



University
of Glasgow

Polychroniou, Anna (2014) *The SSPNet-Mobile Corpus: from the detection of non-verbal cues to the inference of social behaviour during mobile phone conversations*. PhD thesis.

<http://theses.gla.ac.uk/5686/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

THE SSPNET-MOBILE CORPUS: FROM
THE DETECTION OF NON-VERBAL CUES
TO THE INFERENCE OF SOCIAL
BEHAVIOUR DURING MOBILE PHONE
CONVERSATIONS.

ANNA POLYCHRONIOU

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
Doctor of Philosophy

SCHOOL OF COMPUTING SCIENCE
COLLEGE OF SCIENCE AND ENGINEERING
UNIVERSITY OF GLASGOW

JULY 2014

© ANNA POLYCHRONIOU

Abstract

Mobile phones are one of the main channels of communication in contemporary society. However, the effect of the mobile phone on both the process of and, also, the non-verbal behaviours used during conversations mediated by this technology, remain poorly understood. This thesis aims to investigate the role of the phone on the negotiation process as well as, the automatic analysis of non-verbal behavioural cues during conversations using mobile telephones, by following the Social Signal Processing approach. The work in this thesis includes the collection of a corpus of 60 mobile phone conversations involving 120 subjects, development of methods for the detection of non-verbal behavioural events (laughter, fillers, speech and silence) and the inference of characteristics influencing social interactions (personality traits and conflict handling style) from speech and movements while using the mobile telephone, as well as the analysis of several factors that influence the outcome of decision-making processes while using mobile phones (gender, age, personality, conflict handling style and caller versus receiver role).

The findings show that it is possible to recognise behavioural events at levels well above chance level, by employing statistical language models, and that personality traits and conflict handling styles can be partially recognised. Among the factors analysed, participant role (caller versus receiver) was the most important in determining the outcome of negotiation processes in the case of disagreement between parties. Finally, the corpus collected for the experiments (the SSPNet-Mobile Corpus) has been used in an international benchmarking campaign and constitutes a valuable resource for future research in Social Signal Processing and more generally in the area of human-human communication.

Acknowledgements

I strongly want to thank my parents for the endless emotional and moral support they offered me during this truly challenging three year effort. This thesis would not have been accomplished without them.

‘Μαμά και Μπαμπά, σας ευχαριστώ πολύ για την ανεκτίμητη ηθική και ψυχολογική υποστήριξη και για την συνεχή συμπαράσταση σας.
Δεν θα κατάφερα να ολοκληρώσω επιτυχώς αυτό το δύσκολο εγχείρημα χωρίς εσάς.’

I feel grateful to my supervisors Dr. Alessandro Vinciarelli and Dr. Rod Murray-Smith for giving me the opportunity to become a researcher.

I would like to express my thankfulness to my examiners Dr. Hayley Hung and Dr. Maurizio Filippone for helping me to improve my research perspectives and skills by pointing out well-aimed comments.

I would like to express my gratitude to my colleague, and my best Swiss friend, Dr. Hugues Salamin for his objective and acute view on critical issues of this research and for offering generously his expertise on high-level statistics and on technical issues occurred. Finally, for keeping a positive, honest and supportive attitude.

Last, but not least, my friend and colleague Rebecca for her help and support.

To myself.

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Thesis Statement and Research Goals	3
1.3	Thesis Structure	5
1.4	List of Publications	8
2	Non-verbal Behaviour	9
2.1	Introduction	9
2.2	How Non-Verbal Behaviour Is Expressed	9
2.3	Methods To Observe Non-Verbal Communication	14
2.4	The Method Of This Thesis	15
3	State-of-the-Art	18
3.1	Reviewing Telephone-Corpora	19
3.2	Reviewing Technology-Mediated Corpora	20
3.3	Reviewing Corpora on prediction of social information.	22
3.3.1	Corpora on Group Interaction	22
3.3.2	Corpora on Various-Type Interaction	24
3.4	Requirements of Corpus Collection	26
3.5	The SSPNet-Mobile Corpus.	28
3.6	Conclusions	29
4	The SSPNet-Mobile Corpus: Acquisition of the Data.	31
4.1	Introduction	31
4.2	Subjects	31
4.3	Experiment	34
4.4	Psychological Tests	37
4.4.1	Conflict Handling Style: the Rahim Conflict Inventory-II (ROCI-II)	38
4.4.2	Personality Traits: The Big Five Inventory-10 (BFI-10)	41
4.5	Recording Human Behaviour	42
4.5.1	Sensors and Signals	45
4.5.2	Synchronisation	47
4.6	Conclusions	48
5	The SSPNet-Mobile Corpus: Annotation of the Data.	50
5.1	Introduction	50

5.2	Annotation model	50
5.2.1	Behavioural Events	51
5.2.2	Topics	52
5.3	Annotated Behavioural Events	52
5.3.1	Speaking Time	52
5.3.2	Laughter	53
5.3.3	Overlapping Speech	54
5.3.4	Back Channel	55
5.3.5	Fillers	56
5.3.6	Silence	56
5.4	Conclusions	57
6	The “Caller-Receiver Effect” on Negotiations Using Mobile Phones.	58
6.1	Introduction	58
6.2	Previous Work	58
6.2.1	Negotiation and Interpersonal Relationships	58
6.2.2	Technology and Behaviour Change	59
6.3	The <i>Caller-Receiver</i> Effect	60
6.3.1	Gender Effects	62
6.3.2	Age Effects	63
6.3.3	Personality Effects	63
6.3.4	Conflict Handling Style Effects	65
6.4	Conclusions	66
7	Personality traits and conflict handling style recognition from audio and motor activation data.	68
7.1	Introduction	68
7.2	The Data	69
7.3	Experiments and Results	69
7.3.1	Speech Features	69
7.3.2	Motor Activation Features	69
7.3.3	Recognition	70
7.3.4	Results	70
7.3.5	Further experimentation	72
7.4	Conclusions	73
8	Automatic detection of Laughter and Fillers.	75
8.1	Introduction	75
8.2	Previous work	75
8.2.1	The Classification Problem	76
8.2.2	The Segmentation Problem	77
8.3	The data	79
8.4	The model	79
8.4.1	Hidden Markov Model	80
8.4.2	Features and Model Parameters	81

8.5	Experiments and Results	81
8.5.1	Performance Measures	82
8.5.2	Detection Results	84
8.5.3	ComParE Interspeech 2013 Challenge	85
8.6	Conclusions	88
9	Conclusions	89
9.1	Introduction	89
9.2	Results and Contributions of the Thesis	89
9.3	Data efficacy and Future Improvements	90
9.4	Future Work	91
9.5	Final Remarks	92
A	Protocol and Scenario	93
B	Consent Form	97
C	The BF-10 Questionnaire	99
D	The Conflict Handling Style Questionnaire	101
	Bibliography	103

List of Tables

3.1	Past corpora collected for the investigation of various aspects of behaviour during meetings of small groups or dyads.	30
4.1	The table reports the number and the percentages of participants per gender, educational background and nationality.	34
4.2	Motion Sensors	46
6.1	Gender effects. The table reports the gender composition of the subjects (“Total” column) as well as the persuasiveness of male and female subjects at both call and item level. According to a two-tailed binomial test, the p -value is higher than 0.42 for all persuasiveness figures.	63
7.1	The values of variance v for PCA, gamma γ that correspond to the model giving the best performances after cross-validation. The costs (regularization term) C for SVM is equal to 0.01 for all cases.	71
7.2	Accuracy for personality traits (upper part) and conflict handling styles (lower part) using speech, phone movements and their combination (S+M). Bold values are higher than the a-priori probability of the most frequent class (second column from the left) to a statistically significant extent (p-value at 5%).	71
8.1	The table reports the details of the main recent works on laughter detection presented in the literature. The following abbreviations are used: A=Audio, V=Video, C=Classification, S=Segmentation, Acc=Accuracy, EER=Equal Error Rate.	78
8.2	HMM performance over the 5-fold setup.	82
8.3	HMM performance over the challenge setup.	82
8.4	Confusion Matrix for the 2-gram language model with $\lambda = 100$. The rows correspond to the ground truth and the columns to the class attributed by the classifier. Each cell is a time in seconds	84

8.5 The table reports the approaches, the extracted features and the performances of each participant in the *Social Signal Sub-Challenge* of the Interspeech 2013 *Computational Paralinguistics Challenge*. The measurements are the Area Under Curve (AUC) for laughter and filler separately and the Unweighted Average of Area Under Curve (UAAUC) of the two classes and are presented in the three columns on the right. The first line corresponds to the results presented by the organisers of the ComParE. The last line presents the results of the approach presented in this thesis. The numbers in the column of Features correspond to certain features 1: Intensity contour, 2: Pitch contour, 3: Timbral contour, 4: Rhythmic patterns, 5: Spectral tilt, 6: Duration, 7: Length of preceding and following pauses. For further explanation of 1 to 4 see [Oh et al., 2013] and for a full description of phonetic features see [Wagner et al., 2013].

List of Figures

1.1	The figure depicts the steps of this research (following the SSP approach) in two parallel levels of the psychological and physical measurements. Both measurements are the input to train models that result in performance measurements, developing automated approaches of social information prediction.	7
2.1	The Thinker, Auguste Rodin, 1881	12
2.2	Non-verbal behavioural cues as head direction, close distance, holding hands etc. constitute social signals of affection, love, concern, happiness etc. Just by looking at the silhouettes of two individuals, strong information about their type of relationship can already be inferred.	13
2.3	The four methods of research on non-verbal communication. The x-axis corresponds to the type of setting (controlled-naturalistic), while the y-axis corresponds to the presence or not of a manipulation. The red circle indicates the point where the setting of SSPNet-Mobile Corpus is located in the figure. It belongs to the upper left quartile, since the setting does not apply manipulation (upper half) and the participants are aware of taking part in an experiment (left half). However, the value in the negative x-axis, which corresponds to the controlled setting, is small, because the setting approximates to a high extent a real-world activity. In contrast, the value of the positive y-axis which corresponds to no manipulation, is high.	16
4.1	The distribution of conversation length across the database of 60 conversations.	32
4.2	The plot shows the age distribution of the 120 subjects. The median is 23.5 years.	33
4.3	The distribution of 120 scores that measure the obliging conflict handling style per participant.	38
4.4	The distribution of 120 scores that measure the avoiding conflict handling style per participant.	39
4.5	The distribution of 120 scores that measure the dominating conflict handling style per participant.	39
4.6	The distribution of 120 scores that measure the intergrating conflict handling style per participant.	40
4.7	The distribution of 120 scores that measure the compromising conflict handling style per participant.	40
4.8	The distribution of 120 scores that measure the extraversion personality trait per participant.	42
4.9	The distribution of 120 scores that measure the agreeableness personality trait per participant.	43

4.10	The distribution of 120 scores that measure the conscientiousness personality trait per participant.	43
4.11	The distribution of 120 scores that measure the neuroticism personality trait per participant.	44
4.12	The distribution of 120 scores that measure the openness personality trait per participant.	44
4.13	The sensing method: the mobile phone N900 and the SHAKE device attached on the phone.	45
4.14	The SHAKE: physical features and axes directions.	46
5.1	The figure depicts the manual annotation model. The two audio streams of each stereo file that correspond to each speaker appear separated. A topic or a turn is determined spatially (see the red vertical lines) at two different tiers and, hence, each topic/turn is assigned to a particular duration and position in the course of the conversation. Then the topic/turn is annotated with a label in order to code the conversational topic/behavioural event that occurs at that moment.	51
6.1	The plot shows the percentage of calls where there was a disagreement on each item.	61
6.2	The percentages of times the “Caller” and “Receiver” impose their opinion, win the conflict and persuade the other. The <i>Receiver</i> is significantly more persuasive than the <i>Caller</i> in both item and call level.	62
6.3	The upper histogram shows the age distribution across the 120 participants, the lower histogram shows the age difference distribution across the 60 calls.	64
6.4	The plot shows the percentage of times the subject with the higher score along a given trait is the most persuasive at the item and call level. The <i>p</i> -value is always above 0.3 (according to a two-tailed binomial test).	65
6.5	The plot shows the percentage of times the subject with the higher score along a given conflict handling style is the most persuasive at the item and call level. The only case where the <i>p</i> -value is below 0.005, according to a two tayed binomial test, is the integrating style at the item level.	66
8.1	The plots show how F_1 Score, Precision and Recall change as a function of the parameter λ , the weight adopted for the Language Model. The four plots at the top have been obtained for five-fold protocol and the four plots at the bottom for the challenge protocol.	83
8.2	The plot show how the function of Precision and Recall for Laughter class, changes with the parameter λ , the weight adopted for the Language Model, ranging in $(0, 200)$	86
8.3	The plot show how the function of Precision and Recall for Filler class, changes with the parameter λ , the weight adopted for the Language Model, ranging in $(0, 200)$	87

Chapter 1

Introduction

1.1 Introduction

Mobile phones have become one of the main means we adopt in order to interact with others. The International Telecommunication Union (ITU) estimates that the number of subscriptions to mobile phone services will correspond to 96% of the earth's population in 2014 [ICT, 2013]. Assuming that every subscription corresponds to a mobile phone possessor, we can conclude that people communicate and socialise massively using “*socially unaware*” devices, i.e. technologies that neglect non-verbal behaviour and the social context accompanying conversations [Pentland, 2005]. A question that arises is whether such a ubiquitous technology-medium in social interactions, can influence social interaction at all, since it is essentially socially unaware?

Socially intelligent technologies, i.e. automatic approaches capable of sensing the social landscape in the same way as people do in their everyday life [Pentland, 2005], is an inspiring concept that demands humane-centred purposes and motivations. A socially intelligent phone could indicate to autistic people social signals during face-to-face or phone conversations. A virtual counsellor [Kang et al., 2012] could offer a more accessible approach to psychological support to a larger proportion of people. These examples demonstrate how the development of social intelligence for agents, under the perspective of facilitating human needs, can provide better life quality.

This thesis is motivated under similar perspectives and shares goals and motivations with the Social Signal Processing (SSP) domain [Vinciarelli et al., 2012b]. SSP aims at developing *socially intelligent* systems. “*Social intelligence is the delivery and comprehension of Social Signals, the signals that inform about an ongoing interaction, or a social relation-ship, an attitude taken or an emotion felt toward another person.*” (p.185) [Poggi and D’Errico, 2011]. Vinciarelli et al. [2009a] explains that Social Signals, such as empathy, affection, agreement, hostility etc., are combinations of *non-verbal behavioural cues* such as facial expressions, gestures, postures, prosody etc. and convey information about affective or cognitive states. SSP aims at developing systems able to detect non-verbal behavioural cues, understand social signals, infer social phenomena and react to them appropriately.

Initially, this thesis examines the role of the phone in social interaction in terms of whether and how it could influence a negotiation. Furthermore, following the SSP approach, this thesis aims to develop systems able to automatically detect vocal non-verbal behavioural cues and to recognise personality traits and conflict handling style in mobile phone conversations

using speech and motion data. Personality is defined by Funder as *individuals' characteristic patterns of thought, emotions and behaviour, together with the psychological mechanisms – hidden or not – behind those patterns* (p.198) Funder [2001] and, therefore, is the key information to collect during a human behavioural experiment. Furthermore, since this work investigates interaction, it employs an eliciting scenario of engaging into a conversation and stating arguments. Therefore, we were anticipating disagreement and conflicts to occur. Section 4.4 presents a further explanation on the considerations taken into account to develop the protocol of the experimental process to collect SSPNet-Mobile Corpus.

To sum up, this thesis focuses on three sequential steps in the detection of cues, prediction and influence of social phenomena. The development of approaches that could construct an artificial ability includes the first two steps, that theoretically approximate the mechanisms of social intelligence. Specifically, the detection of vocal non-verbal behavioural cues support the development of the key ability of social intelligence, e.g., the detection and recognition of different events during speech. In a similar vein, the prediction of social phenomena via physical measurements e.g., audio signals, takes one step further the development of artificial critical thinking about a social interaction. Finally, the determination of factors that may influence the outcome of a social interaction and that are related to the presence of the agent can assist to understand human interaction with agents to a further extent. All three steps may support research on human-computer interaction and, consequently, on socially intelligent agent development.

This thesis aims to investigate the effect of the phone on the conversation and to develop approaches to automatically detect personality (personality traits and conflict handling styles). Hence, it requires the collection of a new corpus of phone calls between strangers, having the same role in their interaction and no specialised knowledge about the conversational topic. This is to investigate phone calls comparable to each other in terms of prior interpersonal relationships, roles and skills. To the best of our knowledge, previous research could not provide data to fulfil the conditions of the above requirements (see Chapter 3 for a detailed state-of-the-art review). Hence, this thesis collects the SSPNet-Mobile Corpus with phone calls between two people without a prior social relationship. That is to ensure that interpersonal relationships will not dominate the discussion, but mainly personality (personality traits, conflict handling styles) will drive the interaction. This, theoretically, allow personality traits and conflict handling styles to be feasible for detection and prediction using automated approaches. Furthermore, two other factors are taken into account to eliminate possible effects on the negotiation outcome: prior knowledge regarding the conversation subject and the role of the interlocutors during the interaction.

To address the two requirements this thesis uses a tailored survival scenario such as the Winter Survival Task (WST) (see Section 4.3), that does not require specialised knowledge, and assigns the two interlocutors to the same role, a rescue team member. Furthermore, the WST is an eliciting-opinion scenario. The latter sets topics of conversation able to engage the two strangers into a conversation by making them cooperate and state their opinion. First, it elicits the personal Boolean opinion (*yes* or *no*) of each individual about a topic. Then, the interlocutors discuss the topic. Conflicting opinions engage them in negotiation, since they always have to reach a consensus at the end.

This thesis aims to collect phone calls with the minimum trade-off in *ecological validity* of the phenomenon, meaning whether the phenomenon occur in the real world as it is [Brewer, 2000](see Section 2.4). The scenario described above is part of the experimental setting, and approximates real-life conditions, since it is typical to discuss on the phone with a stranger

and state an opinion, or negotiate during participation in a survey, or to resolve an issue with public services etc. The participants make or answer a phone call while alone in a room, a typical office. That is a familiar everyday action to the majority of the participants in SSPNet-Mobile Corpus (see Section 4.2) since they are students or academics, people who work or spend some time of their day in an office. Therefore, the experimental process can highly approximate a real-life action and a real-life setting (see Section 2.3).

1.2 Thesis Statement and Research Goals

The above motivations result in the following **thesis statement**:

The goal of this thesis is to improve the social intelligence of mobile phones by investigating their influence on social interaction, and by using the Social Signal Processing approach to investigate to what extent it is possible to detect speech events (e.g., silence, laughter, fillers, etc.) with sensors available on standard mobile phones (microphones and gyroscopes) and infer from them social phenomena (e.g., personality traits, conflict handling styles, etc.).

The research goals of the statement were addressed after collecting a corpus including 120 subjects talking in dyads with smartphones, the SSPNet-Mobile Corpus. The subjects of each dyad are unacquainted and the conversations follow the WST protocol. The following data was collected for each conversation:

- Physical measurements of audio and motor activation data. Using the audio signals we can investigate non-verbal cues such as prosody (pitch, tempo, energy), turn-organisation etc. The motor activation data refers to upper body fidgeting in combination with head motion. The data is captured by two sensors of different modalities: microphones and gyroscopes, both embedded into the phones.
- Psychological measurements of personality traits and conflict handling style. Two psychological questionnaires, one for each phenomenon, was filled by all participants to assess themselves on personality traits and conflict handling style. (see Section 4.4.)

In order to extract non-verbal behavioural cues from the data, two methods were applied:

1. signal processing to automatically extract features from both audio and motor activation data (see Chapter 7 for a detailed description of the extracted features), and
2. manual annotation of the audio signals in terms of behavioural events and conversation topic (see Chapter 5 for detailed explanation of the annotation model).

The corpus was used to address the Research Questions (RQ) presented below. In addition, the results of applying the approaches of this thesis to address the three RQ are reported.

- **RQ1: What are the factors that influence decision making outcomes in the Winter Survival Task?**

To address RQ1 we set the protocol of the experimental process in the way described briefly above and in details in Section 4.3, which can ensure the explicit detection of cases of agreement and disagreement over conversational topics within interlocutors. Due to the nature of the WST, the counterparts have to make a positive (*Yes*) or negative (*No*) decision as to whether specific items are useful or not for the survival of a group of passengers after a plane crash. In the case of disagreement a conflict situation arises, meaning different goals (*Yes*: useful or *No*: not useful) over a conversation topic (every item accounts for a conversation topic). The interlocutors have no prior interpersonal relationship with one another, and are asked to follow the same experimental protocol while one is randomly asked to make the call and the other to answer it.

The results show that factor of the negotiation outcome appears to be relate with the role of *Caller* and *Receiver* of the phone call. The *Receiver* is up to 70% of the times one to persuade their interlocutor. Other factors such as personality traits, handling conflict styles, age, and gender seem to have insignificant influence on the phenomenon.

- **RQ2: Can we infer personality traits and conflict handling styles from audio and motor activation signals?**

To address this, RQ2 features are extracted from the microphone and gyroscope signals and fed to an SVM classifier to assign each subject to a predefined classes of personality traits and conflict handling styles. The classes (two per personality trait and conflict handling style) are defined based on the median of the scores derived from the psychological questionnaires. Figure 1.1 shows the consecutive steps of the method to address RQ2.

The results show the best recognized personality traits (*Neuroticism*) and conflict handling styles (*Dominating* and *Obliging*) were those which were inevitably favoured, at least to some extent, by the WST and the experimental protocol (both described in details in Chapter 4). Sensors available nowadays on any standard smartphone can be used to detect speech features and to measure motor activation and therefore could predict personality traits and conflict handling styles.

- **RQ3: Can we detect laughter, fillers, silence and speech automatically?**

To address RQ3 we apply an approach which uses Hidden Markov Models (HMMs) in order to segment audio clips into four speech events: laughter, filler, silence, speech. The approach takes into account the sequence of the events and estimates the probability of the speech event of interest to be labelled as laughter, filler, silence and speech and includes language models to predict the sequence of the non-verbal events.

The results shows that the application of language models for the automated detection of laughter, filler, silence and speech events during conversations can significantly improve the performance of purely acoustic models.

Taking into account the work on the collection of a new corpus in order to address the RQ presented earlier, and summarizing the results briefly presented above, the novelties of this thesis are:

- the collection of a new corpus with 60 phone calls using mobile phones. The phone calls are similar –and comparable– in terms of conversational topics; however they are real-world calls, since the speakers interact spontaneously. They express personal experience and opinions simulating an everyday action, i.e, talking on the phone while being alone in a typical office.
- the revelation that the use of the phone as caller or receiver is an unexpected influential factor on decision making process. Particularly, the different role effects on the disagreement about conflicting opinions, meaning opposite opinions on a subject. The receiver of the call is significantly more persuasive.
- the exploration of a new approach for the prediction of personality traits through audio and motion cues captured by sensors available on smartphones.
- the experimentation with a new approach for the prediction of laughter, filler, speech and silence using the language model, in addition to the acoustic model.

Finally, addressing the above RQ, this thesis aims to contribute to the overall SSP goals of detecting non-verbal behavioural cues and inferring social phenomena from them.

1.3 Thesis Structure

The rest of the dissertation is structured as follows:

Chapter 2 introduces how non-verbal behaviour is developed and expressed, and defines the psychological terms of non-verbal behaviour, non-verbal behavioural cues, and social signals. Furthermore, it overviews the recommended methods to investigate non-verbal-communication. Finally, the chapter explains the methods that have been used in this thesis.

Chapter 3 reviews literature to justify the need to collect a new corpus in order to analyse social phenomena during phone calls using mobiles. Specifically, it provides a comparison of our research with relative work on human interactions using phones (landlines or not), it reviews technology-mediated interactions towards face-to-face interactions, and the effect of different communication methods on social interaction. Next, it compares corpora on social interactions investigating automated methods of social information prediction regarding the experimental setting. It draws the fundamental requirements to collect a corpus for the investigation of social interactions, in order to develop automated approaches to predict social information. Finally, it describes the SSPNet-Mobile Corpus and the gaps it aims to fill in.

Chapter 4 describes the SSPNet-Mobile Corpus in terms of how it has been collected, the experimental processes that were applied, the background of the subjects, the sensing method that has been applied and the collected data. The corpus has been collected specifically for research on detection of non-verbal behavioural cues and prediction of social phenomena, e.g. personality, during telephone conversations with smartphones. The work of this chapter was presented at the *International Conference on Language Resources and Evaluation, 2014* [Polychroniou, Salamin, and Vinciarelli, 2014].

Chapter 5 explicitly describes the annotation scheme of the audio data coding into behavioural events and conversational topics. The chapter also explains the importance of each behavioural event that has been chosen to be included in the annotation, and indicates the behavioural information that the event may convey. The work of this chapter was presented at the *International Conference on Language Resources and Evaluation, 2014* [Polychroniou, Salamin, and Vinciarelli, 2014].

Chapter 6 reports the research on the influence of the phone on the outcome of a negotiation. Other factors such as personality traits, conflict handling styles and gender are examined. The only factor that seems to influence the negotiation process is the role of the interlocutors in terms of being a caller or receiver while using the phone. This chapter presents work published in the *Journal of Cognitive Computation* [Vinciarelli, Salamin, and Polychroniou, 2014].

Chapter 7 presents the results on prediction of personality traits and conflict handling styles. Features are extracted from audio and motor activation data from 120 subjects. The SSPNet-Mobile Corpus provides the data and the manual annotation in speech events of the audio signals. An SVM classifier is fed with the features in order to predict personality traits and conflict handling style. This chapter presents work presented at the *International Workshop on Image and Audio Analysis for Multimedia Interactive Services, 2013* [Salamin, Polychroniou, and Vinciarelli, 2013b].

Chapter 8 reports on the detection of laughter and fillers in audio data and the work on detection of four different speech events: laughter, fillers, silence and speech in audio clips, using data from SSPNet-Mobile Corpus. This chapter presents work presented at *Interspeech, 2013* [Björn et al., 2013] and *International Conference on Systems, Man and Cybernetics, 2013* [Salamin, Polychroniou, and Vinciarelli, 2013a].

Chapter 9 summarises the results and conclusions of this dissertation and suggests ideas for future work.

Appendix A contains the scenario, the Winter Survival Task, and the protocol of the SSPNet-Mobile Corpus experimental task.

Appendix B contains the consent form for participation in the experiment of SSPNet-Mobile Corpus.

Appendix C includes the Big Five Inventory - 10, the questionnaire that measures the personality traits.

Appendix D contains the of the Rahim Conflict Inventory-II, the questionnaire that measures the conflict handling styles.

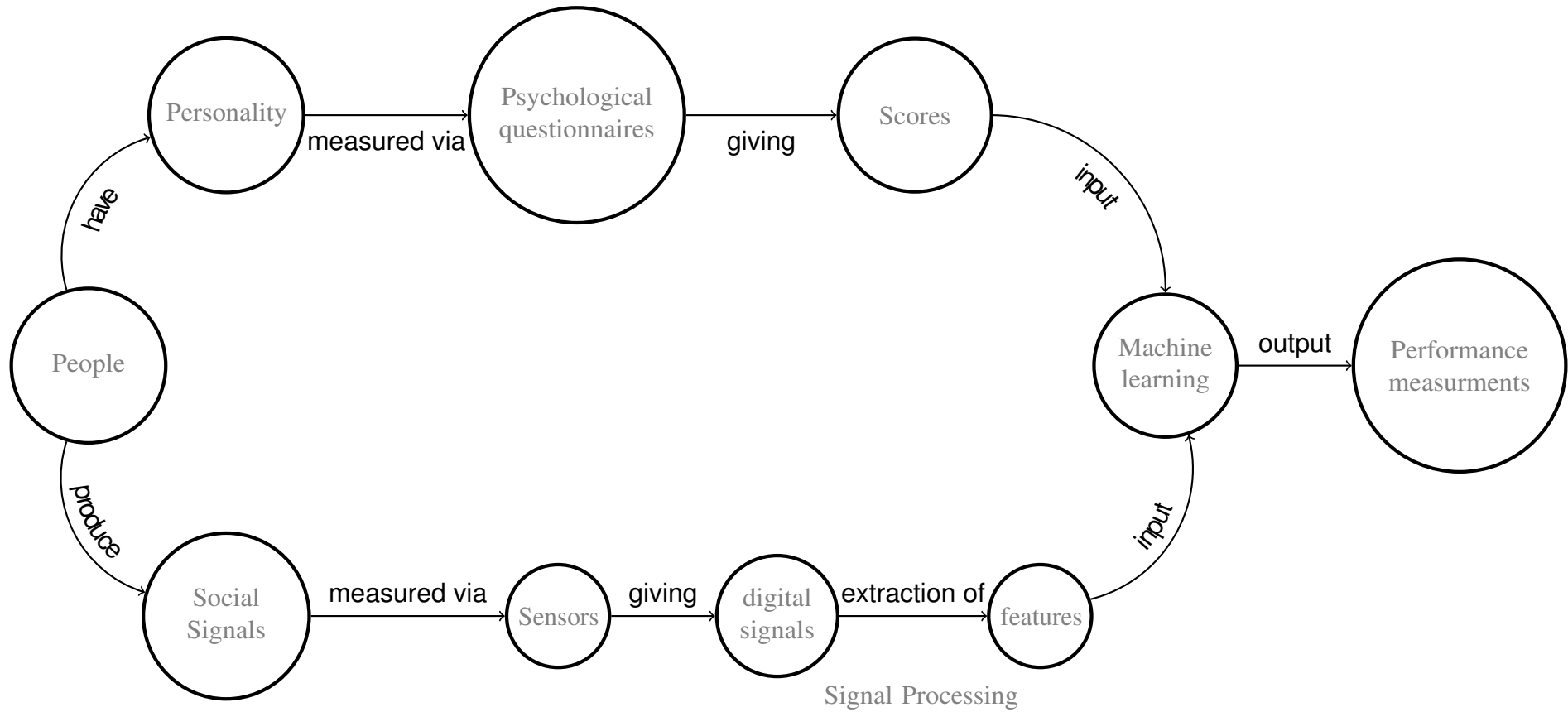


Figure 1.1: The figure depicts the steps of this research (following the SSP approach) in two parallel levels of the psychological and physical measurements. Both measurements are the input to train models that result in performance measurements, developing automated approaches of social information prediction.

1.4 List of Publications

The work of this thesis was published in the following papers:

- A. Polychroniou, H. Salamin and A. Vinciarelli, “**The SSPNet-Mobile Corpus: Social Signal Processing over mobile phones**”, Proceedings of the Language Resources and Evaluation Conference, to be presented, 2014. This work is included in Chapters 4 and 5.
- A. Vinciarelli, H. Salamin and A. Polychroniou, “**Negotiating over Mobile Phones: Calling or Being Called Can Make the Difference**”, Cognitive Computation, Vol. 6, no. 1, pp. 1-12, 2014. This work is included in Chapter 6.
- H. Salamin, A. Polychroniou and A. Vinciarelli, “**Automatic Recognition Of Personality And Conflict Handling Style in Mobile Phone Conversations**”, Proceedings of International Workshop on Image and Audio Analysis for Multimedia Interactive Services, pp. 1-4, 2013. This work is included in Chapter 7.
- H. Salamin, A. Polychroniou and A. Vinciarelli, “**Automatic Detection of Laughter and Fillers in Spontaneous Mobile Phone Conversations**”, Proceedings of IEEE International Conference On Systems, Man and Cybernetics, pp. 4282-4287, 2013. This work is included in Chapter 8.
- B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente and S. Kim, “**The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism**”, Proceedings of Interspeech, 2013. This work is included in Chapter 8.

Related Publications:

- F. Bonin, C. De Looze, S. Ghosh, E. Gilmartin, C. Vogel, A. Polychroniou, H. Salamin, A. Vinciarelli and N. Campbell, “**Investigating Fine Temporal Dynamics of Prosodic and Lexical Accommodation**”, Proceedings of Interspeech, 2013.
- M. Campo, A. Polychroniou, H. Salamin, M. Filippone and A. Vinciarelli, “**Towards Causal Modeling of Human Behavior**”, in “*Neural Nets and Surroundings*”, B. Apolloni, S. Bassis, A. Esposito, F.C. Morabito (eds.), Springer Verlag, pp. 337-344, 2013.
- A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi and A. Origlia, “**From Non-verbal Cues to Perception: Personality and Social Attractiveness**”, in “*Cognitive Behavioral System*”, A.M. Esposito, A. Vinciarelli, R. Hoffmann, V.C. Mueller (eds.), Lecture Notes in Computer Science, Vol. 7403, Springer Verlag, pp. 60-72, 2012.

Chapter 2

Non-verbal Behaviour

2.1 Introduction

This thesis investigates non-verbal behaviour during telephone conversations. Therefore, Section 2.2 describes non-verbal behaviour in terms of how it is adopted and expressed. Section 2.3 briefly reports on the methodology that is usually adopted to investigate non-verbal communication and, finally, Section 2.4 explains the method that this thesis applied to collect data.

2.2 How Non-Verbal Behaviour Is Expressed

According to Kendon [2013], as soon as two humans (who are co-present) recognise one another's presence, they engage in mutual adaptation. The engagement can be accomplished in two ways: verbally and non-verbally. Words and speech can convey an explicit opinion, information, desires that can be described only verbally. On the other hand, non-verbal behavioural cues and social signals (see below), in association with speech, might be less explicit, but according to Vinciarelli et al. [2009a] convey information about affective/attitudinal/cognitive states.

Verbal and non-verbal behaviour typically reinforce one another. For example Anna says to Maria : "I am happy" and Maria understands that Anna is happy, but also indirectly through vocalisations or prosodic features, e.g. if Anna is giggling and/or her voice sounds "happy", then Maria can infer that Anna is actually happy. Therefore, while people talk, they simultaneously communicate both linguistic information and indications about how they feel through speech rate, pauses, loudness etc. Non-verbal behaviour, in the cases where words and non-verbal cues disagree with each other, proved to be more honest [Pentland, 2008]. In other words, non-verbal behaviour can leak information that words try to hide.

According to Ekman and Friesen [1981] three factors must be taken into account to ensure integral perception of the meaning that non-verbal behaviour conveys:

- the *usage* refers to the conditions under which a non-verbal act usually occurs.
- the *origin* refers to the reason and the way an act becomes part of a person's non-verbal

behaviour.

- the *coding* refers to the building up of the correspondence between the act and its meaning.

Ekman and Friesen [1981] explains how non-verbal behaviour is deployed and expressed physically. This work studies Ekman and Friesen [1981] work in order to understand and describe how non-verbal behaviour is displayed facially and, mainly, bodily in order to predict personality. It is out of the scope of this thesis to adopt one of the conflicting theories of psychology on whether behaviour is the cause or causality regarding emotions, although, to the best of our knowledge, Baumeister et al. [2007] reviews recent work on how emotions cause behaviour directly or not, and discusses the cognitive factor that may interject.

Ekman and Friesen [1981] has classified non-verbal behaviour counting the usage, origin and coding to present five categories of non-verbal behaviour:

- **emblems** are gestures with specific (predefined) meaning, that convey less personal information but a specific meaning from the sender to the receiver intentionally. Emblems are not always universal, and the conveyed message can be different or not understood across different cultures. An example is the open palm gesture in Greece, which is a rude gesture to call someone an idiot, an emblem not understood in other cultures. On the contrary, thumb up means a positive outcome or agreement, an emblem coming from Ancient Rome, but widely diffused across multiple cultures nowadays.
- **illustrators** “*are movements directly tied to speech*” (p. 68) [Ekman and Friesen, 1981] supporting the linguistic part. There are six different types of illustrators that “*can repeat, substitute, contradict or augment the information provided verbally*” (p. 69) [Ekman and Friesen, 1981]. Illustrators usually are less intentional than emblems and convey messages without a conscious attempt to do so. An example is the raising of vocal volume when someone wants to highlight something in speech.
- **affect displays** are representations of primary affect states, universal to mankind such as happiness, surprise, fear, sadness, anger, disgust and interest with facial expressions, and subsequently, through body posture and movements that “*can repeat, qualify or contradict a verbally stated affect, or be a separate, unrelated channel of communication*” (p.77) [Ekman and Friesen, 1981]. Facial expressions directly affect displays and the rest of the body conveys the behavioural consequences, for more discussion on the relationship of body movement to affect see [Ekman and Friesen, 1967]. Affect displays carry more personal information than emblems or illustrators.
- **regulators** are the acts with which conversations can be managed. They regulate the “*back and forth nature of speaking and listening*” (p.82) [Ekman and Friesen, 1981] and are not intentional but intuitively adopted. Head nods, non-linguistic sounds like “mmm” (called back channel) ensure attention and engagement in the conversation.
- **adaptors** are movements coming from actions initially performed in order “*to satisfy bodily needs,[..] to develop prototypic interpersonal bonds [..] learned usually in childhood*” (p.84) [Ekman and Friesen, 1981]. A movement, part of the original action, is adapted unconsciously to express similar emotional states to the initial state, i.e., that one during which the movement was associated for the first time to the specific

emotional state. Hence, when different situations cause similar states to the initial one, the corresponding movement (or part of it) is displayed through adaptors. The latter are emitted unintentionally, for example lip bites and eye closing during speaking. This type of non-verbal behaviour is more personalized than the others, meaning that it convey more person-specific information.

