Roderick Murray-Smith (Ed.)

LNCS 8045

Mobile Social Signal Processing

First International Workshop, MSSP 2010
Lisbon, Portugal, September 7, 2010
Invited Papers

Springer

Lecture Notes in Computer Science 8045

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Roderick Murray-Smith (Ed.)

Mobile Social Signal Processing

First International Workshop, MSSP 2010
Lisbon, Portugal, September 7, 2010
Invited Papers

Springer

Volume Editor

Roderick Murray-Smith
University of Glasgow
School of Computing Science
18 Lilybank Gardens, Glasgow G12 8RZ, UK
E-mail: roderick.murray-smith@glasgow.ac.uk

ISSN 0302-9743 e-ISSN 1611-3349
ISBN 978-3-642-54324-1 e-ISBN 978-3-642-54325-8
DOI 10.1007/978-3-642-54325-8
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014931095

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Introduction

This edited volume was produced after the First International Workshop on Mobile Social Signal Processing (SSP). The Workshop, chaired by A. Vinciarelli, R. Murray-Smith, and H. Bourlard, brought together the Mobile HCI and Social Signal Processing research communities. The former investigates approaches for effective interaction with mobile and wearable devices, while the latter focuses on modelling, analysis, and synthesis of nonverbal behavior in human–human and human–machine interactions. While dealing with similar problems, the two domains have different goals and methodologies. However, mutual exchange of expertise is likely to raise new research questions as well as to improve approaches in both domains. This volume contains a range of papers invited after the workshop, which represent the diversity of the two fields and areas of overlap.

Conversation is the "primordial site of human sociality and social life" [16]. Thus, it is not surprising to observe that mobile phones, allowing one to talk with virtually anybody at virtually any moment, have pervaded our everyday life more quickly and deeply than any previous technology, and that they empower people and give them new ways to interact with their environment and social network [8]. However, while becoming a preeminent form of social interaction, mobile phone conversations have been the subject of limited investigation from both psychological and technological points of view [1, 7]. The reason is not only that the diffusion of mobile phones is a relatively recent phenomenon, but also that phone conversations have traditionally been considered nothing more than particular cases of face-to-face conversations, characterized by speech being the only information at disposition, in contrast to actual face-to-face conversations where humans are known to exchange not only words, but also a wide spectrum of nonverbal behavioral cues (e.g., facial expressions, postures, gestures, vocalizations, etc.) accounting for social, affective, and relational phenomena [6, 12, 14]. From the design side, the scientific evaluation of mobile phone designs in realistic settings is difficult, especially when the impact of the design on social aspects is an important factor [11, 4].

Mobile Devices and Social Signal Processing

The situation described above makes clear that there is an interesting gap in the research literature where three important phenomena take place in the scientific and technological landscape:

1. Modern mobile devices have moved beyond basic voice and text communications, and are now equipped with significant sensing and processing ability,

e.g., video, GPS, accelerometers, magnetometers, and capacitive touch [2, 3]. Also, the increasing processing power and the potential to use server-side processing allows the use of algorithms previously considered only possible on powerful PCs, capturing, with unprecedented depth and precision, the context and behavior of their users (e.g., position, movement, hand grip behavior, proximity to social network members, gait type, auditory context). This behavior can also potentially be compared with large numbers of other users, to categorize the style of interaction [13].

2. Automatic analysis, synthesis and understanding of verbal and nonverbal communication, typically captured with multiple sensors, is one of the hottest topics in the computing community. This applies in particular to Social Signal Processing (SSP), the new, emerging domain aimed at bringing social intelligence to machines [18, 19]. The use of nonverbal behavioral cues as a physical, machine-detectable evidence of social phenomena that are not otherwise accessible to human perception and machine sensing [17] is supported by several decades of research in social psychology showing that nonverbal communication is the channel through which we perceive social aspects of our interactions [6, 12, 14]. Probabilistic approaches are used to infer these ambiguous states, and in some cases can also synthesize displays of nonverbal communication via artificial faces, vibration, and voices to elicit the appropriate social perceptions in the humans receiving the message.
3. This mobility and diverse usage provides interesting new research opportunities to measure and influence social interactions in ways that would have been extremely difficult only a few years ago. Because modern mobile devices can sense movement, muscle tremor, location, the proximity of other devices and as they can sample audio and video signals and magnetic field disturbances this gives us opportunities to record in greater detail than ever before human activity, including that of social interactions [10, 9]. It also allows us to design experiments that can stimulate users in specific contexts, allowing a trade-off between realistic conditions and experimental control [15]. Potential benefits for Mobile HCI research from the SSP community include the use of techniques to help infer emotional consequences for users of different mobile interaction designs.

Overview of Chapters

Vinciarelli opens the book with a chapter that provides an overview of the way in which mobile devices can sample the social context, and the SSP and psychology literature associated with such analysis.

Chapter 2, by Favre, presents approaches for automatically recognizing the roles people play in a wide range of interaction settings. The methods are tested on one of the biggest data sets ever used in literature for this task – over roughly 90 hours of material, composed of broadcast material and meeting recordings.

In Chapter 3, Valente and Vinciarelli explore the *Speaker Diarization* problem, which aims at inferring who spoke when in an audio stream and involves two

simultaneous unsupervised machine learning tasks: the estimation of the number of speakers, and the association of speech segments to each speaker. When the roles of the people involved are known, it can lead to a significant improvement in accuracy.

Chapter 4 is also about inferring a user's identity, but this time for authentication purposes, to ensure privacy and security. Authentication, as traditionally achieved by means of a shared secret, is effortful and deliberate. Frequent and repeated authentication easily becomes a hurdle, an annoyance, and a burden. These behavioral biometrics propose using non explicit patterns such as keystroke dynamics, use patterns, and voice analysis techniques to create a multimodal biometric authentication mechanism. These behavioral biometrics take advantage of tasks that the user already performs, thereby reducing the need for explicit authentication by more traditional means, and in many ways mirror the nonverbal communication channels that are the focus of the SSP community.

Two chapters in the book, Chap. 5 by Harper and Chap. 6 by Williamson and Brewster, are informed to a significant degree by a sociological perspective. Harper in this chapter, and in his recent book [5], worries that basing technical solutions on cybernetic theories and Bayesian reasoning is the wrong way to go, because they miss the moral values involved, and that interaction with a machine cannot have these values. He highlights the limitations of many information processing models of humans. While it could be argued that these are merely limitations of incomplete or superficial models, he does highlight the complex nature of human behavior, and, for example, the need to know that the reason someone is reading may have very little, in the short term, to do with the process of information transfer from the newspaper they hold in their hands.

Chapter 6 by Williamson and Brewster explores ways to make the context and activity levels visible to others, and explore this from a performative aspect. The chapter investigates how participants might choose to perform multimodal interactions in real-world settings, examine the social acceptability of that performance, and understand more about the user experience of performing within an application context. Participants were required to generate multimodal input in situ in public and private locations using a mobile remote awareness application with a partner over repeated trials. However, although the application in this study was based on remote awareness, the purpose of this application was not concerned with the meaning or intention behind communications. The application was designed to support divergent multimodal inputs, create the experience of performing in different settings and participate as a distant audience member for a familiar other's performances.

One of the ways mobile interaction designers have attempted to address the reduction in social signals in mobile phone conversations is by augmenting the interaction via other channels. We have three chapters that use vibrotactile and visual channels in different ways.

In Chapter 7 Trendafilov et al. investigate negotiation models for mobile tactile interaction. The chapter describes an experiment with a multimodal implementation that allows users to engage in a continuous interaction with each

other by using capacitive touch input, visual and/or vibro-tactile feedback, and to perform a goal-oriented collaborative task of target acquisition. The participants found this new form of interaction interesting and engaging, and believed it could encourage communication with people, which opens new possibilities for the development of richer social interactions. However, the significant increase in overall workload and decrease in performance, associated with the tactile modality, opens up the need to explore new approaches for future "in pocket" interaction studies.

The feelabuzz system described in Chapter 8 by Tünnerman et al. provides a direct tactile coupling between mobile phones, based on accelerometer readings. This can be used for implicit context communication, i.e., the background monitoring of the natural movements of the users themselves or their environments, as well as for direct voluntary and symbolic communication.

In Chapter 9 Crossan et al. use a multimodal contact list to allow people to express their moods and status in a tactile manner. The Multimodal Contact List provides a mechanism to browse context information and communicate with friends in a contact list both visually and through touch. Each contact can share with their friend group selected information on their current context such as mood and availability. Users can close the loop with their conversation partners not only with the standard audio link, but also via touch and visual feedback or a combination of all three. A user can then progressively probe the contact for more detailed information, eventually allowing the user to open a real-time multimodal voice and tactile communication channel to the contact for verbal or discreet tactile communication. The paper presents an initial two-stage evaluation of this concept, which demonstrates how designers must take care when combining unusual combinations of feedback channels.

As mobile phones are one of the most important instruments of our social life, the cross-pollination between Mobile HCI and Social Signal Processing is likely to foster on one hand a better understanding of the way people interact via phone and, on the other hand, of how to make mobile phones more centered on social interaction. We hope that this book provides a useful starting point to help the different research communities begin to interact more.

