.9 | 83.8 | 3.5 | 57.5 | 81.8 |
| HMM + 2-gram | 80.5 (9.0) | 97.8 | 12.3 | 83.9 | 21.1 | 59.6 | 79.5 |
| HMM + 3-gram | 80.5 (8.3) | 97.8 | 16.5 | 82.7 | 23.5 | 57.5 | 77.9 |
| HMM + 4-gram | 81.0 (8.0) | 97.8 | 16.7 | 84.6 | 22.0 | 58.9 | 77.5 |
| Bayes | 82.5 (6.9) | 98.0 | 3.6 | 97.8 | 8.0 | 64.6 | 79.9 |
| Results over C2 | | | | | | | |
| HMM | 73.5 (10.3) | 60.1 | 88.6 | 78.2 | N/A | 22.1 | 72.0 |
| HMM + 1-gram | 83.8 (6.1) | 74.4 | 91.9 | 91.9 | N/A | 25.8 | 32.4 |
| HMM + 2-gram | 81.3 (8.2) | 70.0 | 88.4 | 90.4 | N/A | 22.2 | 9.8 |
| HMM + 3-gram | 83.3 (8.2) | 70.1 | 89.5 | 90.1 | N/A | 58.3 | 27.9 |
| HMM + 4-gram | 82.1 (7.1) | 67.5 | 88.9 | 89.7 | N/A | 47.6 | 24.4 |
| Bayes | 82.6 (6.8) | 75.0 | 88.3 | 91.6 | N/A | 18.3 | 6.7 |

**Tab. 3.9:** Role recognition performance based on probabilistic sequential models over C1 and C2. The table reports both the overall accuracy and the accuracy for each role. The overall accuracy is accompanied by the standard deviation $\sigma$ of the performance achieved over the single recordings.

| | all ($\sigma$) | PM | ID | ME | UI |
|---|---|---|---|---|---|
| Results over C3 | | | | | |
| HMM | 43.2 (26.2) | 60.5 | 27.0 | 26.2 | 44.0 |
| HMM + 1-gram | 40.5 (25.6) | 56.2 | 27.7 | 27.0 | 36.3 |
| HMM + 2-gram | 41.5 (25.3) | 58.1 | 27.3 | 24.5 | 38.9 |
| HMM + 3-gram | 38.5 (23.1) | 52.4 | 28.9 | 17.6 | 35.3 |
| HMM + 4-gram | 38.0 (22.6) | 51.9 | 29.8 | 13.9 | 37.1 |
| Bayes | 43.5 (23.9) | 75.3 | 15.1 | 40.0 | 15.1 |

**Tab. 3.10:** Role recognition based on probabilistic sequential models over C3. The table reports both the overall accuracy and the accuracy for each role. The overall accuracy is accompanied by the standard deviation $\sigma$ of the performance achieved over the single recordings.

not bring any information and the role observed at turn $k$ depends at most on the last two preceding roles.

The performance tends to be higher for those corpora where the *Perplexity* $PP$ of the language models is lower:

$$PP = [\prod_{k=1}^{N} p(r_k|r_{k-1}, r_{k-2}, \dots, r_{k-n+1})]^{-\frac{1}{N}} \tag{3.24}$$

where $N$ is the length of role sequence $R = \{r_1, \dots, r_N\}$. The $PP$ values are reported in Table 3.11, together with the ratio $PP/|\mathcal{R}|$ of the $PP$ to the number of roles of each corpus.

| | C1 | | C2 | | C3 | |
|---|---|---|---|---|---|---|
| | $PP$ | $PP/\|\mathcal{R}\|$ | $PP$ | $PP/\|\mathcal{R}\|$ | $PP$ | $PP/\|\mathcal{R}\|$ |
| 1-gram | 5.5 | 0.9 | 3.3 | 0.7 | 4.0 | 1.0 |
| 2-gram | 2.1 | 0.4 | 2.5 | 0.5 | 3.0 | 0.8 |
| 3-gram | 1.9 | 0.3 | 2.0 | 0.4 | 2.9 | 0.7 |
| 4-gram | 1.9 | 0.3 | 2.0 | 0.4 | 2.9 | 0.7 |
| 5-gram | 1.9 | 0.3 | 2.0 | 0.4 | 2.9 | 0.7 |
| 6-gram | 1.9 | 0.3 | 2.0 | 0.4 | 2.9 | 0.7 |

**Tab. 3.11:** $PP$ stands for the perplexity measure of the different $n$-gram and $PP/|\mathcal{R}|$ is the proportion of the dictionary that has a probability higher than 0 to produce the $n$-gram sequence.

The $PP$ is the inverse of the geometric mean of $p(r_k|r_{k-1}, \ldots, r_{k-n+1})$ along a sequence $R$, and can be interpreted as the number of roles that, at each step $r_k$ of $R$, have a probability of appearing significantly higher than 0 [Rosenfeld 00]. Thus, when $PP$ is low, this probability is, on average, high and roles from $r_{k-n+1}$ to $r_{k-1}$ influence significantly role $r_k$. The consequence is that only few roles can have probability significantly higher than 0 of appearing immediately after $r_{k-1}$. This corresponds to say that the roles are formal, that is the direct interaction (i.e., adjacency in $R$) between roles is more constrained. Thus, *the Perplexity appears to be a measure of how much a role set is formal, i.e. of how much the interactions between its roles are constrained.*

This is the first work that provides a quantitative measure of how formal a role set is, i.e. of how much the roles under consideration constrain the interaction behavior of the persons. This is important to assess how effectively a role recognition approach can work in different interaction settings. Moreover, the perplexity can be applied each time roles underly a sequence of events (like the speaker turns in the case of this work).

The role recognition accuracies suggest that the roles in the meeting corpus C3 are harder to model than those of broadcast data C1 and C2. This was already observed with the previous approach in Section 3.4.1, and was explained by the fact that meeting roles are *informal.* In fact, the meeting roles correspond to a position in a given social system, and are not associated to stable behavioral patterns as in the *formal* roles typical of broadcast material.

According to the Kolmogorov-Smirnov Test [Massey Jr. 51], the difference between the performance achieved with HMMs and the one achieved with the Bayesian classifier described in Section 3.4.1 is not statistically significant.

However, the two classifiers show a significant degree of *diversity*, i.e. they make different decisions

| C1 | HMM C | HMM W |
|---|---|---|
| Bayes C | 78.0 | 2.2 |
| Bayes W | 4.5 | 15.3 |

| C2 | HMM C | HMM W |
|---|---|---|
| Bayes C | 79.4 | 3.9 |
| Bayes W | 3.2 | 13.5 |

| C3 | HMM C | HMM W |
|---|---|---|
| Bayes C | 22.3 | 11.3 |
| Bayes W | 15.9 | 50.5 |

**Tab. 3.12:** Diversity assessment. The table reports the accuracy of the percentage of data where the two approaches are both correct (C), both wrong (W), or one wrong and the other correct.

over the same sample in a relatively high percentage of cases (see Table 3.12). In particular, probabilistic sequential approaches tend to improve the recognition of less frequent roles that are typically penalized by Bayesian classifiers certainly because of their low *a-priori* probability. *This suggests that the combination of the two approaches is likely to lead to significant performance improvements.* The highest possible performance deriving from a combination corresponds to the sum of the cases where at least one of the two approaches is right. This corresponds to 84.7% for C1, 86.6% for C2, and 49.5% for C3. In all of the cases, this would represent a statistically significant improvement with respect to the best of the approaches.

## 3.6 Combination of Interaction and Lexical Patterns

Both approaches presented in this work for the role recognition task, i.e. the role recognition approach assigning a role to each person using Bayesian classifiers (see Section 3.2) and the role recognition approach taking into account sequential information (see Section 3.5.1), show limitations on the meeting recordings (C3).

One possible explanation of the lower role recognition performance over the C3 corpus may be due to the experimental setup of the C3 corpus itself. In fact, C3 is composed by acted interactions and not real interactions (see Section 3.3.1 for a precise definition of the C3 database).