Moreover, non-verbal behaviour is classified based on the physical medium of expression. In particular, non-verbal behaviour is expressed through [Vinciarelli et al., 2008]:

- facial expressions and gaze behaviour
- speech
- gestures and postures
- physical appearance
- distance from others, and location in space

Figure 2.1 depicts “The Thinker”, one of the most famous sculptures of Auguste Rodin. The statue represents the power of thought and the capacity of creativity. Rodin explained to a journalist that “*he conceived of the Thinker to be [...] , a naked man, seated upon a rock, his feet drawn under him, his fist against his teeth, he dreams. The fertile thought slowly elaborates itself within his brain. He is no longer dreamer, he is creator.*” (p.394) [Alhadeff, 1966].

The direction of the gaze, the arched back, the fist, the muscles, and the seated position of “The Thinker” are non-verbal behavioural cues (NvBC). NvBC can be described as “*sets of temporal changes in neuromuscular and physiological activity that last for short intervals of time (milliseconds to minutes) [...]*” (p.1062) [Vinciarelli et al., 2008], a blink, a nod, a smile, a sigh, a pause, a change in speaking rate. A cluster of combined NvBC constitutes a social signal. Vinciarelli et al. [2009a] explain that “*social signals and social behaviours are the expression of one’s attitude towards social situation and interplay, and they are manifested through a multiplicity of non-verbal behavioural cues including facial expressions, body postures and gestures, and vocal outbursts like laughter.*” (p.2).

Social signals are classified in Poggi and D’Errico [2011] based on presence or absence of intention to convey information. Specifically, *communicative* signals are intentionally sent from the sender to the receiver in order to provide information. However, *informative* signals can be unintentional in terms of a lack of the sender’s intention to convey a message. The same signal can be communicative or informative at the same time, depending on the receiver. For example, the couple in Figure 2.2 are holding hands; the two individuals are walking very closely to each other, they look at one another directly, their body orientation is similar, their direction is common and their gait appears to be synchronised. The corresponding social signals are affection, love, happiness, intimacy, showing that the two individuals are a couple. The signals are communicative to one another, but to random observers the same signals could be informative.

In the specific case of vocal non-verbal behaviour, i.e. everything that is vocally produced besides words, behavioural information can be conveyed through:

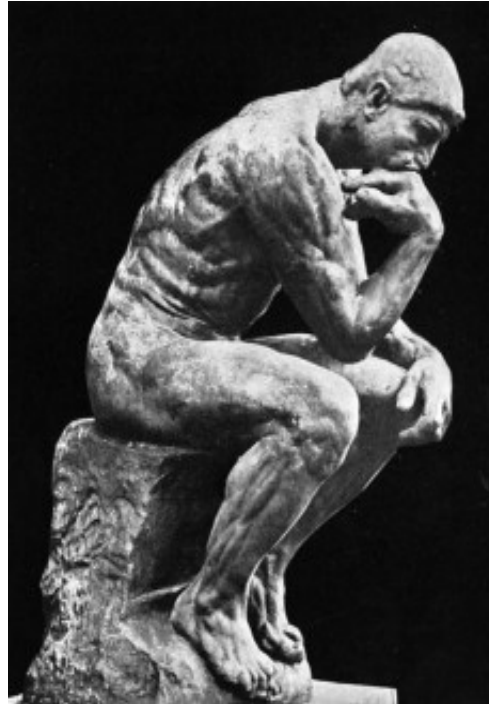


Figure 2.1: The Thinker, Auguste Rodin, 1881

- *Prosody*: the way something is said, which consists of:
 - pitch: the fundamental frequency of the voice
 - tempo: the speaking rate
 - energy: the intensity of voice
- *Turn-taking patterns*: the patterns describe the way *turns* are organised, namely, who speaks for how long, who follows, how many speakers speak at the same time (overlapping speech) etc. The key term is the *turn*, and is defined as the time when only one person speaks. The organisation of the turn-taking patterns could indicate coordination, mutual attention or familiarity.
- *Linguistic vocalisations*: are sounds like “uhm” which substitute for actual words due to uncertainty or indecisiveness.
- *Non-linguistic vocalisations*: are vocal outbursts such as laughter, crying, sighing, and mainly express affective states such as joy, sadness, melancholy or an attitudinal state, e.g., laughter at a speaker can express a listener’s scorn.

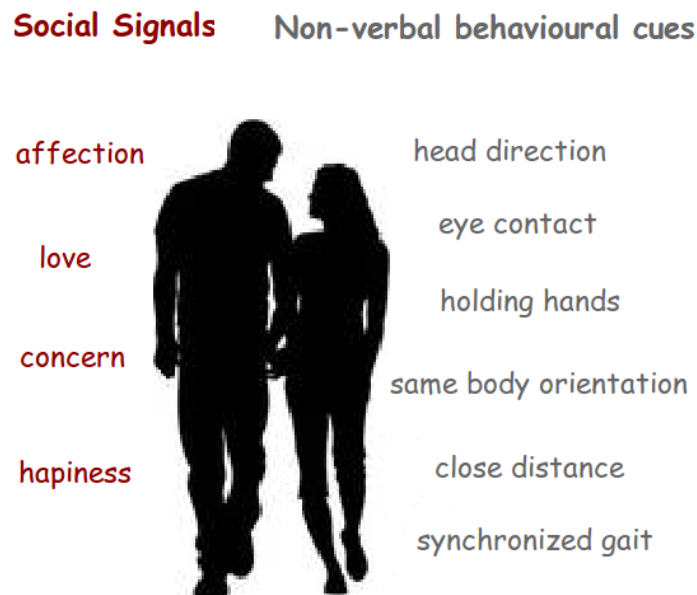


Figure 2.2: Non-verbal behavioural cues as head direction, close distance, holding hands etc. constitute social signals of affection, love, concern, happiness etc. Just by looking at the silhouettes of two individuals, strong information about their type of relationship can already be inferred.

- *Silence*: when no sound is produced. This can convey details about the interaction such as hesitation, etc.

According to Poggi and D’Errico [2011], “*social intelligence is a set of skills that include understanding of other people’s feelings, seeing things from their point of view and giving them effective responses*” (p.185) and designates a substantial fraction of human intelligence. Rephrasing for the needs of this work’s point of view, one could describe *social intelligence* as the ability to perceive, recognise and express social signals. According to Pentland [2005] social signals “*are particularly powerful for analysing and predicting human behaviour*” (p.33) and, furthermore, social outcomes on major aspects of life, such as marital and professional success. Hence, social intelligence can critically affect an individual’s quality of life.

2.3 Methods To Observe Non-Verbal Communication

It is important to underline the methods used to conduct research on non-verbal behaviour due to the fact that the way in which behaviour is observed might have an influence on it. In other words, intrusive environments can disturb the *ecological validity* of phenomena. The term ecological validity refers to whether the phenomenon occurs in the real world as it is [Brewer, 2000]. That includes the methods and the setting of the experimental process. Hence, the preservation of ecological validity of a study ensures that the experimental process approximates real-world situations. Therefore, the setting must be chosen in order to address research questions, and at the same time preserve the realism of the phenomenon of interest. According to Guerrero et al. [1999] there are four different methods to investigate non-verbal behaviour. The classification is based on the combination of two factors:

- **setting**: in this context, it refers to the environment in which the experiment is taking place, and it can be either **naturalistic** or **controlled**.

The naturalistic type includes settings that are part of the “*real world*”. Therefore, this type of setting corresponds to observations that have taken place outside of laboratories, like public places or cafés etc. [Guerrero et al., 1999].

In controlled settings the researcher manipulates the conditions in order serve the needs of his research questions. Usually, this type of setting takes place in laboratories [Guerrero et al., 1999].

- **manipulation**: refers to whether a researcher needs to keep the **same conditions** across all experimental sessions **or not**. A change of one or more conditions of the experiment might cause a different behaviour or outcome. Hence, by comparing the two situations, a potential influence (of conditions on the outcome) may be detected [Guerrero et al., 1999].

Therefore, the combination of naturalistic or controlled settings with manipulation or not, leads to four different methods of research on non-verbal behaviour [Guerrero et al., 1999]:

- **Laboratory Experiments**: these are experiments conducted in a controlled setting with manipulation. This method provides the researcher with the opportunity to focus on the behaviour of interest. The researcher can organise the setting in order to observe the behaviour they want to investigate. The advantage is that behaviours that are difficult to observe in naturalistic settings can be captured in laboratory experiments. The disadvantage of this method, and in general of the methods that include controlled settings, is that they are typically non realistic [Guerrero et al., 1999].
- **Controlled Observations**: in this case the setting is controlled and no manipulation is applied. The researchers select the activity or interaction of the participants, depending on the behaviour of interest, but, to avoid manipulation, they always preserve exactly the same conditions of the experimental process across all sessions. This includes, for example, the different roles that participants might have to play or the arrangement of the space [Guerrero et al., 1999].
- **Field Experiments**: in this type of experiment there is manipulation in a naturalistic setting. In this case, there is the possibility that the behaviour of interest might not be

adequately observed. On the other hand, the naturalistic settings are realistic, unlike those considered above [Guerrero et al., 1999].

- **Naturalistic Observation:** this method encompasses observations that have been conducted in naturalistic settings without manipulation. The advantage is that the observed behaviour can be described in depth, but the disadvantage is that the cause might not be inferred [Guerrero et al., 1999].

There is a trade-off between observing realistic behaviour and having the control (in terms of observing the behaviour of our interest). Naturalistic settings might provide realism, but it is difficult to observe a specific behaviour. It is also very difficult to acquire details about the people who have been observed. Finally, there are legal constraints that limit the extent to which human interaction can be recorded without the consent of the subjects [Guerrero et al., 1999].

2.4 The Method Of This Thesis

In this thesis' case, the goal is to investigate phone calls regarding the role of the phone in social interaction and regarding detection of non-verbal behavioural cues expressed, in order to create an automatic approach of behaviour (personality traits and conflict handling styles) understanding. As mentioned in Section 1.1, this thesis shares common interests and methods with the Social Signal Processing domain. The SSP methods, such as signal processing, presuppose recordings of physical activity and speech, and use sensors to record them. However, SSP and –especially– this work aim to collect and use real-world data, such as the Canal9 Corpus [Vinciarelli et al., 2009b], SSPNet Conflict Corpus [Kim et al., 2012] and SSPNet Speaker Personality Corpus [Mohammadi and Vinciarelli, 2012].

However, this thesis chose not to use real-world phone calls besides the fact that the collection of this data raises privacy issues. To investigate the possible influence of phones on social interaction we needed phone calls comparable to one another. Also, due to the fact that we wanted to predict behaviour (personality traits and conflict handling styles) we needed to have phone calls where behaviour would dominate the interaction instead of prior interpersonal relationships, different roles assigned or specialised knowledge about the conversational topic. Using real-world phones calls we probably could not have preserved all these conditions. Therefore, the need occurred for a new corpus collection. To preserve the conditions described we constructed WST the scenario which is fully described in Section 4.3.

This thesis introduced a method to collect social interaction of investigation (phone calls using mobile phone) by setting as the recording method the medium of the social interaction of investigation, the mobile phone. Inevitably, the use of a mobile device as recording method for the purposes of research on human behaviour during telephone conversations reduces the level of the *Controlled* setting. The use of mobile phones, concerning the level of intrusiveness this technology brings into the experimental setting, is estimated to be from minimal to zero. Users are highly familiar with the most ubiquitous and adapted human-computer interaction technology today [Vinciarelli et al., 2012b]. Furthermore, the spatial setting includes a typical office instead of a laboratory, and leaves the participant alone in the room (see Section 4.3 for

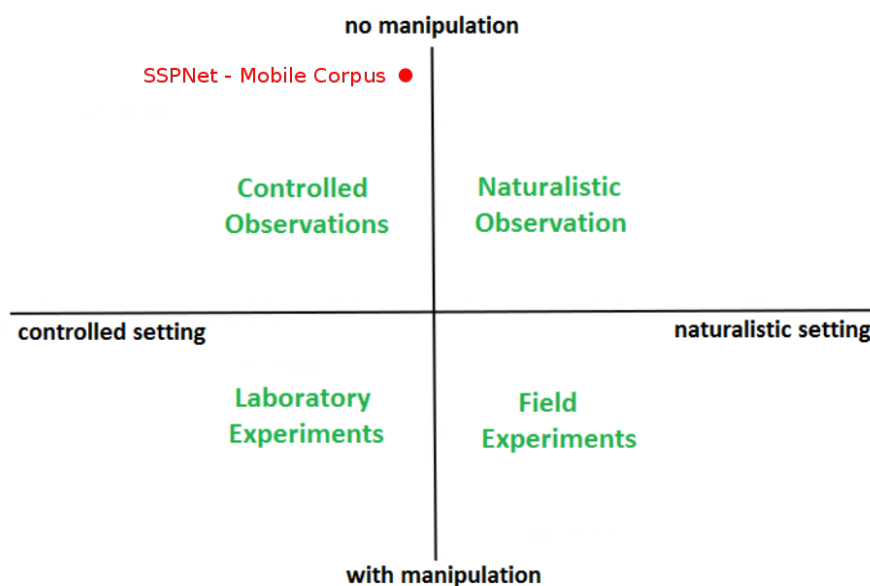


Figure 2.3: The four methods of research on non-verbal communication. The x-axis corresponds to the type of setting (controlled-naturalistic), while the y-axis corresponds to the presence or not of a manipulation. The red circle indicates the point where the setting of SSPNet-Mobile Corpus is located in the figure. It belongs to the upper left quartile, since the setting does not apply manipulation (upper half) and the participants are aware of taking part in an experiment (left half). However, the value in the negative x-axis, which corresponds to the controlled setting, is small, because the setting approximates to a high extent a real-world activity. In contrast, the value of the positive y-axis which corresponds to no manipulation, is high.

a full description of the setting). The experimental scenario leads the participants to talk on the phone, an everyday activity, in order to take part in a conversation with a stranger (another participant) (see Section 4.3 for a full description of the scenario). This scenario could be a real-life situation, since people very often talk to strangers on the phone to resolve an issue, to be offered a service, or to take part in a survey. Phone calls like this are usually recorded and people are aware of that condition.

Hence, the setting and the scenario this work uses, could be a real-life situation and, therefore, highly approximates a *Naturalistic* setting, meaning an actual phone call. At this point we could comment that the *Controlled* setting of this work is, perhaps, as unobtrusive as it could be for social interaction using mobile phones, in order to collect vocal and motor activation data.

Consequently, we conclude that our method includes a *Controlled* setting, since the participants are aware that they are participating in research, *without manipulation*, meaning a *Controlled Observation* which highly approximates a *Naturalistic Observation* (see p.14, Section 2.3).

Figure 2.3 depicts the four most applicable methods for non-verbal communication research, in orthogonal axes of *setting type* (x-axis) and *manipulation* (y-axis). Hence, the mark of this thesis experimental setting belongs to the upper left quartile, close to positive y-axis and far from the negative x-axis.

Chapter 3 reviews related work on compiling corpora that support human interaction investigation. Our work will be compared to various experimental settings and will be compared to other corpora collected to support SSP research.

Chapter 3

State-of-the-Art

This chapter aims at justifying the need to collect a new corpus for the analysis of social phenomena during phone calls with mobile phones and, also, of effects on social interaction caused by the use of the phone per se. As mentioned, this thesis sets three specific conditions for data collection: no prior interpersonal ties, equal roles and conversational topics that do not require expertise. The rest of the chapter indicates the lack of corpora providing these conditions.

Initially, it reviews corpora that include telephone calls, and indicate why could not be used for the purposes of this work. Reviewing work on the effect of the phone on social interaction, previous work on the comparison of face-to-face interactions towards telephone interactions and audio-only technology mediated interactions regarding how social interaction is presented. After that, the chapter places emphasis on previous work on face-to-face negotiation processes since negotiation is the phenomenon affected by phone use presented in this thesis (see Chapter 6). Next, the chapter reviews previous work on the automated prediction of social information during interactions. A literature review with reference to corpora collected to support research on meeting, debates, etc., so as to develop automated approaches for the prediction of emerging social phenomena such as roles, leadership, emotions etc. To the best of our knowledge, comparing SSPNet-Mobile to these corpora, the novelty of our corpus focuses not only on the new type of interaction (phone calls using smartphones) that it is under investigation, but, mainly, on the new experimental setting which is set for the collection of the corpus. The setting highly approximates a naturalistic setting where the phone –and its role which are under investigation– is set as only the sensing method. That eliminates effects that other technology used as sensing methods could bring into the interaction, and at the same time provides an experimental setting that approximates a naturalistic setting, since the mobile phone is one of the most ubiquitous technologies today. Hence, the comparison between corpora on the prediction of social information and SSPNet-Mobile Corpus focuses on the experimental setting.

The rest of the chapter is organised as follows: Section 3.1 provides a comparison of our research with relative work on human interactions using phones (landlines or not), Section 3.2 reviews technology-mediated interactions towards face-to-face interactions, and the effect of different communication methods on social interaction, Section 3.3 compares corpora on social interactions investigating automated methods of social information prediction, and focuses on the experimental setting. Section 3.4 draws the fundamental requirements to collect a corpus for the investigation of social interactions, in order to develop automated approaches to predict social information. Finally, Section 3.5 describes the SSPNet-Mobile

Corpus and the gaps it aims to fill in.

3.1 Reviewing Telephone-Corpora

Corpora on telephone calls have been collected in the past such as the Switchboard Corpus [Godfrey et al., 1992] an extended corpus with actual phone calls of 475 participants, without details about the relationship between the interlocutors, talking about random topics. The CSLU Telephone Speech Corpus [Cole et al., 1994] included several shorter corpora with calls in twenty different languages. The calls were through cellular phones or landlines, and involved questions about personal information (name etc.), census information, to tell a story about them-selves or just to recite the alphabet. Hence, in the data there was an inadequate amount of dialogue and dyad interaction, but more answering questions or monologues.

Another corpus that included phone calls is the Spoken Dutch Corpus (CGN) [Oostdijk, 2000]. Specifically, that corpus was collected to support computational linguistics, language and speech technology. It involved dialogues or multilogues: face-to-face and telephone conversations, interviews, business transactions, discussions, debates, meetings, lectures and monologues (descriptions of pictures, spontaneous commentary, news reports, current affairs, programs, news, commentary, lectures, speeches, read aloud text). The recordings were been accomplished by a variety of different ways.

The three corpora offer rich data on telephone conversations. However, they do not provide information about the type or level of prior interpersonal relationship between the interlocutor, or psychological information about the speakers, such as personality traits etc. Finally, the data might omit decision-making processes, and the absence of scenarios or conversational topics might lead to conversations with massive differences due to different roles or skills regarding the conversational topic, making them incomparable, at least for the purposes of this work. Hence, the conditions that the scenario of the SSPNet-Mobile Corpus preserve, such as no prior social ties, no expertise required, same roles (see 2.4), could not be fulfilled by those or similar corpora. Also, the phones used for the phone calls were not only mobile phones but various devices, such as landlines.

Another interesting work was presented by Miritello et al. [2013], who used a database consisting of over 20 million users and 9 billion calls to investigate the amount of time people dedicated to each of their social ties. For the purposes of this work, only the source and the destination identification numbers, and the duration of calls were collected. However, this corpus not only lacks behavioural information, but also does not include the actual conversation due to privacy issues.

To the best of our knowledge, recent literature has investigated human-human interaction using mobile phones collecting data regarding location, social network and reality mining purposes [Eagle, 2005; Eagle and Pentland, 2005, 2006]. However, to the best of our knowledge, limited research has been done on the investigation of human-human interaction via telephones in terms of how the use of mobile phone influences the social interaction per se. The next section reviews previous work that compares phone and face-to-face interaction to investigate the effects on the social outcome.

3.2 Reviewing Technology-Mediated Corpora

To the best of our knowledge, research on telephone versus face-to-face communication has been mainly investigated to the extent of survey completion, or for medical needs, with less regard paid to the effect of the different communication methods (face-to-face, phone mediated) on the behaviour expressed during the interaction.

Numerous studies have attempted to explain the differences between using phones and face-to-face interactions, regarding the effect of the different interaction modes on surveys' completion about drug use or sexual matters, etc. [Aquilino, 1991; Bajos et al., 1992; Hox and De Leeuw, 1994; Nebot et al., 1994; Galán et al., 2004; Herzog and Rodgers, 1988; Holbrook et al., 2003] reports of alcohol consumption Greenfield et al. [2000], or data collection methods comparing mail, telephone and face-to-face interaction [De Leeuw, 1992; De Leeuw et al., 1996]. In a similar vein, several studies have been made to evaluate medical attention, counselling or dietary success via telephone communication instead of face-to-face interaction [Fox et al., 1992; Weinberger et al., 1994; Fenig et al., 1993; Palmer et al., 2002; Brustad et al., 2003].

To the best of our knowledge, Ten Bosch et al. [2004], compares spontaneous face-to-face and telephone dialogues in terms of turn duration using the Spoken Dutch Corpus. Ten Bosch et al. [2004] have developed a study on turn durations of face-to-face communication and telephone dialogues. This study has shown that the pauses in telephone conversations are shorter, and proposes two possible explanations. The first is the lack of facial cues perception that indicate, e.g. a thinking state to the speaker, that is not provided during a telephone conversation. The second explanation is that usually the task of a telephone interaction is simpler than that of a face-to-face interaction, where multiple activities may emerge, and cause delays between responses. Fowler and Wackerbarth [1980] argue that audio-only communication may be a disadvantage regarding visual non-verbal feedback, and lead to *depersonalised discussions* focussed on the conversational task solution. Bavelas et al. [2008] have investigated the gesturing rate in both cases, and conclude that gesturing differs in rate and in mode during face-to-face and telephone dialogues.

Audio-only computer-mediated communication is compared to face-to-face interactions in this paragraph. To the best of our knowledge, the limited literature has investigated the differences between audio-only computer-mediated towards face-to-face dyadic interactions from the behavioural perspective. During recent years, computers and the internet have been adopted in all types of human communication (formal, informal, personal, professional etc.). There has been an increasing amount of literature on computer-mediated communication [Walther, 1996; Rice and Love, 1987; Joinson, 2001; Hiltz, 1986; Herring, 1995; Kiesler et al., 1984] and comparisons with face-to-face communication [Bordia, 1997; Flaherty et al., 1998; Frohlich and Oppenheimer, 1998; Olaniran, 1994]. Where computer-mediated communication was investigated, poor comparisons were accomplished regarding audio-only communication, and neither visual or other types of communication using computers. In the case of audio-only computer-mediated research, Fowler and Wackerbarth [1980] performed a comparison on conferencing via audio or face-to-face communication across related literature. However, it included group conferencing and not dyadic interaction. Escalera et al. [2012] investigated informal dyadic conversations through online video cameras –but not audio-only– so as to estimate information about social ties.

The rest of this section surveys previous work on the relationship between negotiation and various communication modes such as face-to-face interaction, phone communication and

various computer-mediated communication media (email, chat). This is due to the fact that this work contributes results on the effect of the phone on the negotiation outcome during disagreement, a phenomenon described in Chapter 6.

The main constructs underlying the effect of communication media on negotiations are *social awareness* (“*the degree of consciousness and attention to the other(s) in a social interaction*” [McGinn and Croson, 2004]) and *richness* (“*feedback capability, communication channels utilized, language variety and personal focus*” [Suh, 1999]). Therefore, several works investigate how negotiation outcomes change when using communication media of increasing richness [Sheffield, 1995; Valley et al., 1998; Purdy et al., 2000; Suh, 1999]. This cannot be measured quantitatively, but there is a consensus on the following ordering (from lowest to highest richness): handwritten messages, e-mail, chat, telephone, video-conference and face-to-face interactions.

The experiments by Sheffield [1995] involved 55 dyads (110 subjects in total) acting in a buyer-seller scenario, where the goal was to find an agreement about the price of several commodities. Some dyads negotiated face-to-face (with and without a computer at disposition), others via telephone or chat. The results show that the joint profit increases when moving from a purely textual (chat) to a verbal communication medium (the phone), but does not increase further when adding the visual channel (face-to-face). Furthermore, adding the visual channel is an advantage only if both negotiators are cooperative, i.e. they try to maximise the joint profit of the dyad. The proposed explanation is that the possibility of speaking reduces uncertainty about the task (hence the increase in joint profit when adding the verbal channel) while the possibility of watching one another only increases awareness of the orientation of the counterpart (hence the absence of an effect on joint profit when there are non-cooperative negotiators).

Media richness appears to have an effect in Purdy et al. [2000] as well. The experiments in this work involved 150 subjects bargaining with one another via four different media (face-to-face, videoconference, telephone, chat). The results show that the richer the medium, the higher the tendency to collaborate and avoid competition. Similar results were reported in Valley et al. [1998], where 166 subjects participated in a bargaining game with asymmetric information. In these experiments, the subjects possessing more information were more likely to keep it secret in low-richness media (often resulting in major advantages) than in high-richness ones. The explanation proposed in both Valley et al. [1998] and Purdy et al. [2000] and is that higher richness, in particular the inclusion of a video channel, makes it easier to detect non-cooperative strategies and adopt countermeasures.

While all the experiments above seem to confirm that media richness plays an important role in determining the negotiation outcomes, the results obtained in Suh [1999] contradict such a finding. In this work, 316 subjects were involved in a zero-sum game to be performed face-to-face, as well as via telephone and videoconference. The results show that the communication medium has no effect on both the quality of the decisions and the time needed to achieve them. However, a possible explanation is that the findings have been obtained in a non-Western setting (South Korea) and it is not possible to exclude cultural differences.

The results above suggest that the phone allows one to access sufficient information about the content of the negotiation (there are no breakdowns as in the case of chat and e-mail), but hides attitude and behavioural strategy of the counterpart. As a result, it is easier for non-cooperative negotiators to maximise their individual profit. The findings surveyed in this section have been obtained with landline phones, but they are likely to apply to mobile phones also, because the richness of the two media is the same. Hence, it can be expected

that maximising individual profit should be easier over mobile phones as well.

3.3 Reviewing Corpora on prediction of social information.

There is a large volume of published corpora that collect a wide range of social interactions in order to support research on speech recognition, emotion recognition, role recognition etc., work that could eventually also support social aware agents as SSPNet-Mobile Corpus. Such corpora initially investigate social face-to-face interactions such as group meetings, one-to-one interviews, debates etc., to develop automated systems for the prediction of social information.

SSPNet-Mobile Corpus investigates social interaction with mobile phones, which contributes to the global literature on human social interaction. However, this thesis presents not only a corpus with phone calls, but with results on the influence of the phone on the interaction, specifically on the negotiation outcome, a phenomenon detailed described in Chapter 6. Furthermore, it aims to support the development of predicting social information (personality traits, conflict handling styles) and behavioural events (laughter, fillers etc.) automatically.

Table 3.1 presents corpora aiming at the analysis of social interaction and the prediction of social information. The corpora are presented in chronological order, providing information about participation, number of trials, type of data, type of sensors used, research goals, cultural background of participants and type of experimental setting. The corpora were collected between 2000 and 2012, when the use of mobile phones increased dramatically [ITU, 2014]. However, to the best of our knowledge, an insufficient amount of corpora were collected, accordingly, to investigate specifically the influence of mobile phones on social interactions. This section reviews such corpora as previous works on social interactions and, furthermore, it compares them to this work, regarding the setting in terms of naturalistic or controlled type. That is to indicate the low level of control in our setting, even if it is classified as *Controlled Observation* (see Section 2.4).

The corpora are grouped in two categories: *Corpora on Group Interaction*, corpora focusing on group social interactions such as meetings and *Corpora on Various-Type Interaction*, corpora focusing on other types of social interaction such as one-to-one interviews, one-to-one interaction with agents, one-to-one daily routine social interactions of an individual such as phone calls, face-to-face encounters etc.

3.3.1 Corpora on Group Interaction

Meetings are social interactions that in real-life usually take place in a room. Depending on the type of the meeting, the room might accommodate sensing methods such as microphones, computers, etc. Hence, this type of interaction can be recorded in a relatively low-level controlled setting regarding the space, i.e., a room with microphones or video cameras, following a detailed description of corpora collected to investigate social phenomena during group meetings. The setting usually includes several types of microphones and video cameras in rooms specifically accommodated to support the recording of real-life meetings, or meetings

organised for the purposes of corpus collection. The corpora described here have similarities regarding the setting and the sensing methods. That is why they are compared as a whole to the thesis corpus collection at the end of this section. The ISL (Interactive System Laboratories) Meeting Corpus [Burger et al., 2002; Burger and Sloane, 2004] was collected to investigate the way speech in terms of turn length, speed and disfluencies, is influenced by the type of meeting. The interest of this research justifies the various meeting scenarios that are applied. The corpus includes audio-visual recordings, lapel, table and wireless microphones and three video cameras, of face-to-face *natural or artificial* meetings about work-related or just an open topic for chatting or playing a board, card or role-playing game.

The ICSI Meeting Corpus [Janin et al., 2003] is collected in order to support research on automatic speech recognition. It includes “*Natural*” meetings, meaning meetings that would have taken place anyway. The audio was recorded by lapel or head-mounted microphones and six table microphones, one for each participant on average.

In a similar vein, the NIST Meeting Room Pilot Corpus [Garofolo et al., 2004] and the VACE Corpus [Chen et al., 2006] were developed in order to investigate interaction among speech, gesture, posture and gaze in meetings, and to predict language metadata events such as floor control change, sentence boundaries etc. It includes various types of meetings, whether *Real* meetings which would have occurred anyway, or *Scenario-driven* meetings, where participants were given an artificial task to carry out. The NIST records the meetings by 5 cameras, 2 wireless microphones for each participant, 4 microphones in the centre and at the ends of the conference table, and 3 linear arrays at the front and side wall of the room.

The AMI and AMIDA Meeting Corpus [Mccowan et al., 2005; Carletta et al., 2006; Hain et al., 2008] is collected for the development of a meeting browsing technology, and, was designed to be useful for a wide range of research areas besides the AMI project, which focusses on meetings. The meetings were held in similar rooms, slightly different as to acoustic properties and placement or presence of recording instruments. Multiple microphones, cameras, data projector, meeting browsers and electronic white boards captured time-synchronised audio and video data, slides and textual information. Two-thirds of the data were collected using a meeting elicitation scenario, according to which the participants held four different roles, the same four for all meetings; the project manager, the marketing expert, the user interface designer and the industrial designer, all employees in an company with the goal to develop a prototype of a remote control over the course of a day, while no specific knowledge was required. The remaining data involves naturally occurring meetings in a range of conversational topics.

The CHIL Corpus [Mostefa et al., 2007] is collected to support the analysis of meetings and lectures. The corpus includes recordings of meetings held in five different rooms located in five different institutes-partners of the CHIL Project, in Greece, Germany, Spain, USA and Italy, involving a presenter. The rooms were medium sized and equipped with audio-visual sensors and computers, following a simple recording resource protocol of microphones of four different types, 6 cameras for different capturing angles, and a data-capture computer network, although additional sensors could be utilised. The scenario could be a lecture or an interactive seminar, including a presenter while the interaction with the audience ranges between low to high.

The Mission Survival Corpus-1 & 2 [Pianesi et al., 2007b; Mana et al., 2007] were collected to investigate meetings. The MSC-2 was collected in order to improve the weaknesses of MSC-1 that were due to short duration of the meetings, absence of any kind of overall meeting quality evaluation and lack of personality measurements. This corpus utilized the same scenario

about the plane crash for all 13 meetings. The meetings took place in the CHIL room in Trento, Italy (ITC-irst) [Mostefa et al., 2007], a room equipped with close talk microphones, 6 table and 7 T-shaped arrays microphones, 5 cameras and 4 web cameras and a round shaped table without restrictions about the position and movements of the participants.

The Multimodal Corpus of Multi-Party Meetings [Mana et al., 2007] is collected to support research on automated systems that are capable of analysing social behaviours and predicting personality traits using audio-visual cues. This work collects information on social behaviour and personality traits for each participant. The experimental setting is a lab setting equipped with cameras and microphones in the CHIL room [Mostefa et al., 2007] and the topic is the Winter Survival Scenario, a decision-making disastrous scenario.

The ELEA Corpus [Sanchez-Cortes et al., 2010, 2011, 2012] aims to detect emergent leadership in a meeting using non-verbal behavioural cues. It includes meetings and investigates casual social interactions. The recordings include audio and video collected with two different set-ups: one static and one portable. In total, 27 meetings were recorded with a portable audio-visual setup, 10 with a static setup, and 3 meetings only with microphones.

The corpus includes four questionnaires for each participant: personality traits, a form which describes behaviour under power dominance and leadership, perceived interaction score, which measures perception from every other person in the group in terms of perceived leadership, perceived dominance, perceived competence, perceived liking, and a form with additional information about age, experience in outdoor activities e.g. winter sports, and optional comments on their feelings about the process.

The settings of the corpora described above include cameras, microphones etc. The naturalistic setting of a face-to-face meeting is usually a room which may include recording methods to facilitate other type of meetings such as presentations etc. This setting approximates the real-life setting of a meeting; however, in real-life, besides the fact that face-to-face meetings are usually not recorded, also not all people are familiar with being recorded during a meeting. On the contrary, the setting of SSPNet-Mobile Corpus approximates a naturalistic setting, not only in terms of space arrangement, but also regarding the familiarity of the participants with the sensing method, which is also part of the subject under investigation, i.e. social interaction during phone calls.

3.3.2 Corpora on Various-Type Interaction

This section presents previous work on collected corpora that analyse different types of social interaction such as radio interview, debates, human-computer interaction or random social interactions in everyday life or the work place. They aim to predict social information such as personality traits, emotions, friendship etc. They are compared to the SSPNet-Mobile Corpus regarding the type of experimental setting.

The Electronically Activated Recorder (EAR) Corpus [Mehl et al., 2001] aims to collect real-life data using a “*modified portable audio-recorder that periodically records brief snippets of ambient sounds*” (p.249) [Mehl and Holleran, 2007]. Raters listen to the recordings and select the sections with psychological constructs of interest such as personality traits [Mehl and Holleran, 2007]. The coding of the data takes place using SECSI [Mehl and Pennebaker, 2003] to capture a person’s location, activity, interaction or mood. In comparison to our work, EAR collects data using a device that must be attached to the participants. The sections of conversation are short, and whole conversations are not fully provided. This work includes phone calls, but only unilaterally meaning the sound of the user when he speaks on the phone,

and not the responses of the interlocutor. However, it is a very impressive work in collecting real-life data, but can not cover the conditions that SSPNet-Mobile Corpus sets.

The Canal9 Corpus [Vinciarelli et al., 2009b] consists entirely of real world debates. The corpus was collected to support the analysis of social phenomena during competitive face-to-face interactions. There are similarities with our work in terms of focusing on conflicting opinions about a subject.

The SSPNet Speaker Personality Corpus (or Swiss Radio Corpus) [Mohammadi et al., 2010a; Mohammadi and Vinciarelli, 2012] also includes real-world data from Swiss Radio news broadcasts, so as to support the automatic attribution of personality traits based on non-verbal vocal behavioural cues. The assessments of the personality traits were performed by human judges using the Big Five Inventory.

The SEMAINE Corpus [McKeown et al., 2010, 2012] focuses on developing agents that “sense” different emotions. Specifically, SEMAINE aims at building agents able to engage a person in a conversation using the Sensitive Artificial Listener (SAL). This technique involves a conversation between a human and an agent that either is or appears to be a machine. The experimental process includes two roles: the Operator, who plays the role of four SAL characters and the User, who is encouraged to interact as naturally as possible with the SAL, but not permitted to ask questions. The operator and the user are located in different rooms and could see each other on teleprompters, having the sensation of looking each other in the eye. The setting includes audio-visual recordings from 5 cameras and 4 microphones.

Here, the participants interact either with other humans or with automated programs; however, in both cases, a social interaction is developed. The SAL is described as an induction technique that generates reasonably naturalistic data [McKeown et al., 2010, 2012]. However, the interaction is limited, since the User cannot ask questions, and the experiment takes place in a highly controlled setting.

The SocioMetric Badges (SMB) Corpus [Lepri et al., 2012] was collected to investigate social interactions so as to support research on the investigation, reconstruction and prediction of actual social behaviour in complex organizations, such as research institutions information, but also of individual characteristics such as personality traits, and social context data, such as social networks of the participants. The participants had to wear the Badge when entering the institution and take it off on the way out. Accelerometer, speech, infra-red and blue-tooth recordings gave information about motion, speech, face-to-face interaction incidents, proximity etc. A number of questionnaires were used before the experiences, during and after the experiment about personality, dispositional affectivity, network ties, affect and loneliness, creativity and productivity and situational items to describe the interactional context. The wearable badges are devices, such as mobile phones or mp3s, or small devices that people, carry in everyday life. The data is real-world; therefore, the setting of this experiment may approximate a naturalistic setting.

The literature presented above includes efforts on collecting real-world data as the EAR and SMB corpus by applying discrete sensors in a *Naturalistic setting*. The sensors are described as discrete because they could be a part of a real-life situations. The social badges in SBM or the sensors in EAR, similarly to, e.g., mobile phones could be carried on every-day basis. However, participants are aware of the recordings, therefore, the setting could not be *Naturalistic* as defined by Guerrero et al. [1999]. Examples of corpora with such setting are the Canal9 and Swiss Radio Corpora, which provide collections of pure real-world data captured in a radio or a tv studios under real-life conditions, political debates and interviews. Our case is also another example of a *Naturalistic setting* approximation. However, it might

have an advantage compared to the EAR and SBM corpora, since it provides a rich collection of complete actual phone calls between two acquainted individual. Furthermore, for the same reasons it provides a novelty, at least to the best of our knowledge.

3.4 Requirements of Corpus Collection

This section explains the key issues to be taken into account in order to collect a corpus that collects raw data by setting an experimental procedure so as to investigate human behaviour, such as the SSPNet-Mobile Corpus. The theoretical background about naturalistic and controlled settings explained in Section 2.3 and examples of the corpora presented earlier are used to justify, both in theory and in practice, the fundamental considerations that drive the collection of corpora to support behavioural analysis. Specifically, the experimental setting, the protocol, the recruitment requirements, the psychological information collection and the technological facilities are aspects to be considered.