January 2014 Roderick Murray-Smith

References

1. Arminen, I., Weilenmann, A.: Mobile presence and intimacy–Reshaping social actions in mobile contextual configuration. Journal of Pragmatics 41(10), 1905–1923 (2009)
2. Bellotti, V., Back, M., Edwards, W.K., Grinter, R.E., Henderson, A., Lopes, C.: Making sense of sensing systems: five questions for designers and researchers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves, CHI 2002, pp. 415–422. ACM, New York (2002)
3. Benford, S., Schnädelbach, H., Koleva, B., Anastasi, R., Greenhalgh, C., Rodden, T., Green, J., Ghali, A., Pridmore, T., Gaver, B., Boucher, A., Walker, B., Pennington, S., Schmidt, A., Gellersen, H., Steed, A.: Expected, sensed, and desired: A framework for designing sensing-based interaction. ACM Trans. Comput.-Hum. Interact. 12, 3–30 (2005)
4. Crossan, A., Murray-Smith, R., Brewster, S., Musizza, B.: Instrumented Usability Analysis for Mobile Devices. In: Handbook of Research on User Interface Design and Evaluation for Mobile Technology (2007)
5. Harper, R.H.R.: Texture: human expression in the age of communications overload. MIT Press, Boston (2010)
6. Knapp, M., Hall, J.: Nonverbal Communication in Human Interaction. Harcourt Brace College Publishers (1972)
7. Licoppe, C.: Recognizing mutual "proximity" at a distance: Weaving together mobility, sociality and technology. Journal of Pragmatics 41(10), 1924–1937 (2009)
8. Lumsden, J.: Handbook of Research on User Interface Design and Evaluation for Mobile Technology. IGI Global (2008)
9. Morrison, A., Bell, M., Chalmers, M.: Visualisation of spectator activity at stadium events. In: 2009 13th International Conference on Information Visualisation, pp. 219–226 (2009)
10. Murray-Smith, R., Ramsay, A., Garrod, S., Jackson, M., Musizza, B.: Gait alignment in mobile phone conversations. In: Proc. of Mobile HCI 2007, pp. 214–221 (2007)
11. Oulasvirta, A., Tamminen, S., Roto, V., Kuorelahti, J.: Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile hci. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2005, pp. 919–928. ACM, New York (2005)
12. Poggi, I.: Mind, Hands, Face and Body: A Goal and Belief View of Multimodal Communication. Weidler Buchverlag Berlin (2007)
13. Raento, M., Oulasvirta, A., Eagle, N.: Smartphones: an emerging tool for social scientists. Sociological Methods & Research 37(3), 426 (2009)
14. Richmond, V., McCroskey, J.: Nonverbal Behaviors in interpersonal relations. Allyn and Bacon (1995)
15. Roto, V., Oulasvirta, A., Haikarainen, T., Kuorelahti, J., Lehmuskallio, H., Nyyssonen, T.: Examining mobile phone use in the wild with quasi-experimentation. Helsinki Institute for Information Technology (HIIT), Technical Report, vol. 1 (2004)
16. Schegloff, E.: Analyzing single episodes of interaction: An exercise in conversation analysis. Social Psychology Quarterly 50(2), 101–114 (1987)
17. Vinciarelli, A.: Capturing order in social interactions. IEEE Signal Processing Magazine 26(5), 133–137 (2009)

18. Vinciarelli, A., Pantic, M., Bourlard, H.: Social Signal Processing: Survey of an emerging domain. Image and Vision Computing Journal 27(12), 1743–1759 (2009)
19. Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A.: Social Signal Processing: State-of-the-art and future perspectives of an emerging domain. In: Proceedings of the ACM International Conference on Multimedia, pp. 1061–1070 (2008)

Organization

Executive Committee

Co-chairs

Alessandro Vinciarelli	University of Glasgow, UK, and Idiap Research Institute, Switzerland
Roderick Murray-Smith	University of Glasgow, UK
Hervé Bourlard	Idiap Research Institute and EPFL, Switzerland

Program Committee

Marco Cristani	University of Verona, Italy
Anind Dey	Carnegie Mellon University, USA
Thomas Hermann	University of Bielefeld, Germany
Rob Jenkins	University of Glasgow, UK
Matt Jones	Swansea University, Wales
Juha K. Laurila	Nokia Research Center, Lausanne
Dirk Heylen	University of Twente, The Netherlands
Eamonn ONeill	University of Bath, UK
Antti Oulasvirta	HIIT, Finland
Jean-Marc Odobez	Idiap Research Institute/EPFL
Isabella Poggi	Universitá Roma Tre, Italy
Steve Renals	University of Edinburgh, UK
Christ Schmandt	MIT, USA
Fabio Valente	Idiap Research Institute, Switzerland

Sponsoring Institutions

This workshop was supported in part by the European Community's Seventh Framework Programme (FP7/2007-2013), under grant agreement no. 231287 (SSPNet), and in part by the Swiss National Science Foundation via the National Centre of Competence in Research IM2 (Information Multimodal Information management).

Table of Contents

Mobile Phones and Social Signal Processing for Analysis and Understanding of Dyadic Conversations

Alessandro Vinciarelli

University of Glasgow
Sir A.Williams Building, G12 8QQ Glasgow, UK
Idiap Research Institute
CP592, 1920 Martigny, Switzerland

Abstract. Social Signal Processing is the domain aimed at bridging the social intelligence gap betweeen humans and machines via modeling, analysis and synthesis of nonverbal behavior in social interactions. One of the main challenges of the domain is to sense unobtrusively the behavior of social interaction participants, one of the key conditions to preserve the spontaneity and naturalness of the interactions under exam. In this respect, mobile devices offer a major opportunity because they are equipped with a wide array of sensors that, while capturing the behavior of their users with an unprecedented depth, are still invisible. This is particularly important because mobile devices are part of the everyday life of a large number of individuals and, hence, they can be used to investigate and sense natural and spontaneous scenarios.

1 Introduction

The number of mobile phone users in the world has been recently estimated to be around 3.5 billions, more than 50% of the current world population [13]. The diffusion changes significantly depending on the country: while in Papua New Guinea only 0.44 percent of the population subscribes to a mobile telephony service, the same figure is 154 percent in the case of Luxemburg (more than one phone per person). In the developed countries (in particular Europe and the Americas) virtually everybody holds a mobile subscription, but the penetration is high and growing in the developing world as well (300 millions new users are expected in India in the next few years) [12]. The same variability across countries can be observed for what concens the amount of time spent on the phone, ranging between 22 and 800 minutes per month [13].

A mere 15 years ago it was hard to predict the impressive figures above. Even in a country like Italy, where the density of mobile phones is today among the highest in the world, sociologists used to observe prevailing negative feelings in surveys about the acceptance of mobile technologies [11]. The main change since then is that mobile phones are no longer an instrument for professional or emergency calls only (as it used to be at the beginning of their diffusion), but

R. Murray-Smith (Ed.): MSSP 2010, LNCS 8045, pp. 1–8, 2014.

one of the main channels through which we get involved in social interactions. Mobile phones provide the possibility of starting a conversation, the "primordial site of human sociality and social life" [27], at virtually every moment of the day, almost independently of where we are and what we do. Furthermore, mobile phones extend our opportunities for social contacts well beyond conversations to include the exchange of text messages (roughly 2×10^5 SMS per second have been exchanged worldwide in 2010 [12]) as well as the access to popular social media (e.g., Facebook, LinkedIn, etc.). In this respect, mobile phones seem to be a key support for our social life and an ideal response to the needs of the "social animal" [35].

Thus, it is not surprising to observe that both social scientists and computing researchers have identified mobile phones or, more generally, mobile and wearable devices as an instrument to access social life with at an unprecedented depth and scale [23]. This applies in particular to naturalistic settings difficult to observe in the laboratory, whether this means to identify daily routines in the life of social groups [9], to look for personality traces in everyday speaking behavior [20], or to sense the overall behavior of an organization [22], just to name a few examples. In all cases above, mobile devices have been used as an unobtrusive, but ubiquitous and pervasive sensor that can be carried without effort and, to a certain extent, without awareness in the most natural settings of our everyday life (see [21] for an example of how unobtrusiveness is assessed).

In such a perspective, mobile phones have a major advantage with respect to other wearable devices because they are an everyday object and are carried spontaneously, in contrast with any other device designed for sensing and collecting data. Furthermore, standard mobile phones are now equipped with an increasingly wider range of sensors (magnetometers, GPS, accelerometers, etc.) that reduce the sensing capability gap with respect to devices explicitly designed for scientific experiments.

For the reasons above, mobile phones appear to be particularly suitable for research in Social Signal Processing (SSP), the domain aimed at automatic understanding of social interactions via modeling, analysis and synthesis of nonverbal behavior (see Section 2 for more details) [34]. In fact, the sensors of a standard mobile phone allow one to capture not only nonverbal speech aspects (prosody, vocalizations, pauses, etc.), but also non verbal cues related to body movement (via accelerometers, gyroscopes and magnetometers) that are typically difficult to capture otherwise in ecologically valid settings, but still carry socially relevant information [15].