Another possible explanation of these results could be that the social interaction based role recognition approaches developed in this thesis are not well suited for informal roles and less constrained conversations such as the ones represented in the C3 corpus. We were not able to verify this assumption by applying our

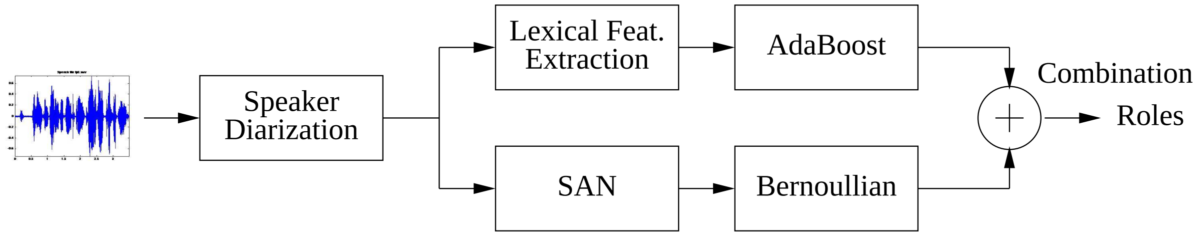

**Fig. 3.8:** Overview of the role recognition approach combining two types of behavioral cues: interaction and lexical patterns. The two parallel paths produce separate decisions that are combined at the end of the process.

automatic role recognition approaches over another scenario of conversations that were not constrained by specific tasks. In fact, when I started my thesis, no other roles labeled spontaneous conversations were available.

However, to assess the role recognition problem over the C3 meetings, we developed a new role recognition system in which we added lexical content to our interaction features.

This section presents the new role recognition approach for the meetings C3 which combines two behavioral cues. The first behavioral cue is the *interaction pattern*, i.e. the patterns representing the tendency of each actor $a$ to interact with certain persons rather than others in a certain proximity in time. These features are extracted from the Affiliation Networks exactly as previously detailed in Section 3.1.3, and are represented by binary n-tuple $\mathbf{x}_a$. The second behavioral cue is the *lexical choice*, i.e. the use of certain words rather than others in the interventions of each person. A full description of how the lexical features have been transcribed from the meetings can be found under [Hain 06].

An overall scheme of the approach is depicted in Figure 3.8: the first step is the application of a speaker diarization approach that identifies the time intervals where each persons talks (see Section 3.1.2). The subsequent steps follow two parallel paths corresponding to the two behavioral sources of information mentioned above.

The lower path corresponds to the interaction pattern modeling and it includes two stages: extraction of the interaction features using a Social Affiliation Network (see Section 3.1.3), and assignment of roles to the features representing each person using a Bernoulli distribution [Bishop 06].

The upper path describes the modeling of the lexical choice and it includes two stages as well: extraction of the lexical features from the automatic speech transcriptions [Hain 06], and mapping of the lexical features into roles using the BoosTexter text categorization approach [Schapire 00].

The next sections present the SAN based role recognition approach (see Section 3.6.1), the lexicon based role recognition approach (see Section 3.6.2), the combination approach (see Section 3.6.3), and finally Section 3.6.4, shows the experiments performed and the role recognition results achieved.

### 3.6.1 Social Affiliation Networks Based Role Recognition

This role recognition approach is based on the Affiliation Networks (see upper part of Figure 3.2 in page 26) [Wasserman 94] described in Section 3.1.3. We have seen previously in Section 3.2.1 that the most natural way of modeling binary features is to use independent Bernoulli discrete distributions:

$$\mathrm{p}(\mathbf{x} \mid \overrightarrow{\mu}) = \prod_{j=1}^{D} \mu_j^{x_j} (1-\mu_j)^{1-x_j} \tag{3.25}$$

where $D$ is the number of events used to capture the interaction patterns, and $\overrightarrow{\mu} = (\mu_1, \ldots, \mu_D)$ is the parameter vector of the distribution. A different Bernoulli distribution is trained for each role. The maximum likelihood estimates of the parameters $\mu_r$ for a given role $r$ are as follows [Bishop 06]:

$$\mu_{rj} = \frac{1}{|A_r|} \sum_{a \in A_r} x_{aj} \tag{3.26}$$

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

Each actor will thus be assigned the role $r^*$ according to:

$$r^* = \arg\max_{r \in \mathcal{R}} p(\mathbf{x} \mid \overrightarrow{\mu}_r) \tag{3.27}$$

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

### 3.6.2 Lexicon Based Role Recognition

The role recognition approach based on lexical features recognizes the roles of the persons using the lexical content of their utterances. The rationale behind this approach is that the meeting content is correlated with the roles of its participants. Thus the lexical cues related to the topics can be useful for determining persons roles. As an example, the person leading the discussion (i.e the Project manager (PM)), can use phrases to re-center the conversation to the main topic or can use phrases to shift to another topic.

Similarly to the SAN based approach (see Section 3.6.1), the goal of the role recognition task is to assign a role to each speaker in every meeting.

For classifying the persons into one of the possible roles, we use BoosTexter, a multi-class classification tool. Boosting aims to combine *weak* base classifiers to come up with a *strong* classifier [Schapire 00]. 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 an uniform distribution of the roles, $D_1(i, r)$, over training examples from the meeting participants, $i$, and over labels (i.e., participant 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 improve the most the classification rule. To represent every example $i$ (i.e. every meeting participant in the training corpus), we use as features, word $n$-grams ($n = 1, 2,$ and 3) from all the turns of a same participant in a meeting.

The weak classifiers check the presence or absence of word $n$-grams in the participant's turns, and can therefore be used for analysis purposes. The final strong classifier is a linear combination of the individual weak classifiers. We used a $k-fold$ cross-validation method [Bishop 06] to compute the optimum number of iterations for the classifier. The classifier outputs a probability for the presence of each class for each person.

If $\mathbf{d}_i$ is the tuple representing the transcription of the interventions of meeting participant $i$, then the BoosTexter approach estimates the probability $p(\mathbf{d}_i \,|\, r)$ of the participant playing role $r$ by combining the weak classifiers described above. The participant $i$ is assigned the role $r^*$ according to:

$$r^* = \arg\max_{r \in \mathcal{R}} p(\mathbf{d}_i \,|\, r) \tag{3.28}$$

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

### 3.6.3 Combination Approach

Both role recognition approaches described above (see Sections 3.6.1 and 3.6.2) estimate the probability of a meeting participant playing a role $r$. The combination is performed by multiplying the two estimates

| approach | all | PM | ME | UI | ID |
|---|---|---|---|---|---|
| SAN | 43.1 | 75.7 | 16.4 | 41.2 | 13.4 |
| lex. | 67.1 | 78.3 | 71.9 | 38.1 | 53.0 |
| SAN+lex. | 67.9 | 84.0 | 69.8 | 38.1 | 50.1 |

**Tab. 3.13:** Role recognition results when combining interaction features (SAN) and lexical features over the meetings C3.

as follows:

$$\begin{aligned} r^* &= \arg\max_{r \in \mathcal{R}} p(\mathbf{x}, \mathbf{d} \,|\, r, \overrightarrow{\mu}_r) \\ &= \arg\max_{r \in \mathcal{R}} \beta \log p(\mathbf{d} \,|\, r) + (1-\beta) \log p(\mathbf{x} \,|\, \vec{\mu}_r) \end{aligned} \tag{3.29}$$

where the factor $\beta$ ensures that both terms are of the same order of magnitude and contribute to the final decision. The $\beta$ value is selected through cross validation (see next section).

### 3.6.4 Experiments and Results

The role recognition approach presented in this section has been developed to improve the performance over the AMI corpus (referred as C3 in this thesis). The description of the C3 database can be found in Section 3.3.1.

The training of the role recognition system is performed using a leave-one-out approach (see Section 3.3.2), i.e. using the same experimental setup as with the other role recognition approaches presented previously in this thesis. All the recordings composed the training set (i.e. for training the role's models) with the exception of one that is used as test set. The hyperparameter $D$ is set through cross-validation over the training set (see Section 3.3.2). The other hyperparameters of the system (number of AdaBoost iterations for the lexicon based approach, and $\beta$ factor for the combination) are tuned over a subset of 20 meetings randomly selected in the training set.

The performance is measured with the *accuracy* $\alpha$, i.e. the percentage of data time correctly labeled in terms of role. Table 3.13 reports the accuracies obtained by using only Social Affiliation Network Analysis, only lexical choices, and the combination of the two. The results are reported for the overall meetings, as well as for the single roles separately.

The lexical choice appears to be a more reliable cue for the recognition of the role for the AMI meetings. The overall accuracy of the lexicon based system is significantly higher (67.1% against 43.1%).

A possible explanation is that the AMI corpus is particularly suitable for lexical analysis, while it is rather unfavorable to the application of SAN. On one hand, the content of the interventions is constrained by the role and this helps the former approach, on the other hand, the similar interaction patterns of the participants may limit significantly the latter approach, as the social networks are not able to distinguish between the roles.