- *Experimental setting*

The setting refers to the arrangement and ambience of the space, the sensing method and the scenario of the experiment. The subject under investigation drives the researcher to set a naturalistic or controlled setting. This is due to the fact that field experiments provide real-life data and information, but the subject of interest might be difficult to be captured to a substantial extent. On the other hand, a controlled setting can ensure the observation of the subject of interest, but with a trade off in the ecological validity of the phenomenon.

However, when the subject of interest takes place in real life in a setting which can easily accommodate sensing techniques or make use of technology during the behaviour under investigation, then the trade off in ecological validity can be reduced. That is because the real life setting already accepts the presence of technology, the major factor to constitute a controlled setting; hence a similar capturing method can included “naturally” in an experimental setting. For example, group meetings in real life usually take place in rooms that include microphones, cameras etc., to accommodate e.g., skype meetings or to record presentations. The real-life setting already includes the presence of sensing technique. This argument justifies the fact that a large amount of corpora mentioned earlier focus on meetings. Moreover, rooms have been developed to facilitate this research, such as the IDIAP smart meeting room. The naturalistic setting of a real-life meeting is approximated though, a room transformed into a fully equipped room that can record synchronised audio, video and textual data through lapel and microphone arrays, high quality CCTV cameras, and a beamer to capture presentation material (slides, video etc.). Other examples of naturalistic data are the SSP Conflict Corpus and SSP Speaker Corpus, where actual debates and radio broadcasts are recorded. In both cases the real world interaction already includes the capturing methods, e.g., video cameras and microphones.

In the case of SSPNet-Mobile Corpus, the priority is to collect actual phone calls to investigate phone-mediated social interaction. Hence, this work implements a setting

where the capturing method is the phone, which is also part of the subject under investigation, and therefore eliminates the presence of a more complex controlled setting and, consequently, the effect on the ecological validity of the interaction.

- *Protocol*

The protocol secures the uniformity of the experimental process by standardising the order of activities before, during and after the experiment, so as to eliminate effects on the subjects under investigation from random factors.

For example, in a naturalistic setting such as the Sociometric Badges Corpus Lepri et al. [2012], the data should always be captured when people are in the institution, so the badges are given to participants at the entrance and are taken off on their way out, and also they have to prove their constant presence in the meantime by completing a survey. These steps constitute part of the protocol of the corpus. In a controlled setting such as the MAHNOB- HCI Corpus Soleymani et al. [2012] a multi-modal database is recorded in response to affective stimuli aims to support emotion recognition and implicit tagging research. Recordings of electroencephalograms, resistance of the skin, electrocardiogram, respiration amplitude and skin temperature are investigated, hence the participants should always be attached to the same electrodes, keeping the same protocol for each trial.

In SSPNet-Mobile Corpus according to the protocol, the participants were accommodated separately from 2 researchers at the same time. Always, the same researcher was accommodating the receiver of the call and the other one the caller. The interaction of the caller with a specific researcher might be a factor of the “*Caller-Receiver*” effect. To eliminate that possibility the two researchers switched their positions to each other for half of the trials.

- *Questionnaires*

In the case of corpora such as SSPNet-Mobile social or behavioural information need to be obtained. The means to collect that kind of information is adopted from psychological method that employs psychometric constructs developed to measure behaviour etc. However, there are two types of measurements: self-assessments and external rating from observers. Previous research has investigated the validity of self-assessment, i.e., to what extent self-assessment gives accurate results Fox and Dinur [1988], the *positive illusions* phenomenon on the human tendency to perceive ourselves in a favourable way [Brown, 1986; Taylor and Brown, 1988] or how self-assessment can differ from external coding [Dunbar and Burgoon, 2005]. Discussion of the advantages and disadvantages of the two modalities are beyond the scope of this thesis. This work only takes into account self-assessments. However, in Chapter 9, this work suggests the use of both perspectives for more complete behavioural information collection.

Furthermore, surveys, forms or any kind of questionnaire could collect information relative to the subject under investigation regarding the perspective of the research. For example, as mentioned above, Lepri et al. [2012] used forms to prove the presence of participants in the university, since the research analysed interaction in the university area. In ELEA Corpus Sanchez-Cortes et al. [2010] used an outdoor survival scenario, and also collected information about experience in outdoor activities. SSPNet-Mobile collects similar information since it also applies a decision-making scenario aimed at surviving a disastrous plane crash.

- *Participants recruitment*

The gender, cultural, educational backgrounds of the people who take part in the experimental process, or any other background possible relative to the subject under investigation, must be taken into account for the recruitment process. This is to avoid effects or bias due to different skills or other effects.

For example in SSPNet-Mobile Corpus, recruitment process required only English native speakers, to avoid misunderstanding due to different languages that could lead to poor engagement in interaction.

- *Technological facilities and budget available*

The concept of the experimental process needs to aim at the best implementation according to given technological or financial limitations. The feasibility of the means has to be considered.

For example, SSP collected Canal9 to investigate social phenomena in competitive interactions. When the subject of investigation focused on personality trait prediction, shorter clips were extracted from the corpus to collect the SSPNet Conflict Corpus, and the clips were rated from external observers in terms of personality traits. This was an efficient and feasible way to collect behavioural information about the data by employing online rating services.

3.5 The SSPNet-Mobile Corpus.

In SSPNet-Mobile Corpus the subject of interest is to investigate social interaction using mobile phones. Specifically, the investigation focuses on the influence of the phone on the interaction, and the prediction of behavioural information (personality traits and conflict handling styles) through speech and motion activation data captured by the phone (see Section 4.5). We wanted to investigate the interaction without involving prior personal relationships, different roles or expertise skills that could produce unbalanced dialogues regarding the speaking time or the decision-making process. Hence, the interaction was isolated from other interpersonal relationship factors, and the personality dominated the social interaction. The scenario chosen was the Winter Survival Task (WST) (see Section 4.3), a disaster decision-making scenario that favours the development of rich conversations in opinion stating and argumentation by employing common sense and not specific knowledge or skills. Also, the WST engages the participants in an unusual experience that is unlikely to be a real experience, and therefore avoids favouring someone with similar experience in the past. In this work it was a priority to ensure that the setting approximated a naturalistic scene as much as possible. The participants were people who work or have worked in offices (see Section 4.2), who, while alone in a typical office, make or answer a phone call and negotiate with a stranger. The setting highly approximates a naturalistic environment.

The corpus included 60 calls between dyads of unacquainted individuals (120 subjects in total). The cultural background of the participants was uniform (118 subjects out of 120 held a British passport). That was the only requirement in order to avoid misunderstanding and different cultural bias in the interaction (see Section 4.2). Most of the participants had a

university education (the most represented subjects were Psychology and Computing Science) and were recruited at the University of Glasgow. The psychometric questionnaires used in this work were driven by the need to have information about the key factor that drives behaviour, personality and how people handle conflicts, since this work was anticipating conversations with opposite opinion statements (employing WST) so as to acquire rich dialogues (see Section 4.4).

3.6 Conclusions

The novelties of this corpus, to the best of our knowledge, starting from the most major one, are the collection of a new corpus with phone calls that:

1. sets an experimental procedure that can reveal the influence of the mobile phone on social interaction, specifically on the negotiation process, the *Caller-Receiver* effect (see Chapter 6).
2. approximates a naturalistic setting, providing data that includes whole conversations, synchronised audio and motor activation recordings and participants' behavioural information (personality traits, conflict handling styles),
3. always uses the same apparatus (N900 smartphone).

This work claims to provide a good corpus on social interaction using mobiles and observation of the influence of the phone on social interaction which, to the best of our knowledge, has not been reported before. This justifies reasonable motivations and solid implementation and, consequently, the importance of compiling a new corpus to investigate the research interest of this work.

Table 3.1: Past corpora collected for the investigation of various aspects of behaviour during meetings of small groups or dyads.

Corpus	Subjects	Trials	Data Type	Task	Cultural Mix	Naturalistic Setting
EAR, 2001	52	4 days/subject	audio	capturing real-world conversations to analyse ongoing behaviour	n/a	No
ISL, 2002	600	104 trials	audio-visual	investigate effects of meeting type on speaking style	Yes	No
ICSI - MRDA, 2001-03	53 unique	75 trials	audio	automated speech recognition	Yes	No
NIST, 2001-03 & VACE, 2006	115 unique	19 trials	audio-visual	recognition of human interaction in meetings	n/a	No
AMI Corpus, 2005-06	213	196 meetings	audio-visual, textual	dialogue acts, group activity recognition	Yes	No
CHIL, 2007	n/a	86 meetings	audio-visual	support realistic human interaction	Yes	No
MSC, 2007	52	13 meetings	audio-visual	classification of behaviour	No	No
Canal9, 2009	190 unique	70 debates	audio-visual	analysis of social phenomena during competitive interactions	Yes	Yes
Swiss Radio, 2010-12	322	96 news bulletins	audio	personality traits recognitions form speech	n/a	Yes
SEMAINE, 2010-12	150	959 conversations	audio-visual	improve HCI	No	No
SMB, 2012	53	6-weeks study	motion, audio, proximity etc.	prediction of actual social behaviour	Yes	No
ELEA, 2011	148	40 meetings	audio-visual	emergent leader recognition	No	No

Chapter 4

The SSPNet-Mobile Corpus: Acquisition of the Data.

4.1 Introduction

The previous chapter surveyed corpora collected in order to support research on emotions, group dynamics etc. The comparison of the corpora under the prism of our research interests on telephone communication highlights a gap that needs to be filled. In particular, the need to collect a corpus, especially for the investigation of conversations with mobile telephones, emerges. The SSPNet-Mobile Corpus was collected to support our research on mobile communication. Our interest in capturing both physical and psychometric measurements has driven the construction of the experimental process. Hence, we used sensors to capture audio and motor activation, psychometric questionnaires to record the social phenomena. The protocol was selected in order to keep the control of the conversation and, also, to provide us with metadata about decision-making processes and their outcomes.

This chapter describes the collection of the SSPNet-Mobile Corpus. The rest of the chapter is organised as follows: Section 4.2 describes the subjects, Section 4.3 describes the experimental task and the protocol of the experiment, Section 4.4 refers to the psychological questionnaires that we used to measure personality traits and conflict handling style and Section 4.5 reports the sensing method, the synchronisation problem and the unobtrusive nature of the method.

4.2 Subjects

The corpus aims to investigate *behavioural events* as laughter, fillers, silence, back channel etc., (full description in Chapter 5) during conversations using mobile phones. Therefore, involving subjects with the same mother tongue and uniform cultural background avoids potential misunderstandings that could influence the flow of conversation. SSPNet-Mobile Corpus limits participation to British cultural background to avoid massive diversities due to different cultural experiences and mother tongues. Hence, the only essential requirement of the experiment was to be native English speakers, and of British nationality. British nationality includes Scottish, English, Irish and Welsh origin; however these differences are minor in

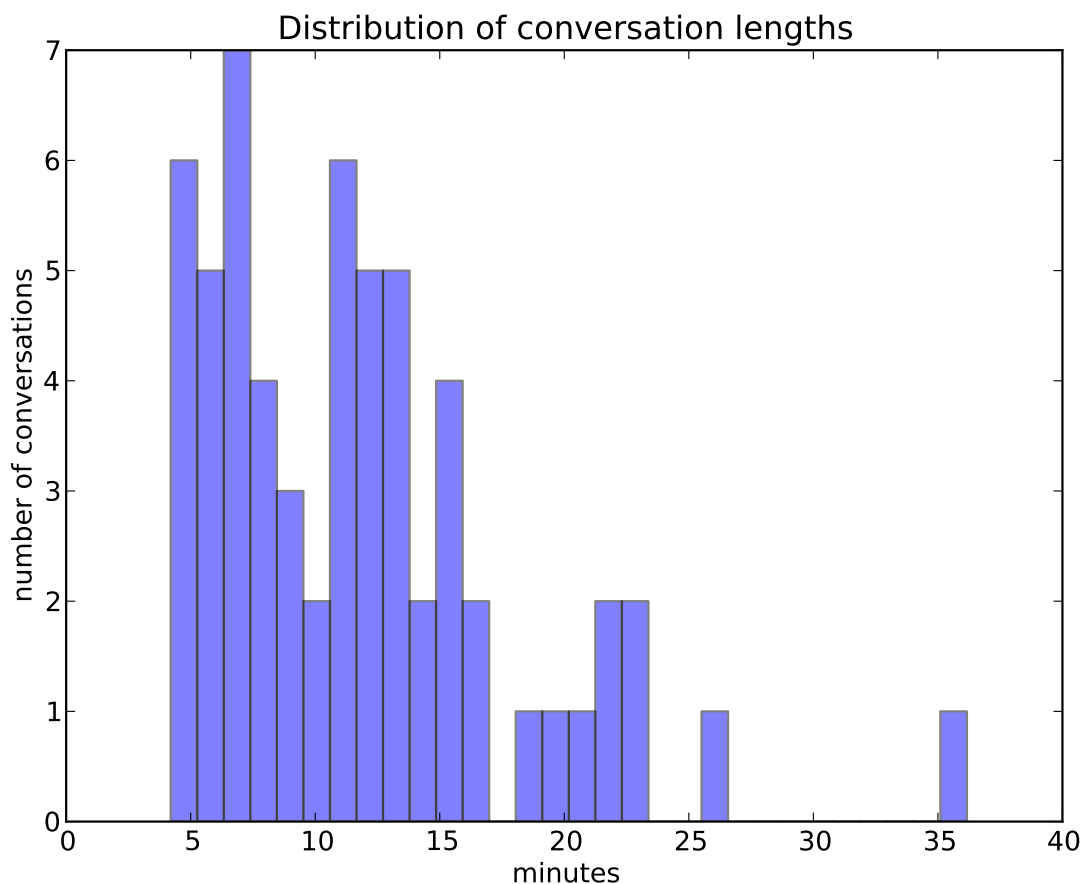


Figure 4.1: The distribution of conversation length across the database of 60 conversations.

causing language shortcomings or massive cultural differences, at least to the extent of this research's goals.

In total, 120 different subjects participated in the experiment. All of them were native English speakers and unacquainted to their phone call counterpart. The main source of participants was the subject pool of the School of Psychology (SPs) at the University of Glasgow (UG). The latter is an official list operated by the SPs, and includes people interested in participating in research experiments. The list can be accessed through a website that provides facilities to filter subjects according to gender, cultural background, mother tongue, age etc.

The subjects were contacted via e-mail. The messages did not provide details about the experiments, to keep the subjects unbiased. The experimental procedure was described as “the phone call experiment” and the given information was explaining that:

- it involves a phone call via mobile telephones
- a questionnaire has to be filled before the phone call
- the duration is approximately 1 hour
- the participation is paid with £6, but extra rewards are possible
- their presence is mandatory at the School of Computing Science

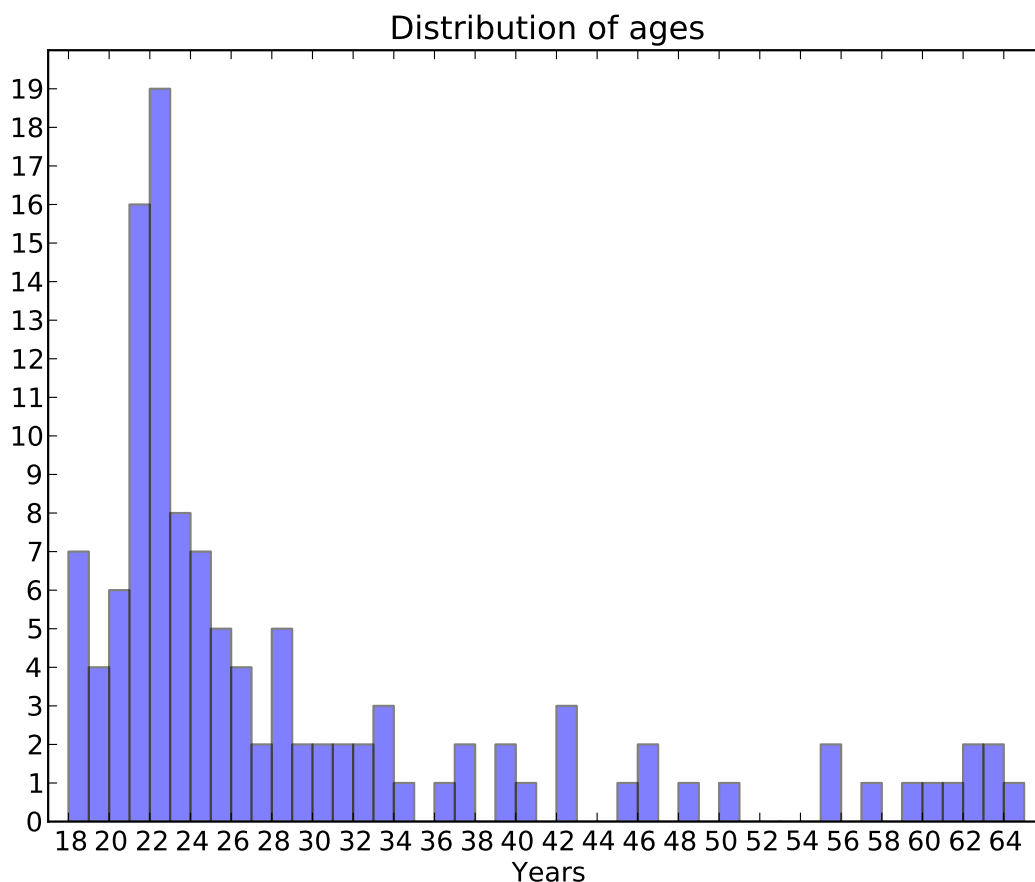


Figure 4.2: The plot shows the age distribution of the 120 subjects. The median is 23.5 years.

Also, posters in the area of the university and an advertisement on Gumtree (<http://www.gumtree.com/glasgow>) in the volunteering section, enabled us to reach the desired number of 120 participants.

Figure 4.1 presents the distribution of phone call duration in minutes. The calls range from 4 to 35 minutes, while 85% of the calls are limited between 5 and 16 minutes. The 60 phone calls sum to a total of 11 hours and 48 minutes or (710 min) with average conversation length around 11 min 48 sec.

Figure 4.2 shows the age distribution with age range from 18 to 64 years old, and median age 23.5 years. 63.3%, 76 individuals, are between 18 and 26 years old, while the remaining 36.7%, 44 individuals, are between 27 and 64 years old.

Table 4.1 shows the distribution of subjects in terms of gender, occupation and nationality. There were 63 females and 57 males, 78 undergraduate students at SPs and School of Computing Science (65%), 3 Masters students (2.5%), 14 PhD students (11.67%), 6 research assistants (5%), 3 people from the academic staff (2.5%) of the School of Computing Science and 16 externals (13.3%), i.e. people who were not students or staff members at the time of their participation, although all of them used to be students or staff of the University of Glasgow in the past. All of the subjects were native English speakers, 98.3% of them were British and only 1.7% were not.

Features of Subjects		Number of Subjects
Feature	Description	
Gender	Female	63 (52.5%)
	Male	57 (47.5%)
	Total	120 (100%)
Background	Bs & Ms Students	81 (67.5%)
	PhD, Research & Academic Staff	23 (19.2%)
	Externals	16 (13.3%)
	Total	120 (100%)
Nationality	British	118 (98.3%)
	American	1 (0.83%)
	Cypriot	1 (0.83%)
	Total	120 (100%)

Table 4.1: The table reports the number and the percentages of participants per gender, educational background and nationality.

4.3 Experiment

The experiments were held in the offices of the School of Computing Science (SCS), University of Glasgow (UG). They were conducted by two experimenters, and lasted approximately 8 months. Two rooms were utilised for the purposes of the experiment, which were chosen based on the strong reception of the phone signal. The rooms were working offices which weren't in use during that period. The space organisation was not identical between the two offices, although the recording method could not be affected since it was independent of the spatial setting.

The task of the experiment is based on the Winter Survival Exercise [Volkema and Ronald, 2002], a decision-making scenario about the rescue of an undefined number of survivors after a plane crash in Northern Canada. The participants are the rescue team members, and they have to rank 12 items according to their usefulness. The Winter Survival Exercise is a scenario usually used to check group dynamics, NASA used it to evaluate potential astronauts [Pianesi et al., 2007a], to capture interpersonal cohesion in a group [Rogelberg and Rumery, 1996], to develop strategic skills in students [Joshi et al., 2005], to measure group performance [Durham et al., 2000], or to compare the performance of groups to individuals [Miner, 1984]. It involves the participants in an unusual situation which it is unlikely that they have experienced. The experience of a disastrous incident, such as a plane crash, can ensure reactions based on common sense and excludes the need of special skills or knowledge. Moreover, the scenario promotes cooperation and the development of a rapport between participants.

In our case, the needs of our experiment led to a modification of the Winter Survival Exercise. Hereinafter, our version will be referred to as the Winter Survival Task (WST). The modifications of the WST are the following:

- We were interested in the interaction of two people talking through telephones instead of a group interacting face-to-face. Hence, the participants were assigned to play the

role of rescue team members and not the role of rescue team members, so as to simulate a more realistic situation of people making decisions about survival through telephones after a plane crash.

- We were interested in the explicit opinion of each individual, so we set the task to be a selection of an undefined number of items, instead of ranking the items. The ranking order of the items, in terms of usefulness to a disastrous incident, is provided from the creators of the Winter Survival Exercise as the solution of the exercise. Hence, in our version i.e. the WST, we set the top five ranked items to be the “useful” ones and the rest not.
- We were interested in the explicit opinion of each individual, *before and after* the interaction, so we added one more column (see p3, Appendix A), to be filled in before the phone call, hereinafter *personal opinion*.
- To motivate the participants to stand for their personal opinion in case of disagreement, we were rewarding the times a “useful” item was marked with a “Yes” by giving an extra £3, but we penalised the times a not “useful” item was marked with a “Yes”, by reducing the final amount by £3. However, we set a minimum threshold of reward equal to £6 in order to motivate people to participate in our research.
- To avoid possible lack of attention regarding the consensus part, we highlighted the information about the consensus column. It should be identical for both participants at the end of the phone call, and the penalty was £3 for every different answer at that list.

To sum up, the unacquainted participants of each dyad had to talk using a mobile telephone. They were asked to decide together whether 12 specific items, found on the plane crash field, could be useful or not for the rescue of the survivors. They had to go through the list of the items twice, giving their personal opinion and then their consensus for each item. Thus the WST provides an explicit way to detect how many times in the course of the interaction a person is the winner or the loser, in case of disagreement about an item.

Setting the conversational topics using a 12-object list, functions as a “regulator” regarding the engagement of the interlocutors to the interaction. The topics-objects engage both participants in the conversation, since they both have to express opinion so as to reach a consensus. Hence, the topics ensure the observation of –interesting for our work– human communication with agreements, disagreements, conflicts etc., to occur. At the same time, the WST and the topics about the usefulness of each object assign equal roles to the participants that merely require common sense. Finally, the participants are free to develop their interpersonal interaction from scratch with an unacquainted person, and are free to talk as long as they want. Summing up, it is important to make a comment on the overall function of the WST. The latter ensures the observation of rich social interaction (such as engagement and conflicts) and, at the same time, preserves the spontaneity of the interactions.

In order to collect the raw data of the SSPNet-Mobile Corpus Corpus, the participants had to fill in two psychological tests that measure *personality traits*, the Big Five Inventory-10 (BFI-10) [Rammstedt and John, 2007a], and *conflict handling style*, the Rahim Conflict Inventory- II (ROCI-II) [Rahim, 1983] (see Section 4.4 for a detailed description) and, then, to take part in a phone call so as to discuss the WST with another participant, unacquainted to one another.

The psychological tests, as mentioned in the Introduction and explained further in the following section, were chosen to record behavioural information. Specifically, the psychological tests measure personality traits as personality is the key factor that drives behaviour[Funder, 2001] and, conflict handling style because the decision-making WST is a scenario that provokes stating-opinions on conflicting conversational topics and anticipates conflicts.

The experiment was performed by sixty different dyads, following the same protocol:

- First, they had to fill the Big Five Inventory-10 (BFI-10) [Rammstedt and John, 2007a] and the Rahim Conflict Inventory- II (ROCI- II) [Rahim, 1983] questionnaires online. This task was completed before the phone call, to shorten the duration of the procedure where subjects should be present.
- They had to attend the SCS on a prefixed date, in different meeting points so as to avoid possible encounters between the participants.
- Once the participants were ready, they were given an N900 mobile phone, information on how to use the phone, the scenario, the item list and the consent form (see Appendix A, B).
- After that, they had to read the scenario and give their *personal opinion*. Writing “Yes” or “No” at the left of the item list (see p3, Appendix A) standing for a “useful” item or not, correspondingly.
- Once both participants had filled the column, the experimenters assigned the roles of *Caller* and *Receiver*. The selection was random.
- The *Caller* had to phone the *Receiver* and discuss the items in the order of their appearance on the list, starting from the top. After reaching a consensus for the first item they had to fill the second column, at the right of the list (see p.3, Appendix A) using “Yes” or “No” as before. Then they had to do the same for the next item and so on.
- After the call, they had to declare in writing whether they had similar experiences in the past or had developed survival skills and to fill in the consent form (see Appendix B). Additionally, they were receiving £6 and in case of extra reward, they had to come back.

Consequently, after the completion of a session, we were collecting for each participant:

- 1 text file, including the BFI-10 and ROCI-II responses
- 1 WST solution document including personal decisions and consensus
- 1 form to describe previous outdoor experience
- 1 filled consent form
- 7 channels of audio-motional data (see Section 4.5).

All responses were stored digitally.

4.4 Psychological Tests

Having as a goal to derive behavioural information in reference to personality traits and attitude during conflicts, we selected two psychological tests: the Big Five Inventory-10 [Rammstedt and John, 2007a] and the Rahim Conflict Inventory-II [Rahim, 1983], to be included among the other tasks of the experimental process.

A psychological test is essentially an objective and standardised measure of a sample of behaviour [Anastasi et al., 1982]. The objective and standardised measure preserves the uniformity and the similarity of administration, scoring and interpreting, excluding the judgement of any experimenter. Also, the measuring value of any psychological test is, ultimately, human behaviour [Thambirajah, 2004].

Allport [1937] believes that human behaviour is the outcome of two components: the *adaptive* aspect, the attitude towards a current situation (skills, intentions) and the *expressive* aspect, i.e. the personality and the temperament. In the SSPNet-Mobile Corpus study, the human behaviour to be investigated includes conversation with another person. The conversation aims to reach a consensus twelve times. This situation leads to a *conflict*, which is defined as the presence of parties with contradictory goals, and does not presuppose shouting, fighting or misbehaving; on the contrary, it can be observed in the course of a calm and positive interaction. In the case of the SSPNet-Mobile Corpus study, the contradictory goals were to take or to leave an item. Hence, the possibility of a conflict to emerge was high. The Rahim Conflict Inventory-II [Rahim, 1983] was included in the experimental process to measure conflict handling style of the subject.

The expressive aspect of behaviour, i.e. personality, is measured by the Big Five Inventory-44 [Rammstedt and John, 2007a]. This construct has been commonly used for measuring personality traits. The SSPNet-Mobile Corpus includes a shorter version of 10 questions (extensive description follows). Also, the requirement to include only British participants limits the cultural background to be the same for all conversations.

Both tests include clusters of questions focusing on one personality trait or conflict handling style. Every question aims to evaluate the presence of trait or style in numbers. All of the questions can be addressed by the same 5 options:

- Disagree strongly
- Disagree a little
- Neither agree nor disagree
- Agree a little
- Agree strongly

The options are assigned to scores which range from [-2, 2]. For positive questions the score -2 corresponds to “Disagree strongly” and 2 to “Agree strongly”. For negative questions the opposite assignment is valid, i.e. the score -2 corresponds to “Agree Strongly”. The sum of the corresponding numbers by cluster of questions equals the score for every trait and style. Hence, the score for every personality trait (2 questions per trait) ranges from -4 to 4 values, and the score for every handling conflict style (7 questions per trait) ranges from -14 to 14 values. The number of questions that correspond to each trait and style and a fully description

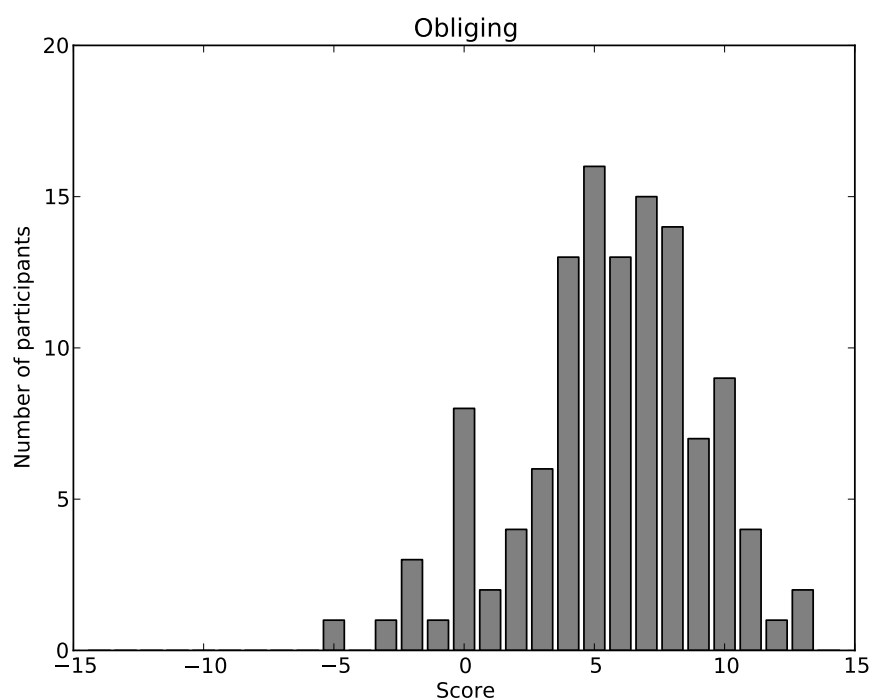


Figure 4.3: The distribution of 120 scores that measure the obliging conflict handling style per participant.

of the psychological tests following. Graphs depicting the distribution of participants' scores per personality traits and conflict handling styles are also presented.

4.4.1 Conflict Handling Style: the Rahim Conflict Inventory-II (ROCI-II)

The ROCI-II, measures the conflict handling style of a person during a conflict. The taxonomy takes place in the basis of two different dimensions: concern for self and for others to a high and low degree [Rahim, 1983] to measure five different styles to cope with a conflict [Rahim, 2010]:

- *Obliging*, low concern for self and high concern for others: tendency to accept the position of others.
- *Avoiding*, low concern for self and low concern for others: tendency to avoid conflict.
- *Dominating*, high concern for self and low concern for others: tendency to impose one's views in case of disagreement.
- *Integrating*, high concern for self and high concern for others: tendency to attract others towards own positions.
- *Compromising*, intermediate concern for self and for others: tendency to find a compromise between different positions or interests.

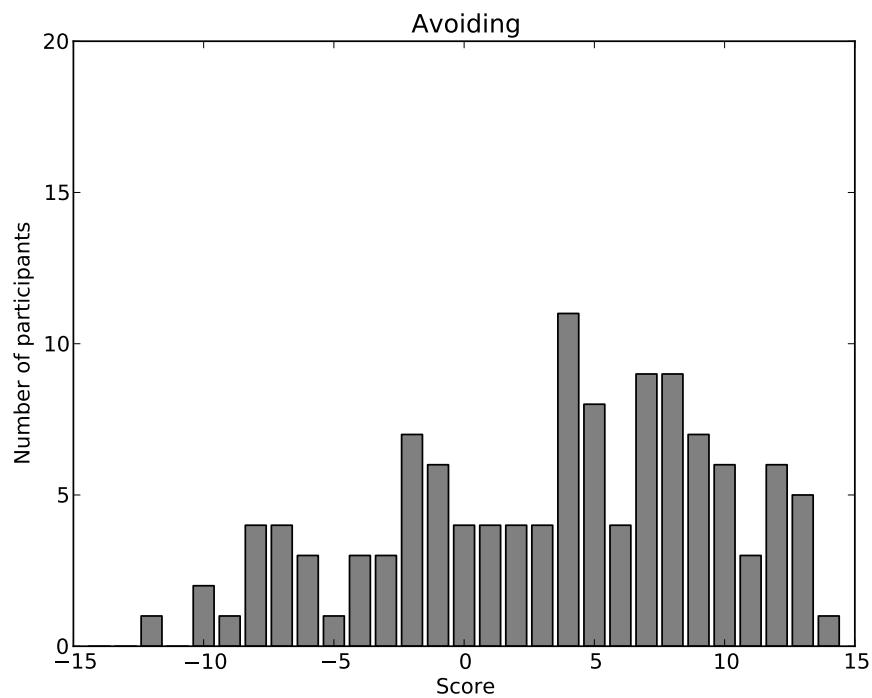


Figure 4.4: The distribution of 120 scores that measure the avoiding conflict handling style per participant.

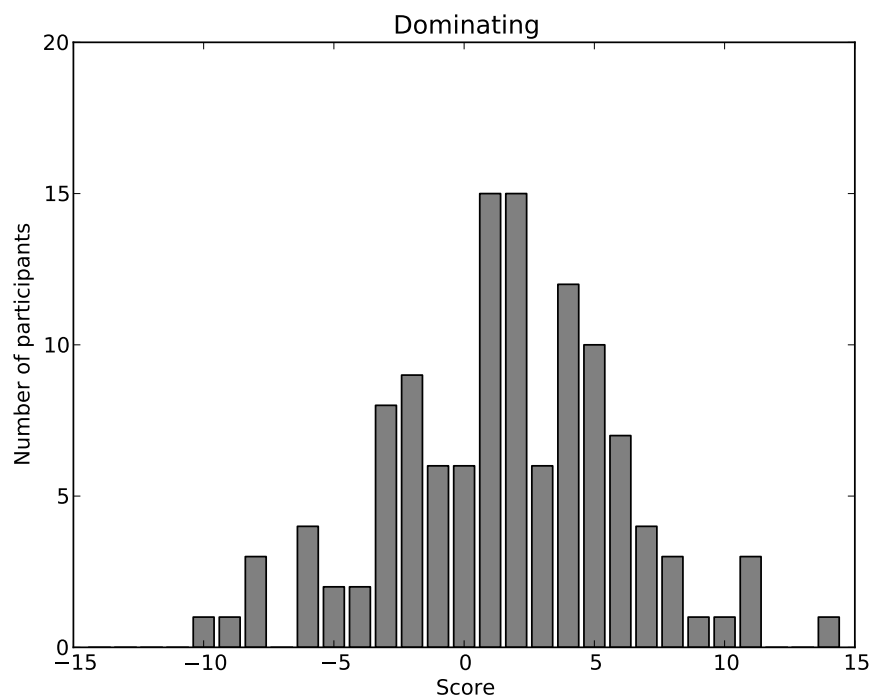


Figure 4.5: The distribution of 120 scores that measure the dominating conflict handling style per participant.

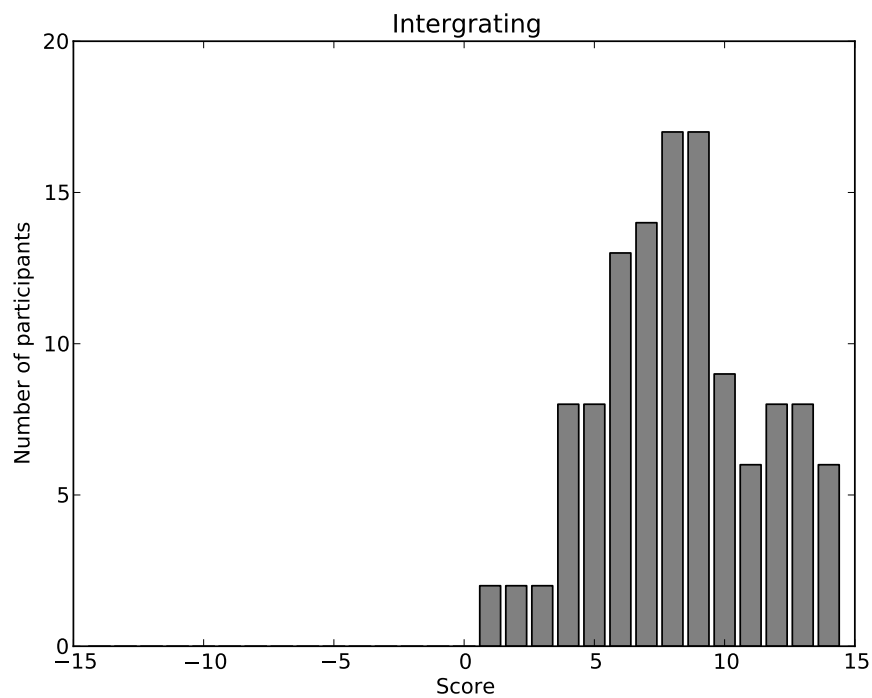


Figure 4.6: The distribution of 120 scores that measure the intergrating conflict handling style per participant.