In particular, SSP appears to be one of the most suitable paradigms to develop approaches for automatic analysis and understanding of dyadic conversations, an interaction scenario that, despite its primacy and frequency (phones are used most of the times to call even though the younger generations tend to favor the use of SMS), has been so far neglected from both a technological and psychological points of view. As a result, mobile phones could reduce the social intelligence gap with respect to their users [35], support the effectiveness of task oriented calls (e.g., moderating people talking too much or deflating conflicts), activate

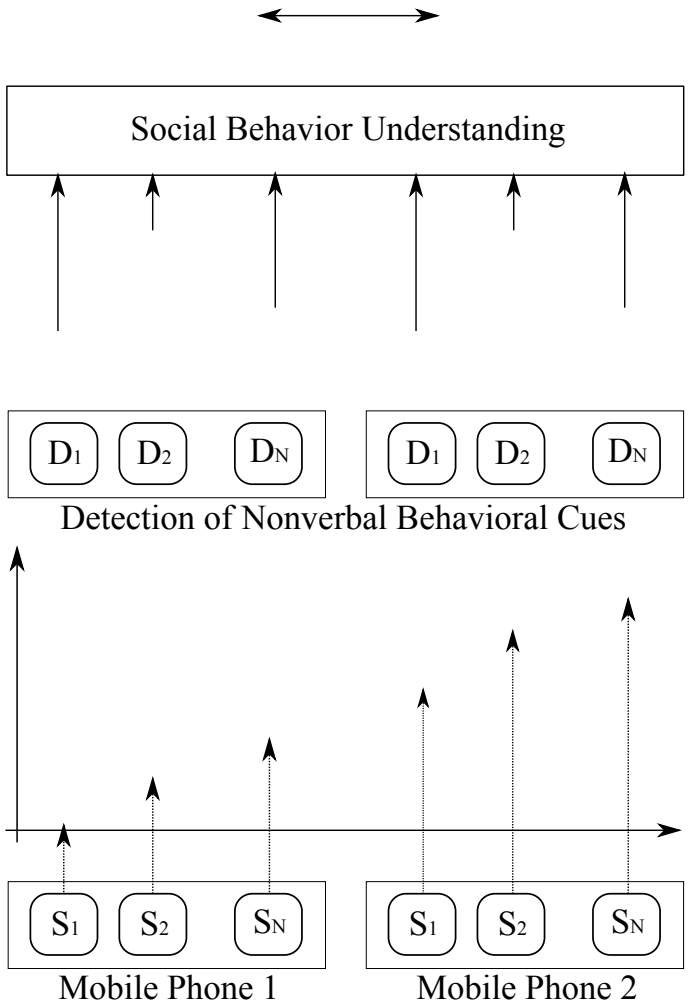


Fig. 1. Overall scheme of an SSP approach applied to mobile phone conversations. The signals captured with the sensors of the two phones $(S_1, \ldots, S_N)$ are fed to nonverbal cues detectors $(D_1, \ldots, D_N)$. The output of these latter is then automatically interpreted to identify the social signals being exchanged between the speakers.

services appropriate to the social context (e.g., by canceling background noise in case of formal conversations), etc.

The rest of this paper shows how mobile phones can be used to perform SSP research (Section 2) and what are the main challenges facing the application of SSP in mobile environments (Section 3). The final Section 4 draws some conclusions.

2 Social Signal Processing and Mobile Phones

Social interactions are accompanied by a wide spectrum of nonverbal behavioral cues (facial expression, vocalizations, gestures, postures, etc.) [15,24] that add layers of meaning, typically related to social and affective aspects of an interaction, to the words being said [36]. While our attention tends to focus on what people say, a number of cognitve processes (typically taking place outside conscious awareness) interpret nonverbal behavior of others in terms of socially relevant cues, including values, beliefs, emotions, goals, intentions, etc. [30]. These processes take place independently of any actual need or will for them taking place, but they influence to a large, sometimes dominant extent our social behavior, especially in the earliest stages of an interaction [31].

Social Signal Processing (SSP) relies on the phenomenon above and proposes to use nonverbal communication as a physical, machine detectable evidence of social signals, the perceivable stimuli (including nonverbal behavioral cues) that are produced during social interactions and "[...] *play a part in the formation and adjustment of relationships and interactions* [...] *or provide information about the agents; and that can be addressed by technologies of signal processing and synthesis*"[1] (see [34,35] for an extensive survey of the domain). The choice of nonverbal behavior as a privileged cue for understanding social phenomena results from several decades of investigations in psychology, anthropology and other human sciences (see [15,24] for extenisve monographies about the subject): "*thin slices of behavior*" [2], short samples of nonverbal behavior collected during a social interaction, appear to be sufficient to provide accurate social judgments in a large number of situations [1,6].

Figure 1 shows how the SSP paradigm can be applied in the case of mobile phone conversations. When two people are involved in a phone conversation, they naturally make use of a number of sensors embedded nowadays in a large number of standard phones available on the market. Besides microphones, without which phone calls would be obviously impossible, the most common sensors available on a phone include accelerometers, magnetometers, Global Positioning Systems, gyroscopes, etc. Thus, each of the phones can be thought of as an array of sensors $(S_1, \ldots, S_N)$ capturing signals that, potentially, carry information about the nonverbal behavior of their users.

[1] The quote comes from the "Belfast Declaration", the document issued by the Social Signal Processing Network (European Network of Excellence on SSP). The document is available for download at the following link: `http://sspnet.eu/about/`.

The main difference with respect to sensing approaches commonly applied in SSP is the lack of cameras, essential to capture fundamental nonverbal cues such as facial expressions and gaze behavior. However, this should not represent a major problem for two main reasons: the first is that approaches based on vocal behavior (accessible via the phone microphones) tend to achieve, at least in the SSP works presented so far in the literature, satisfactory performances [35]. The second is that the lack of visual information about interlocutors corresponds to the actual condition of people talking on the phone. Hence, the lack of cameras portraying the interactants simply reflects the condition of the users. Furthermore, many phones allow one to perform video-calls and such an opportunity, not particulary exploited today, might extend the analysis of face and gaze behavior to mobile phone based interaction scenarios.

Once the signals have been captured, it is possible to detect nonverbal behavioral cues using different approaches (identified as D_i in Figure 1) depending on the particular sensor. The extraction of vocal cues from speech signals is the subject of a large number of works in the literature, in particular when it comes to emotion recognition (see [4,26,28] for psychological research and [32] for technological approaches), inference of social information from turn-organization (see [33] for an introduction to the problem and [35] for an extensive survey), and analysis of traits (see [29] for an exhaustive description of cues currently extracted from speech).

The other sensors available on the phone (accelerometers, magnetometers, etc.) have not been used extensively in SSP, at least for what concerns face-to-face scenarios. SSP works aimed at the analysis of large social networks (see, e.g., [9,22]) generally make use of proximity detectors (e.g., bluetooth and RFID) to identify direct interactions between people, but do not consider accelerometers. In contrast, accelerometers have been used extensively in the ubiquitous computing community, especially to recognize the "context" (see [7] for a definition of what it is meant by this) and the actions being performed by users (see [10,16] for extensive surveys). Furthermore, accelerometers have been used to improve interaction with machines (e.g., in a gesture based design system [14]), or computer mediated communication (e.g., in a system aimed at sharing information about travels [25]).

3 Main Challenges

From a technological point of view, Mobile SSP faces the same challenges as any other SSP investigation (see [35] for an extensive survey), including fusion of multiple modalities where behavioral cues take place at different time-scales, modeling of annotation variability in judgmental studies involving multiple raters, definition of continuous rather than categorical descriptors of social and psychological phenomena, etc. However, two main challeges are specific of the application of SSP in mobile conversations, namely the modeling of principles and laws underlying phone mediated conversations and the redefinition of the concept of privacy. The rest of this section will focus on these.

Phone conversations tend to be considered as a specific case of face-to-face interaction where visual cues are not available. However, such a view does not consider that talking through a phone does not simply eliminate the visual channel, but it constrains the array of cues that people can use to convey social meaning. Therefore, communication practices must undergo significant changes to accomplish simple social goals like, e.g., the communication of immediacy [3] and proximity [18]. Furthermore, people participating in mobile phone conversations are often immersed in contexts where they are interacting with other, co-located individuals and this induces further changes in the social needs to be addressed [8,19]. Taking into account this type of issues is a crucial step towards the improvement of Mobile SSP technologies.

In a context where personal data is considered "*the new oil of the internet and the new currency of the digital world*" [37], mobile SSP can attract significant interest. On one hand, the analysis of nonverbal communication respects the privacy because it does not take into account what people say. On the other hand, recent work on social media shows that privacy protected information can be effectively inferred from publicly available cues [17]. In other words, the very concept of privacy should be redesigned in light of mobile SSP progresses. This is a major issue that can make the difference between SSP technologies being accepted or not by the users.

4 Conclusions

This article has outlined research opportunities and challenges that can emerge from the cross-pollination between Social Signal Processing - the domain aimed at modelling, analysis and synthesis of nonverbal behavior in social interactions - and mobile Human-Computer Interaction. The increasingly wide array of sensors embedded on standard mobile devices is transforming these latter in a laboratory for human behavior analysis [23]. However, technologies capable of analyzing social and psychological phenomena at the level of one-to-one conversations might become a significant threat for the privacy of people.

The identification of a correct tradeoff between the two conflicting phenomena above is beyond the scope of this article and, in any case, it requires a large societal debate [5]. From a strictly scientific point of view, the analysis of mobile phone conversations in a laboratory context, where subjects are aware of being recorded, promises to bring significant progress in domains like understanding of human behavior, development of new sensors, and improvement of automatic behavior analysis techniques. In other words, SSP can contribute to make mobile phones, one of the main infrastructures of our social life, more socially intelligent.