The combination of the two systems does not improve significantly the performance of the best system (see Table 3.13). The main reason is probably that the performance of the SAN approach is too close to the chance (around 25%) for at least two roles (ME and ID). Thus, the SAN 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 SAN system over the PM is significantly better than the chance because the PM plays a formal role as we have seen previously in Section 3.4.1.

In conclusion, the interaction patterns are not enough reliable cues, and lexical content is necessary to obtain an effective role recognition system in the AMI meetings (C3 corpus). We are not certain about the limitation of the use of interaction features extracted with Social Affiliation Networks. In fact, we are not able to state whether this is the proposed interaction features which are not meaningful (because they are similar), or whether this is the C3 corpus which dos not contain relevant interaction patterns (simulated data and not real spontaneous interactions).

Another way to improve the role recognition performance in meetings would be to use also cues extracted from the video channel in combination with the cues extracted from the audio channel. This possibility has not been addressed in this thesis.

## 3.7 Role Recognition Discussion

This chapter has presented automatic approaches for the recognition of roles in multiparty recordings.

The proposed approaches have been tested over roughly 90 hours of material, composed of broadcast material and meeting recordings. This is one of the biggest data sets ever used in literature for this task. Moreover, to the best of our knowledge, the data set used in this work is the only one that includes different interaction settings and different role sets, i.e. both *informal* and *formal* roles (See Section 2.2 for the difference between the two types of role). This is important in order to show how the role typology influences the effectiveness of the recognition, and thus how easily an approach can be ported from one interaction setting to another. Furthermore, the thesis has identified a quantitative measure (the Perplexity) of how formal a role set is, i.e. of how much the roles under consideration constrain the interaction behavior of the persons (see Section 3.5.2). The perplexity measure can be applied each time roles underly a sequence of events (like the speaker turns in the case of this thesis).

Another novelty of the presented approaches is to use the interaction between the persons as features. The Social Affiliation Networks (SAN) [Wasserman 94] allows one to extract these features, which represent the evidence of interactions in terms of proximity in time, from the co-occurence turn-taking patterns structuring the conversations. The rationale behind the SAN is that the persons speaking in the same time intervals are likely to interact with each other.

Furthermore, this chapter has compared approaches based on Bayesian classifiers and approaches based on probabilistic sequential models. The former assigns a specific role to each person involved in the recordings (see Section 3.2). The latter considers the sequence of persons talking during a conversation, and aligns the sequence of their turns with a sequence of roles (see Section 3.5). For both approaches, the results show that the role recognition accuracy is higher than 80% 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 between the different types of data sets. The first, and probably most important, is that broadcast data include formal roles, while meetings include informal ones. Formal roles are easier to model because they impose constraints on the behavior of the persons that can be detected. In contrast, informal roles do not necessarily constrain behavior and so automatic recognition is more difficult through approaches like the ones presented in this thesis, at least for the aspect of behavior used as role evidence in this work, i.e. *who talks with whom and when.*

The second reason is that the broadcast data is real, while the meeting data is acted. The meetings do not involve persons 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 should be sufficient to browse effectively the data, or at least could help it. In fact, users should quickly find segments corresponding to a given role because the mismatch between the ground truth and the automatic output rarely exceeds a few seconds. In the case of meeting recordings, the approach is effective only to identify the Project Manager. However, this should allow one to effectively follow the progress of the meeting as the PM plays the chairman role and, as such, is responsible for following the agenda through her/his interventions.

In order to improve the role recognition performance in the meeting recordings, we have proposed another approach combining lexical patterns to the interaction patterns. The role recognition performance is improved to 67.9%, but this is mainly due to the lexical features. In fact, the combination of the lexical features with the interaction features significantly improves the performance for the Project Manager role only. To our knowledge, this is the first attempt to combine approaches based on both lexical and interaction features.

# Chapter 4

# SEMANTIC SEGMENTATION

The content of this chapter can be found in the following papers:

- "Semantic Segmentation of Radio News Using Social Network Analysis and Duration Distribution Modeling", A. Vinciarelli, F. Fernandez and S. Favre, in Proceedings of the 2007 IEEE International Conference on Multimedia and Expo (ICME), pages 779-782

- "Broadcast News Story Segmentation Using Social Network Analysis and Hidden Markov Models", A. Vinciarelli and S. Favre, in Proceedings of the 15th international conference on Multimedia (ACM), 2007, pages 261-264

Role recognition can have different applications (see Section 1.1) and this chapter shows how one of the approaches presented in the previous chapter has been used to perform semantic segmentation, i.e. to split audio recordings into segments which are meaningful from a user point of view [Koumpis 05]. In particular, roles have been used to detect the structure of a particular radio program composed of two parts (a news bulletin and a talk-show), and to segment the same program into stories (segments during which single specific topic is addressed). The rationale is that different roles might characterize different semantic segments, thus recognizing roles might allow one to identify semantic segments.

Achieved performances have not been compared to other similar semantic tasks. In fact, this chapter shows how to use roles in an application scenario and thus presents really simple approaches for segmentation. To perform role applications was a secondary task during my thesis, and for time reasons, such approaches have not been further developped.

The main contributions of this chapter with respect to the rest of literature are as follows:

- This work is probably the first approach that does not take into account the content of the data (what is said), but the structure of social interactions as described by automatically extracted social networks.

- We believe that this work is one of the first attempts of using role recognition in an application scenario.

The rest of this chapter is organized as follows: Section 4.1 details an approach for segmenting a particular radio program into two parts corresponding to a news bulletin and a talk-show respectively, Section 4.2 presents an approach segmenting a particular radio program into stories, and Section 4.3 draws some conclusions on the presented semantic segmentation approaches.

## 4.1 Structure Detection

The experiments have been performed over the recordings of C2 (see Section 3.3.1 for more details on the concerned database).

C2 is composed of 27 one hour long recordings that can be split into two distinct parts: the first is referred to as *news* and the second is referred to as *talk-show*. The former consists of news subjects presented one after each other, whereas the latter consists of discussions about specific subjects by invited speakers.

This section shows how specific roles, i.e the anchormen, can be used to identify the two distinct parts in C2. In fact, C2 involve two anchormen with the particularity that one talks all along the program, whereas the other talks only during the news part. In this way, identifying the two anchormen allow us to identify the transition between the news and the talk-show parts.

We propose two approaches for the semantic segmentation: the first is based on Social Network Analysis (SNA) (Section 4.1.1) and uses the roles to split the radio programs in two parts. The second approach, called Duration Distribution Modeling (DDM) in the following, is based on the duration of single stories (Section 4.1.2), and is used for results comparison.

The rest of this section is organized as follows: Section 4.1.1 presents the SNA based approach, Section 4.1.2 presents the DDM approach, Section 4.1.3 concludes with the experiments performed and the achieved results.

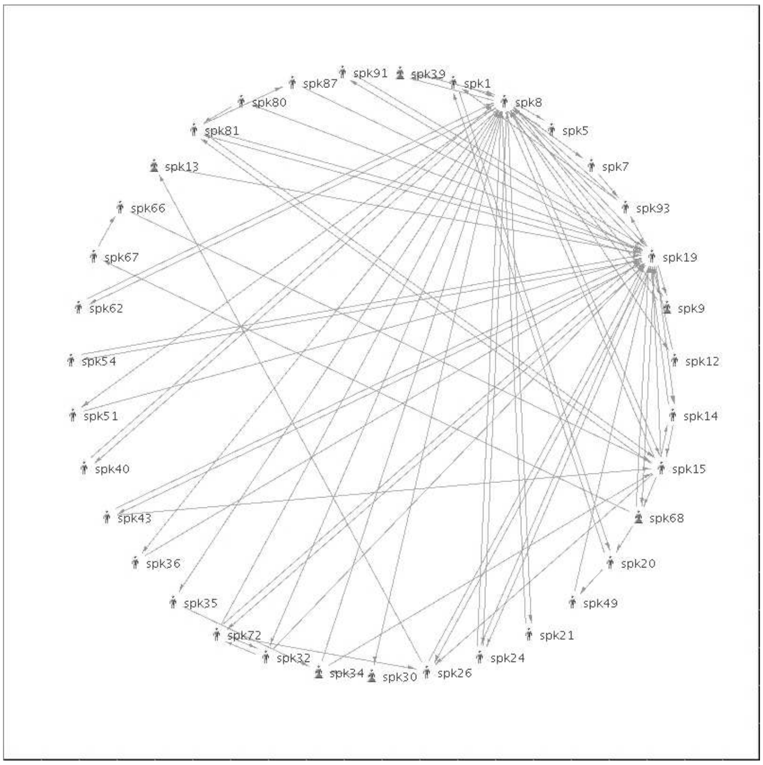

**Fig. 4.1:** Social Network: this figure shows the Social Network extracted from one of the recordings in our collection. We can see that the two speakers labeled as spk19 and spk8 are the two main central speakers.