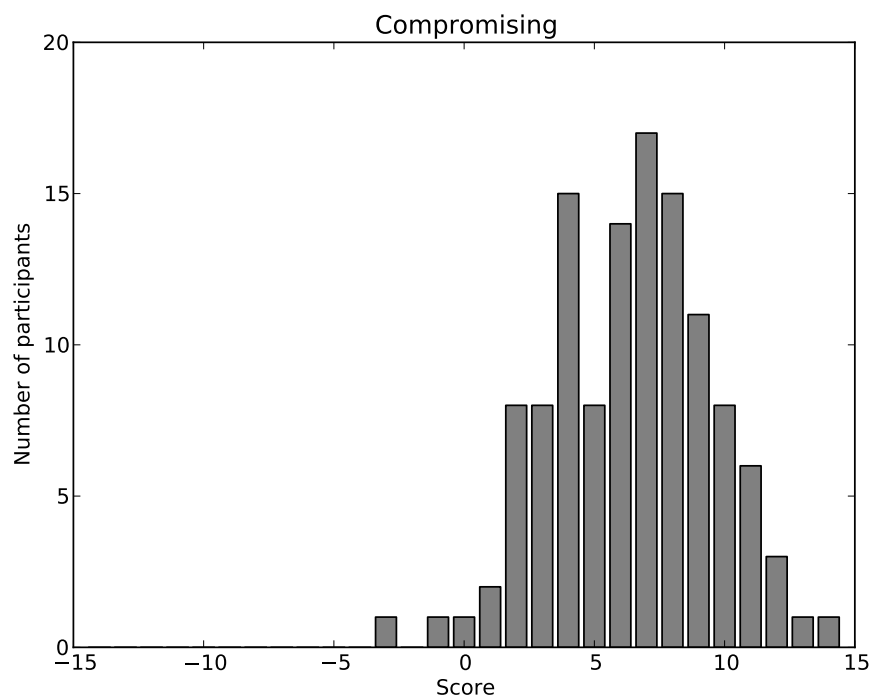


Figure 4.7: The distribution of 120 scores that measure the compromising conflict handling style per participant.

The ROCI-II consists of thirty-five questions, seven questions per style. The five scores (one per conflict handling style) measure how much the behaviour of the participant corresponds to each conflict handling style. The graphs in Figures 4.3 to 4.7 depict the distribution of 120 scores by style. Comparing the figures, one could separate them into two groups: Figure 4.3, Figure 4.6 and Figure 4.7 in the left skewed group, while Figure 4.4 and Figure 4.5 in the group of more wide-spread distributions. In the left skewed group we observe the *Obliging*, *Intergrating* and *Compromising* conflict handling styles to be scored with positive scores, through the answers participants gave while they assessed themselves. The negative scores, corresponding to the absence of the style, are rare. These styles are related to “good” attitude. In contrast, the *Avoiding* and *Dominating* handling conflict styles, related to “bad” attitude, present wider spread score-distributions. The latter include not only positive but, also, negative scores. We observe that the participants tend to see themselves positive and assign their behaviour more often to “good” behaviour and attitude. The *positive illusions* phenomenon reflects the tendency of humans to perceive themselves in a favourable way [Brown, 1986; Taylor and Brown, 1988].

About the information on relevant experience in outdoor survival situation none of the participants declared themselves to have developed such skills in the past.

4.4.2 Personality Traits: The Big Five Inventory-10 (BFI-10)

Personality (as mentioned in the Introduction) is defined as *individuals’ characteristic patterns of thought, emotions and behaviour, together with the psychological mechanisms – hidden or not – behind those patterns* (p.198) [Funder, 2001]. To measure personality we follow the Big Five Trait Taxonomy which “*captures, at a broad level of abstraction, the commonalities among most of the existing systems of personality description and provides an integrative descriptive model for personality research*” (p.45) [Oliver and Srivastava, 1999]. The Taxonomy classifies words-adjectives in five different clusters, each one representative of an essential personality trait. The traits are the following, and in parenthesis are given some of the most representative words of each cluster. The following list demonstrates the five personality traits accompanied by the corresponding cluster:

- *Extraversion* (Talkative, Assertive, Active, Energetic, Outgoing, Outspoken)
- *Agreeableness* (Sympathetic, Kind, Appreciative, Affectionate, Soft-hearted, Warm)
- *Conscientiousness* (Organised, Thorough, Planning, Efficient, Responsible, Reliable)
- *Neuroticism* (Tense, Anxious, Nervous, Moody, Worrying, Touchy)
- *Openness* (Wide interest, Imaginative, Intelligent, Original, Insightful, Curious)

Therefore, personality can be described by five representative scores, which correspond to how well the personality matches the words of each cluster [Mohammadi et al., 2010b]. The Big Five Taxonomy can be formed in a questionnaire, the Big Five Inventory, which originally consists of 44 short-phrase items. In order to decrease the time to fill in the questionnaire, considering the time for filling the 35-questions ROCI-II test, we used the BFI-10 [Rammstedt and John, 2007a], which is a 10-question shorter version of the BFI. The BFI-10 predicts personality traits based on two questions per trait, one proportional and the other inversely proportional versus the trait (see Appendix C).

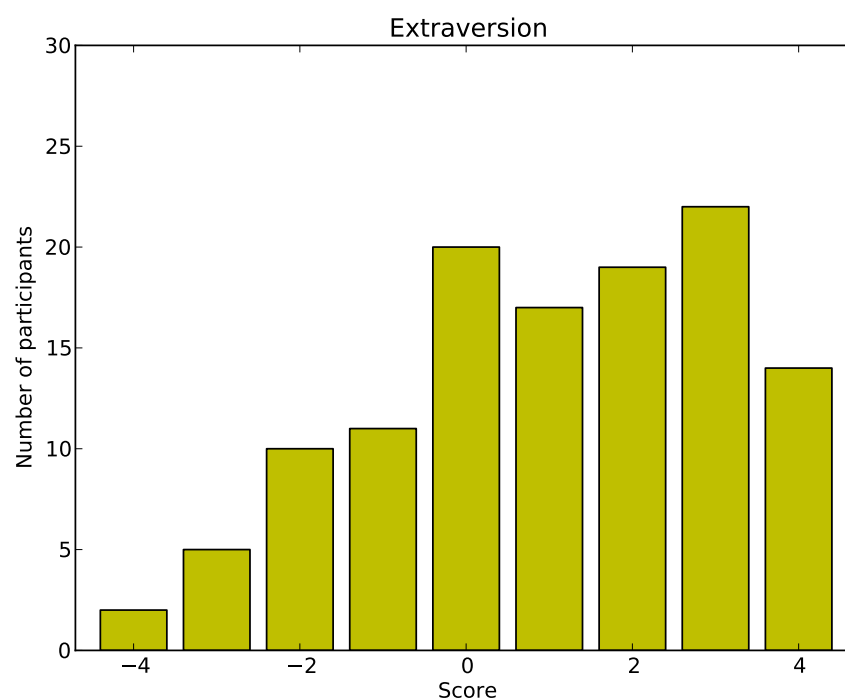


Figure 4.8: The distribution of 120 scores that measure the extraversion personality trait per participant.

The graphs in Figures 4.8 to 4.12 depict the distribution of 120 scores by trait. As mentioned in the previous subsection in discussion of the conflict handling style graphs, we observe the effect of the “positive illusion” in personality traits as well. The *Neuroticism* trait is the only one that has a skewed distribution to the right. That corresponds to more negative scores than positive. This trait is the only one that is related to a negative personality, while the remaining four (*Openness*, *Agreeableness*, *Conscientiousness*) and *Extraversion* are characteristics of personality that are positive, and positive scores (left skewed distributions) correspond to the presence of the trait.

The next trait with the most negative score, after *Neuroticism*, is the *Extraversion* trait and follows the *Openness* trait. The *Agreeableness* and *Conscientiousness* traits includes relatively many neutral scores (zero value), both more than or equal to 25 participants. Both present much fewer negative scores in comparison to positive scores. This fact may also follow the “positive illusion” phenomenon since people prefer to be neutral than to be negative while assessing themselves.

4.5 Recording Human Behaviour

Two types of sensors are used to capture speech and motor activation: microphones and gyroscopes (in combination with accelerometers and magnetometers). Both sensors are attached to the phones. To the best of our knowledge, we applied one of the most unobtrusive

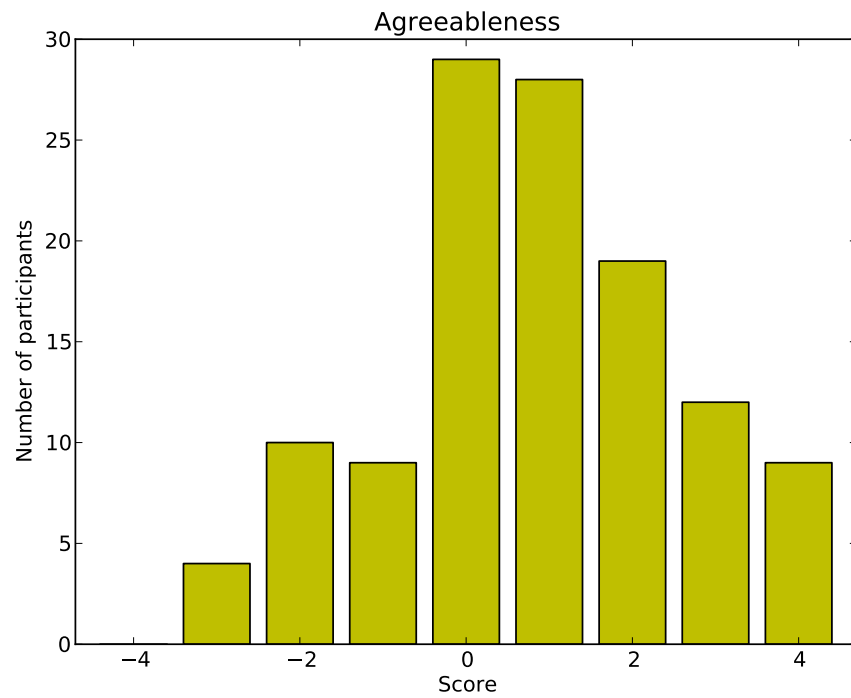


Figure 4.9: The distribution of 120 scores that measure the agreeableness personality trait per participant.

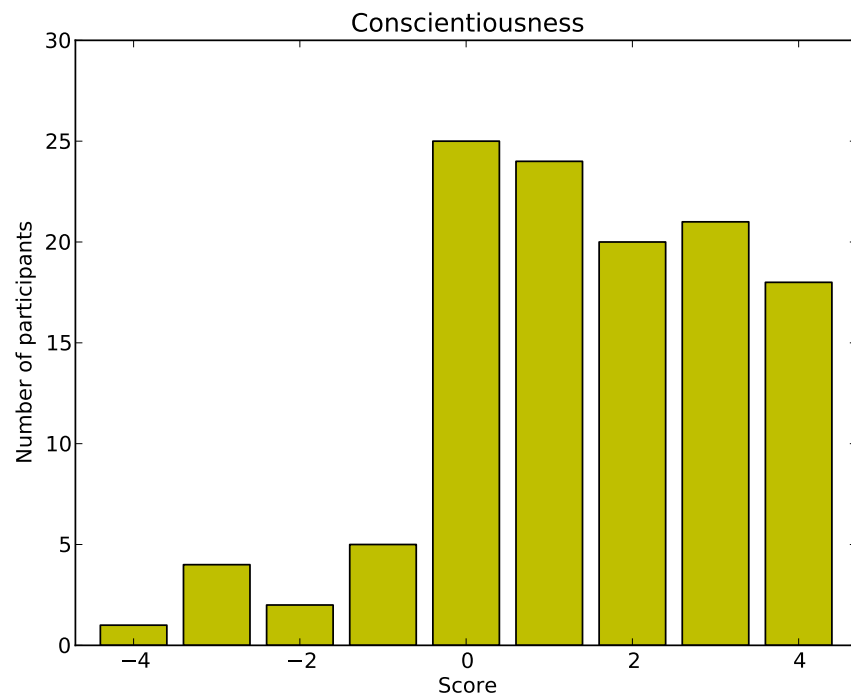


Figure 4.10: The distribution of 120 scores that measure the conscientiousness personality trait per participant.

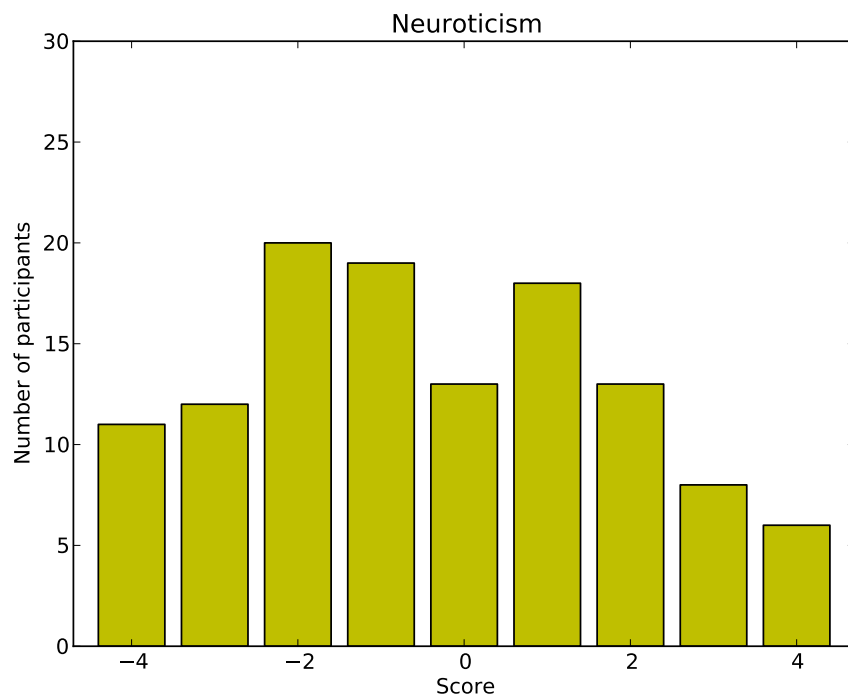


Figure 4.11: The distribution of 120 scores that measure the neuroticism personality trait per participant.

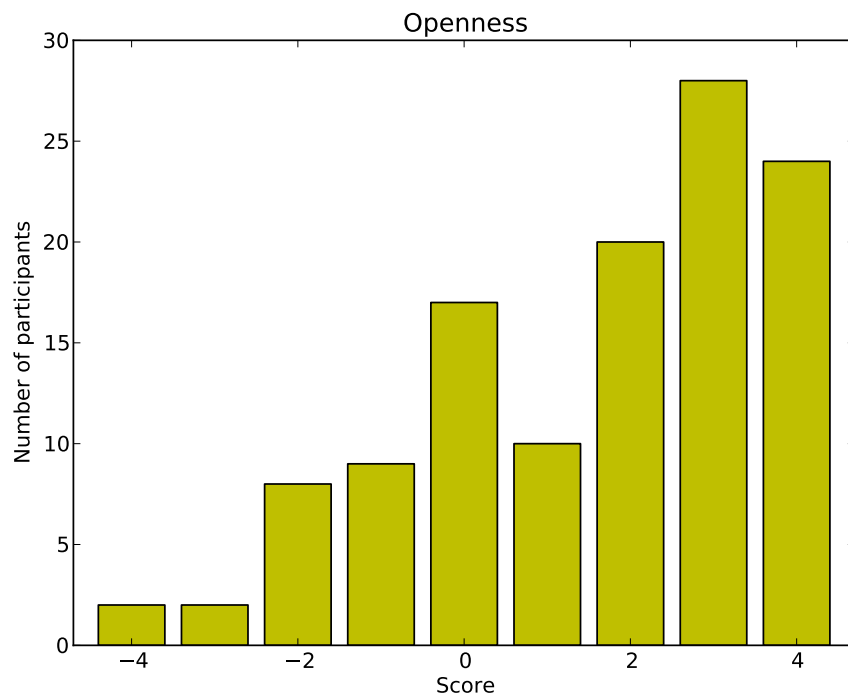


Figure 4.12: The distribution of 120 scores that measure the openness personality trait per participant.

recording methods used to collect a corpus of social interactions.

4.5.1 Sensors and Signals

We used two mobile phones N900 and two SHAKE multi-sensor devices (see Figure 4.13). The telephone microphones could provide adequate quality of audio data and no further audio recording methods are required. A SHAKE device is attached on the back surface of the N900, without impeding the hold of the phone. The SHAKEs record the motor activation (the combined movement of head, hand and the upper body fidgeting) indirectly, via accelerometer, magnetometer and gyroscope.



Figure 4.13: The sensing method: the mobile phone N900 and the SHAKE device attached on the phone.

AUDIO

The phone microphones transmit the audio and, at the same time, GStreamer and PulseAudio [Maemo, 2012] sound systems record it. Hence, each N900 records two audio channels and, after automated audio processing, generates one stereo file:

- the microphone input: what the user says on the microphone
- the speaker output: what the user hears through the speaker

Consequently, two stereo files (four channels in total), correspond to one conversation.

The sampling rate of the audio recording is 44.8 kHz, as it is considered as a standard frequency for audio recording. This is justified by the Nyquist-Shannon sampling theorem, namely, that half of the sample rate should be higher than the maximum frequency of the signal to be sampled, in combination with the fact that human hearing ranges from 20 Hz to 20 kHz.

Table 4.2: Motion Sensors

Sensor	Recordings of phone's:
accelerometer	acceleration
magnetometer	orientation
gyroscope	rotation

MOTION

The motion sensors, accelerometer, magnetometer and gyroscope or angular rate sensor (see Table 4.2) are incorporated in the Sensing Hardware Accessory for Kinaesthetic Expression device (SHAKE) [Hughes, 2010; Hughes and O'Modhrain, 2006], see Figure 4.14.

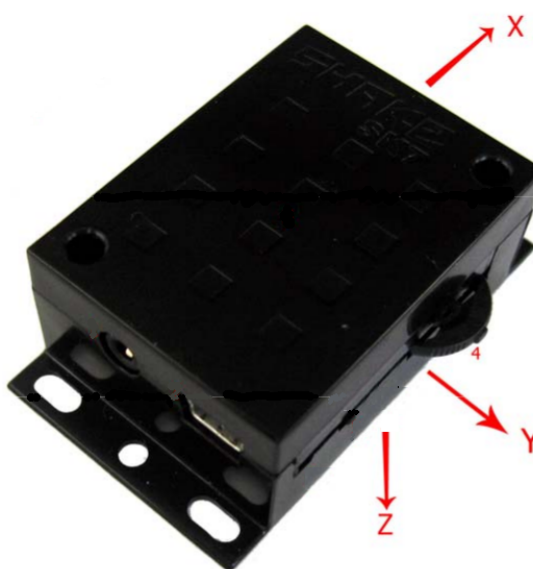


Figure 4.14: The SHAKE: physical features and axes directions.

Accelerometer: Triaxial accelerometers are sensors “that return a real valued estimate of acceleration along x , y and z axes from which velocity and displacement can also be estimated” [Ravi, N. and Dandekar, N. and Mysore, P. and Littman, 2005]. Acceleration stands for the rate at which the velocity of a mass changes during time, or simply a measure of how the velocity of an object changes, and can be positive, with increase of velocity magnitude, or negative, with decrease of velocity magnitude.

The SHAKE accelerometer can measure acceleration magnitudes ranging from $-6g$ to $6g$, with $0.001g$ resolution [Hughes, 2010]. The latter g stands for the acceleration of gravity, and is equal to 9.8 m/s^2 . Vertical acceleration higher than $5g$ can cause loss of consciousness, hence a range $(-6g, 6g)$, with 0.001 resolution is adequate to describe the acceleration of a person in a sitting position talking on the phone.

Magnetometer: Magnetometers are devices that sense changes of magnetic field strength. Earth's magnetic field, creates measurable magnetic disturbances, called magnetic flux densities. The field ranges between approximately from 0.25G to 0.65G, where G stands for a Gauss which measures magnetic field intensity (or magnetic flux density). By comparison, a strong refrigerator magnet has a field of about 100 G [Mag-Lab]. The SHAKE magnetometer can sense magnetic field ranging from -2 to 2 G, with a resolution of 1mG. Therefore, it senses the magnetic field strength in the immediate vicinity of the device, and the changes in it due to magnetic objects, and can be used as a compass to indicate orientation.

Gyroscope: Gyroscopes are devices used to measure angular velocity, the angular displacement of an object in time (degrees/sec), and are primarily utilised for navigation. Hence, triaxial gyroscopes capture the rate of rotation around the axes, called roll for x axis, pitch for y axis and yaw for z axis. Usually, they are used in combination with triaxial accelerometers in order to describe adequately the position and orientation of a body in space, providing a six degree-of-freedom¹ motion tracking system. The SHAKE gyroscope measures angular velocity ranging from to -900deg/s to 900deg/s, with 0.1 deg/s resolution [Hughes, 2010].

The accelerometer records through three different channels the magnitude of acceleration at the three axes (x, y, z) separately. The data format consists of four numbers, representing the magnitude value of x, y, z components and the time stamp. Correspondingly, the magnetometer functions in the same way. The gyroscope also uses three channels to capture roll, pitch and yaw and follows the same data output format.

The aforementioned sensors are embodied in the SHAKE. The sampling frequency for the SHAKE data is adjusted at 68Hz. The latter is the maximum frequency that a SHAKE device can capture. According to the Nyquist-Shannon sampling theorem we can sample movement of frequency equal or less than 34Hz. The work in [Skogstad et al., 2013] investigates the determination of a reasonable cut-off frequency in order to design filters for real-time motion capture application. The work focuses on systems that interpret hand motion for musical interaction, and results that a 15Hz frequency can capture “*all main feature of both rapid and normal motion*” (p.145) [Skogstad et al., 2013]. Therefore, the sampling frequency in our experiment is adequate to capture rapid and normal motion as motor activation of people sitting and talking on the phone.

4.5.2 Synchronisation

Two types of synchronisation issues have arisen: the synchronisation of the two phones and the synchronisation of the different sensing devices (phone and SHAKE). The motion sensors of the SHAKE were already synchronized with one another. Hence, to align all the signals (audio and motion), i.e. to set them start at the same time, we computed the delay of the starting time of recording between the two phones and the delay of the starting time of recording between the phone and the SHAKE.

Two pairs of two audio signals (four signals in total) correspond to the same conversation. The two pairs (that correspond to the same speaker recorded from different phones) include the start-of-recording delay between the phones and, also, the transmission delay, a typical phenomenon in conversations via mobile phones. We cannot eliminate the transmission delay, since it is part of the conversation, however we must remove the start-of-recording delay. To

¹the number of freedom degrees is the number of the independent parameters that define a system in space.

compute the latter we need to take into account the transmission delay as well. First, we calculated the optimal shift to synchronise (using a function that compute mutual correlation) the two audio signals that correspond to the same speaker and are recorded from different phones. Then we calculated their average value, which helps to neglect the presence of transmission delay in our calculation, since we assume it is a symmetric delay at both signals. That means that we assume that we have the same transmission delay in the two pairs of similar signals, but in the first case is added and the second case is subtracted. By calculating the average, we add opposite numbers of same value, so they are mutually eliminated. We shifted one of the signals by the computed delay, that corresponds to the start-of-recording delay between phones, to align the audio signals recorded from the two phones. The audio signals' alignment was crucial, because it provides the way to align the motion signals from both SHAKE devices to the audio signals. The synchronisation of the audio data and the SHAKE data is achieved due to log files for each pair of SHAKE and phone, that provides the starting time and timestamps of the recordings, measured by the same clock. To sum up, the synchronisation of all signals (signals from any phone or SHAKE towards to others) have been achieved. However, we observed two sources of delay that could not be avoided:

- The N900 and the SHAKE device interact through Bluetooth. The interaction between the Bluetooth and the phone at the beginning of the phone call induces a delay. The SHAKE starts recording with a delay from the N900, that ranges to 40 ms.
- The internal delay of the N900, which stands for the time from the action of giving the command to start audio recording, and the actual first instant of recording. This type of delay is not the same for all recording and ranges up to 60 ms.

It is estimated that both delays aggregate a total of 100 ms maximum between audio and motion data. This delay stands for the total delay in the synchronisation of the bimodal recording method this research uses. However, a delay of 1/10 of a second cannot dramatically affect the alignment between the motion and audio data.

4.6 Conclusions

The SSPNet-Mobile Corpus includes an annotated collection of audio and motor activation recordings via smartphones. The 120 subjects were separated into 60 unacquainted dyads and discussed a predefined decision-making scenario. The experimental procedure of the corpus aims at the observation of rich social interaction in engagement, opinion stating and conflicts. However, it also aims at avoiding a trade-off in the spontaneity of the interactions. Besides, there is the fact that the participants are free to interact with the other person without limitations besides the given conversational topics. Furthermore, the corpus has been developed as having a priority to capture human behaviour in the most naturalistic way as possible. The mobile phone is the technology that this work is interested in, in terms of how and whether it influences social interaction. Hence, the setting should focus on that technology, to limit the possibilities of a possible effect to come from another factor. Also, at the same time, we set the most naturalistic setting possible for the needs of our research, meaning the need to set a decision-making scenario. Finally, the corpus prefers a relatively uniform

cultural background, adequate to absorb misunderstandings due to different languages and ease the flow of verbal interaction while using mobile phones.

Moreover, the experimental task was setting the subjects to have a conversation using the phone, while sitting on a chair. This setting simulates an everyday situation of contemporary life. The conversation is with a stranger, which also happens in the real world, and discusses a plane crash, an experience not familiar to the majority of the people. However, this could be a hypothetical conversation topic in real life or a game. The use of the WST is to avoid the need of specific knowledge to participate in the experiment. However, we collected information on previous similar experience that developed survival skills. No participant declared themselves to have developed such skills. The topic of the conversation was predefined, a decision-making scenario. We wanted to collect data following the protocol described in the chapter instead of natural conversations. The latter could provide real-world data, but it is not given that it would always contain argumentation, disagreements, decision-making and, in general, rich social interaction, but mainly, we could not guarantee either equal roles or prior existence of an interpersonal rapport between the interlocutors. These are factors that do not set equal standards for comparison between the conversations. WST provides us with the benefit of collecting similar conversations, that allow the comparison of different behaviours under the prism of the same topic. The SSPNet-Mobile Corpus has developed a simple, but at the same time, unobtrusive recording method. To the best of our knowledge, it is one of the most non-intrusive experimental methods, with minor trade-off in data quality.

However, this work suggests the use of embedded motion sensors in the mobile phone device for the development of similar experimental procedures in the future. That improvement would sidestep technical issues of the synchronisation between the audio and motion data as this work has dealt with (see Section 4.5.2).

Another possible change in our protocol would be to give a specific number of items. We advised the participants to select as much as they believed was adequate, but also, we gave a hint that the less the better. The purpose of this advice was to provoke more disagreements on the usefulness of the items and, consequently, to collect longer conversations rich in opinion and argument. However, instead of the expected effect we observed participants to mark most of the items as useful. They were not worried about the penalties of selecting more or wrong objects; instead they were selecting more objects so as to “save” the survivors. Perhaps a specific number of items would help more people to have longer arguments.

A final suggestion for future work, could be the online release of the two psychological tests. Hence, significantly more samples could be collected, and one could estimate the effect of “positive illusion” phenomenon on this work.

The next chapter describes the annotation scheme of the audio signals collected during the phone calls.

Chapter 5

The SSPNet-Mobile Corpus: Annotation of the Data.

5.1 Introduction

The previous chapter reported how the data of the SSPNet-Mobile Corpus was collected. This chapter, in Section 5.2 describes the annotation of the conversational content of the audio data by topics of conversation, and the annotation of non-linguistic speech by non-verbal behavioural cues. Section 5.3 explains the non-verbal behavioural events that are annotated, and Section 5.4 draws the conclusions.

5.2 Annotation model

The annotation procedure involves the segmentation of audio files into labelled time intervals. The annotation that is described in this chapter has been performed manually by the author of this thesis. Each audio file accounts for a *dialogue* among a dyad and consists of two audio signals (mono channels), derived from the microphone and the speaker of the same phone (see Section 4.5). The segmentation determines intervals accounting for the section of dialogue during which behavioural events occur or *Topics* are discussed and, at the same time, provides temporal length and “position” (i.e. where in the conversation) of the intervals.

Figure 5.1 presents the annotation procedure of the audio signals. I annotated the two streams for each speaker of the phone call at the same time. The timer at the bottom provides the temporal information in the course of the conversation. I used a tier to mark the time-stamps of behavioural events, and a separate tier to time-stamp the *Topics* (see Figure 5.1).

We chose to annotate both audio files that corresponded to the same conversation (see p.45, AUDIO). Hence, our corpus provides annotated audio streams that correspond to the microphone input for both interlocutors and, also, ensures a standardised audio quality for the data. Nonetheless, the double annotation could, additionally, support future investigation of the effect of the transmission delay on human communication with mobile telephones. As mentioned in Section 4.5.2, there is a transmission delay between the devices, which is a common phenomenon on phone calls with mobiles. The SSPNet-Mobile provides

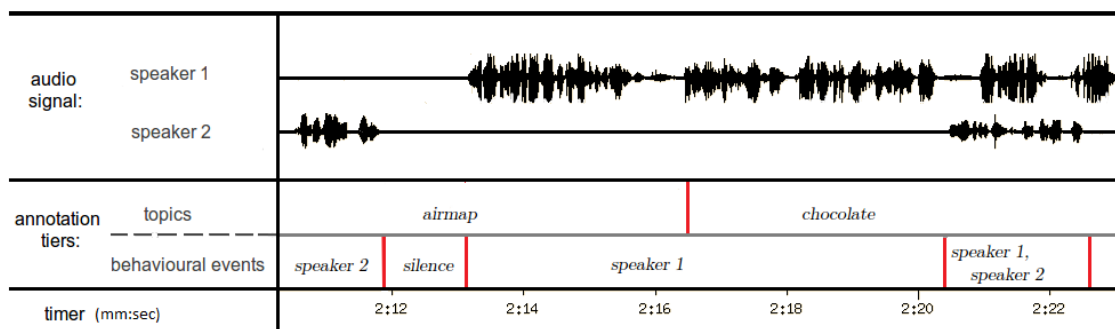


Figure 5.1: The figure depicts the manual annotation model. The two audio streams of each stereo file that correspond to each speaker appear separated. A topic or a turn is determined spatially (see the red vertical lines) at two different tiers and, hence, each topic/turn is assigned to a particular duration and position in the course of the conversation. Then the topic/turn is annotated with a label in order to code the conversational topic/behavioural event that occurs at that moment.

synchronised and annotated audio and motor activation data. The latter could be used for the investigation of a technical failure of one of the most ubiquitous technologies – mobile phones– and its (the failure) effect on human communication.

5.2.1 Behavioural Events

The annotation labels behavioural events after listening, studying and understanding of the phone calls. The events that occur in all conversations, and those that occur most often are annotated. Hence, we annotate the behavioural events that carry the same behavioural content as it is defined below. The behavioural events stand for different events that occur in the course of a dialogue, such as as speaking, not-speaking, laughing, overlapping speech etc. and are different from dialogue acts, since they do not annotate speech acts, for example questions, but non-verbal events occurring during speech. Specifically, the events that are annotated in the SSPNet-Mobile corpus are the following:

- **SPEAKING ACTIVITY:** time interval while only one speaker is talking.
- **LAUGHTER:** time interval during while one or both speakers produce vocalisations, like giggling or laughter.
- **OVERLAPPING SPEECH:** time interval while both speakers are talking.
- **BACK CHANNEL:** time interval during which the listener indicates (with words or vocalizations) that s/he pays attention or s/he is agreed, with no intention to interrupt the speaker.
- **FILLER:** time interval while a speaker indicates uncertainty for any reason, filling the speech with linguistic vocalisation like “*uhm*”, “*eh*”, “*ah*” etc.
- **SILENCE:** time interval where no speech nor vocalisation is occurred.

Laughter is annotated when it is expressed from one speaker or from both simultaneously as a common behavioural event, and corresponds to non-linguistic vocalisations. Back

channel and fillers correspond to linguistic vocalisations. Consecutive behavioural events that occur between two silences and constantly involve the same person, i.e. SPEAKING ACTIVITY or LAUGHTER and FILLER of one speaker, constitute *turns* of that speaker (see p.13). Behavioural events that involve both speakers, i.e. OVERLAPPING SPEECH, BACK CHANNEL or LAUGHTER and FILLER of both speakers, determine *turn-organisation patterns* (see p.13). The annotated behavioural events are extensively described in the following section.

It was remarkable to observe a substantial amount of laughter in phone calls between unacquainted people. A second annotation took place just to annotate laughter, since it was occurring very often, and it could not be neglected from the *Behavioural Event* annotation.

5.2.2 Topics

The *Topics* account for subjects of discussion. The WST (see Section 4.3 and p:1-2, Appendix A) sets twelve subjects of discussion in specific order, the twelve items of the list (see p.3, Appendix A). Specifically, the conversational topics are:

- ITEMS: time intervals while, one out of the 12 items is discussed. The interval is labelled by the name of the item: “steel wool”, “axe”, “pistol”, “butter”, “newspaper”, “lighter”, “clothing”, “canvas”, “airmap”, “whisky”, “compass” or “chocolate”.
- OTHER: time interval while the participants talk about anything other than ITEMS and labelled as “Other”. In four out of sixty conversations, there were discussions about planning a strategy. These intervals were labelled as “Strategy” instead of “Other”.

As mentioned in Chapter 4, the conversation includes 12 topics, one for each object. The segmentation into topics provides the SSPNet-Mobile corpus with conversations where each one can be segmented into 12 discussions, with a potentially conflicting conversational topic. The corpus also provides the personal opinion of the interlocutors about the topic-object before the interaction, and, additionally, the final outcome of each discussion (see Section 4.3 for a description). Hence, the SSPNet-Mobile corpus includes data that could be interesting for a wide range of research on social interactions and the prediction of their outcome.

5.3 Annotated Behavioural Events

This section presents a detailed description of the behavioural events that are annotated in the SSPNet-Mobile corpus. It explains their function in social communication, their meaning and the reasons they are annotated in this work.

5.3.1 Speaking Time

According to Sacks et al. [1974] turn distribution in small groups is an index of power, influence etc. Mast [2001]’s work investigates the dominance in relation to speaking time across genders, and “*results of this study showed that the amount of time talked was positively associated with being perceived as dominant in all-men as well as in all-women groups and*

those associations did not differ in all-men compared to all-women groups.” (p.553). Mast [2002] presented a detailed investigation of the relationship between emergent leadership and speaking time. The study concludes that *“speaking time is usually not the only source of information about dominance [..]. The strong relationship between speaking time and dominance seems to suggest that speaking time may be the most important factor in expressing and inferring dominance.”* (p.445). Finally, Bales [1970] makes an interesting comment about speaking time and persuasion, and explains that the person who talks most may not always be the winner for his ideas in an interaction; however, he prevents others from explaining their thoughts to the same extent as him.

In SSPNet-Mobile Corpus the term “speaking activity” refers to the total speaking time of an individual. The annotation into turns provides the calculation of the total speaking time per speaker.

5.3.2 Laughter

Laughter is defined as a non-verbal vocalisation, accompanied by an associated facial expression [Vettin and Todt, 2004]. Morreall [1982] reviews theories of laughter and reports that it expresses a range of affective states such as amusement, joy, embarrassment, scorn. The same work argues that the oldest theory considers laughter as *“an expression of a person’s feelings of superiority over others”* (p.243) [Morreall, 1982], deriving from Plato [a,b,c,d] and Aristotle [a,b]. At this point it would be useful to remind the reader that this work examines and annotates laughter, and not *humour*. Provine [1993] describes the typical acoustic structure of laughter thus: *“the simple, highly stereotyped acoustic structure of laughter is characterized by one or more forcibly voiced, acoustically symmetric, vowel-like notes (75 ms duration) separated by regular intervals (210-218 ms), and a decrescendo”* (p.291).

Coates refers to past work by Jefferson et al. [1978] and explains that laughter is considered as *“an official conversational activity”* (p.44) [Coates, 2007] and not *“just an accompaniment to talk. It is talk.”* (p.44) [Coates, 2007]. Adelswärd [1989] describes the functional aspect of laughter as *“an integrated phenomenon in spoken interaction, with a special function of modifying the meaning of utterances”* (p.108), and emphasises the role of laughter, and the reason why it has to be taken into account during a conversational analysis.

Laughter is a social event, namely it involves more than one person. In the case of one laugh alone, the stimuli of laughter comes from an external source, an incident of which the laugher is an observer, e.g., a random sound [Vettin and Todt, 2004]. Conversational laughter is a social act *“by which you can display an attitude towards your interlocutor as well as towards what you are talking about”* (p.129) [Adelswärd, 1989] and it can be mutual or unilateral, conveying a different message [Provine, 1993; Adelswärd, 1989]. Specifically, *“mutual laughter is a sign of rapport and consensus, unilateral laughter often used to modify verbal expressions or attitudes”* (p.129) [Adelswärd, 1989]. Also, *“sharing laughter displays mutual co-orientation towards the laughable object, action or utterance and, also, affiliation of the laughs with each other”* (p.139) [Glenn, 1991]. In addition, mutual laughter maintains *“the flow of interaction and the interest and attention of the conversational partner”* (p.108) [Vettin and Todt, 2004]. Therefore, the importance of mutual laughter occurrences during a conversation is that they can give a indication of coordination according to O’Donnell Trujillo and Adams [1983], and intimacy, according to Coates [2007].

In the SSPNet-Mobile Corpus, we have annotated as laughter events parts of the conversation, during which linguistic or non-linguistic outbursts were performed expressing amusement,

joy, embarrassment or scorn, and following the common sound of laughing or giggling.