References

1. Ambady, N., Bernieri, F., Richeson, J.: Towards a histology of social behavior: judgmental accuracy from thin slices of behavior. In: Zanna, M.P. (ed.) Advances in Experimental Social Psychology, pp. 201–272 (2000)

2. Ambady, N., Rosenthal, R.: Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. Psychological Bulletin 111(2), 256–274 (1992)
3. Arminen, I., Weilenmann, A.: Mobile presence and intimacy - reshaping social actions in mobile contextual configuration. Journal of Pragmatics 41(10), 1905–1923 (2009)
4. Bachorowski, J.-A.: Vocal Expression and Perception of Emotion. Current Directions in Psychological Science 8(2), 53–57 (1999)
5. Bauman, Z., Lyon, D.: Liquid Surveillance. Polity Press (2013)
6. Curhan, J.R., Pentland, A.: Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. Journal of Applied Psychology 92(3), 802–811 (2007)
7. Dourish, P.: What we talk about when we talk about context. Personal and Ubiquitous Computing 8(1), 19–30 (2004)
8. Dourish, P., Bell, G.: Divining a digital future: mess and mythology in ubiquitous computing. MIT Press (2011)
9. Eagle, N., Pentland, A.: Reality mining: sensing complex social systems. Personal and Ubiquitous Computing 10(4), 255–268 (2005)
10. Figo, D., Diniz, P.C., Ferreira, D.R., Cardoso, J.M.P.: Preprocessing techniques for context recognition from accelerometers data. Personal and Ubiquitous Computing 14(7), 645–662 (2010)
11. Fortunati, L., Manganelli, A.M.: The social representation of communications. Personal and Ubiquitous Computing 12(6), 421–431 (2008)
12. ITU. The World in 2010: ICT Facts and Figures. Technical report, International Telecommunication Union (2010)
13. Kalba, K.: The Global Adoption and Diffusion of Mobile Phones. Technical Report December, Center for Information Policy Research Harvard University (2008)
14. Kela, J., Korpipää, P., Mäntyjärvi, J., Kallio, S., Savino, G., Jozzo, L., Di Marca, S.: Accelerometer-based gesture control for a design environment. Personal and Ubiquitous Computing 10(5), 285–299 (2006)
15. Knapp, M.L., Hall, J.A.: Nonverbal Communication in Human Interaction. Harcourt Brace College Publishers (1972)
16. Knight, J.F., Bristow, H.W., Anastopoulou, S., Baber, C., Schwirtz, A., Arvanitis, T.N.: Uses of accelerometer data collected from a wearable system. Personal and Ubiquitous Computing 11(2), 117–132 (2007)
17. Kosinski, M., Stillwell, D., Graepel, T.: Private traits and attributes are predictable from digital records of human behavior. Proceeedings of the National Academy of Sciences 110(15), 5802–5805 (2013)
18. Licoppe, C.: Recognizing mutual 'proximity' at a distance: Weaving together mobility, sociality and technology. Journal of Pragmatics 41(10), 1924–1937 (2009)
19. Ling, R.: New tech, new ties. MIT Press (2008)
20. Mehl, M.R., Pennebaker, J.W., Crow, D.M., Dabbs, J., Price, J.H.: The Electronically Activated Recorder (EAR): a device for sampling naturalistic daily activities and conversations. Behavior Research Methods, Instruments and Computers 33(4), 517–523 (2001)
21. Mehl, M.R., 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)

22. Olguin Olguin, D., Waber, B.N., Kim, T., Mohan, A., Ara, K., Pentland, A.: Sensible organizations: technology and methodology for automatically measuring organizational behavior. IEEE Transactions on Systems, Man and Cybernetics Part B 39(1), 43–55 (2009)
23. Raento, M., Oulasvirta, A., Eagle, N.: Smartphones: An emerging tool for social scientists. Sociological Methods & Research 37(3), 426–454 (2009)
24. Richmond, V.P., McCroskey, J.C.: Nonverbal Behavior in Interpersonal Relations. Allyn and Bacon (2000)
25. Robinson, S., Eslambolchilar, P., Jones, M.: Exploring casual point-and-tilt interactions for mobile geo-blogging. Personal and Ubiquitous Computing 14(4), 363–379 (2010)
26. Russell, J.A., Bachorowski, J.A., Fernandez-Dols, J.M.: Facial and Vocal Expressions of Emotion. Annual Reviews in Psychology 54(1), 329–349 (2003)
27. Schegloff, E.A.: Analyzing Single Episodes of Interaction: An Exercise in Conversation Analysis. Social Psychology Quarterly 50(2), 101–114 (1987)
28. Scherer, K.R.: Vocal communication of emotion: A review of research paradigms. Speech Communication 40(1-2), 227–256 (2003)
29. Schuller, B.: Voice and Speech Analysis in Search of States and Traits, pp. 233–258. Springer (2011)
30. Uleman, J.S., Newman, L.S., Moskowitz, G.B.: People as flexible interpreters: Evidence and issues from spontaneous trait inference, vol. 28, pp. 211–279. Elsevier (1996)
31. Uleman, J.S., Saribay, S.A., Gonzalez, C.M.: Spontaneous inferences, implicit impressions, and implicit theories. Annual Reviews of Psychology 59, 329–360 (2008)
32. Ververidis, D., Kotropoulos, C.: Emotional speech recognition: Resources, features, and methods 48(9), 1162–1181 (2006)
33. Vinciarelli, A.: Capturing Order in Social Interactions. IEEE Signal Processing Magazine 26(5), 133–137 (2009)
34. Vinciarelli, A., Pantic, M., Bourlard, H.: Social Signal Processing: Survey of an emerging domain. Image and Vision Computing Journal 27(12), 1743–1759 (2009)
35. Vinciarelli, A., Pantic, M., Heylen, D., Pelachaud, C., Poggi, I., D'Errico, F., Schroeder, 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 (2012)
36. Wharton, T.: The Pragmatics of Non-Verbal Communication. Cambridge University Press (2009)
37. World Economic Forum. Personal data: the emergence of a new asset class. Technical report, World Economic Forum (2011)

Turns Analysis for Automatic Role Recognition

Sarah Favre

Idiap Research Institute
CP 592, 1920 Martigny, Switzerland
Ecole Polytechnique Federale de Lausanne
1015 Lausanne, Switzerland

Abstract. This article presents approaches for recognizing automatically the roles people play in a wide range of interaction settings. The proposed role recognition approach includes two main steps. The first step aims at representing the individuals involved in an interaction with feature vectors accounting for their relationships with others. This step includes three main stages, namely segmentation of audio into turns (i.e. time intervals during which only one person talks), conversion of the sequence of turns into a social network, and use of the social network as a tool to extract features for each person. The second step uses machine learning methods to map the feature vectors into roles. The experiments have been carried out over roughly 90 hours of material. This is not only one of the largest databases ever used in literature on role recognition, but also the only one, to the best of our knowledge, including different interaction settings. In the experiments, the accuracy of the percentage of data correctly labeled in terms of roles is roughly 80% in production environments and 70% in spontaneous exchanges (lexical features have been added in the latter case).

1 Introduction

The computing community has shown a significant interest for the analysis of social interactions in the last decade. Different aspects of social interactions have been studied such as dominance, emotions, conflicts, etc. However, the recognition of roles has been neglected whereas these are a key aspect of social interactions. In fact, sociologists have shown not only that people play roles each time they interact but also that roles shape behavior and expectations of interacting participants:

> *"People do not interact with one another as anonymous beings. They come together in the context of specific environments and with specific purposes. Their interactions involve behaviors associated with defined statuses and particular roles. These statuses and roles help to pattern our social interactions and provide predictability"* [17].

Recently, the role recognition problem has attracted more and more interest and has been addressed by different groups in the computing community (see e.g. [2][9]

R. Murray-Smith (Ed.): MSSP 2010, LNCS 8045, pp. 9–21, 2014.

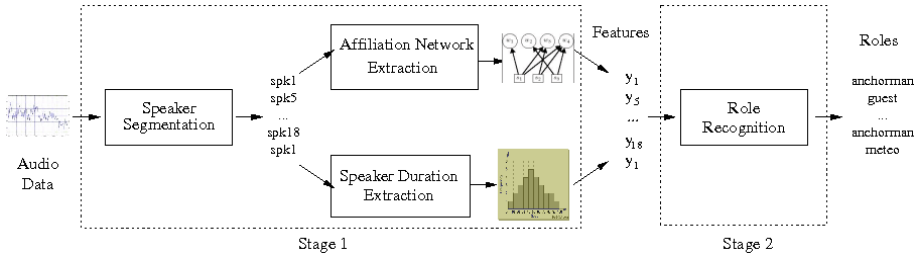


Fig. 1. Role recognition approach. The picture shows the two main stages of the approach: the features extraction and the actual role recognition.

[18][20][12] and [8]). The work presented in this article is part of this effort to tackle the role recognition problem.

The next section (Section 2) presents approaches for the automatic detection of the roles of the persons interacting in different situations, such as production environment contexts (e.g., news and talk-shows) and spontaneous exchanges (e.g.m meetings).

2 Role Recognition

Even if the concept of *role* is one of the most popular ideas in the social sciences, a formal definition is hard to find. In our approaches, we have considered roles defined as the following:

> "Role theory concerns one of the most important features of social life, characteristic behavior patterns or *roles*. It explains roles by presuming that persons are members of *social positions* and hold *expectations* for their own behaviors and those of other persons" [3].

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

The presented role recognition approach includes two main stages (see Figure 1): the first is the *feature extraction* and it involves the automatic construction of a Social Affiliation Network (SAN) [19] as well as its conversion into features that represent each person in terms of their interactions with the others. The second stage is the *role recognition*, i.e. the mapping of the features extracted in the first stage into roles belonging to a predefined set.