### 4.1.1 Social Network Analysis for Structure Detection

The SNA approach developed here relies on the fact that the radio programs in C2 involve two anchormen: the first one talks all along the program, while the second talks only during the first part. By identifying the two anchormen is then possible to identify the transition between first and second part. In fact, the transition can be detected as the last intervention of the second anchorman, i.e. the one that stops talking before the end of the program.

In order to extract the social networks describing the interactions between the persons, we first apply a speaker diarization system to obtain the sequence of persons interventions (turn-taking). This step is not detailed here as it has been done previously in this report (see Section 3.1.1).

The result of the speaker diarization process is that the audio data is converted into a sequence of speaker ID codes $a_i$, with $i \in \{1, \ldots G\}$ ($G$ is the total number of detected speakers in the speaker diarization process described in the previous Section 3.1.1).

We use as interaction evidence between two individuals $a_i$ and $a_j$ the fact that $a_i$ talks immediately before $a_j$ at least once. The use of the ordering includes the temporal information involved in the sequence resulting from the speaker diarization process. This allows to build the so-called *sociomatrix* $X$, i.e. a matrix where the element $x_{ij}$ is the number of times speaker $a_i$ talks immediately before speaker $a_j$. For each sociomatrix there is an associated directed graph where each node corresponds to a speaker and each edge corresponds to the interaction between the connected speakers: such a graph is called *Social Network* (SN) and it is shown, for one of the recordings in our data set, in Figure 4.1. Sociomatrices and SNs encode *relational data*, i.e. the interaction patterns involving the speakers participating in each recording.

In this work, the most important information is the speakers *centrality* [Wasserman 94], i.e. the inverse of the average geodesic distance between a given individual and the others (the geodesic distance between two nodes is the number of edges to be traversed to go from a node to the other):

$$C(a_i) = \frac{G-1}{\sum_{j \neq i} d(a_i, a_j)} \tag{4.1}$$

where $d(a_i, a_j)$ is the geodesic distance between $a_i$ and $a_j$ and $G$ is the total number of speakers. The reason for the name centrality is that such index is a measure of how much individuals are close to the others on average and then of how much they are central in the interaction pattern.

In the performed experiments, we show that the two anchormen are the individuals with the highest centrality. In other words, the extraction of the Social Network and the calculation of the centrality index enable one to find the anchormen $a_i^*$ and $a_j^*$ as follows:

$$a_i^*, a_j^* = \arg \max_{a_i, a_j \in (1, \ldots, G)} C(a_i) + C(a_j) \tag{4.2}$$

If $\tau(a_k)$ is the time at which the last intervention of speaker $a_i$ ends, then the approach described in this section identifies the transition time $t^*$ between news and talk-show as follows:

$$t^* = \arg \min_{a_k \in \{a_i^*, a_j^*\}} \tau(a_k) \tag{4.3}$$

in other words, the transition is considered to take place at the end of the last intervention of the

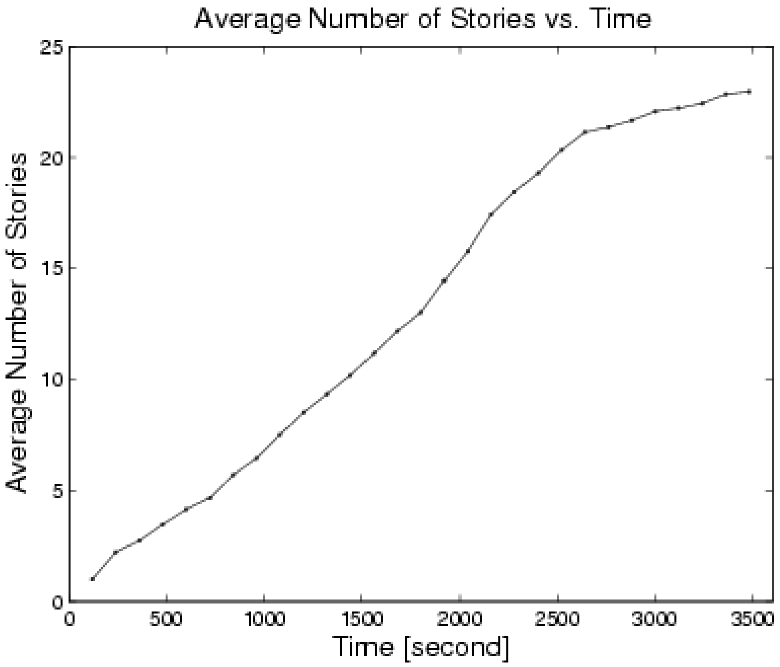

**Fig. 4.2:** Average number of stories vs time. This plot shows the average number of stories (estimated at two minutes long time steps) as a function of the time.

anchorman that disappears first from the program.

### 4.1.2 Duration Distribution Modeling for Structure Detection

The rationale behind such approach is that our data can be considered as a sequence of stories and that the transition points between consecutive stories follow a Poisson Stochastic Process (PSP) [Papoulis 91]. This can be seen by observing the following: given a recording $p$ in the collection, consider the *staircase* function $f_p(t)$ which gives the number of story transitions that took place between time 0 and time $t$. Such function is called staircase because it increases by one each time there is a transition and then it remains stable until there is another transition. The average number of transitions $n(t)$ in the data set at a given time $t$ can be estimated as follows:

$$n(t) = \frac{1}{P}\sum_{p=1}^{P} f_p(t) \tag{4.4}$$

where $P$ is the total number of recordings in the data set. The function $n(t)$ is plotted in Figure 4.2 and it consists of two linear pieces that can be expressed as $n_1(t) \simeq \lambda_1 t$ and $n_2(t) \simeq \lambda_2 t$. This shows that the transitions actually follow a PSP and that the PSP underpinning the transitions changes at a certain point of the program. The change of slope corresponds to the transition between the news and the talk-show: the segmentation process can be thought of as finding the story in correspondence of which the underlying PSP (and the corresponding $\lambda$ parameter) changes.

Since the transition points are supposed to follow a PSP, the probability of a story being long $\tau$ can be written as follows [Papoulis 91]:

$$p(\tau|\lambda) = \lambda e^{-\lambda\tau} \tag{4.5}$$

and the likelihood of a sequence $T = \{\tau_1, \ldots, \tau_S\}$ of story durations in a given recording can be expressed as follows:

$$p(T|\lambda_1, \lambda_2) = \prod_{k=1}^{n} p(\tau_k|\lambda_1) \prod_{l=n+1}^{S} p(\tau_l|\lambda_2) \tag{4.6}$$

where $n$ is the index of the story where the PSP underlying the story transitions changes, i.e. the index of the story where the news end and the talk-show starts. The value of $n$ can be found by maximizing the logarithm of the likelihood:

$$\begin{aligned} n = \arg\max\nolimits_p p \log \lambda_1 + (S-p)\log \lambda_2 - \\ \lambda_1 \textstyle\sum_{k=1}^{p} \tau_k - \lambda_2 \sum_{k=p+1}^{S} \tau_k \end{aligned} \tag{4.7}$$

The last problem to be solved is the estimation of the parameters $\lambda_1$ and $\lambda_2$. This is performed using a leave-one-out approach, i.e. by using all recordings except the one used for testing the algorithm (see Section 3.3.2 for more details on the leave-one-out approach). Given a set of recordings for which $n$ is known, the $\lambda_i$ values are those that maximize the likelihood of all the $T$ sequences observed in the training set:

$$\lambda_i = \frac{S_i}{\sum_{k=1}^{S_i} \tau_k^{(i)}} \tag{4.8}$$

| Approach | $\alpha$ |
|---|---|
| SNA | 94.5% |
| DDM | 99.8% |

**Tab. 4.1:** Structure detection results. The table reports the accuracy (percentage of time correctly labeled in terms of semantic class) obtained using SNA and DDM approaches.

where $S_i$ is the total number of stories following a stochastic process with parameter $\lambda_i$ and $\tau_k^{(i)}$ is the $k^{th}$ story following the same stochastic process.