5.3.3 Overlapping Speech

Overlapping Speech is defined as talking activity by “*more than one at a time*” (p.11) [Schegloff, 2000], i.e., speech that consists of simultaneous outbursts coming from different speakers. In the course of conversation, different instances of overlapping speech could emerge, such as rude interruption, a delayed completion of a sentence by the prior speaker, a sentence’s co-completion, a simultaneous consensus to a conclusion, a simultaneous linguistic or non-linguistic expression of surprise or reminiscence, a simultaneous attempt to grab the floor after a silence or a listener’s outburst, whether linguistic or not [Sacks et al., 1974; Lerner, 1989; Goldberg, 1990; Lerner, 1996a,b, 2002]. Apparently, based on the utilisation of overlapping speech that the literature indicates above, simultaneous talk may be destructive or constructive to the conversation and, also, it can be more than an attempt to claim a turn. According to Goldberg [1990], the hallmark to label an instance of overlapping speech as interruption is the intention of gaining the floor from the prior speaker. Furthermore, in the same research, interruption phenomena are grouped in power-oriented and rapport-oriented categories. Power interruptions involve speakers in divergent goal orientation, regardless of their partner’s interest, i.e., the speaker wants to be listened to, ignoring listener needs, or the listener wants to break off the speaker, by inserting incoherent remarks. Rapport interruptions involve the speakers in mutual, shared overlapping goal orientation, whether or not they have the same approach to reach the common goal, i.e., the speaker wants to be interesting to the listener, the listener wants to make cohesive comments. Also, Goldberg assigns power-interruptions to a new topic introduction, and rapport interruption to maintenance of the same topic.

Bennett [1981], referring to public lectures of Schegloff, defines the distinction between overlap and interruption, as it is driven by completion point: “*By overlap we tend to mean talk by more than a speaker at a time which has involved that a second one to speak given a first was already speaking, the second one has projected his talk to begin at a possible completion point of the prior speaker’s talk. [...] If it’s projected to begin in the middle of a point for the turn, then we speak of it as an interruption*” (p.172) [Bennett, 1981]. Duncan [1972], describes the completion point as a constellation of turn-yielding cues, and claims that it be displayed as “*a set of six discrete behavioral cues*” (p.286) [Duncan, 1972], the following: intermediate pitch level at the end of phonemic clause, drawl on the final syllable, termination of any hand gesticulation, appearance of stereotyped expression, e.g. “*but uh*”, drop of paralinguistic pitch/loudness and completion of grammatical clause, e.g., subject-predicate combination.

Overlapping speech constitutes a break-down of a turn-taking system, but can indicate precision tracking of the emerging course of an utterance [Goodwin and Heritage, 1990]. For example, shifting turns or choral co-production can exhibit understanding, affiliation and agreement among speakers or reciprocal reciprocity in the course of talking [Lerner, 2002]. According to Robinson and Reis [1989], interruption – including all types of overlapping speech which involve a statement– leads to negative personality traits, such as lack of sociability and assertiveness. Also, Goldberg [1990], reports that power interruptions are “*heard as rude, impolite, intrusive and inappropriate; conveying the interruptor’s antipathy, aggression, hostility etc.*” (p.890) [Goldberg, 1990], while rapport interruptions “*are generally understood as expressions of open empathy, affection, interest, concern etc.*” (p.890) [Goldberg, 1990]

and can be perceived as displays of collaboration, cooperation and mutual orientation by assisting the interrupter with immediate feedback or extra information. Interruptions are always overlapping speech, but overlapping speech is not always an interruption [Lerner, 1989].

A very common instance of overlapping speech is back channel. This phenomenon has been described as *listener response* [Dittmann and Llewellyn, 1968; Duncan, 1972] and is defined as the linguistic, e.g. “yes”, “right”, or non-linguistic, e.g., “uh huh”, outburst of the listener, indicating attention and interest *without any intention* to grab the floor.

To sum up, overlapping speech conveys both positive and negative meaning. In the case of indications by the prior speaker for turn completion, and/or the goal of the speaker-listener is mutually oriented, overlapping speech is constructive. On the contrary, claiming the floor without any turn-yielding cues and/or no parallel conversational goals of the dyad, could convey lack of cooperation, and express negativity and low level interpersonal interaction. Back channel phenomena are discussed more extensively in the following paragraph.

In SSPNet-Mobile Corpus, overlapping speech annotated events stand for parts of conversation where both speakers speak at the same time. Moreover, events which involve both speakers while not both performing speaking activity, e.g. if one is laughing and one is speaking, or one is laughing and one is producing a filler, are not taken into account as overlapping speech. In a similar vein, with the rest of behavioural events this work focuses on the phenomenon of simultaneous speech production from both speakers, and does not take into account the linguistic information.

This paragraph explains the different functions of overlapping speech, such as overlap and interruption, and underlines the value of overlapping speech during conversations. However, the annotation of this work groups both events, overlaps and interruptions, into overlapping speech phenomena and proposing further distinction, as future work.

5.3.4 Back Channel

Back channel refers to short messages such as “yes” and “uh huh” which the person who has the turn receives without relinquishing his/her turn, and is considered as a response to the speaker, conveying attention and willingness to sustain the listener’s role, with no bid to grab the floor [Maynard, 1986; Yngve, 1970]. A back channel continuer is always an acknowledged move, e.g., “yes”, “right” and as such it does not constitute a turn [Clark and Schaefer, 1989; Cathcart et al., 2003].

The function of the back channel phenomenon is highlighted in Ward [1996], where “*one important component of responsiveness is back channel feedback*” (p.1728) [Ward, 1996] and investigates when it is appropriate. Back channel can express attention, encourage the speaker to continue, a positive mood, acknowledgement, interest but also an indication of wanting to speak [Duncan, 1972; Ward, 1996; Ward and Tsukahara, 2000; Benus et al., 2007; Gravano and Hirschberg, 2009]. Drummond and Hopper [1993]’s work separates back channel continuers into two groups; one of back channel responses which indicates *speakership incipiency* expressed by “yeah”, and one group of back channel indications with *passive reciprocity* function, through “uh huh” and “mm hm” utterances. Therefore, back channel is a cue which not only conveys intention to interrupt, but in specific forms is adequate to inform the speaker about the listener’s mood when the latter is ready to receive the speakership.

Back channel is a responsive action and as such, should follow some signs in order to occur. In

the course of conversation, the speaker signals possible back channel opportunities, producing unfilled or filled (fillers) pauses or dropping the pitch across final syllables [Duncan, 1972]. In SSPNet-Mobile Corpus, back channel events are annotated as the linguistic, i.e. “yes”, “yeah” or non-linguistic, i.e. “mm mhm”, “uh huh” utterances of the listener during the speaker’s turn. We observed incidents of back channels and filled pauses (fillers) to occur in parallel, as the literature argues [Duncan, 1972]. In the next section fillers are discussed more extensively.

5.3.5 Fillers

Fillers, or filled pauses, are non-linguistic outbursts like “uh”, “um”, “ehm” etc., which substitute words expressing uncertainty, but are also willing to sustain the floor [Wennerstrom and Siegel, 2003]. They are also called filled pauses, because their presence constitutes of a pause, meaning no presence of words, filled by a non-linguistic sound instead of a word [Clark and Treeb, 2002]. Moreover, the same study argues that due to the fact that “uh” and “um” stand for English words, filler constitutes a proper and adequate naming for the aforesaid outbursts.

Fillers hold a twofold function: to keep the floor without speaking [Sacks et al., 1974; Wennerstrom and Siegel, 2003] and to signal processing problems or high cognitive load experiencing [Clark and Wasow, 1998; Maatman et al., 2005]. Therefore, unfilled pauses usually follow filler occurrences. Clark and Treeb [2002] argue that “uh” suggests a minor delay before speaking, while on the contrary, “um” suggests that a major delay precedes speaking.

In SSPNet-Mobile Corpus, fillers are annotated utterances like “uh”, “um”, “ehm”, “ah”, which assist the speaker to gain time for thinking, securing the floor at the same time. As mentioned before, frequently in our corpus, fillers occurred at the same time at back channels events. Obviously, the speaker wants to hold the floor, and the listener accords, performing a back channel, setting himself again as a listener. This type of synchronised interaction during a conversation indicates mutual attention, coordination and positive rapport among the interlocutors, who behave respectfully towards the goals of one another (see overlapping speech section) [Goldberg, 1990]. In addition, filled pauses are signals to the listener that the speaker is experiencing processing problems, and frequently elicit “take your time” feedback from the listener, which is expressed by the listener with back-channels [Cassell et al., 2000; Maatman et al., 2005].

5.3.6 Silence

Silence is defined as the absence of sound. However, even if the definition of silence is straightforward, different types of silent events have been detected in past conversational analysis studies. Silence is classified according to the status of the silence-ensuing speaker. Therefore, if the speaker who precedes and follows the silent event is the same, the event is a pause. On the contrary, if the silence-ensuing speaker has changed, the silent event is classified as a gap, short silence, or a lapse, a longer silence [Heldner and Edlund, 2010; Sacks et al., 1974].

A silent event could occur in a conversation intentionally or unintentionally [Kurzon, 1995]. In combination with Verschueren [1985]’s work, Kurzon [1995] classifies the causes of unintentional silences as indecisiveness on what to say next, incapability out of amazement, grief, or other strong emotion, an indifferent speaker and the case that speaker has forgotten what

s/he was going to say. On the contrary, intentional silence could appear due to “temperamental” disinclination of the speaker to talk, or the speaker concealing something. The case of nothing to say could be classified either as intentional or intentional silence.

The meaning of the presence of silence in conversations is strongly influenced by the context. Saville-Troike [1985], considers silence to be “more context-embedded than speech”, while Newman [1982] correlates silence with relationship context. For example, occurrences of silence during a conversation between acquaintances could mean “interpersonal incompatibility and awkwardness”; however, in more intimate relationships, silences do not induce the same perception [Newman, 1982].

In the SSPNet-Mobile Corpus, the silent parts of conversation between turns, where it is explicit that the floor do not belong to any speaker, are annotated as silence. In the case of unclear possession of the floor during a silent part, the latter is included in the turn of the speaker who precedes it. To compare our annotation scheme to Sacks et al. [1974], pauses are not annotated in our corpus, as we considered them to be part of a speaking turn, while gaps and lapses stand for the definition of silence in our study.

5.4 Conclusions

This chapter described the annotation scheme according the occurrences of behavioural events in the audio recorded phone calls of SSPNet-Mobile Corpus. In addition, due to the WST scenario, the segmentation of the conversations in *Topics* results in 12 different discussions that end in consensus. The following chapter explains how our annotation scheme supports our research in factors that influence decision making and persuasiveness.

Chapter 6

The “Caller-Receiver Effect” on Negotiations Using Mobile Phones.

6.1 Introduction

Previous chapters describe how the WST scenario induces a decision-making process (see 4.3) and how the annotation scheme supports conversational topic coding. This chapter investigates the effect of the use of mobile phones on the social interaction. Specifically, regarding how the phone use influences the outcome of the negotiations in cases of disagreement during the decision-making process. The rest of the chapter is organized as follows: Section 6.2 presents previous work on the effect of gender, personality and conflict handling style on negotiation outcome and behaviour change phenomena meaning the effect on behaviour when technology is included in an activity consciously as “*persuasive technologies*” [Fogg, 2002] or unconsciously as the medium of communication, e.g., only audio, only text etc. Section 6.3 reports the investigation on *Caller-Receiver* role as the only influential factor on the final outcome and persuasiveness and Section 6.4 shows some conclusions of the chapter.

6.2 Previous Work

Section 3.2 reviews previous work on how negotiation outcomes change when using different communication media such as handwritten messages, e-mail, chat, telephone, video-conference and face-to-face interactions. The rest of this section considers in particular works on the effect of gender, personality and conflict handling style, three factors analysed in this work as well. Furthermore, it reviews previous work on the phenomenon of behaviour change due to technology presence in long-term activities, such as healthier every-day activities, or in short-term interaction such as persuasiveness using different medium of communication.

6.2.1 Negotiation and Interpersonal Relationships

Since negotiation is an inherently social activity - it cannot take place unless there are at least two parties - several investigations were dedicated to the effect of interpersonal factors, i.e. to

“the ways that negotiators’ behaviour and outcomes depend upon the presence of the other party or parties [...] and the dyadic aspects of negotiation behaviour” [Thompson et al., 2010].

Power differences between negotiators, real or perceived, were the subject of Magee et al. [2007]. In the experiments of such work, 38 participants acting in a bargaining scenario, people in a higher power position were shown to have higher propensity to initiate a negotiation. Furthermore, other experiments presented in the same work (involving 62 subjects), show that the tendency to make the first move is beneficial from a bargaining point of view. Hence, negotiating with lower power counterparts appears to be an advantage. The perception of power differences seems to explain some observed gender effects as well (negotiators tend to accept the stereotypical view of women possessing less power than men) [Kray et al., 2004]. However, extensive meta-analyses contradict such a result and show that gender, overall, has a limited effect on negotiation outcomes [Walters et al., 1998; Stuhlmacher and Walters, 1999]. The analysis of Barry and Friedman [1998] focused on negotiators’ Big-Five personality traits, in particular Extraversion, Agreeableness and Conscientiousness (see [Rammstedt and John, 2007b] and Section 6.3.3 for more details). The experiments involved 184 dyads (386 subjects in total) working on a distributive bargaining task, i.e. on the distribution of a finite amount of resources across the dyad members. The results show that higher Extraversion and Agreeableness tend to be associated to lower gains, i.e. to lower effectiveness in maximizing personal advantages. Furthermore, Conscientiousness appears to have no significant effect. The way participants frame the interaction with the other parties is one of the main factors influencing negotiation outcomes [Bazerman et al., 2000]. The experiments of Pinkley and Northcraft [1994], performed over 75 dyads (150 subjects in total), show that the outcomes tend to be less satisfactory for people that frame negotiations in terms of relationship (focus on interpersonal concerns) and winning (focus on maximization of personal profit). Vice versa, negotiation outcomes tend to be more favourable for individuals framing their interactions in terms of task (focus on material aspects of the negotiation) and cooperation (focus on maximization of joint profit). The effect of conflict handling style (see [Rahim, 1983] and Section 6.3.4 for more details) was investigated in Cheung et al. [2006]. The analysis of questionnaires gathered from 70 construction professionals indicate that adopting an integrating conflict style (tendency to maximize joint satisfaction of all parties) tends to achieve functional outcomes, while the adoption of a compromising style (tendency to find a trade-off between conflicting outcomes) is helpful to resolve disputes.

The findings above concern face-to-face negotiations and do not take into account the effect of the communication media. According to the works surveyed in Section 3.2, the phones tend to hide attitude and behaviour of negotiation participants. Hence, it can be expected that personal characteristics like personality and conflict handling style, typically manifested through attitude and behaviour, should not have a major influence on phone mediated negotiation outcomes.

6.2.2 Technology and Behaviour Change

Users react to technologies displaying human-like behaviour as if these were actually human [Nass and Brave, 2005]. The reason is that unconscious cognitive processes evolved in absence of technology and, therefore, cannot distinguish between natural and artificial human behaviour [Nass and Min Lee, 2001]. The phenomenon, known as *“Media Equation”* [Reeves and Nass, 1996], is the basis for the development of *“persuasive technologies”* [Fogg, 2002],

i.e. machines that aim at changing beliefs, attitude and behaviour of their users towards a desired, predefined direction.

Mobile phones, with their ubiquitous presence in everyday life, are an ideal platform for persuasive technologies [Lathia et al., 2013]. Several works (e.g., [Aharony et al., 2011; Chiu et al., 2009; Gasser et al., 2006]), show that mobile applications can help people to adopt healthier lifestyles not only by suggesting health oriented practices, but also by supporting social pressure mechanisms, one of the main techniques psychologists adopt to foster behavior change [Abraham and Michie, 2008]. In Aharony et al. [2011], experiments involving 130 subjects show that people can be persuaded to perform more physical activity if other subjects, connected through a mobile application, pay for their lack of movement. In Chiu et al. [2009], mobile phones were used to measure the water intake of 16 subjects and to perform a game where drinking more water allows one to achieve better scores. Furthermore, the same mobile phones support the competition with other players towards better water drinking practices. The results show that the subjects actually drink more water especially when the game is social, i.e. it includes competition with others. In the same vein, the experiments of Gasser et al. [2006] show that mobile applications help 40 people to adopt healthy nutrition styles through social facilitation supported via the phones.

Other works, focused on the way communication technologies change the interaction between people [Dourish and Bell, 2011; Harper, 2010; Ling, 2008] and modify, to a measurable extent, interaction outcomes [Bradner and Mark, 2002; Cardy, 2005; Mohammadi et al., 2013]. The results of Bradner and Mark [2002] show that people collaborating at distance via different communication technologies (e.g., videoconferencing and Instant Messaging) can be persuaded more easily when they think that their interlocutor is geographically closer. In the case of Mohammadi et al. [2013], the experiments show that people are more persuasive when they communicate via video than when they do it via audio or text. In the case of Cardy [2005], the experiments show that political propaganda does not persuade electors when these are reached via phone.

The findings surveyed in this section illustrate the interplay between persuasiveness and technology, whether this means to produce persuasive artefacts or to change the persuasiveness of people that adopt technology to interact. The experiments of this work focus on the latter aspect and it is possible to expect that the very use of phones produces changes in persuasiveness.

6.3 The Caller-Receiver Effect

Chapter 4 describes the experimental process applied for the collection of SSPNet-Mobile Corpus. At the end of the process two WST-solution documents are collected from each dyad (see Appendix A) which provide three types of information about decision-making:

1. the personal decision of the dyad members, i.e. the initial opinion about the solution of the WST *before the call* (one decision per member per item) and
2. the consensual decision about the solution of the WST resulting from the discussion *during the call* (one decision per item).

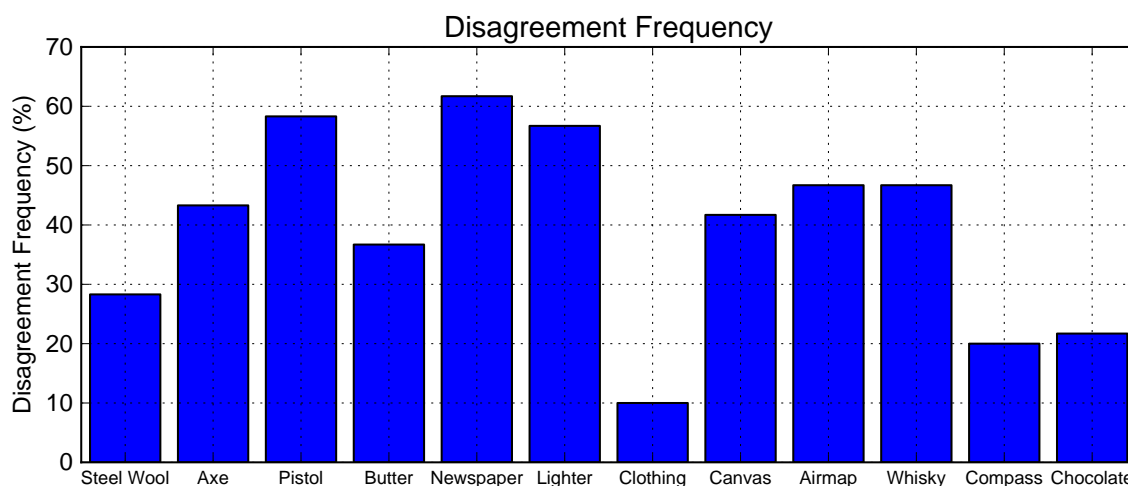


Figure 6.1: The plot shows the percentage of calls where there was a disagreement on each item.

The comparison of the personal decisions of the dyad members, can show explicitly whether there was agreement or disagreement on each item. In case of agreement, meaning same personal decisions, a tie is reported. In case of disagreement, meaning different personal decisions, the winner of the discussion about an item is the person whose personal decision is the same as the consensual one. Hereafter we will use the term winner and loser to refer to the description above.

The dataset includes 60 phone calls and, in each of them, the subjects must make 12 decisions (one per item). Thus, the total number of decisions made during the experiment is $60 \times 12 = 720$. In 437 cases (60.7% of the total), the subjects have made the same decision when they performed the task individually before the call (third step of the protocol outlined above). In these cases, the subjects briefly discuss the item just to confirm their decision. In the remaining 283 cases (39.3% of the total), the subjects have made different decisions before the call and, hence, they need to discuss in order to reach consensus. In these cases, one of the two subjects must necessarily persuade the other. The persuasiveness can be measured at two levels. At the *item level*, we consider the decision on each conflicting item as the variable of interest. We report the percentage of items where subjects of one category convince the others about their decision. At the *call level*, we aggregate the result for each call. The most persuasive subjects are those that convince the other in the largest number of items in the call. The subjects get the same information (see Appendix A), undergo the same protocol (see Section 4.3), use the same phone model and, by scenario design, hold similar competences and skills about the WST. However, at the item level, the subjects that *receive the call* persuade those who *make the call* significantly more frequently than the other way around, namely 59.0% of the times (p -value 0.003 according to a two-tailed binomial test). The effect is even more evident at the call level, the subjects receiving the call are the most persuasive in 70.0% of the 51 conversations that do not end with a tie (p -value 0.005 according to a two-tailed binomial test). In Figure 6.2 the results are depicted by level (item, call) as cumulative bars of win percentages per *Caller* and *Receiver*.

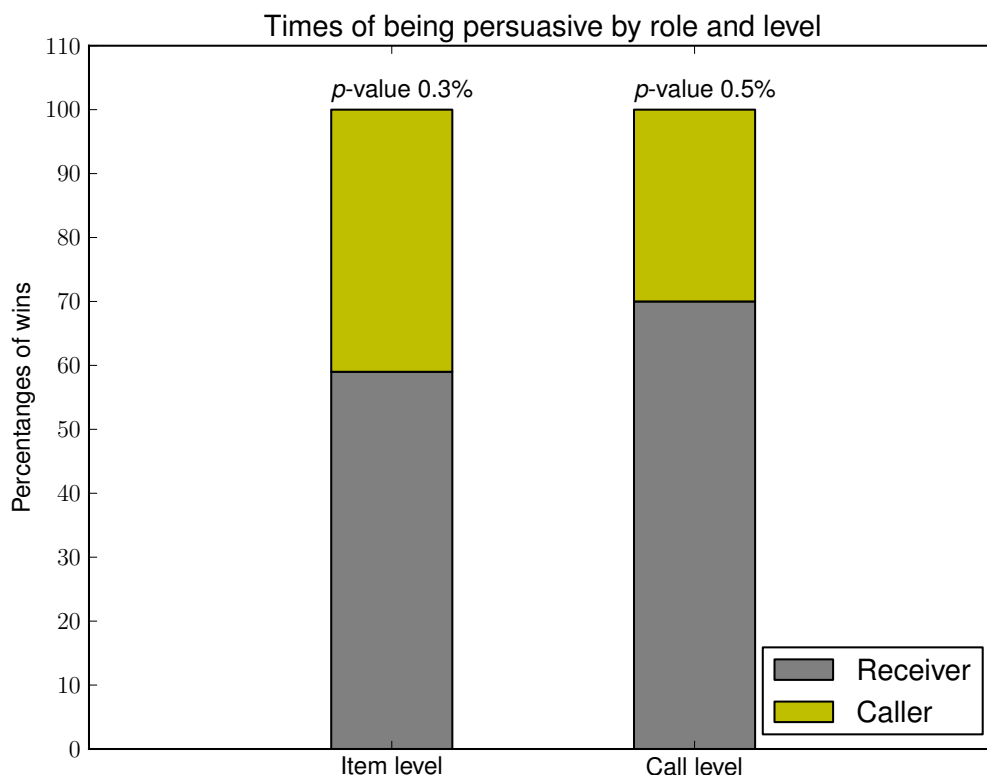


Figure 6.2: The percentages of times the “Caller” and “Receiver” impose their opinion, win the conflict and persuade the other. The *Receiver* is significantly more persuasive than the *Caller* in both item and call level.

According to the experimental protocols adopted in most negotiation studies [Purdy et al., 2000; Sheffield, 1995; Suh, 1999; Valley et al., 1998], each subject participates only in one call. Therefore, it is not possible to verify whether a change in role for the same subject (e.g., from caller to receiver) corresponds to a change of persuasiveness as well. However, this ensures that the subjects are not familiar with the task, a condition that eliminates effects and biases difficult to assess.

The results seem to be in line with the indications of Section 3.2 showing that the phones make it easier for non-cooperative negotiation participants to maximize their individual profit (there is one winner in 51 calls out of the total 60). However, this does not explain why receivers should be more effective than callers. Therefore, it is necessary to verify whether the result is the effect of other factors known to make a difference in negotiations (age, gender, personality and conflict handling style).

6.3.1 Gender Effects

The data includes 31 calls where participants have different gender and the results of this section (see Table 6.1) apply to them. In these conversations, men persuade women 52.0% of the 125 times that consensus about an item must be reached through discussion (p -value 0.72 according to a two-tailed binomial test). At the call level, men appear to be most persuasive in 60.0% of the cases, but the p -value of such an observation is 0.42 (according to a two-tailed

	Total	Item Level	Call Level
Female	63 (52.5%)	48.0%	60.0%
Male	57 (47.5%)	52.0%	40.0%

Table 6.1: Gender effects. The table reports the gender composition of the subjects (“Total” column) as well as the persuasiveness of male and female subjects at both call and item level. According to a two-tailed binomial test, the p -value is higher than 0.42 for all persuasiveness figures.

binomial test), well above the acceptance level of 0.05. In other words, *gender effects, if any, are too moderate to produce observable effects on discussion outcomes and persuasiveness*. The result is in line with previous findings of the literature (see Section 6.2.1) showing that gender, overall, does not influence significantly the outcome of negotiations [Stuhlmacher and Walters, 1999; Walters et al., 1998]. Furthermore, the results show that gender cannot explain the persuasiveness difference between callers and receivers because men and women are evenly distributed across these two conditions. In other words, gender cannot be considered an explanation of the persuasiveness difference between callers and receivers observed in the experiments.

6.3.2 Age Effects

The upper plot of Figure 6.3 shows the age distribution of the experiment participants. The subjects are between 18 and 64 years old (average and standard deviation are 28.9 and 12.2, respectively). In 35 calls, the age difference is lower or equal to 10 years (see lower plot of Figure 6.3), but in the other cases the age difference goes up to 44 years. Since the subjects are unacquainted and do not meet before the call, it is probably difficult for them to estimate how old is their counterpart. However, the age difference might still have an effect on the negotiation outcomes.

At the item level, the younger participants of each call win in 54.1% of the cases (p -value 0.999 according to a two-tailed binomial test). At the call level, the percentage of successes for the younger participant is 51.0% (p -value 0.190 according to a two-tailed binomial test). Therefore, *age effects, if any, are too moderate to produce observable effects on the outcomes of the negotiation*. Furthermore, in the 55 calls where the participants are of different age, the older subject is the receiver in 52.7% of the cases (p -value 0.787 according to a two-tailed binomial test). In other words, older and younger subjects of each call are distributed evenly across callers and receivers. Therefore, the age difference cannot be considered an explanation for the higher persuasiveness of the receivers.

6.3.3 Personality Effects

Differences in personality might explain persuasiveness differences between the individuals that call and those that receive, especially if people with certain traits tend to be more frequent in one of the two conditions. For this reason, before the experiment, all subjects filled a personality questionnaire [Rammstedt and John, 2007b] based on the Big-Five model, a personality representation relying on five traits known to capture most of the individual differences:

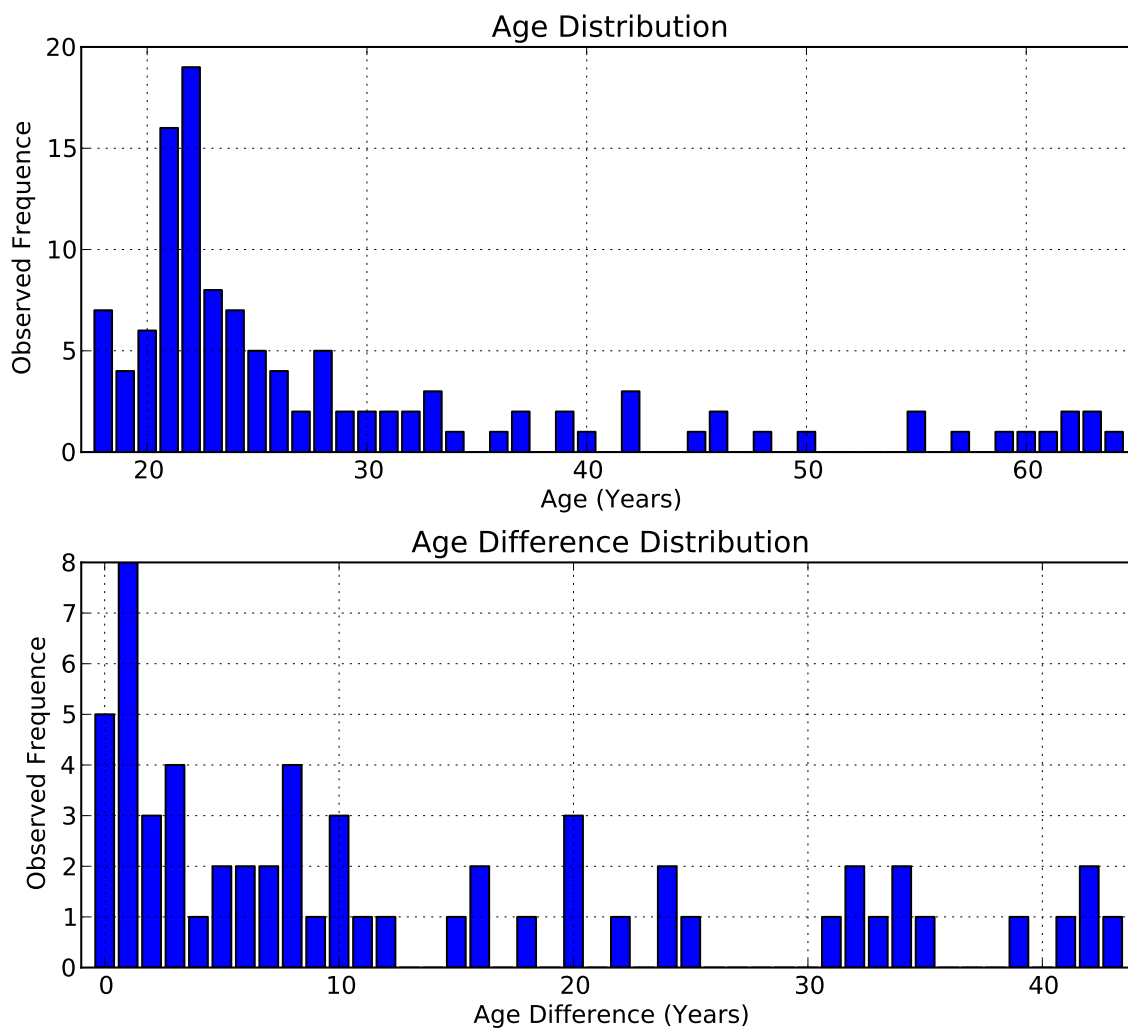


Figure 6.3: The upper histogram shows the age distribution across the 120 participants, the lower histogram shows the age difference distribution across the 60 calls.

- *Extraversion*: Active, Assertive, Energetic, etc.
- *Agreeableness*: Appreciative, Kind, Generous, etc.
- *Conscientiousness*: Efficient, Organized, Thorough, etc.
- *Neuroticism*: Anxious, Self-pitying, Tense, etc.
- *Openness*: Artistic, Curious, Imaginative, etc.

The analysis of the questionnaires provides five scores, one per trait, that account for how well the adjectives above describe the personality of each subject. If a trait influences the outcome of the discussions, the subjects that score higher along such a trait - in a pair of persons involved in the same call - should tend to persuade (or to be persuaded by) their interlocutor significantly more frequently. The plot of Figure 6.4 shows, for each trait, how frequently the dyad members with the higher trait score are more persuasive at item and call level. In all cases, the p -value is higher than 0.3 (according to a two-tailed binomial test). Hence, *the effect of personality traits, if any, is too small to produce observable consequences in terms*

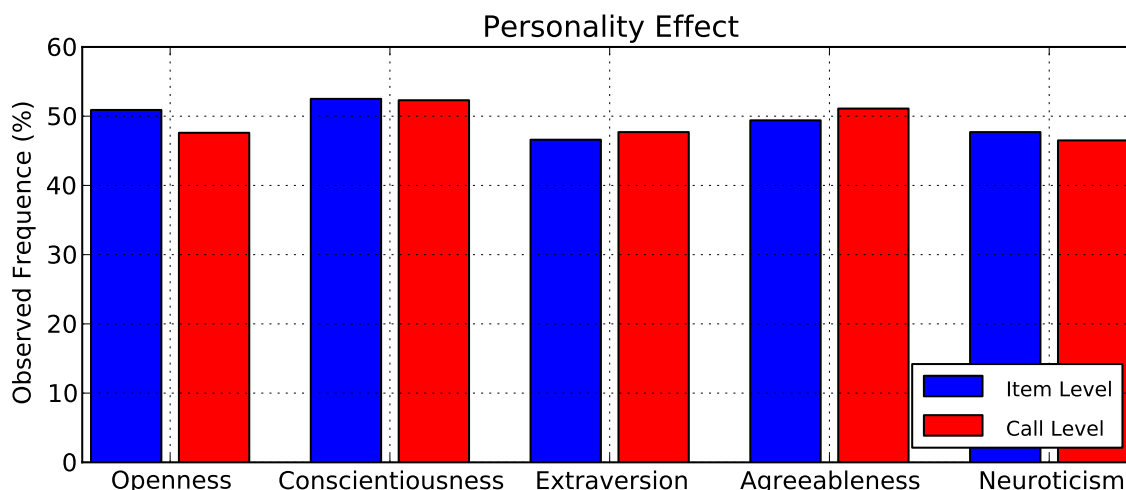


Figure 6.4: The plot shows the percentage of times the subject with the higher score along a given trait is the most persuasive at the item and call level. The p -value is always above 0.3 (according to a two-tailed binomial test).

of persuasiveness. Furthermore, people that score higher on a given trait distribute evenly across callers and receivers and possible effects would still not explain the persuasiveness gap between the two conditions.

6.3.4 Conflict Handling Style Effects

Conflict is a “*mode of interaction [where] the attainment of the goal by one party precludes its attainment by the others*” [Judd, 1978]. When there is disagreement, the two participants of each call start from opposite decisions, namely “Yes” and “No”. Since the decision has to be consensual, if one of the two positions wins, the other necessarily loses. In other words, the subjects involved in one call pursue, in case of disagreement about an item, incompatible goals and are in conflict. For this reason, the subjects filled, before the experiment, a questionnaire about their conflict handling style, i.e. the behavioral strategy they tend to adopt when involved in competitive discussions [Rahim, 1983]. In fact, the way subjects deal with conflict might influence the discussion outcomes and explain the persuasiveness differences observed in the experiment (see Section 6.2.1). The questionnaire provides five scores that measure how well the behavior of a subject matches one of the following tendencies:

- *Avoiding*: individuals tend to accept any, possibly unfavourable outcomes proposed by others to avoid conflict or unpleasant interaction.
- *Compromising*: individuals tend to find a trade-off between all, possibly incompatible outcomes proposed by different actors.
- *Dominating*: individuals tend to impose the outcomes they propose while rejecting those proposed by other actors.
- *Integrating*: individuals tend to find outcomes favourable and satisfactory for most, possibly all actors involved in the discussion.
- *Obliging*: individuals tend to address needs and suggestions by others at the cost of accepting, if necessary, unfavourable outcomes.

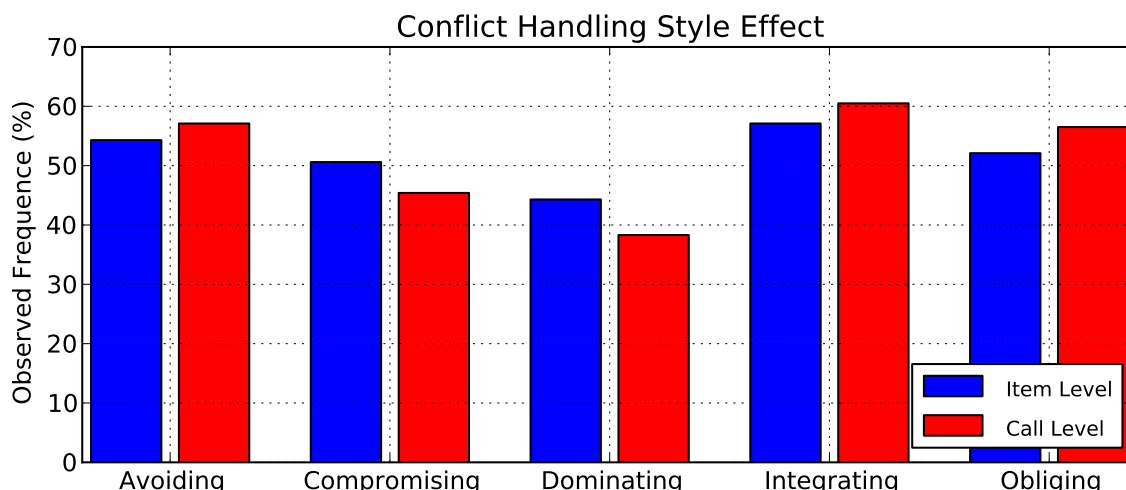


Figure 6.5: The plot shows the percentage of times the subject with the higher score along a given conflict handling style is the most persuasive at the item and call level. The only case where the p -value is below 0.005, according to a two tailed binomial test, is the integrating style at the item level.