2.1 Feature Extraction

This section presents the feature extraction stage aimed at extracting and representing the interaction patterns of each person (see first stage in Figure 1).

The feature extraction stage includes three steps: the first is the segmentation of the conversations into single speaker segments. This detects the persons

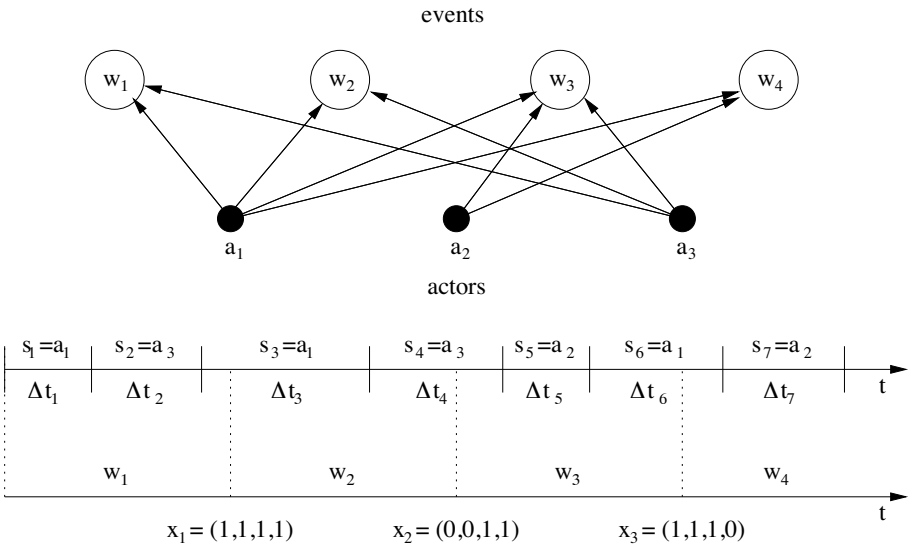


Fig. 2. Social Affiliation Network extraction. The events of the network correspond to the segments w_j and the actors are linked to the events when they talk during the corresponding segment.

involved in the conversations and the sequence of their interventions, i.e. the turn-taking informing on who talks when and how much (see left side of Stage 1 in Figure 1). The second stage is the extraction of a Social Affiliation Network (SAN) [19] from the resulting turn-taking. The SAN represents each person in terms of their interactions with the others (see upper part of right side of Stage 1 in Figure 1). The third step is the extraction of the fraction of time a person is talking, computed from the resulting turn-taking obtained at the first step (see lower part of right side of Stage 1 in Figure 1).

In our experiments, we have considered two kinds of data: broadcast material where there is a single audio channel, and meeting recordings [11], where each participant wears a headset microphone. This requires the application of different speaker diarization techniques: in the first case (single audio channel), an unsupervised speaker diarization technique identifies the voices of the different persons involved in the conversations (see [1] for a full description). In the second case (headset microphones), the diarization splits the channel of each microphone into speech and non-speech segments (see [5] for a detailed description).

The result of the speaker diarization process is that each recording is split into a sequence of turns, i.e. into a sequence $S = \{(s_k, t_k, \Delta t_k)\}$, where $k \in \{1, \ldots, N\}$, s_k is the label corresponding to the voice detected in the k^{th} turn, t_k is the beginning of speaker s_k intervention, and Δt_k is the duration of the k^{th} turn. The label s_k belongs to the set $A = \{a_1, \ldots, a_G\}$ of G unique speaker labels as provided by the speaker diarization process (see lower part of Figure 2). G is the total number of speakers in the conversation. The sequence of turns S

extracted from the speaker diarization can be used to extract a Social Affiliation Network (SAN), capturing the interaction patterns between the speakers. A SAN is a bipartite graph with two types of nodes: the *actors* and the *events* [19]. Actors can be linked to events, but no links are allowed between nodes of the same type, following the definition of bipartite graphs (see upper part of Figure 2). In our experiments, the actors correspond to the persons involved in the conversations, detected during the diarization process. The events correspond to uniform non-overlapping segments spanning the whole length of the recordings (see lower part of Figure 2), thus capturing the proximity in time of the persons interventions. Each recording is thus split into a number of D uniform, non-overlapping events.

One of the main advantages of this representation is that each actor a can be represented by a n-tuple $\boldsymbol{x}_a = (x_{a1}, \ldots, x_{aD})$, where D is the number of events and the component x_{aj} accounts for the participation of the actor a in the j^{th} event. Component x_{aj} is 1 if the actor a talks during the j^{th} event and 0 otherwise (the corresponding n-tuples are shown at the bottom of Figure 2).

We have also considered the fraction τ of the total time of a recording attributed to each voice as features. In this way, each actor a is represented by a pair $\boldsymbol{y}_a = (\boldsymbol{x}_a, \tau_a)$.

2.2 Role Recognition Approach Based on Bayesian Classifiers

The work presented in this section is further detailed in paper [15].

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

The previous section (see Section 2.1) has shown that each actor corresponds to a pair $\boldsymbol{y}_a = (\boldsymbol{x}_a, \tau_a)$. Thus, given the set of observations $Y = \{\boldsymbol{y}_a\}_{a \in A}$ and the function $\varphi : A \to \mathcal{R}$, the problem of assigning a role to each actor can be formulated as the maximization of the *a-posteriori* probability $p(\varphi|Y)$. By applying Bayes Theorem, and by taking into account that $p(Y)$ is constant during recognition, this problem is equivalent to finding $\hat{\varphi}$ such that:

$$\hat{\varphi} = \arg \max_{\varphi \in \mathcal{R}^A} p(Y|\varphi) p(\varphi) \tag{1}$$

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

In order to simplify the problem, some assumptions are made: the first is that the observations are mutually conditionally independent given the roles. The second is that the observation $\boldsymbol{y}_a$ of actor a only depends on its role $\varphi(a)$ and not on the role of the other actors. The last one is that the speaking time τ_a and the interaction n-tuples $\boldsymbol{x}_a$ of actors a are statistically independent given the role $\varphi(a)$. The equation to solve thus becomes:

$$\hat{\varphi} = \arg \max_{\varphi \in \mathcal{R}^A} p(\varphi) \prod_{a \in A} p\left(\boldsymbol{x}_a|\varphi(a)\right) p(\tau_a|\varphi(a)) \tag{2}$$

The three next sections show how $p(\boldsymbol{x}_a|\varphi(a))$, $p(\tau_a|\varphi(a))$, and $p(\varphi)$ have been estimated in the experiments.

Modeling Interaction Patterns. This section shows how the probability $p(\boldsymbol{x}_a|\varphi(a))$ is estimated. As the components of the n-tuple $\boldsymbol{x}_a$ are binary, i.e. $x_{aj} = 1$ when actor a talks during event j and 0 otherwise, the most natural way of modeling $\boldsymbol{x}_a$ is to use independent Bernoulli discrete distributions:

$$p(\boldsymbol{x}|\overrightarrow{\mu}) = \prod_{j=1}^{D} \mu_j^{x_j}(1-\mu_j)^{1-x_j} \tag{3}$$

where D is the number of events used to capture the interaction patterns in the SAN, and $\overrightarrow{\mu} = (\mu_1, \ldots, \mu_D)$ is the parameter vector of the distribution. A different Bernoulli distribution like the one in equation 3 is trained for each role. The maximum likelihood estimates of the parameters $\overrightarrow{\mu}_r$ for a given role r are as follows [4]:

$$\mu_{rj} = \frac{1}{|A_r|} \sum_{a \in A_r} x_{aj} \tag{4}$$

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

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

$$\mu_r = \frac{1}{|A_r|} \sum_{a \in A_r} \tau_a \tag{5}$$

$$\sigma_r^2 = \frac{1}{|A_r|} \sum_{a \in A_r} (\tau_a - \mu_r)^2 \tag{6}$$

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

Estimating Role Probabilities. We assume that the roles are independent and thus that $p(\varphi)$ is simply the product of the a-priori probabilities of the roles assigned through φ to the different actors:

$$p(\varphi) = \prod_{a \in A} p(\varphi(a)) \tag{7}$$

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

$$p(\varphi(a)) = \frac{|A_r|}{G} \tag{8}$$

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

Using the above approach, (2) boils down to

$$\hat{\varphi} = \arg \max_{\varphi \in \mathcal{R}^A} \prod_{a \in A} p(\boldsymbol{x}_a|\varphi(a))p(\tau_a|\varphi(a))p(\varphi(a)) \tag{9}$$

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

2.3 Role Recognition Approach Based on Probabilistic Sequential Models

The work presented in this section is further detailed in paper [6].

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

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

Section 2.1 has shown that each actor corresponds to a pair $\boldsymbol{y}_a = (\boldsymbol{x}_a, \tau_a)$ of dimension $D+1$. We have reduced the dimensionality of the tuples representing the interaction patterns through Principal Component Analysis (PCA). The application of PCA to the $\boldsymbol{y}_a$ tuples results into L-dimensional projections $\boldsymbol{w}_a$, where $L \leq D+1$. Therefore, each recording can be represented through a sequence of tuples $W = (\boldsymbol{w}_a 1, \ldots, \boldsymbol{w}_a N)$, where N is the number of turns detected at the speaker diarization step, and $\boldsymbol{w}_a k$ is the tuple representing the actor a talking at turn k.