### 4.1.3 Experiments and Results

The performance is measured in terms of *accuracy* $\alpha$, i.e. in terms of the percentage of time where the semantic class (news or talk-show) is assigned correctly. Since each recording contains only two segments, $100 - \alpha$ expresses the distance (in terms of percentage with respect to the total duration of the recording) between the actual transition point and the transition point detected automatically. In other words, if the accuracy in a recording is 95%, then the difference between the real transition and the detected transition accounts for 5% of the total duration of the recording.

The results of the experiments are reported in Table 4.1. The method based on the story transitions performs better than the other, but such a performance is overestimated. In fact, the results are obtained over a manual segmentation, i.e. the story transitions have been detected by a human assessor. The process is then not fully automatic.

On the contrary, the results obtained using the SNA based approach are realistic because the process does not involve any manual intervention. The speaker segmentation (see Section 3.1.1) is automatic as well as the analysis of the resulting Social Network. The average distance between the actual transition and the transition detected automatically is around 3 minutes. This means that a potential user does not need to listen to more than 6 minutes (3 minutes before and 3 minutes after the detected point) in order to find the actual transition between news and talk-show. This reduces by roughly 40% the variability range observed in our data (the transition point is between $\simeq 35$ and $\simeq 45$ minutes), then it decreases the amount of time needed for an operator to find the real transition point.

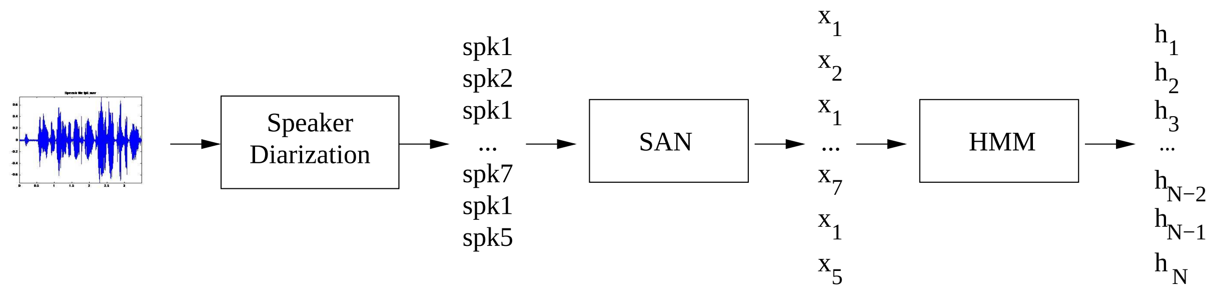

**Fig. 4.3:** Story Segmentation approach. This figure shows the three stages of the story segmentation approach: the first splits the audio into a sequence of single speaker segments (i.e. turn-taking), the second converts the turn-taking into features using SAN, the third maps the features into stories using HMMs.

## 4.2 Story Segmentation

Broadcast news data are structured by the specific issues that are presented one after each other along a news bulletin. It implies that in the case of broadcast news, the segmentation is typically performed in terms of *stories*. The stories play in broadcast news the same role that the articles play in newspapers. The stories can be thought of as the main building block of broadcast news: any news bulletin can be split into stories and, vice-versa, a sequence of stories can form a news bulletin.

This section presents a new approach for segmenting broadcast news into stories. The main rationale behind the presented approach is that persons involved in the same story have a high degree of mutual interaction. This means that the stories can be identified by grouping the persons that have a high degree of mutual interaction.

The proposed approach is composed of three major stages: the first performs a speaker diarization and splits the audio into segments corresponding to a single voice (Section 3.1.1). The goal of this stage is to detect the persons involved in the radio program and the sequence of their interventions. The second stage is the representation of social interactions by means of a SAN (see Section 3.1.3, to identify individuals with high mutual interaction. The third step is the application of HMM [Rabiner 89] and Statistical Language Models (SLM) [Rosenfeld 00] to map social interactions into stories.

The following two sections present the story segmentation approach (Section 4.2.1), experiments and results achieved (Section 4.2.2).

### 4.2.1 Story Segmentation Approach

This section presents in details the story segmentation approach depicted in Figure 4.3. The first stage of the process is the speaker diarization, which is fully described in Section 3.1.1. The result of the speaker diarization process is that each recording is split into a sequence of turns $S = \{(s_k, t_k, \Delta t_k)\}$, where $k \in \{1, \ldots, N\}$, $s_k$ is the label corresponding to the voice detected in the $k^{th}$ turn, $t_k$ is the beginning of speaker $s_k$ intervention, and $\Delta t_k$ is the duration of the $k^{th}$ turn. The label $s_k$ belongs to the set $A = \{a_1, \ldots, a_G\}$ of $G$ unique speaker labels as provided by the speaker diarization process (see lower part of Figure 3.2 in page 26).

The second stage of the approach, as depicted in Figure 4.3, uses the sequence of turn-taking $S$ extracted from the speaker diarization, to build a SAN which represents the interactions between the speakers. A SAN is a graph with two types of nodes: the *actors* and the *events* [Wasserman 94]. A complete description of how to build SAN can be found in Section 3.1.3. The important thing is that SANs extract the evidence of interactions in terms of *who talks to whom and when*, and thus capture the mutual degree of interaction between the participants.

The events are defined using the proximity in time (see lower part of Figure 3.2 in page 26): the news bulletins are split into $D$ uniform non-overlapping segments, called events $e_j$. An actor $a_i$ is said to participate in event $e_j$ when he/she talks during it. In this way, each actor $a$ is represented by a n-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. Thus, component $x_{aj}$ is 1 if the actor $a$ talks during the $j^{th}$ event and 0 otherwise (the corresponding n-tuples are shown at the bottom of Figure 3.2 in page 26). Since the number of events can be rather high (up to 20 in this work), the dimensionality of the tuples $\mathbf{x}$ representing the interaction patterns is reduced through PCA [Bishop 06]. The amount of variance to be retained after PCA has a minimum set to 70%. The application of PCA to the tuples $\mathbf{x}_a$ results into L-dimensional projections $\mathbf{u}_a$, where $L < D$. Therefore, each news bulletin can be represented through a sequence of L-dimensional tuples $U = (\mathbf{u}_a 1, \ldots, \mathbf{u}_a N)$, where $N$ is the number of turns detected at the speaker diarization step, and $\mathbf{u}_a k$ is the tuple representing the actor $a$ talking at turn $k$.

The goal of the story segmentation is to assign each tuple $\mathbf{u}_i$ a label $h_i$ which corresponds to the number of a story (e.g. *story 2*, or *story 7*). Thus, given the sequence of observations $U$, the story segmentation problem can be formulated as finding the story sequence $H^*$, satisfying the following expression:

$$H^* = \arg\max_{H \in \mathcal{H}} p(U, H)p(H) \tag{4.9}$$

where $H = (h_1, \ldots, h_N)$ is a sequence of stories of length $N$, and $\mathcal{H}$ is the set of all possible story sequences $H$. In our experiments, the joint probability $p(U, H)$ was estimated with a fully connected, ergodic, HMM [Rabiner 89] with $S$ states, where $S$ is the maximum number of stories that can be observed. The emission probability function associated to each state are Gaussians.

The *a-priori* probability $p(H)$ was estimated using a 3-gram statistical language model [Rosenfeld 00]:

$$p(H) = \prod_{k=3}^{N} p(h_k | h_{k-1}, h_{k-2}) \tag{4.10}$$

### 4.2.2 Experiments and Results

The experiments of this work have been performed over a corpus of 27 one hour long news bulletins referred to as C2 previously in this report and in the following. The bulletins are managed by two anchormen that starts and stops the stories by giving the floor to different persons. C2 is fully described in Section 3.3.1.

The story segmentation results are presented in terms of *purity* $\pi$, a performance metric commonly applied in segmentation problems. The purity compares the repartition in time of the stories segments obtained with the automatic approach and the actual (i.e. groundtruth) story segmentation.