Figure 6.5 shows the percentage of times that, in a given discussion, the individuals with higher scores are more persuasive, at both item and call level. The only statistically significant effect (p -value 0.035 according to a two-tailed binomial test) is that the most “integrating” subjects of each pair persuade their interlocutor 57.0% of the times at the item level (in line with the findings in [Cheung et al., 2006]). However, these subjects appear the same number of times among both those who call and those who are called. Therefore, the effect does not explain the difference in persuasiveness between callers and receivers. Furthermore, the effect of the integrating strategy is not statistically significant at the call level (p -value = 0.14 according to a two-tailed binomial test). In this case as well, *the conflict handling style does not produce effects that can explain the persuasiveness difference between subjects that call and subjects that are called.*

6.4 Conclusions

This chapter, based on the decision-making experimental procedure and the topic annotation investigated possible influences on the determination of negotiation outcome, i.e. the determination of the winner in case of disagreement during the item-discussions of the WST. In the scenario adopted for the experiments of this work, calling or being called appears to make a significant difference in terms of persuasiveness. This work identified a statistically significant dependency of the consensual decision on the role, i.e. *Caller* or *Receiver*. The *Receiver* tends to win more often than the *Caller*, namely 59% of the times at item level and 70% of the times at call level. The effect is statistically significant (p -value <5%). Gender, age, personality and conflict handling style were considered as an alternative explanation, but they appear to have negligible influence on the outcomes of the experiment. Calling or being

called, remains the factor that better explains the persuasiveness differences observed among subjects.

The lack of symmetry between callers and receivers was previously observed in [Fortunati, 1995], where callers were shown to have an advantage in terms of setting the call agenda, defining the tone of the conversation, etc. The results of this work seem to be in contradiction (the receivers appear to be in a more favourable condition). However, this might simply depend on the fact that the experiments take place in a controlled setting where both participants share the same information about the call. Furthermore, the receivers know that they are going to be called and, therefore, there is no surprise effect, something that seems to play a role in the findings of [Fortunati, 1995].

Unlike previous research on interplay between communication media and negotiation (see Section 6.2), this work focused on mobile phones and not on landline ones. The reason is that the number of individuals that subscribe only to mobile telephony services is constantly increasing and the trend is likely to continue in the foreseeable future [ITU, 2013]. However, both types of phone have the same richness (see Section 3.2) and the findings of this work seem to be in line with those obtained over landline phones under at least two important points of view. The first is that both media are *efficient*, i.e. they allow one to exchange sufficient information to complete the negotiation task [McGinn and Croson, 2004]. No breakdowns have been observed like it happens, e.g., using text only chat or written documents. The second is that, via phone, the maximization of the individual profit (the number of won items in this work) tends to be more frequent than the maximization of the joint profit [Sheffield, 1995]. Still, nothing conclusive can be said about the comparison between mobile and landline phones without repeating the experiment with these latter.

To the best of our knowledge, this is one of the few works that measure the effect of mobile phones on one-to-one conversations. Unlike other works in the literature (see Section 6.2), this article does not propose software applications expected to change the behaviour of people when installed on their phones. The experiments of this work simply show that the very use of mobile phones results into a change of persuasiveness. The results seem to be in line with the indications of the literature about the effect of communication media on negotiations. However, the experiments were performed in a controlled setting and it is unclear how much they can be generalized. Hence, future work will aim in particular at investigating whether different, more naturalistic conditions lead to similar findings.

The next two chapters focus on social phenomena inference (Chapter 7) and *behavioural events* prediction (Chapter 8). They investigate to what extent social behaviour can be inferred from the recorded data of the SSPNet-Mobile Corpus and whether behavioural events, as described in Chapter 5, can be detected automatically.

Chapter 7

Personality traits and conflict handling style recognition from audio and motor activation data.

7.1 Introduction

This chapter focuses on the analysis of behaviour during mobile phone calls and, in particular, presents experiments aimed at predicting personality traits and conflict handling style of people talking over the phone.

The inference of personality traits from speech has recently attracted significant attention and it is the subject of international benchmarking campaigns [Schuller et al., 2012]. Most approaches aim at Automatic Personality Perception (see [Mohammadi and Vinciarelli, 2012; Polzehl et al., 2010] for short surveys), i.e. at predicting the personality of speakers as perceived by listeners. This work aims at Automatic Personality Recognition, i.e. at inferring from speech how people assess their own personality. The latter task appears to be more challenging because the relationship between physical characteristics of speech and self-assessed traits tends to be weak, unlike the case of traits assessed by others [Mairesse et al., 2007].

The results of this article are comparable to those obtained in the main works addressing the same problem (see for example [Mairesse et al., 2007; Batrinca et al., 2011; Mohammadi et al., 2013]). However, to the best of our knowledge, the number of subjects involved in the experiments is significantly larger than in the rest of the literature (our corpus includes 120 individuals while the largest corpus we are aware of includes only 96). Furthermore, we address the problem of recognizing the conflict handling style that, to the best of our knowledge, was never addressed before.

The rest of the chapter is organized as follows: Section 7.2 presents the data used for the experiments, Section 7.3 reports on experiments and results, and the final Section 7.4 draws some conclusions.

7.2 The Data

This study uses the SSPNet-Mobile Corpus. The experimental procedure of the corpus was designed to support research on personality and conflict handling style inference from audio and motor activation data. Chapter 4 describes the experimental procedure and the WST scenario (Section 4.3) and the self-assessment psychometric questionnaires that was applied (Section 4.4) and how the audio and motor activation recordings were collected, meaning the sensing method (Section 4.5).

7.3 Experiments and Results

The goal of the experiments was to predict whether a subject was above or below median with respect to the scores observed for each of the Big Five traits and each of the Conflict Handling dimensions. The subjects with a score equal to the median were assigned to one of the two classes taking care of minimizing the imbalance. That is due to the fact that the model used (full description in Section 7.3.3) works better with balanced classes. Hence, for the separation process of the scores in two classes, the score/s equal to the median, was/were classified into the smaller class. The class with scores less than the median were classified as the negative class regarding the personality trait or conflict handling style under investigation. The bigger class (or most frequent) is presented as a % percentage in the first column of Table 7.2 and also corresponds to the accuracy of a trivial baseline system that predicts such a class. For each subject, features were extracted from the speech signal (see Section 7.3.1) and from the angular rates measured with the gyroscopes (see Section 7.3.2). The features were then fed, both separately and in combination, to an SVM for the recognition experiments (see Section 7.3.3).

7.3.1 Speech Features

Speech signals were first segmented into syllables using the approach in Petrillo and Cutugno [2003], then Praat [Boersma, 2002] and the pitch stylization model described in Origlia et al. [2013] were used to extract the following 36 features from syllables nuclei: harmonicity, nucleus length, syllable length, spectral centroid, spectral tilt, spectral skewness, spectral kurtosis, mean energy, jitter, shimmer, mean pitch, mean of first three formants, Bandwidth of first three formants, glissando likelihood, Teager Energy Operator, zero crossing rate, position of the nucleus, the first 13 Mel Frequency Cepstral Coefficients and their differences between consecutive frames. The features representing a speech sample (all the speaking time of a given person in a call) are mean, variance and entropy of the feature values extracted at the individual syllable level. The total number of resulting features is 109 (the number of syllables is included in the feature set).

7.3.2 Motor Activation Features

Angular rates measured with the Shake provide an indirect measurement of head movements and, in general, of the motor activation of the subjects. The assumption is that people tend

to keep the phone close to the ears and, hence, united with the entire head. The sampling rate of the gyroscope was 68 Hz and the data consist of 3 channels encoding the angular rate of pitch, roll and yaw. The three axes were treated as disjoint channels and a Fast Fourier transform was applied to each channel with a window size of 128 samples (roughly 2 seconds) at regular steps of 32 samples (roughly 0.5 seconds). Mean and variance of each energy bin were calculated for a total of 390 features.

7.3.3 Recognition

Recognition experiments were conducted using Support Vector Machines (SVM) with Radial Basis Function (RBF) kernel as implemented in libSVM [Chang and Lin, 2011]. After the normalisation of the data, the first step of the approach is to reduce the dimensionality of the feature vectors by applying Principal Component Analysis (PCA) and then to map the resulting vectors into one of the two classes described above for each personality trait or conflict handling dimension.

We apply PCA in order to reduce the dimensions of the orthogonal system constructed from the features. The PCA considers all features to select those principal components with the higher variability of data possible under the constraint to be uncorrelated (orthogonal) with one another. The variance refers to the variability of the principal component dimensionality to take into account. Percentage of 100% means that all the dimensions are used. The experiments were run using PCA with variance ranging in (0.5, 0.7, 0.8, 0.9, 0.95, 0.97, 0.98, 0.99) and the results are presented in Table 7.2.

This setup has three meta-parameters that need to be set: the amount of variance to retain after the PCA, the γ parameter of the RBF kernel and the regularization term C providing a way to control over fitting for the SVM. The best combination of C and γ is selected by a grid search with exponentially growing sequences of C and γ and variance parameter of PCA v , $C \in \{10^{-2}, 10^{-1}, \dots, 10^1\}$; $\gamma \in \{e^{-10/2}, e^{-9/2}, \dots, e^{-1/2}, e^0\}$; $v \in \{70\%, 80\%, 90\%, 95\%, 97\%, 98\%, 99\%\}$. Each combination of parameter selection is checked using cross validation, and the parameters with best cross-validation accuracy are picked. The final model, which is used for testing and for classifying new data, is then trained on the whole training set using the selected parameters.

Table 7.1 reports the best parameters for each personality trait and conflict handling style, after applying cross-validation to all dataset. All results reported in Table 7.2 were obtained using a Leave-One-Out scheme to estimate the accuracy (percentage of correctly classified subjects). This approach allows us to use the entire corpus for testing while keeping a rigorous distinction between training and test set. Assuming that we have n samples we train the classifier (with the best parameters for each case) using $n - 1$ samples leaving out one as the training set. We repeat the same procedure n times so every sample to become the testing set.

7.3.4 Results

Table 7.2 presents the results obtained in the experiments. The first column is the *a-priori* probability of the most frequent class and, as mentioned before, the accuracy of a trivial baseline system that predicts always such a class. The column “*Speech*” reports the accuracies obtained when using only the features described in Section 7.3.1. The values reported in bold are higher than chance to a statistically significant extent (p-value $< 5\%$ according to a binomial test). The column “*Move*.” reports the same information in the case of the features described in Section 7.3.2. The last column (“*S+M*”) shows the results of the feature

	Speech	Move.	S+M
Agreeableness	$v = 0.8, \gamma = 0.03$	$v = 0.97, \gamma = 0.135$	$v = 0.8, \gamma = 0.01$
Conscientiousness	$v = 0.95, \gamma = 0.004$	$v = 0.8, \gamma = 0.082$	$v = 0.8, \gamma = 0.0025$
Extroversion	$v = 0.8, \gamma = 0.018$	$v = 0.9, \gamma = 0.01$	$v = 0.9, \gamma = 0.0015$
Neuroticism	$v = 0.8, \gamma = 0.01$	$v = 0.98, \gamma = 0.0025$	$v = 0.97, \gamma = 0.004$
Openness	$v = 0.8, \gamma = 0.03$	$v = 0.9, \gamma = 0.0015$	$v = 0.97, \gamma = 0.0009$
Avoiding	$v = 0.8, \gamma = 0.0067$	$v = 0.8, \gamma = 0.607$	$v = 0.8, \gamma = 0.0025$
Compromising	$v = 0.8, \gamma = 0.0025$	$v = 0.95, \gamma = 0.22$	$v = 0.9, \gamma = 0.0015$
Dominating	$v = 0.8, \gamma = 0.082$	$v = 0.8, \gamma = 0.37$	$v = 0.8, \gamma = 0.135$
Integrating	$v = 0.9, \gamma = 0.05$	$v = 0.9, \gamma = 0.2$	$v = 0.8, \gamma = 0.03$
Obliging	$v = 0.9, \gamma = 0.05$	$v = 0.95, \gamma = 0.135$	$v = 0.9, \gamma = 0.03$

Table 7.1: The values of variance v for PCA, gamma γ that correspond to the model giving the best performances after cross-validation. The costs (regularization term) C for SVM is equal to 0.01 for all cases.

	P (%)	Speech	Move.	S+M	Speech	Move.	S+M
$\gamma = e^{\frac{x}{2}}$		$x \in \{-10, 0\}$			$x \in \{-17, 5\}$		
Agreeableness	56.7	56.7	55.0	61.7	48.3	58.3	61.7
Conscientiousness	50.8	48.3	53.3	28.3	55.8	62.5	31.7
Extroversion	54.2	54.1	57.5	59.2	65.0	58.3	59.2
Neuroticism	51.7	62.5	49.2	53.3	58.3	50.8	53.3
Openness	56.7	61.7	55.0	50.0	61.7	56.7	50.0
Avoiding	51.7	40.0	39.2	30.0	36.7	41.7	30.0
Compromising	51.7	50.8	51.6	48.3	51.7	44.2	48.3
Dominating	52.5	32.5	64.2	51.7	45.0	55.8	61.7
Integrating	55.0	54.2	56.6	57.5	53.3	56.7	56.7
Obliging	54.2	49.2	61.7	65.8	45.0	61.7	65.8

Table 7.2: Accuracy for personality traits (upper part) and conflict handling styles (lower part) using speech, phone movements and their combination (S+M). Bold values are higher than the a-priori probability of the most frequent class (second column from the left) to a statistically significant extent (p-value at 5%).

combination obtained by concatenating the feature vectors corresponding to the two modalities (before applying the PCA).

Speech features led to an accuracy higher than chance for *Neuroticism*, the trait that corresponds to how calm or anxious a person tends to be. Given that the scenario concerns a problem of survival (the participants are supposed to help people that crashed with their plane in a polar area and need advice to survive), it is not surprising to observe that such a trait emerges with more evidence with respect to the others.

Motor activation features allowed the recognition of individuals with a *Dominating* conflict handling style, i.e. a tendency to impose their views during discussion with others (in the case of the Winter Survival Task the disagreement takes place each time two subjects have different

opinions about one item). The result is coherent with the indications of the psychological literature showing that the level of motor activation helps to discriminate between dominant and submissive persons [Richmond and McCroskey, 1995] .

The multimodal combination of features improved over the trivial baseline to a statistically significant extent in the case of subjects adopting an *Obliging* conflict handling style. However, it should be observed that the combination works better than the individual modalities, but the difference with respect to the motor activation alone is not statistically significant. This result is probably complementary to the performance observed for the *Dominating* style. In fact, whenever one of the two subjects tends to impose her view, the other is often forced to step back (the scenario imposes a “yes” or “no” decision for every object and no other strategies are probably possible).

7.3.5 Further experimentation

In several cases, the accuracy of the approach is lower than the prior. Extra investigation was conducted to eliminate the possibility of bug in the algorithms, extended the parameters grid search and cross-validation fold number and investigate the assumptions on the behaviour under investigation given the experimental process by referring to literature.

The algorithm includes scaling of the data, PCA analysis with different variance on the principal components (the application of the SVM directly on the data has performed worse than the results in Table 7.2), cross-validation of regularization parameter, gamma parameter and variance of PCA, cross validation with different number of folds (5, 10, 120). Hence, the lower performances cannot be explained by the state-of-the-art methodology this approach uses. In the cases of non-significantly lower accuracies by applying the pair of best parameters resulting from cross-validation process (see Table 7.1) and training a model using all data, the model gives performances equal to the prior probability. That means that the model is not overfitting and leads us to the assumption that the model cannot work better than the prior due to poor information in the data. The model for the specific personality trait or conflict handling style cannot predict the subject of interest from speech or motor activation given the specific experimental process (WST, protocol, sensors etc.). That is probably because features do not carry information about the trait to be predicted.

In the cases of significantly lower performances an extensive examination has been performed over the factors that could influence the SVM model to perform poorly.

The effectiveness of SVM depends on the selection of kernel (RBF in this case), the kernel’s parameters (γ in this case), and the regularization parameter C . Hence, further experiments have been run to test the performance of the model by applying a broader grid search on C and γ parameters and by using less folds regarding the cross-validation process. The grid search was extended: $C \in \{10^{-3}, 10^{-1}, \dots, 10^3\}$; $\gamma \in \{e^{-17/2}, e^{-9/2}, \dots, e^{4/2}, e^{5/2}\}$. The number of folds of cross validation was initially 120, equal to the number of observations, and experiments were repeated with less number of folds (5 and 10).

The performances of the experiments taken place for further examination are presented in the left part of Table 7.2. The results shown that the model work better tuned to C regularization parameter equal to 0.01 for all cases.

The results on left show statistically significant prediction of in 4 cases, two personality traits *Conscientiousness* and *Extroversion*, and two conflict handling styles *Dominating* and *Obliging*. The approach with a wider range of γ parameter performs similar, perhaps slightly better, to the initial experimentation (results on the right Table 7.2). The same two conflict

styles are predicted from both approaches, but two other personality traits are significantly predicted using a wider range in model parameters. The results for conflict handling styles prediction keep up with the assumption about the experimental protocol to favour *Dominating* and *Obliging* conflict styles. The accuracy for *Neuroticism* is still above prior and not far from being statistically significant. However, personality prediction from self-assessment scoring was always a challenging task [Mairesse et al., 2007]. Our assumptions on behaviour for the implementation of the work described in this chapter are further reviewed in the following section that draws the conclusions.

To sum up, this work has extensively examine all possible factors that influence the performance of the SVM classifier in order to eliminate errors at the algorithm of the machine learning approach. Furthermore, an extensive additional experimentation has been conducted. Experiments were run to check more pairs of model parameters and for different folding in cross-validation process. The results, see Table 7.2, shows that the approach performs slightly better giving one more significant performance. The extra analysis concludes that the SVM-RBF with a PCA approach is unlikely to be improved further. Perhaps, different feature selection scheme might be an interesting path to be investigated in the future. The number of subjects, while being high compared to most of the other works in the literature (see [Vinciarelli et al., 2012a] for an extensive survey), might be too low to observe significant effects. This thesis conclude on the fact that other parts of the experimental procedure, such as the task-oriented scenario, the shorter version of self-assessment questionnaires, the audio-only channel of communication could have provide data adequate to predict up to five out of ten behavioural dimensions of interest. The next section explains in details the conclusions of these considerations.

7.4 Conclusions

This chapter presented experiments on automatic recognition of personality traits and conflict handling style via mobile phones. The experiments were performed over the *SSPNet-Mobile Corpus* one of the largest collections (60 calls and 120 subjects) of dyadic conversations annotated with psychometric measurements and behavioural events (see [Vinciarelli et al., 2012a] for an extensive survey). The results show that sensors available nowadays on any standard smartphone can be used to detect speech features (in particular for what concerns prosody and voice quality) as well as to measure motor activation. According to the indications of domains like Social Signal Processing [Vinciarelli et al., 2012a], these features provided machine detectable evidence of social and psychological phenomena like personality and conflict handling. However, some of the personality and conflict handling styles were significantly predicted. Section 7.3.3 presents the algorithm used in details and Section 7.3.4 reports the extensive investigation of our approach to determine possible errors on the model and reach to the conclusion that different kernel should be the next step to consider so as to improve the performances.

A possible explanation of the performances of the model for only some dimensions of conflict handling styles could be the of the experimental protocol of the WST scenario. The latter includes task oriented topics providing such data that could train a model to predict only *Dominating* and *Obliging* handling conflict styles using speech and motor activation.

Specifically, *Dominating* and *Obliging*, the tendency to impose one's own views or to accept the views of others, respectively. Since the scenario imposes to reach consensual "Yes" or "No" decisions when there is disagreement about a given object (see Section 4.3), the two conflict styles above are those most likely to be adopted and to be evident. In fact, for each object one of the two persons has to impose her view while the other has to step back.

The best recognized personality traits are *Neuroticism*, the tendency to be calm or anxious, *Conscientiousness* the tendency to be organised or not, and *Extroversion* the tendency to be talkative or not. However, are not predicted from both approaches. According to [Mairesse et al., 2007] personality traits prediction using scoring from self-assessments can be a challenging task. Moreover, the lack of eye contact and, consequently, the unavailability of visual non-verbal feedbacks (facial expressions, posture, gestures, position in space etc.) cannot be neglected as a factor on the social interaction under investigation. The results of Fowler and Wackerbarth [1980] draw that audio-only communication may be a disadvantage regarding several cues and leads to "*depersonalised, task-oriented discussions*". Findings of Burgoon et al. [2002] work enhance this assumption since they conclude on the fact that interaction aiming at low level task solving processes can be effective using audio only communication but for more complex tasks as judgemental task or collaborative work that depends on group trust and morale, audiovisual non-verbal cues could provide more personalised interaction. That is explained also by Halbe [2011] who argues that people tend to adapt themselves to the telephone mediated communication by change their non-verbal expression patterns such as overlaps, back-channels, turn-taking etc.

The fact that audio-only communication implies a less personalised interaction could be a possible factor of the low performance of our model since the personality traits might not been expressed physically to an adequate extent during the interaction. In combination with the application of the shorter version of Big Five Inventory instead of the full version of 45 questions, that could lead to not efficient enough data to train our model. As the results of this work also agree on the fact that self-assessment questionnaires are more challenging to be automatically predicted, it suggests –for future considerations– the collection not only of self-assessments but also of scoring from the interlocutor or /and from external raters.

Chapter 8

Automatic detection of Laughter and Fillers.

8.1 Introduction

This chapter aims at developing an approach to detect automatically vocal behavioural events of laughter, fillers and silence in audio recordings, as they are annotated in the SSPNet-Mobile Corpus (see Chapter 5). The methodology applied in this chapter introduces a new approach, to the best of our knowledge, which takes into account the sequence of the events (the precedent and the following event) in order to estimate the probability of the event of interest. The SSPNet-Mobile Corpus supported, until now, work on diverse aspects of research, including Computational Paralinguistics Björn et al. [2013] and *lexical and prosodic accommodation* Bonin et al. [2013]. In particular, the *Computational Paralinguistic Challenge* [Björn et al., 2013], a competition aiming at comparing approaches for detection of paralinguistic events of speech such as laughter and fillers (see Section 5.3.5), uses the SSPNet Vocalisation Corpus (SVC). The latter includes 2763 audio clips (11 seconds long each) annotated in terms of laughter and fillers for a total duration of 8 hours and 25 minutes and was extracted from the SSPNet-Mobile Corpus (see Section 8.3).

This chapter presents two sets of experiments: the first separates the speakers in 5-folds to apply cross-validation and the second follows the *Computational Paralinguistics Challenge* setting which separates the speakers in two folds, one test set and the other as training set. The rest of this chapter is organized as follows: Section 8.2 presents an overview of related work on laughter and fillers detection, Section 8.3 describes the data used, Section 8.4 explains the approach, Section 8.5 reports the results and Section 8.6 draws the conclusions on this work.

8.2 Previous work

To the best of our knowledge, no major efforts were done to detect fillers while laughter was the subject of several works (see Table 8.1). Therefore, most of the works presented in this section focus on laughter detection. Two main problems were addressed in the literature. The first, called *classification* hereafter, is performed over collections of audio samples (typically

around one second of length) that include either laughter or other forms of vocal behaviour (speech, silence, etc.). The classification problem consists in correctly discriminating between laughter samples and the others. The second problem, called *segmentation* hereafter, consists in automatically splitting audio recordings into intervals corresponding either to laughter or to other observable behaviours (overlapping speech, hesitations, etc.). The rest of this section surveys the main works dedicated to both problems (see Table 8.1 for a synopsis). For a better understanding of the reviewed literature in this section, it was suggested a description of the following terms:

- Mel-Frequency Cepstral Coefficients (MFCC): are parametric representations of the acoustic data based on the Fourier spectrum [Rabiner and Juang, 1993]. MFCC present the speech spectrum in a compact form [Lee et al., 2009] and because of that have been proven to be very effective in automatic speech recognition [Aucouturier and Pachet, 2003; Rabiner and Juang, 1993].
- Modulation Spectrum (MS) features: Modulation spectral analysis tries to capture long-term spectral dynamics within an acoustic signal and uses a modulation spectral model, a two-dimensional joint “acoustic frequency” and “modulation frequency” representation [Kinnunen, 2006; Atlas and Shamma, 2003]. Acoustic frequency stands for the frequency of conventional spectrogram whereas modulation frequency captures time-varying information through temporal modulation of the signal [Lee et al., 2009].
- Perceptual Linear Prediction (PLP) Coding features: PLP models the human speech based on the concept of psychophysics of hearing [Hermansky, 1990; Xie and Liu, 2006] and discards irrelevant information of the speech improving speech recognition rate. PLP analysis is computationally efficient and yields a low-dimensional representation of speech [Hermansky, 1990].

8.2.1 The Classification Problem

Classification is the most popular task for automatic laughter detection and was first investigated in [Kennedy and Ellis, 2004]. The initial data consisted of 29 meetings with 8 participants recorded using tabletop microphones. Clips of one second length were then extracted and labelled as “laughter” or “non-laughter” based on the number of participants laughing at the same time. That means they use a threshold on the percentage of participation in the laughing event and annotate as non-laughter the laughing events with low participation. Therefore, the approach detected only events where several participants were laughing. The experiments were performed over a corpus of clips including 1926 laughter samples. The samples were represented with MFCC and MS features. The classification was performed with Support Vector Machines (SVM) and the best Equal Error Rate, achieved with MFCC only, was 13%.

In [Truong and Van Leeuwen, 2007a], the experiments were performed over 6838 clips for a total of 3 hours and 38 minutes extracted from meeting recordings. The training set was composed of 5102 clips in English and three test sets where used in order to compare different conditions (same speakers as in training set, different speakers, different languages). Four types of features were investigated: PLP, Pitch and Energy (on a frame by frame basis), Pitch and Voicing (at the clips level) and MS features. The classification was performed with Gaussian Mixture Models (GMM), SVMs and Multi Layer Perceptrons. The main finding was that PLP features used with GMM produce the highest accuracy (82.4% to 93.6%). When

combining the output of several classifiers on different features, the accuracy can be improved up to 97.1% (by combining GMM trained on PLP and SVM trained on pitch and voicing features at the utterance level).

In [Petridis and Pantic, 2011], the authors investigate the joint use of audio (MFCC, and pitch and energy based statistics) and video features (head pose and facial expressions are extracted by tracking landmarks on the face of people) for laughter detection. The dataset is composed of 649 clips (218 laughter events) extracted from meeting recordings and from interactions between humans and artificial agents. The classifiers were Neural Networks and achieved an accuracy of 91.6% when using MFCC, with an improvement to 92.3% when adding pitch and energy. Adding video features brings the accuracy to 94.7%.

The only work on classification of laughter versus other types of non-verbal vocalization we are aware of is in [Schuller et al., 2008]. The data is composed of 2901 clips divided in 5 classes: Breathing, Consent, Garbage, Hesitation and Laughter. The authors investigate three models: Hidden Markov Models, Hidden Conditional Random Fields (hCRF) and SVM. The performances of the three models using MFCC and PLP features were compared. The best performance was achieved with HMMs and PLP features (80.7% accuracy).

Previous work presented so far has focused on the classification of laughter events either among laughter/non-laughter [Kennedy and Ellis, 2004; Petridis and Pantic, 2011], laughter/speech [Truong and Van Leeuwen, 2007a] or among other types of non-verbal events [Schuller et al., 2008]. This thesis presents work on the segmentation problem reviewed in the next section. The segmentation problem includes: first the segmentation of the event (to localise the event and define its duration) and then the classification of the event as laughter, filler, speech or silence.

8.2.2 The Segmentation Problem

The segmentation problem is addressed in this thesis as well and, on average, it is more challenging than the classification one. In segmentation, the input is not split a-priori into laughter and non-laughter and the goal is to identify laughter segments in the data stream. In [Truong and Van Leeuwen, 2007b], the authors extended the work in [Truong and Van Leeuwen, 2007a] and assessed their approach on a segmentation task. The proposed approach adopted PLP features to train Hidden Markov Models with GMM as emission probability distributions. The data set consisted of 29 meetings, with 3 meetings held out as a test set. The Markov model achieved an F_1 score of 0.62 for the segmentation of laughter versus the rest.

The work in [Knox and Mirghafori, 2007; Knox et al., 2008] investigates the use of Multi-Layer Perceptrons for the segmentation of laughter in meetings. The approach used MFCC features, Pitch and Energy. An HMM model fitted on the output of the MLP to take into account the temporal dynamics and a F_1 score of 0.81 was achieved. However, the approach was tested only with the silence segments manually removed. Furthermore, segments where the subjects were laughing and speaking at the same time were manually removed.

In [Laskowski and Schultz, 2008; Laskowski, 2009], the authors presented a system for segmenting the audio of meetings in three classes: silence, speech and laughter. Their approach is based on HMMs and uses MFCC and energy as features. The approach takes into account the state of all the participants when segmenting a meeting and achieves a F_1 score of 0.35 for laughter.

In [Scherer et al., 2012], audio-visual data from 2 meetings involving 4 speakers were segmented into laughter and non-laughter. Modulation Spectrum and PLP features were extracted

Ref.	Dataset	Instances	Mode	Task	Performance
[Kennedy and Ellis, 2004]	ICSI Meetings, 29 meetings, 25 h, 16 subjects	1926 laughter, non-laughter pre-segmented events	A	C	EER 13%
[Truong and Van Leeuwen, 2007a]	ICSI and CGN, 3 h 38', 24 subjects, 2 languages	3574 laughter and speech pre-segmented events (53.3%)	A	C	Acc. 64.0% (different Language) to 97.0% (same speakers)
[Petridis and Pantic, 2011]	7 AMI meetings & SAL Dataset, 16', 25 subjects	218 laughter, 331 non-laughter pre-segmented events	AV	C	Acc. 85.4%–92.3% for A. Acc. 94.7% for combination with V.
[Schuller et al., 2008]	AVIC, 21 subjects, 2291 clips	4 types of pre-segmented events: Breathing, Consent, Garbage, Hesitation, Laughter. 261 laughter events	A	C	5 classes: one for each event. Acc 77.8% – 80.7%
[Truong and Van Leeuwen, 2007b]	ICSI Meetings, 29 meetings, 25h, 16 subjects	1h33' of laughter (6.2%)	A	S	$F_1 \sim 0.62$
[Knox and Mirghafori, 2007; Knox et al., 2008]	ICSI Meetings, 29 meetings, 25 h, 16 subjects	1h33' of laughter (6.2%)	A	S	Silence manually removed, F_1 0.81
[Laskowski, 2009; Laskowski and Schultz, 2008]	ICSI Meetings, 29 meetings, 25h, 16 subjects	16.6' of laughter (2.0% of test set)	A	S	Automatic detection of silence. F_1 0.35 on voiced laughter, 0.44 on unvoiced laughter.
[Scherer et al., 2009, 2012]	FreeTalk Data, 3h, 4 subjects	10% is laughter	AV	S	F_1 0.44 – 0.72. Any laughter segment partially labelled is counted as fully detected for the computation of recall in F_1 .

Table 8.1: The table reports the details of the main recent works on laughter detection presented in the literature. The following abbreviations are used: A=Audio, V=Video, C=Classification, S=Segmentation, Acc=Accuracy, EER=Equal Error Rate.

from the audio. The features extracted from the video did not provide any improvement when used in combination with the audio features. The authors used 3 models: HMM, GMM and Echo State Networks, with the HMM outperforming the other two models with a F_1 score of 0.72. However, the recall was computed by considering fully detected also laughter events that were automatically detected only partially. This leads to an overestimate of recall and F_1 score.

This thesis present work on a new approach to address the segmentation problem. It employs the language model and not only the acoustic model yielding from MMFC like previous work has focused on. The language model which, to the best of our knowledge, has not been taken into account in previously, uses the sequence of words to predict events. In this work the language model is not trained on the sequence of words but on the sequence of behavioural events (filler, laughter, speech, silence).

8.3 The data

To address the segmentation and classification problem we use a new database, the SSPNet Vocalization Corpus (SVC), which is extracted from the conversation database of SSPNet-Mobile Corpus. The amount of fillers and laughter versus speech and silence during an actual conversation is heavily unbalanced. Hence, the segmentation of the conversations into clips is a practical way to focus on the events of interest, i.e. filler and laughter. The SVC includes clips, 11sec long, so as to include not too many or too few behavioural events. Each clip was selected in such a way that it contains at least one laughter or filler event between $t = 1.5$ seconds and $t = 9.5$ and up to three in total. The clips were recorded by the microphones of the phones. Therefore they contain the voice of one speaker only. Clips from the same speaker never overlap. In contrast, clips from two subjects participating in the same call may overlap (for example in the case of simultaneous laughter). However, they do not contain the same audio data because they are recorded with different phones. Overall, the database contains 1158 filler instances and 2988 laughter events. Both types of vocalisation can be considered fully spontaneous.

8.4 The model

The experiments aim at segmenting the clips of the SVC into *laughter*, *filler*, *silence* and *speech*. More formally, given a sequence of acoustic observations $X = x_1, \dots, x_{n+1}$, extracted at regular time steps from a clip, we want to find a segmentation $Y = (y_1, s_1), \dots, (y_m, s_m)$, with $y_i \in \{f, l, p, c\}$ ¹, $s_i \in \{1, \dots, n\}$, $s_1 = 1$ and $s_i < s_{i+1}$. This encodes the sequence of labels and when they start. We set $s_{m+1} = n$ to ensure that the segmentation covers the whole sequence.

¹labels are (f)iller, (l)laughter, s(p)eech, silen(c)e

We will use a maximum a-posteriori approach to find the following sequence \hat{Y} of labels:

$$\hat{Y} = \arg \max_{Y \in \mathcal{Y}} P(X, Y), \quad (8.1)$$

where \mathcal{Y} is the set of all possible label sequences of length between 1 and n .

The rest of the section is organized in two parts. The first describes the model we use for estimating $P(X, Y)$. The second describes the speech features extracted and the different parameters that need to be set in the model.

8.4.1 Hidden Markov Model

A full description of Hidden Markov Models is available in [Rabiner and Juang, 1986]. HMMs can be used to model a sequence of real valued vector observations $X = x_1, \dots, x_n$ with $x_i \in \mathcal{R}^m$. The model assumes the existence of latent variables $H = h_1, \dots, h_n$ and defines a joint probability distribution over both latent variables and observations:

$$P(X, H) = P(x_1 | h_1) \cdot P(h_1) \cdot \prod_{i=2}^n P(x_i | h_i) \cdot P(h_i | h_{i-1}). \quad (8.2)$$

From this joint distribution, the probability of the observations can be computed by marginalizing the latent variables:

$$P(X) = \sum_{H \in \mathcal{H}} P(X, H). \quad (8.3)$$

\mathcal{H} is the set of all possible sequences of latent variables of length n . The sum can be efficiently computed using Viterbi decoding even though the size of \mathcal{H} make direct summation intractable.

We train a different HMM $P_y(X, H)$ for each label $y \in \{f, l, p, c\}$ using sequences of observations extracted from the clips of the training set. The four HMMs, called *acoustic models* hereafter, capture the acoustic characteristics of the four classes corresponding to the labels. The probability of the whole sequence is given by

$$P(X, Y) = P(Y) \cdot \prod_{i=1}^m P_{y_i}(X_{s_i \dots s_{i+1}}), \quad (8.4)$$

where $X_{s_i \dots s_{i+1}} = x_{s_i}, \dots, x_{s_{i+1}}$, and P_{y_i} corresponds to the probability given by the HMM model trained on sequences associated with the label y_i . In previous works, $P(Y)$, usually called the *language model*, was assumed to be uniform and was therefore ignored. Specifically, the language model assigns a probability to a sequence of m events (words) by means of a probability distribution. In this work, the sequence do not refer to words but to behavioural events (filler, laughter, speech, silence) and the $P(Y)$ is modelled explicitly:

$$P(X, Y) = \prod_{i=1}^m P_{y_i}(X_{s_i \dots s_{i+1}}) \left(\prod_{i=2}^m P(y_i | y_{i-1}) \right)^\lambda, \quad (8.5)$$

where $P(y_i | y_{i-1})$ corresponds to a bigram language model, i.e. a model that takes into account one precedent event (an N -gram language model uses only $N-1$ events of prior context). The parameter λ is used to adjust the relative importance of the language model with respect to

the acoustic model. When $\lambda \neq 1$ in Equation (8.5), $P(X, Y)$ is not normalized (includes two factors one of them to the λ , i.e. values of different scales not adjusted –normalized– to a notionally common scale). However, this does not represent a problem because the segmentation process aims at finding the label sequence \hat{Y} maximizing the probability and not the exact value of the probability, as shown in Equation (8.1).

8.4.2 Features and Model Parameters

The experiments were conducted using the HTK Toolkit [Young et al., 2002] for both extracting the features and training the Hidden Markov Models (HMM) used for the segmentation. The parameters for the feature extraction were set based on the current state-of-the-art. For each clip, MFCC were extracted every 10 ms, from a 25 ms long Hamming window. We extracted 13 MFCC using a Mel filter bank of 26 channels. Filter bank analysis is an alternative to obtain the desired non-linear frequency resolution that approximates the human ear processes and have show to improve recognition performance [Young et al., 2002]. The HTK Toolkit provides a simple Fourier transform based filter bank designed to give approximately equal resolution on a mel-scale. To augment the spectral parameters derived from mel-filterbank analysis, an energy term can be included in the feature vector, instead of the zeroth coefficient. The energy term is the log-energy of every window. The feature vectors were extended using first and second order regression coefficients to describe information about the MFCC coefficients over time, which has been proved to increases speech recognition performance [Young et al., 2002]. The resulting feature vectors have dimension 39, 12 MFCC and log-energy expanded with their 1st and 2nd order delta (Δ) regression coefficients.

For the HMMs we used a left-right topology. Four acoustic models were trained, one for each of the four classes (silence, speech, laughter and fillers). Each model has 9 hidden states and each hidden state uses a mixture of 8 Gaussian distributions with diagonal covariance matrix. This topology has been shown to work well in practice. We have further investigated pairs of number of states, 5 to 20, and number of mixtures (3, 4, 5, 7, 8, 9, 11, 13, 15, 17, 19) and observed that in all cases the results were not significantly changed. Furthermore, the topology enforces a minimum duration of 90 ms for each segment since the model has to stay in each hidden state for at least 1 frame.