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

$$R^* = \arg \max_{R \in \mathcal{R}^N} p(W, R)p(R) \tag{10}$$

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

Table 1. Corpora. The table reports the main characteristics of the corpora used in the experiments. From left to right: number of recordings, interaction setting, total time, average recording length, average number of participants. Note that the length is the same (one hour) for all recordings in C2, and the number of participants is constant (four) in C3.

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

In our experiments, the joint probability $p(W, R)$ was estimated with a fully connected, ergodic, HMM [13] where each state corresponds to a role $r \in \mathcal{R}$. The emission probability function associated to each state are Gaussians.

The *a-priori* probability $p(R)$ was estimated using a n-gram ($n \geq 1$) statistical language model [14]:

$$p(R) = \prod_{k=1}^{N} p(r_k | r_{k-1}, r_{k-2}, \ldots, r_{k-n+1}) \tag{11}$$

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

2.4 Experiments and Results

This section describes the data, the experimental setup, and presents the achieved role recognition performances.

The experiments of this work have been performed over three different corpora for a total amount of roughly 90 hours of material (one of the largest databases used for role recognition the literature). The first, referred to as C1 in the following, contains 96 news bulletins broadcasted by *Radio Suisse Romande* (the French speaking Swiss National broadcasting service) during February 2005. The second corpus, referred to as C2 in the following, contains 27 one hour long talk-shows also broadcasted by *Radio Suisse Romande* during February 2005. The third corpus, referred to as C3 in the following, is the AMI meeting corpus [11][3], a collection of 137 meeting recordings. Table 1 summarizes the main characteristics of C1, C2, and C3.

The roles of C1 and C2 share the same names and correspond to similar functions: the *Anchorman* (AM), i.e. the person managing the program, the

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

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

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

Table 2. Role recognition performance based on Bayes classifiers and probabilistic sequential models over C1 and C2. The table reports both the overall accuracy and the accuracy for each role.

	all (σ)	AM	SA	GT	IP	HP	WM
Results over C1							
Bayes	82.5 (6.9)	98.0	3.6	97.8	8.0	64.6	79.9
HMM + 3-gram	80.5 (8.3)	97.8	16.5	82.7	23.5	57.5	77.9
Results over C2							
Bayes	82.6 (6.8)	75.0	88.3	91.6	N/A	18.3	6.7
HMM + 3-gram	83.3 (8.2)	70.1	89.5	90.1	N/A	58.3	27.9

Second Anchorman (SA), i.e. the person supporting the AM, the *Guest* (GT), i.e. the person invited to report about a single and specific issue, the *Interview Participant* (IP), i.e. interviewees and interviewers, the *Headline Reader* (HR), i.e. the speaker reading a short abstract at the beginning of the program, and the *Weather Man* (WM), i.e. the person reading the weather forecasts. However, even if the roles have the same name and correspond to roughly the same functions, they are played in a different way in C1 and C2 (e.g., consider how different is the behavior of an anchorman in news supposed to inform and in talk-shows supposed to entertain). In C3, the role set is different and contains the *Project Manager* (PM), the *Marketing Expert* (ME), the *User Interface Expert* (UI), and the *Industrial Designer* (ID).

The experiments have been performed using a leave-one-out approach. We have thus selected all the recordings of the corpus in the training set (i.e. for training the role's models) with the exception of one that is used as test set. Training and test are repeated as many times as there are recordings in the corpus, and each time a different recording is left out as test set.

Table 2 reports the results achieved over C1 and C2, Table 3 those obtained for C3. The results are reported in terms of *accuracy* α, i.e. the percentage of data time correctly labeled in terms of role. Each overall accuracy value is accompanied by the standard deviation of the accuracies achieved over the different recordings of each corpus. The first row shows the results with the Bayes approach, and the second one shows the accuracy achieved when using HMMs and language models of order 3 (HMM+3-gram). The overall α is above 80% for both C1 and C2, and around 43% for C3.

The roles in meeting data (C3) are harder to model. A probable explanation is that the roles in meetings (C3) correspond to a position in a given social system and do not correspond to stable behavioral patterns like in the case of the roles in broadcast data (C1 and C2). Moreover, the meetings in C3 are not real-world data, i.e. the participants are asked to *act* in a scenario. It can thus happen that the participants have to play roles they are not used to and this might result into non ecologically valid data. Not surprisingly, the only meeting role recognized with a high accuracy is the *Project Manager* (PM). The reason is that the PM

Table 3. Role recognition performance based on Bayes classifiers and probabilistic sequential models over C3. The table reports both the overall accuracy and the accuracy for each role.

	all (σ)	PM	ID	ME	UI
Results over C3					
Bayes	43.5 (23.9)	75.3	15.1	40.0	15.1
HMM + 3-gram	38.5 (23.1)	52.4	28.9	17.6	35.3

Table 4. Diversity assessment. The table reports the accuracy of the percentage of data where the two approaches are both correct (C), both wrong (W), or one wrong and the other correct.

C1	HMM C	HMM W
Bayes C	78.0	2.2
Bayes W	4.5	15.3
C2	HMM C	HMM W
Bayes C	79.4	3.9
Bayes W	3.2	13.5
C3	HMM C	HMM W
Bayes C	22.3	11.3
Bayes W	15.9	50.5

acts as a *chairman*, having a specific task to achieve, and thus having distinct behavioral turn-taking patterns, in opposition with the domain experts ID, ME, and UI which have similar interaction patterns.

According to the Kolmogorov-Smirnov Test [10], the difference between the performance achieved with HMMs and the one achieved with the Bayesian classifier is not statistically significant. However, the two classifiers show a significant degree of *diversity*, i.e. they make different decisions over the same sample in a relatively high percentage of cases (see Table 4). In particular, probabilistic sequential approaches tend to improve the recognition of less frequent roles that are typically penalized by Bayesian classifiers certainly because of their low *a-priori* probability. *This suggests that the combination of the two approaches is likely to lead to significant performance improvements.* The highest possible performance deriving from a combination corresponds to the sum of the cases where at least one of the two approaches is right. This corresponds to 84.7% for C1, 86.6% for C2, and 49.5% for C3. In all of the cases, this would represent a statistically significant improvement with respect to the best of the approaches.

2.5 Combination of Interaction and Lexical Patterns

The work presented in this section is further detailed in paper [7].

Both approaches presented in this work for the role recognition task, i.e. the role recognition approach assigning a role to each person using Bayesian classifiers (see Section 2.2) and the role recognition approach taking into account sequential information (see Section 2.3), show limitations on the meeting recordings (C3).

One possible explanation of the lower role recognition performance over the C3 corpus may be due to the experimental setup of the C3 corpus itself, as C3 is composed by acted interactions and not real interactions. Another possible explanation of these results could be that the social interaction based role recognition approaches developed in this thesis are not well suited for less constrained conversations such as the ones represented in the C3 corpus. We were not able to verify this assumption by applying our automatic role recognition approaches over another scenario of conversations that were not constrained by specific tasks, as no other roles labeled spontaneous conversations were available. However, to assess the role recognition problem over the C3 meetings, we developed a new role recognition system in which we added lexical content to our interaction features.

This section presents the new role recognition approach for the meetings C3 which combines two behavioral cues. The first behavioral cue is the *interaction pattern*, i.e. the patterns representing the tendency of each actor a to interact with certain persons rather than others in a certain proximity in time. These features are extracted from the Affiliation Networks exactly as previously detailed in Section 2.1, and are mapped to roles using Bernoulli distribution. The second behavioral cue is the *lexical choice*, i.e. the use of certain words rather than others in the interventions of each person. The lexical features are mapped into roles using the BoosTexter text categorization approach [16].

Experiments and Results. The role recognition approach presented in this section has been developed to improve the performance over the AMI corpus (referred as C3 in this article). The training of the role recognition system is performed using a leave-one-out approach, i.e. using the same experimental setup as with the other role recognition approaches presented previously in this article.

The performance is measured with the *accuracy* α, i.e. the percentage of data time correctly labeled in terms of role. Table 5 reports the accuracies obtained by using only Social Affiliation Network Analysis, only lexical choices, and the combination of the two. The results are reported for the overall meetings, as well as for the single roles separately.

The lexical choice appears to be a more reliable cue for the recognition of the roles for the AMI meetings. The overall accuracy of the lexicon based system is significantly higher (67.1% against 43.1%). A possible explanation is that the AMI corpus is particularly suitable for lexical analysis, while it is rather unfavorable to the application of SAN. On one hand, the content of the interventions is constrained by the role and this helps the former approach, on the other

Table 5. Role recognition results when combining interaction features (SAN) and lexical features (lex.) over the meetings C3

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

hand, the similar interaction patterns of the participants may limit significantly the latter approach, as the social networks are not able to distinguish between the roles. The combination of the two systems does not improve significantly the performance of the best system (see Table 5). The main reason is probably that the performance of the SAN approach is too close to the chance (around 25%) for at least two roles (ME and ID). Thus, the SAN does not bring useful information in the combination, but simply some random noise. This seems to be confirmed by the case of the PM role, where the combination improves by almost 6% the performance of the best classifier. Not surprisingly, the performance of the SAN system over the PM is significantly better than the chance because the PM plays a formal role as we have seen previously in Section 2.4.

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

3 Conclusion

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

The proposed approaches have been tested over roughly 90 hours of material, composed of broadcast material and meeting recordings. This is one of the biggest data sets ever used in literature for this task. Moreover, to the best of our knowledge, the data set used in this work is the only one that includes different interaction settings and different role sets. This is important in order to show how the role typology influences the effectiveness of the recognition, and thus how easily an approach can be ported from one interaction setting to another.