Given a recording, consider a groundtruth segmentation $S = \{(s_1, \Delta t_1), \ldots, (s_{N_g}, \Delta t_{N_g})\}$ and an automatic segmentation $S^* = \{(s_1^*, \Delta t_1^*), \ldots, (s_{N_a}^*, \Delta t_{N_a})\}$. The purity $\pi$ is:

$$\pi = \left( \sum_{i=1}^{N_g} \frac{\tau(s_i)}{T} \sum_{j=1}^{N_a} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_i)} \right) \cdot \left( \sum_{j=1}^{N_a} \frac{\tau(s_j^*)}{T} \sum_{i=1}^{N_g} \frac{\tau^2(s_i, s_j^*)}{\tau^2(s_j^*)} \right)$$

where $\tau(s_i, s_j^*)$ is the length of the intersection between the time interval corresponding to segment $s_i$ and the time interval corresponding to segment $s_j^*$, $\tau(s_i)$ is the length of the time interval corresponding to segment $s_i$, $T$ is the total length of the segmented recording. In each parenthesis, the first term is the fraction of recording a segment accounts for, and the second term is a measure of how much a given segment is split into smaller fragments. The terms $\tau(s_i)$ at the numerator and $\tau^2(s_i)$ at the denominator

**Tab. 4.2:** Story segmentation performance. The table reports the purity as a function of the number of segments $D$ used to split the news bulletins and capture the interaction patterns, and of the amount of variance retained after PCA.

| | variance fraction | | | |
|---|---|---|---|---|
| D | 70% | 80% | 90% | 100% |
| 10 | 0.74 | 0.76 | 0.76 | 0.78 |
| 12 | 0.74 | 0.76 | 0.76 | 0.78 |
| 14 | 0.74 | 0.76 | 0.76 | 0.77 |
| 16 | 0.76 | 0.74 | 0.78 | 0.78 |
| 18 | 0.74 | 0.78 | 0.78 | 0.79 |
| 20 | 0.75 | 0.77 | 0.78 | 0.79 |

are left explicit for the sake of clarity. The purity value is bounded between 0 and 1, the closer it is to 1, the better it is the segmentation. When the segmentation is perfect, i.e. $S = S^*$, the value of $\pi$ is 1.

The story segmentation process involves two hyperparameters, the first is the number $D$ of segments used to split the recordings, corresponding to the events during which the interaction patterns are captured. The second is the amount of variance retained after the application of the PCA. The experiments have been performed using $D$ values between 10 and 20, and keeping at least 70% of the variance. Table 4.2 shows the performance for different values of the hyperparameters. The achieved purity is always around 0.75 and no major changes are observed when increasing the number $D$ or the amount of retained variance (at least in the observed ranges). This seems to suggest that the system is stable with respect to the choice of the above parameters.

The results of Table 4.2 have been obtained using a leave-one-out approach (see Section 3.3.2), with all the recordings of the corpus in the training set except one which is used as test set.

On average, the number of stories in the bulletins is 25.2, but the average number of stories detected by the system is 16.5. This means that the most common error consists in grouping different stories rather than in splitting singles stories into smaller segments. The main reason is that the persons involved in different stories, but talking in the same event tend to be represented with similar features, thus tend to be attributed to the same story. This apply in particular to shorter stories (less than two minutes) that often follow each other in some specific moments of the bulletins. Another cause of error is that the anchormen tend to talk about different stories in the same intervention and the corresponding story changes can thus not be detected as the system needs a change of speaker to detect a story change.

Table 4.3 shows the effect of the speaker diarization errors over the story segmentation performance. The reported results are obtained with $D = 14$. The first line of Table 4.3 shows the purity achieved

**Tab. 4.3:** Effect of the speaker diarization errors. The reported results have been obtained using D=14, for both manual and automatic speaker segmentations.

| | variance fraction | | | |
|---|---|---|---|---|
| speak. segm. | 70% | 80% | 90% | 100% |
| manual | 0.80 | 0.80 | 0.80 | 0.82 |
| automatic | 0.74 | 0.76 | 0.76 | 0.77 |

using the groundtruth speaker segmentation, the second line shows the performance achieved using the automatic speaker segmentation (see Section 3.1.1). The differences are rather low and the impact of the diarization errors on the story segmentation performance seems to be negligible.

## 4.3 Semantic Segmentation Discussion

This chapter has shown that automatic role recognition approaches can be used to perform semantic segmentation. In fact, a specific role, i.e. anchorman, enables to split radio programs into its two distinct parts with an average error of transition of 3 minutes.

Moreover, a story segmentation has been performed using the degree of mutual interaction to identify the different stories presented along a radio program. On a specific news bulletins corpora, the achieved story segmentation purity is around 0.75. Such a performance can be considered satisfactory for tasks like fast browsing (where the goal is to quickly reach a point of interest in a long recording), or semi-automatic data editing (where the goal is to manually adjust the automatic segmentation in order to achieve fully correct results).

This chapter finally shows that the structure of social interactions, i.e. roles in this work, can be used to perform semantic segmentation. Existing systems rather use the content of the data such as what is said.

Chapter 5

# Conclusion

This chapter summarizes the work presented in this thesis and states potential future research directions.

## 5.1 Conclusions

In this thesis, an investigation of automatic role recognition has been performed. Research on this problem was in its very early stages when I started my thesis and today is one of the main areas in automatic analysis of social interactions [Vinciarelli 09b]. The presented approaches use Social Network Analysis for representing the individuals in terms of their interactions with others, and Machine Learning approaches for assigning roles to the individuals (Bayes classifiers and probabilistic sequential models). Experiments have been performed over one of the largest data sets ever used in literature for role recognition, including for the first time, to the best of our knowledge, different human-human interaction settings, i.e. production environment contexts and spontaneous exchanges.

## 5.2 Future Research Directions

On the short term, this work can be further developed as follows (the list is not exhaustive):

- It will be more natural to assign roles turn by turn rather than assigning roles to the persons. This has been already implemented in the work done by Hugues and al. [Salamin 10], where they assign a role to each turn with Conditional Random Fields in the case of broadcast data, i.e. C1 and C2. The results show that the role recognition performance is improved. This new approach also allows to remove the varying parameter present in this work, i.e the number $D$ of events used in the Social

Affiliation Network to capture the interaction patterns (see Section 3.1.3). Even if it works well, it will be better to avoid such a brute segmentation of the audio for defining the events.

- As shown in Table 3.12, the two role classifiers presented in this work (Bayes classifier and probabilistic sequential models) tend to make different decisions over the same data. It is thus likely that the combination of the two approaches will lead to significant performance improvements.
- The approaches proposed in this work use only the turn-taking patterns as role evidence (except in the case of combination with lexical features in meetings), while other behavioural cues can be extracted from both audio (e.g. prosody, pitch), and video (e.g. gestures) when available as in the case of AMI meetings.
- The role recognition results over meetings show that informal roles are more difficult to recognize. It could be interesting to implement the proposed approaches over other spontaneous data sets.

The obtained role results also suggest other possible future investigations on the long term. The proposed role recognition approaches assume a groundtruth annotation for the roles and make use of supervised machine learning techniques. However, it would be interesting to apply unsupervised approaches in order to detect characteristic patterns of behavior possibly corresponding to roles. To assess the effectiveness of this approach, we can think of using human assessors to evaluate the clusters, i.e. assess to what extent the people grouped in one cluster are playing the same role.

Moreover, the roles recognized in this thesis have characteristic patterns as they are extracted from production environment as well as during professional meetings. Even if we have defined formal and informal roles corresponding to the different data sets, the studied interaction settings are characterized by role constraints. We could imagine to study new human-human interactions in private contexts, where people interact without any constraints. In fact, it would be interesting to study wether such “social” roles have also characteristic interaction patterns and if they are similar and can be extracted through the same kind of approaches than the roles recognized in this thesis.

# BIBLIOGRAPHY

[Ajmera 02] J. Ajmera, H. Bourlard, I. Lapidot & I. McCowan. *Unknown-multiple speaker clustering using HMM.* In International Conference on Spoken Language Processing, pages 573–576, 2002.

[Ajmera 03] J. Ajmera & C. Wooters. *A Robust Speaker Clustering Algorithm.* In Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding, 2003.

[Ajmera 04] J. Ajmera. *Robust Audio Segmentation.* PhD thesis, École Polytechnique Fédérale de Lausanne (EPFL), 2004.

[Banerjee 04] S. Banerjee & A.I. 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.

[Barzilay 00] R. Barzilay, M. Collins, J. Hirschberg & S. Whittaker. *The rules behind the roles: identifying Speaker roles in radio broadcasts.* In Proceedings of American Association of Artificial Intelligence Symposium, 2000.

[Bickmore 05] T. Bickmore & J. Cassell. *Social Dialogue with Embodied Conversational Agents.* In J. van Kuppevelt, L. Dybkjaer & N. Bernsen, editors, Advances in Natural, Multimodal, Dialogue Systems, pages 23–54. Kluwer, 2005.

[Biddle 79] Bruce J. Biddle. Role theory: expectations, identities, and behaviors. New York Academic Press, 1979.

[Biddle 86] B.J. Biddle. *Recent developments in role theory.* Annual Review of Sociology, vol. 12, pages 67–92, 1986.