The language model is a back-off bi-gram model (see [Katz, 1987] for a detailed explanation of a back-off model) with a Good-Turing discounting [Katz, 1987] and addresses the problems associated with sparse training data.

8.5 Experiments and Results

We adopted two experimental protocols. In the first, the clips in the corpus were split into five folds in order to conduct cross-validation. The folds were created such that each speaker is guaranteed to appear only in one fold and such that the folds are of approximately the same size (between 547 and 555 clips). Each fold was used, iteratively, as a test set while the others were used as training set. In the second protocol, the corpus was divided into two parts, using the clips of 105 speakers as training set and the clips of the remaining 15 speakers as test set. Both setups guarantee that no speaker appears in both test and training

Table 8.2: HMM performance over the 5-fold setup.

λ	F ₁ Score					Precision π					Recall ρ				
	0	1	10	50	100	0	1	10	50	100	0	1	10	50	100
Filler	0.49	0.51	0.54	0.58	0.57	0.35	0.37	0.40	0.48	0.51	0.82	0.82	0.81	0.74	0.65
Laughter	0.48	0.53	0.58	0.64	0.63	0.35	0.39	0.45	0.60	0.67	0.79	0.80	0.79	0.69	0.61
Speech	0.77	0.79	0.81	0.83	0.84	0.94	0.95	0.94	0.91	0.89	0.65	0.67	0.70	0.76	0.79
Silence	0.87	0.88	0.87	0.87	0.86	0.82	0.82	0.82	0.82	0.82	0.92	0.93	0.93	0.92	0.91

Table 8.3: HMM performance over the challenge setup.

λ	F ₁ Score					Precision π					Recall ρ				
	0	1	10	50	100	0	1	10	50	100	0	1	10	50	100
Filler	0.50	0.52	0.54	0.59	0.58	0.36	0.39	0.40	0.49	0.64	0.80	0.76	0.79	0.73	0.64
Laughter	0.49	0.54	0.59	0.65	0.63	0.37	0.42	0.49	0.65	0.56	0.74	0.76	0.74	0.66	0.56
Speech	0.77	0.79	0.80	0.83	0.84	0.94	0.94	0.94	0.91	0.80	0.66	0.68	0.70	0.76	0.80
Silence	0.85	0.86	0.86	0.85	0.85	0.80	0.80	0.80	0.80	0.91	0.91	0.93	0.93	0.91	0.91

set. Therefore, all performances reported in Section 8.5.1 are speaker independent. The first protocol aims at using the entire corpus as test set while maintaining a rigorous separation between training and test data. The second protocol corresponds to the experimental setup of the *Computational Paralinguistics Challenge*² [Björn et al., 2013]. This will allow one to compare the results of this work with those obtained by the challenge participants (see Section 8.5.3).

8.5.1 Performance Measures

One of the main challenges of the experiments is that the classes are heavily unbalanced. In the SVC, laughter accounts for slightly less than 3.5% of the time, fillers account for 5.0%, silence represents 40.2% and speech is 51.3% of the total time. Given this distribution, accuracy (the percentage of time frames correctly labelled) is not a suitable performance measure. Therefore, for each class, we use *precision* π , *recall* ρ and *F₁ Score*. For a given class y , we consider as positive all frames in the time intervals labelled as y and as negative frames those of all other intervals. Then, we define *true positive* frames (TP) every positive frame positive classified and *false positive* frames (FP) every negative frame positive classified. Similarly we can define true negative frames (TN) every negative frame negative classified and false negative frames (FN) every positive frame negative classified. We can now define π as follows:

$$\pi = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (8.6)$$

the fraction of samples labelled with a given class that actually belong to such a class. We also define ρ :

$$\rho = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (8.7)$$

the fraction of samples from the class of interest that are correctly classified. The F₁ Score is a single score that takes into account both precision and recall and is defined as follows:

$$F_1 = 2 \cdot \frac{\pi \cdot \rho}{\pi + \rho}. \quad (8.8)$$

²emotion-research.net/sigs/speech-sig/is13-compare

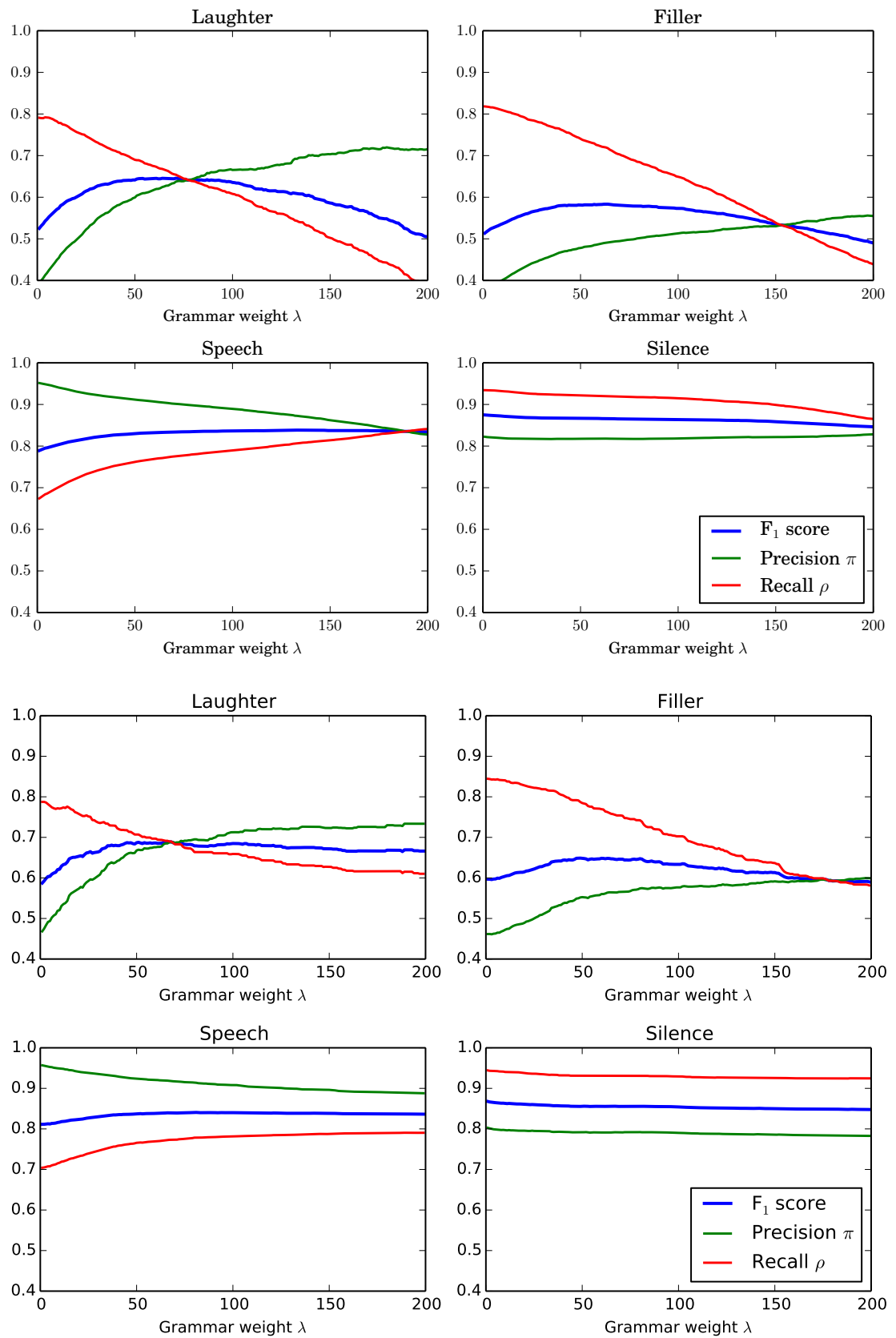


Figure 8.1: The plots show how F₁ Score, Precision and Recall change as a function of the parameter λ , the weight adopted for the Language Model. The four plots at the top have been obtained for five-fold protocol and the four plots at the bottom for the challenge protocol.

The F_1 Score arithmetically is equal to the harmonic mean of *precision* and *recall*. It can be interpreted as a weighted average of the precision and recall, and conveys the balance between the two measurements.

8.5.2 Detection Results

Tables 8.2 and 8.3 present the results for the segmentation task using different values of λ , for both experimental protocols described above. Figure 8.1 shows how π , ρ and F_1 Score change when λ ranges between 1 and 200 to provide a full account of the effect of such a parameter. For low values of λ , the Language Model does not influence the segmentation process and the performances are close to those obtained when using only the HMMs. The reason is that the contribution of the HMM term in Equation (8.5) tends to dominate with respect to the Language Model term. Hence, the value of λ must be increased to observe the actual effect of the bigrams. In fact, when $\lambda = 10, 50, 100$, the F_1 Score shows a significant improvement with respect to the application of the HMMs only. Since the goal of the experiments presented here is to show that the Language Model carries information useful for the segmentation process, no cross-validation is performed to set the λ value leading to the highest F_1 Score. The results are rather reported for several λ values. When λ is too high, the F_1 Score tends to

Table 8.4: Confusion Matrix for the 2-gram language model with $\lambda = 100$. The rows correspond to the ground truth and the columns to the class attributed by the classifier. Each cell is a time in seconds

	filler	laughter	silence	voice
filler	989	8	79	440
laughter	7	601	144	311
silence	83	105	11182	825
voice	584	120	1883	12949

drop for laughter and fillers (see Figure 8.1). The reason is that the Language Model tends to favour the most frequent classes (in this case speech). Hence, when λ is such that the Language Model becomes the dominant term of Equation (8.5), the segmentation process tends to miss laughter events and fillers. This phenomenon appears clearly when considering the effect of λ on Precision and Recall for the various classes. For laughter and fillers, π tends to increase with λ while ρ tends to decrease. In the case of speech, the effect is inverted, while for silence no major changes are observed (see Figure 8.1).

An interesting insight in the working of the model is given by the confusion matrix in Table 8.4. The Table applies to the case $\lambda = 100$, but different weights lead to similar matrices. Most of the confusions occur between fillers and speech, as well as between laughter and speech. This is due to the fact that sometimes people speak and laugh at the same time, but the corresponding frames were still labelled as laughter. Similarly, fillers and speech both include the emission of voice and are acoustically similar. The confusion between silence and laughter is also significant. This is due to the presence of unvoiced laughter, for which acoustic characteristics are close to silence (absence of voice emission, low energy).

8.5.3 ComParE Interspeech 2013 Challenge

The Interspeech 2013 *Computational Paralinguistics Challenge (ComParE)* [Björn et al., 2013] used SVC database for one of its four Sub-Challenges. As co-organisers of the ComParE we could not take part in the challenge but, also, we were not aware of the exact task of the challenge before the official announcement of the ComParE. The approach to address the segmentation problem was developed before the official announcement of the ComParE. This thesis is interested in the segmentation problem so as to detect non-verbal events occur in speech and not to classify pre-segmented events, the classification problem. Hence, we addressed an 4-class segmentation problem while ComParE participants address a 2-class classification problem. Hence, further analysis was necessary for a comparison of the our approach with the methods developed to address the ComParE *Social Signal Sub-Challenge*. The analysis uses the results presented earlier and leads inevitably to lower performances compared to the ComParE participants methods, due to not optimised classes of laughter and filler. As mentioned earlier, this thesis is interested in the segmentation problem and any

Research Group & Ref.	Approach	Features	Laugh.	Filler	UAAUC
Baseline: Björn et al. [2013]	SVM-SMO	baseline	82.9	83.6	83.3
Gupta et al. [2013]	DNN	baseline	93.3	89.7	91.5
Janicki [2013]	GMM-SVM	MFCC, Δ , Δ^2	90.7	89.0	89.8
Wagner et al. [2013]	SVM-SMO	phonetic + baseline	89.4	85.9	87.7
Oh et al. [2013]	SVM-SMO	20 f. related to 1-4 (Δ & AV)	85.9	84.6	85.3
An et al. [2013]	SVM-SMO	1, 2, 5 & Δ , 6, 7	84.6	85.1	84.9
Our approach	GMM-HMM	MFCCs, Δ , Δ^2	70.0	65.0	67.3

Table 8.5: The table reports the approaches, the extracted features and the performances of each participant in the *Social Signal Sub-Challenge* of the Interspeech 2013 *Computational Paralinguistics Challenge*. The measurements are the Area Under Curve (AUC) for laughter and filler separately and the Unweighted Average of Area Under Curve (UAAUC) of the two classes and are presented in the three columns on the right. The first line corresponds to the results presented by the organisers of the ComParE. The last line presents the results of the approach presented in this thesis. The numbers in the column of Features correspond to certain features 1: Intensity contour, 2: Pitch contour, 3: Timbral contour, 4: Rhythmic patterns, 5: Spectral tilt, 6: Duration, 7: Length of preceding and following pauses. For further explanation of 1 to 4 see [Oh et al., 2013] and for a full description of phonetic features see [Wagner et al., 2013].

further work or experiments would focus on the improvement of our approach. That is to address the segmentation problem since we obtained results that show better performances while using HHMs combined with a weighted language model.

The classification task of the *Social Signal Sub-Challenge* was addressed by the seven participants. The task is the detection and localisation of laughter and filler frames (10ms). The features extracted for the baseline performance were: MFCCs (1 – 12) and logarithmic energy along with first and second order delta regression coefficients, voicing probability,

harmonic-to-noise ratio (HNR), F0 and zero-crossing rate and their first order delta, voicing related low-level descriptors' (LLD) arithmetic mean and standard deviation calculated across the frame itself plus 4 frames before and 4 frames after.

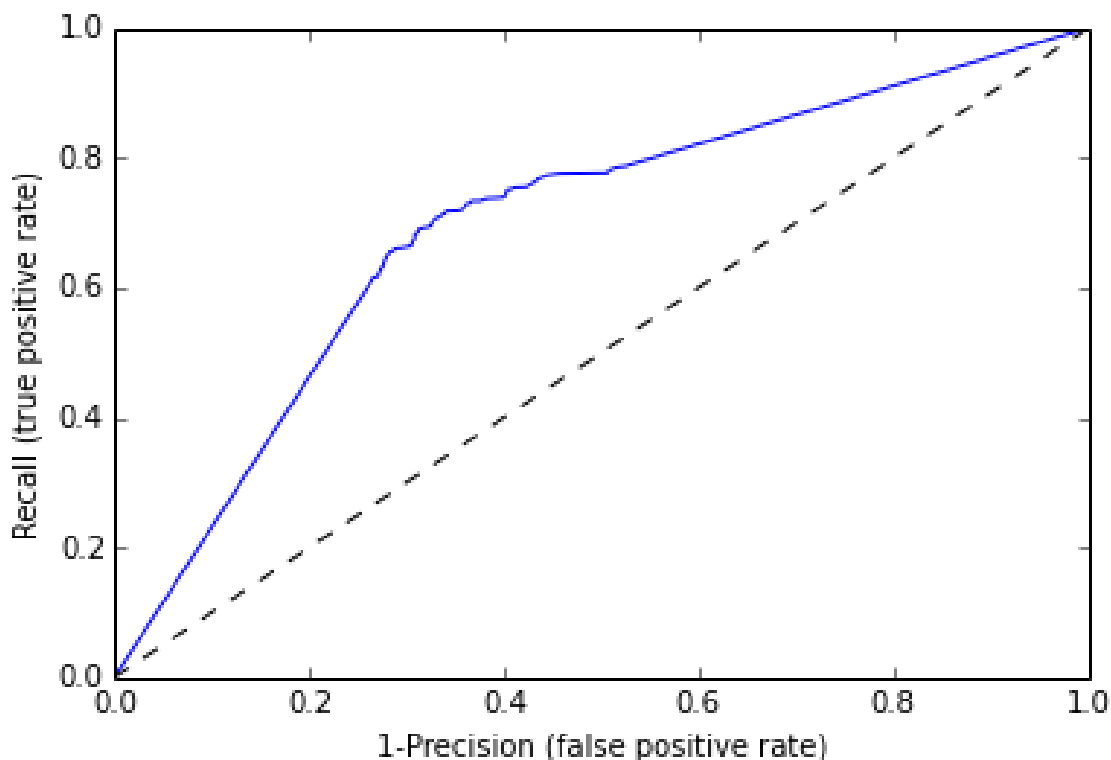


Figure 8.2: The plot show how the function of Precision and Recall for Laughter class, changes with the parameter λ , the weight adopted for the Language Model, ranging in $(0, 200)$.

Table 8.5 reports the features, method and results of each participation and the baseline set by the organisers of the ComParE. The order of the participants in the table follows their performances beginning with the best one.

Gupta et al. [2013] uses a Deep Neural Network (DNN) classifier to calculate the probabilities for each frame to be filler or laughter, and then smoothing is applied to remove the noise in the time series and masking to reduce the false alarm rate during detection. They extract the baseline features. The performances resulting by the DNN system only are very high. They are further improved using along with the DNN, the smoothing and masking techniques, to the best performance of the *Social Signal Sub-Challenge*.

Janicki [2013] uses a hybrid Gaussian Mixture Models -Support Vector machine (GMM-SVM) approach to address the challenge task. The approach train a three-class SVM classifier to find decision boundaries between GMM log-likelihood scores using MFCCs and Δ , Δ^2 parameters.

Wagner et al. [2013], Oh et al. [2013] and An et al. [2013] use the same approach a linear kernel SVM classifier with Sequential Minimal Optimisation (SMO) as the challenge baseline. Wagner et al. [2013] use the baseline features combined with phonetic features which are extracted using an automated phonetic transcription. Oh et al. [2013] focuses on the analysis of laughter in syllabic level. The approach includes the baseline classifier with no further modification. They extract 20 syllabic-level features and calculate Δ and average (AV) on each one of the them so as to combine them with the baseline features to train the classifier. An et al. [2013]

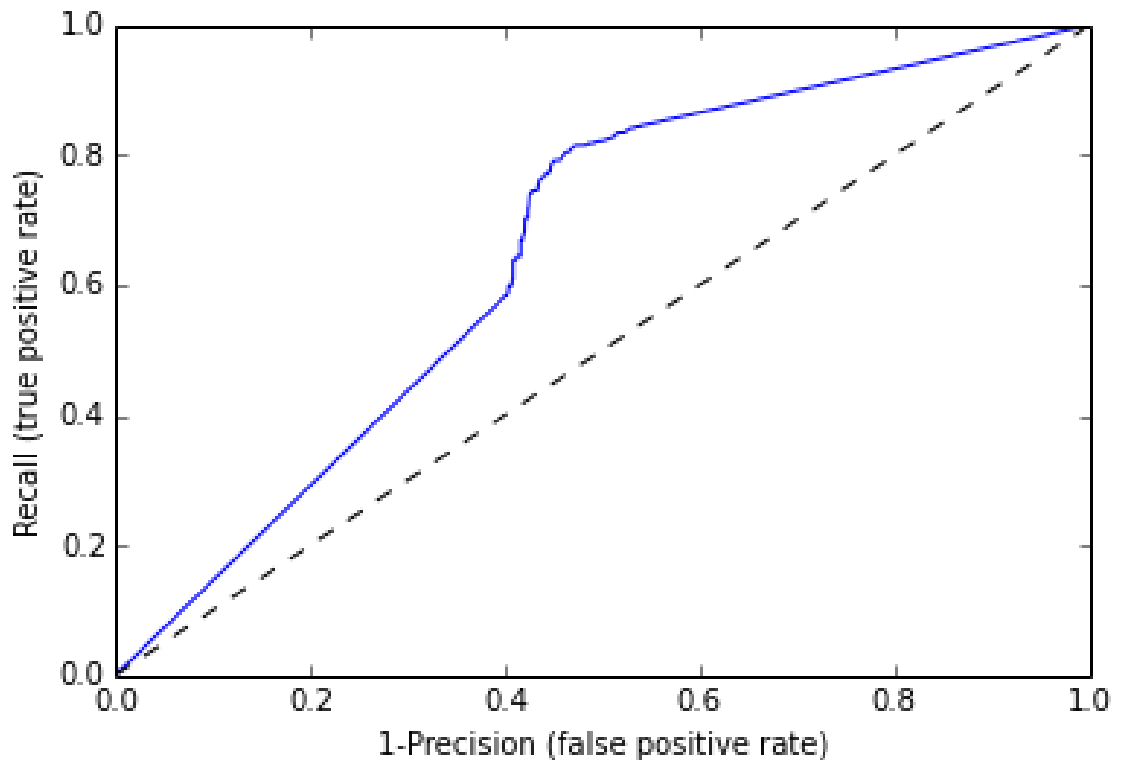


Figure 8.3: The plot show how the function of Precision and Recall for Filler class, changes with the parameter λ , the weight adopted for the Language Model, ranging in $(0, 200)$.

apply a pseudo-syllabification approach to smooth the baseline classifier's (SVM-SMO) first pass scores. They run re-scoring experiments by extract the 9 acoustic features for current, following and previous syllable, plus the position of the syllable to describe acoustic context in frame level. Finally, Krikke and Truong [2013] and Gosztolya et al. [2013] participated in the ComParE. However, they presented results using other measurements than the AUC and UAAUC for filler and laughter separately and therefore are not included in this comparison. The following paragraphs describe the extra analysis that was conducted so as to obtain a fair comparison between the performances of this work on the segmentation problem and the performances participants.

Further analysis of the results has been conducted so as to compare our results with the ComParE challenge. In order to compute Receiver Operator Characteristic (ROC) curve [Hanley and McNeil, 1982] and evaluate the AUC for our approach, we first computed false positive rate (1-Recall) and true positive rate (Recall) for the different values of the λ parameter. We extended the curve by considering randomized version of the classifier with the highest (respectively lowest) true positive rate. The randomized version of the classifier answers with probability p positively (respectively negatively) and with probability $(1-p)$ answers using the underlying classifier. It is easy to see that the graph of performance of this randomised classifier, see Figure 8.2 and Figure 8.3, is a straight line between the underlying classifier and the point $(1,1)$ and respectively between the classifier and the point $(0,0)$. The selection of the underlying classifiers also guarantees that we are estimating a lower bound on the AUC. In a ROC curve the true positive rate (Recall) is plotted in function of the false positive rate (1-Recall) for different cut-off points of a parameter (i.e. family classifiers). Each point on the ROC curve represents a Recall/1-Precision pair corresponding to a particular decision

threshold (e.g. a plane defined by a two-class classifier). The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two groups (filler/laughter). However, in our method the results are obtained using a 4-class method and classifies into 4 groups filler, laughter, speech and silence. That inevitably reduces the performance of our ROC curve and AUC and UAAUC measurements. Furthermore, the parts that correspond to the randomised classifier also lead to inevitably lower the measurements.

Our method could be improved in light of the reported methods above. The best performance on the *Social Signal Sub-Challenge* was accomplished by Gupta et al. [2013] by applying a DNN classifier. We could improve our approach by substituted the emission probability process of our approach, i.e. HMM-GMM, with the DNN approach described in Gupta et al. [2013], but also combined with the weighted language model which, to the best of our knowledge, is the novelty of our method.

8.6 Conclusions

This chapter proposes experiments on automatic detection of laughter and fillers in spontaneous phone conversations. The results were performed over the SSPNet Vocalization Corpus, one of the largest datasets available in the literature in terms of both number of subjects and amount of laughter events and fillers (see Section 8.3). To the best of our knowledge, this is the first attempt to jointly segment spontaneous conversations into laughter, fillers, speech and silence. Such a task can be considered more challenging than the simple classification of audio samples (see Section 8.2) or the application of segmentation processes to audio manually pre-segmented into silence and speech.

The most important innovation of the work is the adoption of Statistical Language Models aimed at estimating the a-priori probability of segment sequences in the data like, e.g., the probability of laughing after a silence and before speaking. The results show that the Language Models can significantly improve the performance of purely acoustic models. However, it is necessary to find an appropriate trade-off (by setting the parameter λ) between the weight of the HMMs and the weight of the bigrams.

Chapter 9

Conclusions

9.1 Introduction

This thesis has addressed the RQs, posed in the Introduction of this thesis, on the application of SSP to the analysis of mobile phone conversations. In particular, the thesis proposed experiments on automatic recognition of personality traits and conflict handling style based on speech and motor features, experiments on automatic detection of behavioural events (laughter, fillers, speech and silence) based on speech features, and experiments aimed at negotiation process understanding. The experiments were performed over a corpus collected during the thesis that includes 60 mobile phone conversations between 120 unacquainted subjects. The corpus was annotated in terms of non-verbal behavioural cues, decision-making events, personality traits (Big-Five) and conflict handling style.

The rest of this chapter is organized as follows: Section 9.2 summarizes the work and results reported in this thesis to show how the RQs stated at the beginning were addressed and the contributions of this research, Section 9.3 discusses the efficacy of the data, and improvements on the data collection. Section 9.4 proposes future work, and how the SSPNet-Mobile Corpus can be further exploited, and Section 9.5 draws some final comments.

9.2 Results and Contributions of the Thesis

As mentioned earlier, this thesis has as a goal to develop automated methods for the prediction of social information during phone calls using mobiles. Therefore, it supports research on SSP on the development of social artificial intelligence, as described in the Introduction. To accomplish these goals, this work sets as a priority the collection of a corpus especially designed for this work. The experimental protocol was designed to include decision-making conditions. We applied a modified version of the WST to ensure 12 decision-making discussions per dyad involved in the experiments. This provided our research with efficient data, 12x60 discussions, to investigate decision-making processes.

We observed a significant influence of mobile phone use on negotiation outcomes in the cases of disagreement. When the *Caller* and *Receiver* have conflicting opinions, the results show an advantage in favour of the *Receiver* that appear to be, on average, more persuasive than

the *Caller*. Chapter 6 describes the investigation on the *Caller-Receiver* effect as well as on gender, age, personality traits and conflict handling styles as potential factors of the effect. No significant influences of the latter factors were observed, and consequently the use of the phone affects the persuasiveness in favour of the *Receiver*, under the conditions of our study. This work addresses RQ1: “*What are the factors that influence decision making outcomes in the Winter Survival Task?*”.

The research on social behaviour prediction used the multi-modal (audio, motion) recordings and the psychological measurements of the self-assessments on personality and conflict handling style. Features were extracted from 2 channels of audio recorded data and 3 channels of motor activation data per participant. The audio features represented prosody and voice quality, and the motion data were processed with Fast Fourier transform to extract energy features. Chapter 7 describes this work and the results show that the personality trait of *Neuroticism* and the *Dominating* and *Obliging* conflict handling styles can be predicted given our experimental conditions. This work addresses RQ2: “*Can we infer personality traits and conflict handling style from audio and motor activation signals?*”.

The annotation of the data in terms of non-verbal behavioural events supported the investigation on automatic detection of laughter, fillers, silence and speech events. The 4 events were successfully detected with *F1*-scores up to 60%. Our approach adopted language models to estimate the probability of a sequence of events for the first time to the best of our knowledge. This work addresses question RQ3: “*Can we detect laughter, fillers, silence and speech automatically?*”.

9.3 Data efficacy and Future Improvements

The data of this work are 60 phone calls about the WST recorded, simultaneously, by two mobile devices N900 to record the audio information and two SHAKE devices attached on the phones to record the motion of the phones. Self-assessments on personality traits and conflict handling styles and manual annotations of the calls in behavioural events and topics are provided.

The work on the prediction of social information (RQ2) reveals that the personality traits and conflict handling styles that have been automatically recognised to a statistically significant extent, are those that have been favoured by the protocol of the experimental setting. Aiming to improve the approach by extracting richer social information the collection of psychological tests provided by the interlocutors and/or external observers in combination with self-assessments, instead of only self-assessments, should be included in the data. Furthermore, the use of the full version of BFI instead of BFI-10 (only two questions per personality trait) could provide richer behavioural information. Finally, more participants could, also, be a possible factor to improve the performance of the method.

The annotations into behavioural events were sufficient for the development of a novel, to the best of our knowledge, approach to address the segmentation problem of audio clips in speech, silence, filler and laughter events (RQ3). Further annotations on laughter events, such as voiced and unvoiced type, could provide sufficient data for the analysis of laughing patterns and the prediction of various social information that laughter carries such as happiness, scorn, embarrassment, politeness etc.

Finally, the data was sufficient for the investigation of phone's influential role in a social interaction (RQ1). The experimental protocol and its identical implementation over 120 experimental trials resulted to the observation of the *Caller-Receiver* effect.

Hence, the data was sufficient to address all three RQs since even in the case of the prediction of social information personality traits and conflict handling style (RQ2) the data gave results which can indicate the limitations of our method, such as the nature of the scenario, etc.

Finally, certain improvements related to the data collection could be suggested for future reference on this work:

- The use mobile apparatus with incorporated motion sensors is highly recommended to avoid synchronization issues.
- A certain number of correct items in the WST should be given to the participants instead of an unlimited number of items to collect. This could increase argumentation and, hence, the collection of longer decision-making and negotiation processes.

9.4 Future Work

The results on behavioural events detection can be considered preliminary and further work is needed to achieve higher performances. In particular, the current version of the approach does not discriminate between voiced and unvoiced laughter, a distinction that has been shown to be important in several works [Bachorowski and Owren, 2001; Owren and Bachorowski, 2003; Bachorowski et al., 2001; Truong and Van Leeuwen, 2007b]. Furthermore, the acoustic features are basic - although they have been shown to be effective in the literature - and can certainly be improved to capture subtle differences between, e.g., speech and fillers or silence and unvoiced laughter. Finally, the work on laughter resulted into a new approach, to the best of our knowledge, that detect the events of laughter in relation with the previous event. Future work can be oriented to the detection of the laughter events not only in short clips (the experiments were performed on short intervals extracted from the SSPNet-Mobile Corpus), but in complete audio conversations employing the DNN classifier suggested by Gupta et al. [2013], the winner of the ComParE *Social Signal Sub-Challenge*, in our approach instead of the HMM-GMM for the emission probability estimation.

The results on personality traits and conflict handling style prediction using the SVM-RBF approach after an extensive grid search on cost C and γ parameters suggest that further investigation for the improvement of the method should be oriented on the extraction of different speech and motor activation features.

Future work could involve examination of the movement of the "free hand". In our work the motion of the free hand is only partially investigated as the arm is part of the upper body. However, the hand is free from holding the phone and is free to move without constraints even when a person is seated. Hence, it might carry richer non-verbal behavioural information from the motor activation. Motion sensors that look like bracelets have already been released in the market and they could be used to set a low-level controlled setting.

Future work can address the transmission delay between the two mobile phones and whether and how this has an impact on human-human communication. The double annotated conver-

sations that correspond to the same interaction recorded from both phones is a database to fulfil research on communication using smartphones.

The SSPNet-Mobile Corpus can contribute further to research on SSP and mobile phones. The corpus can be transcribed and contribute as a language resource. Since the subjects share the same cultural background (the corpus includes only native English speakers holding a British passport in 98% of the cases), it can support ethnographic research on the use of mobile phones.

9.5 Final Remarks

This work has contributed to SSP by investigating non-verbal behaviour during actual phone calls using smartphones. Furthermore, the thesis has shed some light on decision-making processes taking place via the phone. The main results of the thesis are as follows:

- analysis of phone's influence on negotiation outcomes regarding the different use of it from the caller or the receiver. Calling or being called appears to make a significant difference in terms of persuasiveness during negotiations. The *Receiver* tends to win more often than the *Caller*, namely 70% of the times at call level. Gender, age, personality and conflict handling style were considered as an alternative explanation, but they appear to have negligible influence on the outcomes of the experiment.
- prediction of personality traits and conflict handling style through audio and motion data recorded with mobile devices. The results show that sensors available nowadays on any standard smartphone can be used to detect speech features (in particular for what concerns prosody and voice quality) as well as to measure motor activation. *Neuroticism*, personality trait which correspond to the tendency to be calm or anxious and *Dominating* and *Obliging*, handling styles assigned to the tendency to impose one's own views or to accept the views of others are predicted using our approach. The best recognized traits and styles were those which the WST inevitably favours due to its protocol, one imposes view while the other steps back.
- detection of vocal behavioural events with a new approach that takes into account sequential aspects. The approach includes language models to predict the sequence of the non-verbal events which can significantly improve the performance of purely acoustic models.

Last, but not least, the thesis involved the collection of a large database of mobile phone conversations (60 calls for 120 subjects). The data can serve as a basis for further research on the problems above as well as on new themes that can emerge from the analysis of the conversations (e.g., mimicry, interpersonal attraction, voice attractiveness, etc.).

Appendix A

Protocol and Scenario

THE SCENARIO

You are member of a rescue team. Your duty is to provide assistance to any person facing dangerous situations in a large area of Northern Canada. You have just received an SOS call from a group of people that survived a plane crash and report on their situation as follows:

“Both the pilot and co-pilot were killed in the crash. The temperature is -25°C , and the night-time temperature is expected to be -40°C . There is snow on the ground, and the countryside is wooded with several rivers criss-crossing the area. The nearest town is 32,2 km (≈ 20 miles) away. We are all dressed in city clothes appropriate for a business meeting.”

The survivors have managed to extract 12 objects (shown in p.3) from the plane. But they have to leave the site of the accident, carrying only a few objects - the less the better - in order to increase their chances of survival.

THE MISSION

Your mission is to identify the objects most likely to maximize the chances of survival of the plane passengers. The protocol includes two steps:

Step 1 – Individual Step

You receive a table (p.3) showing the 12 items and you have to decide for each one of them whether it is worth carrying or not.

You must write your decision, using YES or NO (YES: they have to carry it, NO: must not carry it), in the column on the left of the table.

Step 2 - Discussion

You will have a telephone conversation with another member of the rescue team in order to decide together which objects must be carried and what objects must be left in the plane.

As the call is a matter of life and death for the survivors, you will follow an emergency discussion protocol:

1. Consider the first object in the list.
2. Discuss with your colleague whether or not the object must be carried until you make a decision. **The decision must be consensual and you can take as much time as you need in order to make the right decision.**
3. Write your decision in the column to the right part of the table (p.3): **the decision must be the same for both participants.**
4. Once you have made a decision, move to the following object and repeat steps 2 and 3.
5. Continue until all objects have been considered and a consensual decision has been made for each one of them.

Please consider the following:

- Discuss one object at a time and move onto the next only after a consensual decision has been made.
- Once a decision has been made, do not go back and change the decision about previous objects.
- Discuss the objects in the order shown on the attached list.
- Do not interrupt the call until all objects have been discussed and all decisions have been made.

At the end of the conversation you have to return the table with the items, completed with “YES” or “NO” decisions for each item. The results must be the same for both you and your colleague. **The phone call will be recorded.**

PLEASE DO NOT USE THE LOUD SPEAKER

REWARDING SCHEME













You will receive £6 for your participation, but you can significantly increase your reward if you make the right decisions. **Some objects are actually necessary and must be carried while others should be left on the crash site:**

- You receive £3 extra, each time you decide to carry an item that must actually be carried (a right item).
- You lose £3, each time you decide to carry an item that must not actually be carried.
- You lose £3 for each decision marked on your list that is different from the one of your colleague.

In any case, a payment of £6 is guaranteed for your participation.

User ID: C001-Y

Table

Your opinion (fill this column BEFORE the call)		Items	Consensus (fill this column DURING the call)
	1.	A ball of steel wool	
	2.	A small axe	
	3.	A loaded 45-caliber pistol	
	4.	Can of butter	
	5.	Newspapers (one per person)	
	6.	Cigarette lighter (without fluid)	
	7.	Extra shirt and trousers for each survivor	
	8.	6m x 6m (≈20 ft x 20 ft) piece of heavy-duty canvas	
	9.	A sectional air map made of plastic	
	10.	750 ml of whisky	
	11.	A compass	
	12.	Family-size chocolate bars (one per person)	

Thank you for your participation!

Appendix B

Consent Form

Consent Form

I hereby acknowledge that I have received £6 as a payment for my participation in the “Quality of Rapport Experiment” led by Dr Alessandro Vinciarelli at the School of Computing Science of the University of Glasgow.

By signing this receipt, I authorize the following:

- Use of the phone recordings for research purposes.
- Distribution of the anonymized phone recordings in the scientific community (only for research purposes). We will remove any identifying information from the audio file before it is shared.
- Use of the questionnaires I have filled for research purposes.
- Distribution of questionnaires in the scientific community (only for research purposes).

Furthermore, by signing this receipt I commit to keep confidential all the details of the experiment, including content of scenario and questionnaires, conversations with Dr Vinciarelli and his collaborators, content of the recordings, etc.