Another novelty of the presented approaches is to use the interaction between the persons as features. The Social Affiliation Networks (SAN) [19] allows one to extract these features, which represent the evidence of interactions in terms of proximity in time, from the co-occurence turn-taking patterns structuring the conversations. The rationale behind the SAN is that the persons speaking in the same time intervals are likely to interact with each other.

This article has compared approaches based on Bayesian classifiers and approaches based on probabilistic sequential models. The former assigns a specific role to each person involved in the recordings (see Section 2.2). The latter considers the sequence of persons talking during a conversation, and aligns the sequence of their turns with a sequence of roles (see Section 2.3). For both approaches, the results show that the role recognition accuracy is higher than 80% in the case of broadcast data, and it is around 45% in the case of meeting recordings.

In the case of the broadcast data, the performance should be sufficient to browse effectively the data, or at least could help it. In fact, users should quickly find segments corresponding to a given role because the mismatch between the ground truth and the automatic output rarely exceeds a few seconds. In the case of meeting recordings, the approach is effective only to identify the Project Manager. However, this should allow one to effectively follow the progress of the meeting as the PM plays the chairman role and, as such, is responsible for following the agenda through her/his interventions.

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

Acknowledgements. This work is supported by the Swiss National Science Foundation The author wishes to thank Alessandro Vinciarelli and Hugues Salamin.

References

1. Ajmera, J., Wooters, C.: A robust speaker clustering algorithm. In: Proceedings of IEEE Workshop on Automatic Speech Recognition Understanding, pp. 411–416 (2003)
2. Barzilay, R., Collins, M., Hirschberg, J., Whittaker, S.: The rules behind the roles: identifying speaker roles in radio broadcasts. In: Proceedings of the 17th National Conference on Artificial Intelligence, pp. 679–684 (2000)
3. Biddle, B.J.: Recent developments in role theory. Annual Review of Sociology 12, 67–92 (1986)
4. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
5. Dines, J., Vepa, J., Hain, T.: The segmentation of multi-channel meeting recordings for automatic speech recognition. In: Proceedings of Interspeech, pp. 1213–1216 (2006)
6. Favre, S., Dielmann, A., Vinciarelli, A.: Automatic role recognition in multiparty recordings using social networks and probabilistic sequential models. In: Proceedings of ACM International Conference on Multimedia, pp. 585–588 (2009)

7. Garg, N., Favre, S., Salamin, H., Hakkani-Tür, D., Vinciarelli, A.: Role recognition for meeting participants: an approach based on lexical information and Social Network Analysis. In: Proceedings of the ACM International Conference on Multimedia, pp. 693–696 (2008)
8. Laskowski, K., Ostendorf, M., Schultz, T.: Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In: Proceedings of the 9th ISCA/ACL SIGdial Workshop on Discourse and Dialogue, pp. 148–155 (June 2008)
9. Liu, Y.: Initial study on automatic identification of speaker role in broadcast news speech. In: Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pp. 81–84 (June 2006)
10. Massey Jr., F.J.: The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 68–78 (1951)
11. McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., Wellner, P.: The AMI meeting corpus. In: Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research, p. 4 (2005)
12. Pianesi, F., Zancanaro, M., Lepri, B., Cappelletti, A.: A multimodal annotated corpus of consensus decision making meetings. Language Resources and Evaluation 41(3-4), 409–429 (2008)
13. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proc. of the IEEE 77, 257–286 (1989)
14. Rosenfeld, R.: Two decades of statistical language modeling: Where do we go from here. Proceedings of the IEEE 88, 1270–1278 (2000)
15. Salamin, H., Favre, S., Vinciarelli, A.: Automatic role recognition in multiparty recordings: Using social affiliation networks for feature automatic role recognition in multiparty recordings: Using social affiliation networks for feature extraction. IEEE Transactions on Multimedia 27(12), 1373–1380 (2009)
16. Schapire, R.E., Singer, Y.: Boostexter: a boosting-based system for text categorization, vol. 39, pp. 135–168 (2000)
17. Tischler, H.L.: Introduction to Sociology. Harcourt Brace College Publishers (1990)
18. Vinciarelli, A.: Speakers role recognition in multiparty audio recordings using Social Network Analysis and duration distribution modeling. IEEE Transactions on Multimedia 9(6), 1215–1226 (2007)
19. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1994)
20. Weng, C.Y., Chu, W.T., Wu, J.L.: Rolenet: Movie analysis from the perspective of social networks. IEEE Transactions on Multimedia 11(2), 256–271 (2009)

Speaker Diarization of Multi-party Conversations Using Participants Role Information: Political Debates and Professional Meetings

Fabio Valente[1] and Alessandro Vinciarelli[1,2]

[1] Idiap Research Institute, Martigny, Switzerland
[2] University of Glasgow, Glasgow, United Kingdom
fvalente@idiap.ch, Alessandro.Vinciarelli@glasgow.ac.uk

1 Introduction

Speaker Diarization aims at inferring *who spoke when* in an audio stream and involves two simultaneous unsupervised tasks: (1) the estimation of the number of speakers, and (2) the association of speech segments to each speaker. Most of the recent efforts in the domain have addressed the problem using machine learning techniques or statistical methods (for a review see [11]) ignoring the fact that the data consists of instances of human conversations.

When humans want to use language to communicate orally with each other, they are faced to a coordination problem. "*Avoidance of collision is one obvious ground for this coordination of actions between the participants. In order to coordinate efficiently and successfully, they will therefore have to agree to follow certain rules of interaction*" [8]. One such rule is that no one monopolizes the floor but the participants take turns to speak. This concept is called *turn-taking*. The computational linguistic literature is rich on the analysis of human conversations; the seminal work of [9] shows that conversations obey to predictable interactions pattern between participants and a speaker turn is related in predictable ways to the previous and next turn and follows a structure similar to a grammar. In between the social phenomena that regulates the turns in a conversation, lot of attention has been devoted to roles. In fact people interact in different ways depending on the context of the environment but "*Their interactions involve behaviors associated with defined statuses and particular roles. These statuses and roles help to pattern our social interactions and provide predictability*" [10].

Only recently it has been shown that the turn-taking behavior can be statistically modeled and used to automatically classify a certain number of characteristics in groups conversations like roles. Examples include the automatic recognition of roles in meetings recordings like CMU or AMIDA recordings [2,4], the recognition of participant seniority (professor, phd or graduate student) in the ICSI meeting data set [6] and the recognition of functional roles in the MSC corpus [3,15]. Typically those studies are based on the use of statistical

R. Murray-Smith (Ed.): MSSP 2010, LNCS 8045, pp. 22–33, 2014.

classifiers trained on a set of automatically or semi-automatically derived audio features including the speaker turn durations, the overlap between speakers and the speaker turn statistics. They assume that the participants interactions and specifically the turn-taking patterns can be statistically modeled and provide enough information for recognizing the role of each speaker in the conversation.

This work investigates whether the use of the statistical information derived from roles can reversely increase the performance of conventional audio processing systems like diarization. In details, this work discusses the use of turn-taking information induced by the roles that participants have in the discussion as prior information in the speaker diarization systems. Previous attempts have used participant interaction patterns to improve the diarization performance, e.g. [5], however this information was not induced by, or put in relation with, any social phenomena. In this work, we make the following hypothesis: 1) the turn-taking patterns are conditioned on the role that each speaker has in the conversation, 2) they can be estimated on an independent development data set.

We propose to model the speaker sequence using N-gram of speaker roles. N-gram models can be then combined with the acoustic information coming from MFCC features. The approach is largely inspired by the current Automatic Speech Recognition (ASR) framework where the acoustic information from the signal, i.e., the acoustic score, is combined with the prior knowledge from the language, i.e., the language model. The most common form of language model is represented by words N-gram. In a similar way, given a mapping speakers to roles, N-gram models can encode the statistical information on how the participants take turns in the conversation.

The investigation is carried on two very different dataset, the first one is composed of political debates recorded with close-talk high quality microphones while the second one is composed of professional meetings recorded with far-field low quality microphones. The use of those datasets aim at studying how those findings generalizes across different types of conversations and different acoustic conditions. Let us briefly describe those datasets in the following.

2 Data Description

The first dataset used for this study consists of political debates [14] that represent an excellent resource for their realism. In contrast with other benchmarks, political debates are real-world data. Debate participants do not act in a simulated social context, but participate in an event that has a major impact on their real life (for example, in terms of results at the elections). Thus, even if the debate format imposes some constraints, the participants are moved by real motivations leading to highly spontaneous social behavior.

Each debate revolves around a yes/no question like “Are you favorable to new laws on education ?”. The participants state their answer (yes or no) at the beginning of the debate and do not change it during the discussion. Each debate involves a moderator and a variable number of guests (four or more). The dataset is annotated in terms the role that each participant has in the discussion, i. e.

moderator or guests. All debates include one moderator expected to ensure that all participants have at disposition the same amount of time for expressing their opinion. Furthermore, the moderator intervenes whenever the debate becomes too heated and people tend to interrupt one another or to talk together. The guests are labeled in terms of groups according to how they answer to the central question of the debate. Participants belonging to the same group agree with one another, while participants belonging to different groups disagree with one another. The dataset is divided in two non-overlapping parts, a development dataset (composed of 25 debates for a total of 17 hours and 2600 speaker turns) and a test dataset (composed of 25 debates for a total of 15 hours and 2500 speaker turns).