[Bishop 95] Christopher M. Bishop. Neural Networks for Pattern Recognition. Oxford University Press, 1995.

[Bishop 06] C.M. Bishop. Pattern recognition and machine learning. Springer Verlag, 2006.

[Bourlard 93] Herve A. Bourlard & Nelson Morgan. Connectionist speech recognition: A hybrid approach. Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[Carletta 07] J. Carletta. *Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus.* Language Ressources and Evaluation, vol. 41, no. 2, pages 181–190, 2007.

[Cassell 04] Justine Cassell. *Towards a Model of Technology and Literacy Development: Story Listening Systems.* Journal of Applied Developmental Psychology, vol. 25, no. 1, pages 75–105, 2004.

[Crowley 06] J. Crowley. *Social Perception.* ACM Queue, vol. 4, no. 6, pages 34–43, 2006.

[Dielmann 07] A. Dielmann & S. Renals. *Automatic meeting segmentation using dynamic Bayesian networks.* IEEE Transactions on Multimedia, vol. 9, no. 1, pages 25–36, 2007.

[Dines 06] J. Dines, J. Vepa & T. Hain. *The segmentation of multi-channel meeting recordings for automatic speech recognition.* In Proceedings of Interspeech, pages 1213–1216, 2006.

[Dong 07] W. Dong, B. Lepri, A. Cappelletti, A. Pentland, F. Pianesi & M. Zancanaro. *Using the Influence Model to Recognize Functional Roles in Meetings.* In Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI), pages 271–278, November 2007.

[Frith 07] C.D Frith & U. Frith. *Social Cognition in Humans.* Current Biology, vol. 17, no. 16, pages 724–732, 2007.

[Gatica-Perez 05] D. Gatica-Perez, I. McCowan, D. Zhang & S. Bengio. *Detecting group interest-level in meetings.* In IEEE International Conference on Acoustics, Speech, and Signal Processing, pages 489–492, 2005.

[Gatica-Perez 09] D. Gatica-Perez. *Automatic Nonverbal Analysis of Social Interaction in Small Groups: a Review.* Image and Vision Computing, vol. To appear, 2009.

[Hain 06] Thomas Hain, Lukas Burget, John Dines, Iain McCowan, Giulia Garau, Martin Karafiat, Mike Lincoln, Darren Moore, Vincent Wan, Roeland Ordelman & Steve Renals. *The development of the AMI system for the transcription of speech in meetings.* In Proceedings of the Conference on Machine Learning for Multimodal Interaction, volume 3869, pages 344–356, 2006.

[Hermansky 90] Hynek Hermansky. *Perceptual linear predictive (PLP) analysis of speech.* The Journal of the Acoustical Society of America, vol. Volume 87, no. Issue 4, pages 1738–1752, 1990.

[Huang 01] X. Huang, A. Acero & H.-W. Hon. Spoken language processing: A guide to theory, algorithm and system development. Prentice Hall, 2001.

[Iacoboni 09] M. Iacoboni. Mirroring people: The science of empathy and how we connect with others. Picador, 2009.

[Jayagopi 08a] D.B. Jayagopi, S. Ba, J.M. Odobez & D. Gatica-Perez. *Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues.* In Proceedings of the International Conference on Multimodal Interfaces, pages 45–52, 2008.

[Jayagopi 08b] D.B. Jayagopi, H. Hung, C. Yeo & D. Gatica-Perez. *Predicting the dominant clique in meetings through fusion of nonverbal cues.* In Proceedings of the ACM International Conference on Multimedia, pages 809–812, 2008.

[Jayagopi 09] D. Jayagopi, H. Hung, C. Yeo & D. Gatica-Perez. *Modeling Dominance in Group Conversations using Non-verbal Activity Cues.* IEEE Transactions on Audio, Speech and Language, to appear, vol. 17, no. 3, pages 501–513, 2009.

[Kirkpatrick 83] S. Kirkpatrick, C. D. Gelatt & M. P. Vecchi. *Optimization by simulated annealing.* Science, vol. 220, pages 671–680, 1983.

[Knapp 72] M.L. Knapp & J.A. Hall. Nonverbal Communication in Human Interaction. Harcourt Brace College Publishers, 1972.

[Kohavi 95] Ron Kohavi. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection.* pages 1137–1143. Morgan Kaufmann, 1995.

[Koumpis 05] K. Koumpis & S. Renals. *Content-Based Access to Spoken Audio.* IEEE Signal Processing Magazine, vol. 22, no. 5, pages 61–69, 2005.

[Kunda 99] Z. Kunda. Social cognition. MIT Press, 1999.

[Laskowski 08] K. Laskowski, M. Ostendorf & 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, pages 148–155, June 2008.

[Lazer 09] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy & M. Van Alstyne. *Computational Social Science.* Science, vol. 323, pages 721–723, 2009.

[Lepri 09] B. Lepri, A. Mani, Alex Sandy Pentland & Pianesi F. *Honest Signals in the Recognition of Functional Relational Roles in Meetings.* In Proceedings of the Association for the Advancement of Artificial Intelligence Symposium(AAAI), pages 31–36, 2009.

[Levine 98] J.M. Levine & R.L. Moreland. *Small groups.* In D. Gilbert & G. Lindzey, editors, The handbook of social psychology, volume 2, pages 415–469. Oxford University Press, 1998.

[Liu 06] Yang 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, pages 81–84, June 2006.

[Massey Jr. 51] F.J. Massey Jr. *The Kolmogorov-Smirnov test for goodness of fit.* Journal of the American Statistical Association, pages 68–78, 1951.

[McCowan 03] I. McCowan, S. Bengio, D. Gatica-Perez, G. Lathoud, F. Monay, D. Moore, P. Wellner & H. Bourlard. *Modelling human interaction in meetings.* In Proceedings of IEEE International Conference on Audio Speech and Signal Processing, 2003.

[McCowan 05a] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma & P. Wellner. *The AMI Meeting Corpus.* In Proceedings of the 5th

International Conference on Methods and Techniques in Behavioral Research, page 4, 2005.

[McCowan 05b] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard & D. Zhang. *Automatic analysis of multimodal group actions in meetings.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pages 305–317, 2005.

[Otsuka 05] K. Otsuka, Y. Takemae & J. Yamato. *A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances.* In Proceedings of ACM International Conference on Multimodal Interfaces, pages 191–198, 2005.

[Pantic 08] M. Pantic, A. Nijholt, A. Pentland & T.S. Huang. *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, pages 168–187, 2008.

[Papoulis 91] A. Papoulis. Probability, random variables, ans stochastic processes. McGraw Hill, 1991.

[Pentland 05] Alex (Sandy) Pentland. *Socially Aware Computation and Communication.* Computer, vol. 38, no. 3, pages 33–40, 2005.

[Pentland 07] A. Pentland. *Social Signal Processing.* IEEE Signal Processing Magazine, vol. 24, no. 4, pages 108–111, 2007.

[Pentland 08] Alex Pentland. Honest Signals, How they shape our world. MIT Press, 2008.

[Pianesi 07] Fabio Pianesi, Massimo Zancanaro, Bruno Lepri & Alessandro Cappelletti. *A multimodal annotated corpus of consensus decision making meetings.* Language Resources and Evaluation, vol. 41, no. 3-4, pages 409–429, 2007.

[Pianesi 08] F. Pianesi, M. Zancanaro, E. Not, C. Leonardi, V. Falcon & B. Lepri. *Multimodal support to group dynamics.* Personal Ubiquitous Computing, vol. 12, no. 3, pages 181–195, 2008.

[Pickles 82] J. Pickles. An introduction to the physiology of hearing. Academic Press, New York, 1982.

[Rabiner 89] Lawrence R. Rabiner. *A tutorial on hidden markov models and selected applications in speech recognition.* In Proc. of the IEEE, volume 77, pages 257–286, 1989.

[Raducanu 09] B. Raducanu, J. Vitrià & D. Gatica-Perez. *You are Fired! Nonverbal Role Analysis in Competitive Meetings.* In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 2009.

[Reiter 07] S. Reiter, B. Schuller & G. Rigoll. *Hidden Conditional Random Fields for Meeting Segmentation.* In Proc. IEEE ICME, pages 639–641, 2007.

[Richmond 95] V.P. Richmond & J.C. McCroskey. Nonverbal behaviors in interpersonal relations. Allyn and Bacon, 1995.

[Rienks 06a] R. Rienks & D. Heylen. *Dominance Detection in Meetings Using Easily Obtainable Features.* In Lecture Notes in Computer Science, volume 3869, pages 76–86. Springer, 2006.