Name: _____ Glasgow, - - 2012

Surname: _____ Signature

Age: _____

Appendix C

The BF-10 Questionnaire

1. I see myself as someone who is reserved
2. I see myself as someone who is generally trusting
3. I see myself as someone who tends to be lazy
4. I see myself as someone who is relaxed, handles stress well
5. I see myself as someone who has few artistic interests
6. I see myself as someone who is outgoing, sociable
7. I see myself as someone who tends to find fault with others
8. I see myself as someone who does a thorough job
9. I see myself as someone who gets nervous easily
10. I see myself as someone who has an active imagination

Every question should be addressed by choosing one of the following options:

- Disagree strongly
- Disagree a little
- Neither agree nor disagree
- Agree a little
- Agree strongly

Appendix D

The Conflict Handling Style Questionnaire

11. I try to investigate an issue with others to find a solution acceptable to us
12. I generally try to satisfy the needs of others
13. I attempt to avoid being “on the spot” and try to keep my conflict with others to myself
14. I try to integrate my ideas with those of others to come up with a decision jointly
15. I give some to get some
16. I try to work with others to find solutions to a problem which satisfy our expectations
17. I usually avoid open discussion of my differences with others
18. I usually hold on to my solution to a problem
19. I try to find a middle course to resolve an impasse
20. I use my influence to get my ideas accepted
21. I use my authority to make a decision in my favour
22. I usually accommodate the wishes of others
23. I give in to the wishes of others
24. I win some and I lose some
25. I exchange accurate information with others to solve a problem together
26. I sometimes help others to make decision in their favour
27. I usually allow concessions to others
28. I argue my case with others to show the merits of my position
29. I try to play down our differences to reach a compromise
30. I usually propose a middle ground for breaking deadlocks
31. I negotiate with others so that a compromise can be reached
32. I try to stay away from disagreement with others
33. I avoid an encounter with others
34. I use my expertise to make a decision in my favour
35. I often go along with the suggestions of others
36. I use “give and take” so that a compromise can be made
38. I am generally firm in pursuing my side of the issue
39. I try to bring all our concerns out in the open so that the issues can be resolved in the best possible way
40. I collaborate with others to come up with decisions acceptable to us
41. I try to satisfy the expectations of others
42. I sometimes use my power to win a competitive situation
43. I try to keep my disagreement with others to myself in order to avoid hard feelings
44. I try to avoid unpleasant exchanges with others
45. I generally avoid an argument with others
46. I try to work with others for a proper understanding of a problem

Every question should be addressed by choosing one of the following options:

- Disagree strongly
- Disagree a little
- Neither agree nor disagree
- Agree a little
- Agree strongly

Bibliography

- Abraham C. and Michie S. A Taxonomy of Behavior Change Techniques Used in Interventions. *Health Psychology*, 27(3):379–387, 2008.
- Adelswärd V. Laughter and Dialogue: The Social Significance of Laughter in Institutional Discourse. *Nordic Journal of Linguistics*, 12(02):107–136, 1989.
- Aharony N., Pan W., Ip C., Khayal I., and Pentland A. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011.
- Alhadeff A. Rodin: A self-portrait in the gates of hell. *The Art Bulletin*, 48(3/4):393–395, 1966.
- Allport G. W. *Personality: a psychological interpretation*. H. Holt and Company, 1937.
- An G., Brizan D.-G., and Rosenberg A. Detecting laughter and filled pauses using syllable-based features. In *INTERSPEECH*, pages 178–181, 2013.
- Anastasi A., Urbina S., et al. *Psychological testing*. New York: Macmillan, 1982.
- Aquilino W. S. Telephone versus face-to-face interviewing for household drug use surveys. *Substance Use & Misuse*, 27(1):71–91, 1991.
- Aristotle. *Rhetoric II*. a.
- Aristotle. *Nicomachean Ethics IV*. b.
- Atlas L. and Shamma S. A. Joint acoustic and modulation frequency. *EURASIP Journal on Applied Signal Processing*, 2003:668–675, 2003.
- Aucouturier J.-J. and Pachet F. Representing musical genre: A state of the art. *Journal of New Music Research*, 32(1):83–93, 2003.
- Bachorowski J.-A. and Owren M. J. Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science*, 12(3):252–257, 2001.
- Bachorowski J.-A., Smoski M. J., and Owren M. J. The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, 110(3):1581–1597, 2001.
- Bajos N., Spira A., Ducot B., Messiah A., et al. Analysis of sexual behaviour in france (acsf): A comparison between two modes of investigation: Telephone survey and face-to-face survey. *Aids*, 1992.

- Bales R. F. *Personality and interpersonal behavior*. New York, 1970.
- Barry B. and Friedman R. Bargainer characteristics in distributive and integrative negotiation. *Journal of Personality and Social Psychology*, 74(2):345–359, 1998.
- Batrinca L. M., Mana N., Lepri B., Pianesi F., and Sebe N. Please, tell me about yourself: automatic personality assessment using short self-presentations. In *Proceedings of the 13th international conference on multimodal interfaces, ICMI '11*, pages 255–262, New York, NY, USA, 2011. ACM.
- Baumeister R. F., Vohs K. D., DeWall C. N., and Zhang L. How emotion shapes behavior: Feedback, anticipation, and reflection, rather than direct causation. *Personality and Social Psychology Review*, 11(2):167–203, 2007.
- Bavelas J., Gerwing J., Sutton C., and Prevost D. Gesturing on the telephone: Independent effects of dialogue and visibility. *Journal of Memory and Language*, 58(2):495–520, 2008.
- Bazerman M., Curhan J., Moore D., and Valley K. Negotiation. *Annual Review of Psychology*, 51:279–314, 2000.
- Bennett A. Interruptions and the interpretation of conversation. *Discourse Processes*, 4(2): 171–188, 1981.
- Benus S., Gravano A., and Hirschberg J. The prosody of backchannels in american english. In *Proceedings of ICPHS*, pages 1065–1068, 2007.
- Björn S., Steidl S., Batliner A., Vinciarelli A., Scherer K., Ringeval F., Chetouani M., Weninger F., Eyben F., Marchi E., Salamin H., and Polychroniou A. e. a. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. *Proceeding INTERSPEECH*, 2013.
- Boersma P. Praat, a system for doing phonetics by computer. *Glott international*, 5(9/10): 341–345, 2002.
- Bonin F., De Looze C., Ghosh S., Gilmartin E., Vogel C., Polychroniou A., Salamin H., Vinciarelli A., and Campbell N. Investigating fine temporal dynamics of prosodic and lexical accommodation. *Proceeding INTERSPEECH*, 2013.
- Bordia P. Face-to-face versus computer-mediated communication: A synthesis of the experimental literature. *Journal of Business Communication*, 34(1):99–118, 1997.
- Bradner E. and Mark G. Why distance matters: effects on cooperation, persuasion and deception. In *Proceedings of the ACM International Conference on Computer Supported Cooperative Work*, pages 226–235, 2002.
- Brewer M. B. Research design and issues of validity. *Handbook of research methods in social and personality psychology*, pages 3–16, 2000.
- Brown J. Evaluations of self and others: Self-enhancement biases in social judgments. *Social Cognition*, 4:353–376, 1986.

- Brustad M., Skeie G., Braaten T., Slimani N., and Lund E. Comparison of telephone vs face-to-face interviews in the assessment of dietary intake by the 24h recall epic soft program—the norwegian calibration study. *European journal of clinical nutrition*, 57(1):107–113, 2003.
- Burger S. and Sloane Z. The isl meeting corpus: Categorical features of communicative group interactions. In *Meeting Recognition Workshop. Proceedings*, 2004.
- Burger S., MacLaren V., and Yu H. The isl meeting corpus: The impact of meeting type on speech style. In *ICSLP. Proceedings.*, volume 2, pages 301–304, 2002.
- Burgoon J. K., Bonito J. A., Ramirez A., Dunbar N. E., Kam K., and Fischer J. Testing the interactivity principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of communication*, 52(3):657–677, 2002.
- Cardy E. An experimental field study of the GOTV and persuasion effects of partisan direct mail and phone calls. *The Annals of the American Academy of Political and Social Science*, 601(1):28–40, 2005.
- Carletta J., Ashby S., Bourban S., Flynn M., Guillemot M., Hain T., Kadlec J., Karaiskos V., Kraaij W., Kronenthal M., et al. The ami meeting corpus: A pre-announcement*. In *Machine Learning for Multimodal Interaction: Second International Workshop*, volume 3869, page 28. Springer, 2006.
- Cassell J. et al. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. *Embodied conversational agents*, pages 1–27, 2000.
- Cathcart N., Carletta J., and Klein E. A shallow model of backchannel continuers in spoken dialogue. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, volume 1, pages 51–58, 2003.
- Chang C. and Lin C. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.
- Chen L., Rose R., Qiao Y., Kimbara I., Parrill F., Welji H., Han T., Tu J., Huang Z., Harper M., et al. Vace multimodal meeting corpus. *Machine Learning for Multimodal Interaction*, pages 40–51, 2006.
- Cheung S., Yiu T., and Yeung S. A study of styles and outcomes in construction dispute negotiation. *Journal of Construction Engineering and Management*, 132(8):805–814, 2006.
- Chiu M.-C., Chang S.-P., Chang Y.-C., Chu H.-H., Chen C. C.-H., Hsiao F.-H., and Ko J.-C. Playful bottle: a mobile social persuasion system to motivate healthy water intake. In *Proceedings of the International Conference on Ubiquitous Computing*, pages 185–194, 2009.
- Clark H. and Schaefer E. Contributing to discourse. *Cognitive Science*, 13(2):259–294, 1989. ISSN 03640213.
- Clark H. and Treeb J. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111, 2002.

- Clark H. H. and Wasow T. Repeating words in spontaneous speech. *Cognitive psychology*, 37(3):201–242, 1998.
- Coates J. Talk in a play frame: More on laughter and intimacy. *Journal of Pragmatics*, 39(1):29–29, 2007.
- Cole R., Fanty M., Noel M., and Lander T. Telephone speech corpus development at csu. In *Proceedings of ICSLP*, volume 94, page 1, 1994.
- De Leeuw E., Mellenbergh G., and Hox J. The influence of data collection method on structural models a comparison of a mail, a telephone, and a face-to-face survey. *Sociological Methods & Research*, 24(4):443–472, 1996.
- De Leeuw E. D. *Data Quality in Mail, Telephone and Face to Face Surveys*. ERIC, 1992.
- Dittmann A. and Llewellyn L. Relationship between vocalizations and head nods as listener responses. *Journal of personality and social psychology*, 9(1):79–84, 1968.
- Dourish P. and Bell G. *Divining a digital future: mess and mythology in ubiquitous computing*. MIT Press, 2011.
- Drummond K. and Hopper R. Back channels revisited: Acknowledgment tokens and speaker-ship incipency. *Research on Language and Social Interaction*, 26(2):157–177, 1993.
- Dunbar N. E. and Burgoon J. K. Perceptions of power and interactional dominance in interpersonal relationships. *Journal of Social and Personal Relationships*, 22(2):207–233, 2005.
- Duncan S. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283–292, 1972.
- Durham C. C., Locke E. A., Poon J. M., and McLeod P. L. Effects of group goals and time pressure on group efficacy, information-seeking strategy, and performance. *Human Performance*, 13(2):115–138, 2000.
- Eagle N. and Pentland A. Social serendipity: Mobilizing social software. *Pervasive Computing, IEEE*, 4(2):28–34, 2005.
- Eagle N. and Pentland A. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.
- Eagle N. N. *Machine perception and learning of complex social systems*. PhD thesis, Massachusetts Institute of Technology, 2005.
- Ekman P. and Friesen W. V. Head and body cues in the judgment of emotion: A reformulation. *Perceptual and motor skills*, 24(3):711–724, 1967.
- Ekman P. and Friesen W. V. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Nonverbal communication, interaction, and gesture*, pages 57–106, 1981.
- Escalera S., Baró X., Vitria J., Radeva P., and Raducanu B. Social network extraction and analysis based on multimodal dyadic interaction. *Sensors*, 12(2):1702–1719, 2012.

- Fenig S., Levav I., Kohn R., and Yelin N. Telephone vs face-to-face interviewing in a community psychiatric survey. *American Journal of Public Health*, 83(6):896–898, 1993.
- Flaherty L. M., Pearce K. J., and Rubin R. B. Internet and face-to-face communication: Not functional alternatives. *Communication Quarterly*, 46(3):250–268, 1998.
- Fogg B. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):5, 2002.
- Fortunati L. *Gli italiani al telefono*. Franco Angeli, 1995.
- Fowler G. and Wackerbarth M. Audio teleconferencing versus face-to-face conferencing: A synthesis of the literature. *Western Journal of Communication (includes Communication Reports)*, 44(3):236–252, 1980.
- Fox S. and Dinur Y. Validity of self-assessment: A field evaluation. *Personnel psychology*, 41(3):581–592, 1988.
- Fox T., Heimendinger J., and Block G. Telephone surveys as a method for obtaining dietary information: a review. *Journal of the American Dietetic Association*, 92(6):729, 1992.
- Frohlich N. and Oppenheimer J. Some consequences of e-mail vs. face-to-face communication in experiment. *Journal of Economic Behavior & Organization*, 35(3):389–403, 1998.
- Funder D. C. Personality. *Annual Review of Psychology*, 52(6):197–221, 2001.
- Galán I., Rodríguez-Artalejo F., Zorrilla B., et al. [telephone versus face-to-face household interviews in the assessment of health behaviors and preventive practices]. *Gaceta sanitaria/SESPAS*, 18(6):440, 2004.
- Garofolo J. S., Laprun C. D., Michel M., Stanford V. M., and Tabassi E. The nist meeting room pilot corpus. In *Proceedings of Language Resource and Evaluation Conference*, 2004.
- Gasser R., Brodbeck D., Degen M., Luthiger J., Wyss R., and Reichlin S. Persuasiveness of a mobile lifestyle coaching application using social facilitation. In IJsselsteijn W., Kort A., Midden C., Eggen B., and Hoven E., editors, *Persuasive Technology*, volume 3962 of *Lecture Notes in Computer Science*, pages 27–38. 2006.
- Glenn P. Current speaker initiation of two party shared laughter. *Research on Language & Social Interaction*, 25(1):139–162, 1991.
- Godfrey J. J., Holliman E. C., and McDaniel J. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE, 1992.
- Goldberg J. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts. *Journal of Pragmatics*, 14(6):883–903, 1990.
- Goodwin C. and Heritage J. Conversation Analysis. *Annual Review of Anthropology*, 19: 283–307, 1990.
- Gosztolya G., Busa-Fekete R., and Tóth L. Detecting autism, emotions and social signals using adaboost. In *INTERSPEECH*, pages 220–224, 2013.

- Gravano A. and Hirschberg J. Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of Interspeech*, pages 1019–1022, 2009.
- Greenfield T. K., Midanik L. T., and Rogers J. D. Effects of telephone versus face-to-face interview modes on reports of alcohol consumption. *Addiction*, 95(2):277–284, 2000.
- Guerrero K. L., Joseph D. A., and Michael H. H. *The nonverbal communication reader : classic and contemporary readings*. Waveland Press Lone Grove, Prospect Heights Ill., 2nd ed. edition, 1999.
- Gupta R., Audhkhasi K., Lee S., and Narayanan S. Paralinguistic event detection from speech using probabilistic time-series smoothing and masking. In *INTERSPEECH*, pages 173–177, 2013.
- Hain T., Burget L., Dines J., Garau G., Karafiat M., Leeuwen D.van , Lincoln M., and Wan V. The 2007 ami (da) system for meeting transcription. In *Multimodal Technologies for Perception of Humans*, pages 414–428. Springer, 2008.
- Halbe D. " who's there?": Differences in the features of telephone and face-to-face conferences. *Journal of Business Communication*, page 0021943611425238, 2011.
- Hanley J. A. and McNeil B. J. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.
- Harper R. *Texture*. MIT Press, 2010.
- Heldner M. and Edlund J. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, 2010.
- Hermansky H. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.
- Herring S. Gender and democracy in computer-mediated communication. In *Computerization and controversy (2nd ed.)*, pages 476–489. Academic Press, Inc., 1995.
- Herzog A. R. and Rodgers W. L. Interviewing older adults mode comparison using data from a face-to-face survey and a telephone resurvey. *Public Opinion Quarterly*, 52(1):84–99, 1988.
- Hiltz S. R. The “virtual classroom”: Using computer-mediated communication for university teaching. *Journal of communication*, 36(2):95–104, 1986.
- Holbrook A. L., Green M. C., and Krosnick J. A. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Quarterly*, 67(1):79–125, 2003.
- Hox J. J. and De Leeuw E. D. A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality and Quantity*, 28(4):329–344, 1994.
- Hughes S. SHAKE Model SK7 Product Guide, 2010.

- Hughes S. and O'Modhrain S. SHAKE-Sensor Hardware Accessory for Kinesthetic Expression. In *Proceedings on Enactive Interfaces*, pages 155–156, 2006.
- ICT. International Telecommunication Union, 2013.
- ITU. Mobile-cellular telephone subscriptions 2000-2012, 2014. URL <http://www.itu.int>.
- ITU. Measuring the information society. Technical report, International Telecommunication Union, 2013.
- Janicki A. Non-linguistic vocalisation recognition based on hybrid gmm-svm approach. In *INTERSPEECH*, pages 153–157, 2013.
- Janin A., Baron D., Edwards J., Ellis D., Gelbart D., Morgan N., Peskin B., Pfau T., Shriberg E., Stolcke A., et al. The icsi meeting corpus. In *Acoustics, Speech, and Signal Processing. Proceedings. IEEE International Conference on*, volume 1, pages I–364. IEEE, 2003.
- Jefferson G., Sacks H., and Schegloff E. A. *Notes on laughter in the pursuit of intimacy*. Multilingual Matters, Clevedon, 1978.
- Joinson A. N. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2):177–192, 2001.
- Joshi M. P., Davis E. B., Kathuria R., and Weidner C. K. Experiential learning process: Exploring teaching and learning of strategic management framework through the winter survival exercise. *Journal of Management Education*, 29(5):672–695, 2005.
- Judd C. Cognitive effects of attitude conflict resolution. *Journal of Conflict Resolution*, 22(3): 483–498, 1978.
- Kang S.-H., Gratch J., Sidner C., Artstein R., Huang L., and Morency L.-P. Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 63–70. International Foundation for Autonomous Agents and Multiagent Systems, 2012.
- Katz S. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(3): 400–401, 1987.
- Kendon A. Communication conduct in co-present interaction. *Social Signal Processing Summer School, Vietri sul Mare, Italy*, 2013.
- Kennedy L. S. and Ellis D. P. Laughter detection in meetings. In *NIST ICASSP 2004 Meeting Recognition Workshop, Montreal*, pages 118–121. National Institute of Standards and Technology, 2004.
- Kiesler S., Siegel J. A., and McGuire T. W. *Social psychological aspects of computer-mediated communication*. Carnegie-Mellon University, Committee on Social Science Research in Computing, 1984.

- Kim S., Filippone M., Valente F., and Vinciarelli A. Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and gaussian processes. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 793–796. ACM, 2012.
- Kinnunen T. Joint acoustic-modulation frequency for speaker recognition. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I–I. IEEE, 2006.
- Knox M. and Mirghafori N. Automatic laughter detection using neural networks. In *Proceedings of interspeech*, pages 2973–2976, 2007.
- Knox M., Morgan N., and Mirghafori N. Getting the last laugh: Automatic laughter segmentation in meetings. In *Proc. INTERSPEECH*, pages 797–800, 2008.
- Kray L., Reb J., Galinsky A., and Thompson L. Stereotype reactance at the bargaining table: The effect of stereotype activation and power on claiming and creating value. *Personality and Social Psychology Bulletin*, 30(4):399–411, 2004.
- Krikke T. F. and Truong K. P. Detection of nonverbal vocalizations using gaussian mixture models: looking for fillers and laughter in conversational speech. 2013.
- Kurzon D. The right of silence: A socio-pragmatic model of interpretation. *Journal of Pragmatics*, 23(1):55–69, January 1995.
- Laskowski K. Contrasting emotion-bearing laughter types in multiparticipant vocal activity detection for meetings. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4765–4768. IEEE, 2009.
- Laskowski K. and Schultz T. Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings. In *Machine Learning for Multimodal Interaction*, pages 149–160. Springer, 2008.
- Lathia N., Pejovic V., Rachuri K., Mascolo C., Musolesi M., and Rentfrow P. Smartphones for large-scale behaviour change interventions. *IEEE Pervasive Computing*, 12(3):66–73, 2013.
- Lee C.-H., Shih J.-L., Yu K.-M., and Lin H.-S. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *Multimedia, IEEE Transactions on*, 11(4):670–682, 2009.
- Lepri B., Staiano J., Rigato G., Kalimeri K., Finnerty A., Pianesi F., Sebe N., and Pentland A. The sociometric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 623–628. IEEE, 2012.
- Lerner G. Notes on overlap management in conversation: The case of delayed completion. *Western Journal of Speech Communication*, 53(2):167–177, August 1989.
- Lerner G. Turn-sharing: The choral co-production of talk-in-interaction. In Ford C., Fox B., and Thompson S., editors, *The language of turn and sequence*, pages 225–256. Oxford University Press, 2002.

- Lerner G. H. Finding “face” in the preference structures of talk-in-interaction. *Social Psychology Quarterly*, 1996a.
- Lerner G. H. On the “semi-permeable” character of grammatical units in conversation: Conditional entry into the turn space of another speaker. *Studies in interactional sociolinguistics*, 1996b.
- Ling R. *New Tech, New Ties. How Mobile Communication is Reshaping Social Cohesion*. MIT Press, 2008.
- Maatman R., Gratch J., and Marsella S. Natural behavior of a listening agent. In *Intelligent Virtual Agents*, pages 25–39, 2005.
- Maemo. Developer Guide, 2012. URL http://wiki.maemo.org/Documentation/Maemo/_5_Developer_Guide/Architecture/Multimedia_Domain.
- Mag-Lab. National high magnetic field laboratory.
- Magee J., Galinsky A., and Gruenfeld D. Power, propensity to negotiate, and moving first in competitive interactions. *Personality and Social Psychology Bulletin*, 33(2):200–212, 2007.
- Mairesse F., Walker M., Mehl M., and Moore R. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.
- Mana N., Lepri B., Chippendale P., Cappelletti A., Pianesi F., Svaizer P., and Zancanaro M. Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection. In *Proceedings of the 2007 workshop on Tagging, mining and retrieval of human related activity information*, pages 9–14. ACM, 2007.
- Mast M. S. Gender differences and similarities in dominance hierarchies in same-gender groups based on speaking time. *Sex Roles*, 44(9-10):537–556, 2001.
- Mast M. S. Dominance as expressed and inferred through speaking time. *Human Communication Research*, 28(3):420–450, 2002.
- Maynard S. On back-channel behavior in Japanese and English casual conversation. *Linguistics*, 24(6):1079–1108, 1986.
- Mccowan I., Lathoud G., Lincoln M., Lisowska A., Post W., Reidsma D., and Wellner P. The ami meeting corpus. In *In: Proceedings Measuring Behavior, International Conference on Methods and Techniques in Behavioral Research*. LPJJ Noldus, F. Grieco, LWS Loijens and PH Zimmerman (Eds.), Wageningen: Noldus Information Technology. Citeseer, 2005.
- McGinn K. and Croson R. What do communication media mean for negotiations? A question of social awareness. In Gelfand M. and Brett J., editors, *The handbook of negotiation and culture*, pages 334–339. Stanford University Press, 2004.
- McKeown G., Valstar M. F., Cowie R., and Pantic M. The semaine corpus of emotionally coloured character interactions. In *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, pages 1079–1084. IEEE, 2010.

- McKeown G., Valstar M., Cowie R., Pantic M., and Schroder M. The semaine database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. *Affective Computing, IEEE Transactions on*, 3(1):5–17, 2012.
- Mehl M. R. and Holleran S. E. An empirical analysis of the obtrusiveness of and participants' compliance with the electronically activated recorder (ear). *European Journal of Psychological Assessment*, 23(4):248–257, 2007.
- Mehl M. R. and Pennebaker J. W. The sounds of social life: a psychometric analysis of students' daily social environments and natural conversations. *Journal of personality and social psychology*, 84(4):857, 2003.
- Mehl M. R., Pennebaker J. W., Crow D. M., Dabbs J., and Price J. H. The electronically activated recorder (ear): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, 33(4):517–523, 2001.
- Miner F. Group versus individual decision making: An investigation of performance measures, decision strategies, and process losses/gains. *Organizational Behavior and Human Performance*, 33(1):112–124, 1984.
- Miritello G., Moro E., Lara R., Martínez-López R., Belchamber J., Roberts S. G., and Dunbar R. I. Time as a limited resource: Communication strategy in mobile phone networks. *Social Networks*, 35(1):89–95, 2013.
- Mohammadi G., Vinciarelli A., and Mortillaro M. The Voice of Personality : Mapping Nonverbal Vocal Behavior into Trait Attributions. In *Proceedings of international workshop on Social signal processing*, pages 17–20. ACM, 2010a.
- Mohammadi G., Park S., Sagae K., Vinciarelli A., and Morency L.-P. Who is persuasive? The role of perceived personality and communication modality in social multimedia. In *Proceedings of the ACM International Conference on Multimodal Interaction (to be presented)*, 2013.
- Mohammadi G. and Vinciarelli A. Automatic personality perception: Prediction of trait attribution based on prosodic features. 3(3):273–284, 2012.
- Mohammadi G., Vinciarelli A., and Mortillaro M. The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In *Proceedings of the 2nd international workshop on Social signal processing*, pages 17–20. ACM, 2010b.
- Morreall J. A new theory of laughter. *Philosophical Studies*, 42(2):243–254, 1982.
- Mostefa D., Moreau N., Choukri K., Potamianos G., Chu S. M., Tyagi A., Casas J. R., Turmo J., Cristoforetti L., Tobia F., et al. The chil audiovisual corpus for lecture and meeting analysis inside smart rooms. *Language Resources and Evaluation*, 41(3):389–407, 2007.
- Nass C. and Min Lee K. Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction and consistency-attraction. *Experimental Psychology: Applied*, 7(3):171–181, 2001.
- Nass C. I. and Brave S. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT Press Cambridge, 2005.

- Nebot M., Celentano D. D., Burwell L., Davis A., Davis M., Polacsek M., and Santelli J. Aids and behavioural risk factors in women in inner city baltimore: a comparison of telephone and face to face surveys. *Journal of Epidemiology and Community Health*, 48(4):412–418, 1994.
- Newman H. The sounds of silence in communicative encounters. *Communication Quarterly*, 30(2):142–149, 1982.
- O'Donnell Trujillo N. and Adams K. Heheh in conversation: Some coordinating accomplishments of laughter. *Western Journal of Speech Communication*, 47(2):175–191, 1983.
- Oh J., Cho E., and Slaney M. Characteristic contours of syllabic-level units in laughter. In *INTERSPEECH*, pages 158–162, 2013.
- Olaniran B. A. Group performance in computer-mediated and face-to-face communication media. *Management Communication Quarterly*, 7(3):256–281, 1994.
- Oliver J. and Srivastava S. The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives. In Previn L. and John O., editors, *Handbook of Personality: Theory and Research*, volume 2, chapter 4, pages 102–138. New York: The Guilford Press, 2nd edition, 1999.
- Oostdijk N. The spoken dutch corpus: overview and first evaluation. In *Proceedings of LREC-2000, Athens*, volume 2, pages 887–894, 2000.
- Origlia A., Abete G., and Cutugno F. A dynamic tonal perception model for optimal pitch stylization. *Computer Speech and Language*, 27(1):190–208, 2013.
- Owren M. J. and Bachorowski J.-A. Reconsidering the evolution of nonlinguistic communication: The case of laughter. *Journal of Nonverbal Behavior*, 27(3):183–200, 2003.
- Palmer R., Birchall H., McGrain L., and Sullivan V. Self-help for bulimic disorders: a randomised controlled trial comparing minimal guidance with face-to-face or telephone guidance. *The British Journal of Psychiatry*, 181(3):230–235, 2002.
- Pentland A. Socially aware, computation and communication. *Computer*, 38(3):33–40, 2005.
- Pentland A. S. *Honest signals: how they shape our world*. MIT Press, 2008.
- Petridis S. and Pantic M. Audiovisual discrimination between speech and laughter: why and when visual information might help. *Multimedia, IEEE Transactions on*, 13(2):216–234, 2011.
- Petrillo M. and Cutugno F. A syllable segmentation algorithm for english and italian. In *Proceedings of Eurospeech*, pages 2913–2916, 2003.
- Pianesi F., Zancanaro M., Lepri B., and Cappelletti A. A multimodal annotated corpus of consensus decision making meetings. *Language Resources And Evaluation*, 41(3):409–429, 2007a. ISSN 1574020X.
- Pianesi F., Zancanaro M., Lepri B., and Cappelletti A. A multimodal annotated corpus of consensus decision making meetings. *Language Resources and Evaluation*, 41(3):409–429, 2007b.

- Pinkley R. and Northcraft G. Conflict frames of reference: Implications for dispute processes and outcomes. *Academy of Management Journal*, 37(1):193–205, 1994.
- Plato. *Republic III*. a.
- Plato. *Republic V*. b.
- Plato. *Laws VII*. c.
- Plato. *Laws XI*. d.
- Poggi I. and D'Errico F. "social signals: A psychological perspective". In "Salah A. A. and Gevers T., editors, *Computer Analysis of Human Behavior*, pages "185–225". "Springer London", 2011.
- Polychroniou A., Salamin H., and Vinciarelli A. The sspnet-mobile corpus: Social signal processing over mobile phones. *Proceedings of the Language Resources and Evaluation Conference*, 2014.
- Polzehl T., Moller S., and Metze F. Automatically assessing personality from speech. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 134–140. IEEE, 2010.
- Provine R. Laughter Punctuates Speech: Linguistic, Social and Gender Contexts of Laughter. *Ethology*, 95(4):291–298, 1993.
- Purdy J., Nye P., and Balakrishnan P. The impact of communication media on negotiation outcomes. *International Journal of Conflict Management*, 11(2):162–187, 2000.
- Rabiner L. and Juang B. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.
- Rabiner L. R. and Juang B.-H. *Fundamentals of speech recognition*, volume 14. PTR Prentice Hall Englewood Cliffs, 1993.
- Rahim A. A measure of styles of handling interpersonal conflict. *Academy of Management journal*, 26(2):368–376, 1983.
- Rahim A. *Managing conflict in organizations*. Transaction Pub, 4th edition, 2010.
- Rammstedt B. and John O. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212, 2007a.
- Rammstedt B. and John O. Measuring Personality in One Minute or Less: A 10-item Short Version of the Big Five Inventory in English and German. *Journal of Research in Personality*, 41(1):203–212, 2007b.
- Ravi, N. and Dandekar, N. and Mysore, P. and Littman M. Activity recognition from accelerometer data. In *Proceedings of the national conference on artificial intelligence*, pages 1541–1546, 2005.

- Reeves B. and Nass C. *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press, 1996.
- Rice R. and Love G. Electronic emotion socioemotional content in a computer-mediated communication network. *Communication research*, 14(1):85–108, 1987.
- Richmond V. and McCroskey J. *Nonverbal behavior in interpersonal relations*. Allyn and Bacon, 1995.
- Robinson L. and Reis H. The effects of interruption, gender, and status on interpersonal perceptions. *Journal of nonverbal behavior*, 13(3):141–153, 1989.
- Rogelberg S. and Rumery S. Gender Diversity, Team Decision Quality, Time on Task, and Interpersonal Cohesion. *Small Group Research*, 27(1):79–90, 1996. ISSN 1046-4964.
- Sacks H., Schegloff E., and Jefferson G. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.
- Salamin H., Polychroniou A., and Vinciarelli A. Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In *Systems, Man, And Cybernetics.(IEEE SCM). IEEE International Conference on*, 2013a.
- Salamin H., Polychroniou A., and Vinciarelli A. Automatic recognition of personality and conflict handling style in mobile phone conversations. *International Workshop on Image and Audio Analysis for Multimedia Interactive Services*, pages 1–4, 2013b.
- Sanchez-Cortes D., Aran O., Mast M. S., and Gatica-Perez D. Identifying emergent leadership in small groups using nonverbal communicative cues. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 39:1–39:4. ACM, 2010.
- Sanchez-Cortes D., Aran O., and Gatica-Perez D. An audio visual corpus for emergent leader analysis. In *Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future*, Workshop, 2011.
- Sanchez-Cortes D., Aran O., Jayagopi D., Schmid Mast M., and Gatica-Perez D. Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition. *Journal on Multimodal User Interfaces*, 6:1–15, 2012.
- Saville-Troike M. The place of silence in an integrated theory of communication. In *Perspectives on silence*, pages 3–18. Ablex Norwood, NJ, 1985.
- Schegloff E. Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(1):1–63, 2000.
- Scherer S., Schwenker F., Campbell N., and Palm G. Multimodal laughter detection in natural discourses. In *Human Centered Robot Systems*, pages 111–120. Springer, 2009.
- Scherer S., Glodek M., Schwenker F., Campbell N., and Palm G. Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. *ACM Transactions on Interactive Intelligent Systems (TiIS)*, 2(1):4, 2012.

- Schuller B., Eyben F., and Rigoll G. Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In *Perception in multimodal dialogue systems*, pages 99–110. Springer, 2008.
- Schuller B., Steidl S., Batliner A., Nöth E., Vinciarelli A., Burkhardt F., Son R. van , Weninger F., Eyben F., Bocklet T., et al. The interspeech 2012 speaker trait challenge. In *Proc. Interspeech*, volume 2012, 2012.
- Sheffield J. The effect of communication medium on negotiation performance. *Group Decision and Negotiation*, 4(2):159–179, 1995.
- Skogstad S. A., Nymoen K., Hovin M., Holm S., and Jensenius A. R. Filtering motion capture data for real-time applications. *Group*, 20:40, 2013.
- Soleymani M., Lichtenauer J., Pun T., and Pantic M. A multimodal database for affect recognition and implicit tagging. *Affective Computing, IEEE Transactions on*, 3(1):42–55, 2012.
- Stuhlmacher A. and Walters A. Gender differences in negotiation outcome: A meta-analysis. *Personnel Psychology*, 52(3):653–677, 1999.
- Suh K. Impact of communication medium on task performance and satisfaction: an examination of media-richness theory. *Information & Management*, 35(5):295–312, 1999.
- Taylor S. and Brown J. Illusion and well-being: a social psychological perspective on mental health. *Psychological bulletin*, 1988.
- Ten Bosch L., Oostdijk N., and De Ruiter J. P. Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In *Text, Speech and Dialogue*, pages 563–570. Springer, 2004.
- Thambirajah M. *Psychological Basis of Psychiatry*. New York:Macmillan, 2004.
- Thompson L., Wang J., and Gunia B. Negotiation. *Annual Review of Psychology*, 61:491–515, 2010.
- Truong K. and Van Leeuwen D. Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158, 2007a.
- Truong K. and Van Leeuwen D. Evaluating automatic laughter segmentation in meetings using acoustic and acoustics-phonetic features. In *Proc. ICPhS Workshop on The Phonetics of Laughter, Saarbrücken, Germany*, pages 49–53, 2007b.
- Valley K., Moag J., and Bazerman M. “a matter of trust”: Effects of communication on the efficiency and distribution of outcomes. *Journal of Economic Behavior & Organization*, 34(2):211–238, 1998.
- Verschueren J. *What people say they do with words: Prolegomena to an empirical-conceptual approach to linguistic action*. Ablex Publishing Corporation, 1985.
- Vettin J. and Todt D. Laughter in Conversation: Features of Occurrence and Acoustic Structure. *Journal of Nonverbal Behavior*, 28(2):93–115, 2004.

- Vinciarelli A., Pantic M., and Bourlard H. Social signal processing: survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009a.
- Vinciarelli A., Pantic M., Heylen D., Pelachaud C., Poggi I., D’Errico F., and Schröder M. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012a.
- Vinciarelli A., Pantic M., Bourlard H., and Pentland A. Social signal processing: state-of-the-art and future perspectives of an emerging domain. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 1061–1070. ACM, 2008.
- Vinciarelli A., Dielmann A., Favre S., and Salamin H. Canal9: A database of political debates for analysis of social interactions. In *Affective Computing and Intelligent Interaction and Workshops. International Conference on*, pages 1–4. IEEE, 2009b.
- Vinciarelli A., Pantic M., Heylen D., Pelachaud C., Poggi I., and D’Errico F. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *Affective Computing, IEEE Transactions on*, 3(1):69–87, 2012b.
- Vinciarelli A., Salamin H., and Polychroniou A. Negotiating over mobile phones: Calling or being called can make the difference. *Cognitive Computation*, 6(1):1–12, 2014.
- Volkema R. and Ronald H. The Influence of Cognitive-based Group Composition on Decision-making Process and Outcome. *Journal of Management Studies*, 35(1):105–121, 2002.
- Wagner J., Lingenfelser F., and André E. Using phonetic patterns for detecting social cues in natural conversations. In *INTERSPEECH*, pages 168–172, 2013.
- Walters A., Stuhlmacher A., and Meyer L. Gender and negotiator competitiveness: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 76(1):1–29, 1998.
- Walther J. B. Computer-mediated communication impersonal, interpersonal, and hyperpersonal interaction. *Communication research*, 23(1):3–43, 1996.
- Ward N. Using prosodic clues to decide when to produce back-channel utterances. In *Proceedings of Spoken Language*, pages 1728–1731, 1996.
- Ward N. and Tsukahara W. Prosodic features which cue back-channel responses in English and Japanese. *Journal of Pragmatics*, 32(8):1177–1207, 2000.
- Weinberger M., Nagle B., Hanlon J. T., Samsa G. P., Schmader K., Landsman P. B., Uttech K. M., Cowper P. A., Cohen H. J., Feussner J. R., et al. Assessing health-related quality of life in elderly outpatients: telephone versus face-to-face administration. *Journal of the American Geriatrics Society*, 42(12):1295, 1994.
- Wennerstrom A. and Siegel A. Keeping the floor in multiparty conversations: Intonation, syntax, and pause. *Discourse Processes*, 36(2):77–107, 2003.
- Xie L. and Liu Z.-Q. A comparative study of audio features for audio-to-visual conversion in mpeg-4 compliant facial animation. In *Machine Learning and Cybernetics, 2006 International Conference on*, pages 4359–4364. IEEE, 2006.

-
- Yngve V. On getting a word in edgewise. In *Chicago Linguistics Society*, pages 567–578, 1970.
- Young S., Evermann G., Gales M., Hain T., Kershaw D., Liu X., Moore G., Odell J., Ollason D., Povey D., et al. The htk book. *Cambridge University Engineering Department*, 3, 2002.