[Rienks 06b] R. Rienks, D. Zhang & D. Gatica-Perez. *Detection and application of influence rankings in small group meetings.* In Proceedings of the International Conference on Multimodal Interfaces, pages 257–264, 2006.

[Rizzolatti 04] G. Rizzolatti & L. Craighero. *The mirror-neuron system.* Annual Reviews of Neuroscience, vol. 27, pages 169–192, 2004.

[Robins 05] B. Robins, K. Dautenhahn, R. te Boekhorst & A. Billard. *Robotic assistants in therapy and education of children with autism: Can a small humanoid robot help encourage social interaction skills?* Access in the Information Society (UAIS), vol. 4, pages 105–120, 2005.

[Rosenfeld 00] R. Rosenfeld. *Two decades of statistical language modeling: where do we go from here?* Proceedings of the IEEE, vol. 88, no. 8, pages 1270–1278, 2000.

[Salamin 10] H. Salamin, G. Mohammadi, K. Truong & A. Vinciarelli. *Automatic Role Recognition Based on Conversational and Prosodic Behaviour.* In Proceedings of the ACM International Conference on Multimedia, pages 847–850, 2010.

[Schapire 00] R.E. Schapire & Y. Singer. *BoosTexter: a boosting-based system for text categorization.* volume 39, pages 135–168, 2000.

[Schuller 07] B. Schuller, R. Müeller, B. Höernler, A. Höethker, H. Konosu & G. Rigoll. *Audiovisual recognition of spontaneous interest within conversations.* In Proceedings of the International Conference on Multimodal Interfaces, pages 30–37, 2007.

[Schuller 09] B. Schuller, R. Müller, F. Eyben, J. Gast, B. Hörnler, M. Wöllmer, G. Rigoll, A. Höthker & H. Konosu. *Being Bored? Recognising Natural Interest by Extensive Audiovisual Integration for Real-Life Application.* Image and Vision Computing, to appear, 2009.

[Tischler 90] H.L. Tischler. Introduction to sociology. Harcourt Brace College Publishers, 1990.

[Tychsen 06] Anders Tychsen. *Role playing games: comparative analysis across two media platforms.* In In proceedings of the 3rd australasian conference on interactive entertainment, pages 75–82, 2006.

[Vinciarelli 07] A. Vinciarelli. *Speakers Role Recognition in Multiparty Audio Recordings Using Social Network Analysis and Duration Distribution Modeling.* IEEE Transactions on Multimedia, vol. 9, no. 6, 2007.

[Vinciarelli 09a] A. Vinciarelli. *Capturing Order in Social Interactions.* IEEE Signal Processing Magazine, vol. 26, no. 5, pages 133–137, 2009.

[Vinciarelli 09b] A. Vinciarelli, M. Pantic & H. Bourlard. *Social Signal Processing: Survey of an emerging domain.* accepted for publication by Image and Vision Computing, vol. 27, no. 12, 2009.

[Waller 08] B. Waller, J. Cray & A. Burrows. *Selection for universal facial emotion.* Emotion, vol. 8, no. 3, page 435, 2008.

[Wang 07] Fei-Yue Carley Kathleen M. Zeng Wang. *Social Computing: From Social Informatics to Social Intelligence.* IEEE Intelligent Systems, vol. 22, no. 2, pages 79–83, 2007.

[Wasserman 94] S. Wasserman & K. Faust. Social network analysis. Cambridge University Press, 1994.

[Weng 09] C.Y. Weng, W.T. Chu & J.L. Wu. *RoleNet: Movie Analysis from the Perspective of Social Networks.* IEEE Transactions on Multimedia, 2009.

[Wrede 03] Britta Wrede & Elizabeth Shriberg. *Spotting "Hot Spots" in Meetings: Human Judgments and Prosodic Cues.* In in Proceedings of Eurospeech, pages 2805–2808, 2003.

[Wrigley 05] S. Wrigley, G. Brown, V. Wan & S. Renals. *Speech and Crosstalk Detection in Multichannel Audio.* IEEE Transactions on Speech and Audio Processing, vol. 13, no. 1, pages 84–91, 2005.

[Zancanaro 06] M. Zancanaro, B. Lepri & F. Pianesi. *Automatic detection of group functional roles in face to face interactions.* In proceedings of International Conference on Mutlimodal Interfaces, pages 47–54, 2006.

[Zhang 06] D. Zhang, D. Gatica-Perez, S. Bengio & I. McCowan. *Modeling individual and group actions in meetings with layered HMMs.* IEEE Transactions on Multimedia, vol. 8, pages 509–520, June 2006.

# Appendix A

# CURRICULUM VITAE

## SARAH FAVRE

| | |
|---|---|
| Le Bleusy | Nationality: Swiss |
| 1997 Haute-Nendaz | Age: 30 years |
| sarah.favre@idiap.ch | Civil status: Married |

## Education

**2010** PhD in Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
Thesis: *Social Network Analysis for Automatic Role Recognition.*

**2005** MSc in Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.
Thesis: *A system for Vertical Jump Evaluation using Accelerometers and Gyroscopes.*

## Professional Experience

**Research Assistant, Idiap Research Institute, Martigny, Switzerland**

**2006 – 2010** Carry out research in the field of role recognition, more generally, social interactions analysis. The aim of my work was to design, implement and test new statistical models inspired by human-human interactions in multi-party recordings for recognizing the roles of the persons interacting. Proposed systems are new to state-of-the art in role recognition and are computationnaly effective.

**2005-2006** **R&D Engineer, Myotest SA, Sion, Switzerland**

Research work on the development of a prototype for a product in the field of sport. It included microprocessor programming, soft development, design of electronic circuits, management of the product's supplies, production's follow-up and after-sales service. This prototype resulted in an invention disclosure and is now available on the market.

**2004** **Research Assistant for Prof. L.K.Von Segesser, Cardiology Service (CHUV), Lausanne, Switzerland**

Design of a LabView interface for cardiac signals' acquisitions of the myocarde muscle during a surgical intervention.

## Skills

**Competencies:** Machine Learning, Pattern Recognition, Social Signal Processing.

**Computer Programming:** C, C++, Matlab, Python, Assembleur.

**Languages:** French (mother tongue), English (good knowledge), German (basics).

## Extra-curricular Activities

**Associations:**

- Delegate for KIDSinfo (A project of the Swiss Association for Women Engineers ASFI).
- Public relations responsible (2003-2005) of ADELE (Association of the Electrical Engineers at EPFL).
- President (2002-2003) of ADELE.

**Sports:** skiing (ski instructor), Volleyball (trainer of Junior B team, Swiss Federal licence).

# Publications

### Journal Papers

∗ 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, November 2009.

### Conference Proceedings

∗ S. Favre, A. Dielmann and A. Vinciarelli, *Automatic Role Recognition in Multiparty Recordings Using Social Networks and Probabilistic Sequential Models*, Proceedings of the 2009 ACM International Conference on Multimedia, pp. 585-588, 2009.

∗ S. Favre, *Social Network Analysis in Multimedia Indexing: Making Sense of People in Multiparty Recordings*, Proceedings of the Doctoral Consortium of the International Conference on Affective Computing & Intelligent Interaction (ACII), pp. 25-32, 2009.

∗ S. Favre, H. Salamin, J. Dines and A. Vinciarelli, *Role Recognition in Multiparty Recordings using Social Affiliation Networks and Discrete Distributions*, International Conference on Multimodal Interfaces (ICMI), pp. 29-36, 2008.

∗ N.P. Garg, S. Favre, H. Salamin, D.H. Tur and A. Vinciarelli, *Role Recognition for Meeting Participants: an Approach based on Lexical Information and Social Network Analysis*, International Conference on Multimedia (ACM), pp. 693-696, 2008.

∗ A. Vinciarelli and S. Favre, *Broadcast News Story Segmentation Using Social Network Analysis and Hidden Markov Models*, International Conference on Multimedia (ACM), pp. 261-264, 2007.

∗ A. Vinciarelli, F. Fernandez and S. Favre, *Semantic Segmentation of Radio News Using Social Network Analysis and Duration Distribution Modeling*, IEEE International Conference on Multimedia and Expo (ICME), pp. 779-782, 2007.

**Technical Reports**

∗ A. Vinciarelli and S. Favre, *Role Recognition in Radio Programs Using Social Affiliation Networks and Mixtures of Discrete Distributions: an Approach Inspired by Social Cognition*, IDIAP Technical Report IDIAP-RR-07-40, 2007.