The second dataset is based on the AMI meeting database [7], a collection of 138 meetings recorded with distant microphones for approximatively 100 hours of speech, manually annotated at different levels (roles, speaking time, words, dialog acts). Each meeting consists of a scenario discussion in between four participants where each participant has a given role: project manager PM, user interface expert UI, marketing expert ME and industrial designer ID. The scenario consists in four employes of an electronic company that develop a new type of television remote controller. The meeting is supervised by the project manager. The dataset is divided in two non-overlapping parts, a development data set (118 meetings) and a test set (20 meetings).

3 Turn-Taking Patterns and Roles

Let us formalize the turn-taking and role informations as follows. For each recording the following triplets are available:

$$T = \{(t_1, \Delta t_1, s_1), ..., (t_N, \Delta t_N, s_N)\} \tag{1}$$

where t_n is the beginning time of the n-th turn, Δt_n is its duration, s_n is the speaker associated with the turn and N is the total number of turns in the recording. The begin of the turn corresponds to the time at which the speaker s_n grabs the floor of the discussion and the length Δt_N corresponds to the time during which s_n holds the floor.

Each participant is labeled according to the role he or she has in the recording and the mapping between each speaker and his/her role is given by the function $\varphi(S) \rightarrow R$. In case of debates the roles are moderator m, or guest g. Guests are furthermore labeled in two groups $g1$ and $g2$ according to their agreement/disagreement thus the space of roles is given by $R = \{m, g1, g2\}$. On the other hand , in case of meetings, the space of roles is given by $R = \{PM, UI, ME, ID\}$.

The sequence of speakers $S = \{s_1, ..., s_n\}$ can be statistically modeled as a first-order Markov chain in which the probability of the participant s_n speaking after the participant s_{n-1} is regulated by their respective roles $\varphi(s_n)$ and $\varphi(s_{n-1})$ (see [13]).

Table 1 represents the conditional probability $p(\varphi(s_n)|\varphi(s_{n-1}))$ of a speaker role conditioned to the role of the previous speaker on the development dataset in case of debates while Table 2 represent the same quantities in case of meeting recordings. Those statistics are obtained disregarding overlapping speech regions (including back-channels).

Table 1. Transition matrix between roles estimated on the debates development data set

	Moderator	Group 1	Group 2
Moderator	0	0.51	0.49
Group 1	0.68	0.06	0.26
Group 2	0.67	0.25	0.08

Table 2. Transition matrix between roles estimated on the meetings development data set

	PM	UI	ME	ID
PM	0	0.34	0.31	0.35
UI	0.39	0	0.30	0.31
ME	0.43	0.28	0	0.29
ID	0.41	0.29	0.30	0

Tables 1 and 2 can be interpreted in straightforward way. In case of debates, the moderator aims at sharing the available time in between the two groups and this is reflected in the fact that $p(g1|m)$ is approximatively equal to $p(g2|m)$ as well as $p(m|g1)$ is approximatively equal to $p(m|g2)$. On the other hand speakers with different opinions are more likely to take turn (on average) after a speaker they disagree with and this explains why $p(g2|g1)$ and $p(g1|g2)$ are considerably higher then $p(g1|g1)$ and $p(g2|g2)$. The probability $p(m|m)$ is equal to zero as there is only one moderator in each debate.

In case of meetings the Program Manager acts as moderator aiming at sharing the time in between the other participants; similarly the probability that a participant will take turn after the Program Manager is higher then the probability of taking turn after a non-chairperson participants.

In other words, the possible speaker sequences $S = \{s_1, ..., s_N\}$ in a conversations are not all equally probable and their probability can be simply estimated as:

$$p(S) = p(s_1, ..., s_n) = p(\varphi(s_1), ..., \varphi(s_n)) = p(\varphi(s_0)) \prod_{i=1}^{N} p(\varphi(s_n)|\varphi(s_{n-1})) \tag{2}$$

where $p(\varphi(s_n)|\varphi(s_{n-1}))$ are elements of the matrix (1) and $p(\varphi(s_0))$ is the probability of the role associated with the speaker that opens the discussion. In the most general case the sequence S can be modeled using an N-gram, i.e.:

$$p(S) = p(s_1, ..., s_n) = p(\varphi(s_1), ..., \varphi(s_n)) =$$
$$= p(\varphi(s_1), ..., \varphi(s_p)) \prod_{n=p}^{N} p(\varphi(s_n)|\varphi(s_{n-1}), ..., \varphi(s_{n-p})) \quad (3)$$

where the probability of a speaker taking the n-th turn is conditioned to the role of the previous p speakers taking turns before him. Those N-gram models will be referred as *speaker role N-gram* and the paper will investigate how this information can be included as prior knowledge in a speaker diarization system.

4 Speaker Diarization System

Speaker Diarization is the task that aims at inferring *who spoke when* in an audio stream. The system used here is a state-of-the-art system described in [12] and briefly summarized in the following.

Acoustic features consist of 19 MFCC coefficients extracted using a 30ms window shifted by 10ms. After speech/non-speech segmentation and rejection of non-speech regions, the acoustic features $X = \{x_1, \ldots, x_T\}$ are uniformly segmented into chunks of 250ms. Then hierarchical agglomerative clustering is performed grouping together speech segments according to a distance inspired from information theory and the clustering stops when a criterion based on Normalized Mutual Information (NMI) is met (see [12] for details). This produces an estimate of the number of participants in the debate and a partition of the data in clusters, i. e., it associates each acoustic vector x_t to a speaker s. As the diarization system classifies silence regions as non-speech, the actual turn-taking can be obtained bridging together consecutive speech segments from the same speaker separated by silence regions. For instance, the turns can simply be obtained bridging the silence regions that separates the three utterances spoken by the first speaker.

We refer this initial segmentation into speakers as T^*:

$$T^* = \{(t_1^*, \Delta t_1^*, s_1^*), ..., (t_N^*, \Delta t_N^*, s_N^*)\} \quad (4)$$

After clustering, the speaker sequence is re-estimated using an ergodic Hidden Markov Model/Gaussian Mixture Model where each state represents a speaker. The emission probabilities are modeled as GMMs trained using acoustic vectors x_t assigned to speaker s. Each state enforces a minimum duration constraint. This step aims at refining the data partition obtained by the agglomerative clustering and improving the speaker segment boundaries [11].

The decoding is performed using a conventional Viterbi algorithm, i. e. the optimal speaker sequence $\mathbf{S}^* = (s_1, s_2, ..., s_N)$ is obtained maximizing the following likelihood:

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} \log\, p(X|S) \quad (5)$$

The emission probability $p(x_t|s_t)$ of the acoustic vector x_t conditioned to speakers s_t is:

$$\log\, p(x_t|s_t) = \log \sum_r w^r_{s_t} \mathcal{N}(x_t, \mu^r_{s_t}, \Sigma^r_{s_t})$$

where $\mathcal{N}(.)$ is the Gaussian pdf; $w^r_{s_t}, \mu^r_{s_t}, \Sigma^r_{s_t}$ are weights, means and covariance matrix corresponding to speaker model s_t. The output of the decoding step is a sequence of speakers with their associated speaking time.

Let us report the performance of this system on the meetings and the debates that compose the test data set. The most common metric for assessing diarization performances is the Diarization Error Rate [1] which is composed by speech/non-speech and speaker errors. As the same speech/non-speech segmentation is used across experiments, in the following only the speaker error is reported. Table 3 reports the speaker error in case of a-priori known number of speakers K. It can be notice from table 3 that the diarization performance is significantly worst in case of meetings because the audio is recorded with far field microphones while in case of debates the audio is acquired using close talk microphones.

Table 3. Speaker Error reported on the test data set in case of debates and meetings

	Debates	Meetings
Speaker Error	6.2%	14.4%

5 Speaker-turns Based Diarization

The decoding step 5 only depends on the acoustic score $p(X|S)$ (see Eq. (5)) and completely neglects the fact that not all speaker sequences S have the same probability. In section 3, we discussed that the roles regulate the way speakers take turns and the probability of a given speaker sequence can be estimated using Eq. (3). It is thus straightforward to extend the objective function (see Eq. 5) in order to include this type of information i. e.:

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} \log\, p(X|S)p(S) = \arg\max_{\mathbf{S}} \log\, p(X|S)p(\varphi(S)) \tag{6}$$

In other words, the optimal speaker sequence (and the associated speaker times) can be obtained combining the evidence from the acoustic score $p(X|S)$ together with the prior probability of a given sequence $p(S)$. This is somehow similar to what is done in Automatic Speech Recognition (ASR) where sentences (i. e. word sequences) are recognized combining acoustic information together with linguistic information captured in the language model. Looking at Eq. (6), it is possible to notice that while the acoustic score $p(X|S)$ is modeled using a probability density function, i. e. a GMM, $p(S)$ is a probability; as in ASR, we

[1] `http://www.itl.nist.gov/iad/mig/tests/rt/`

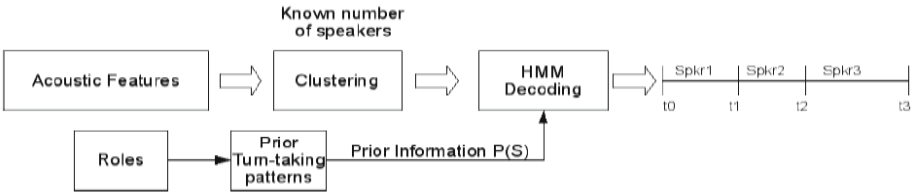


Fig. 1. Schematic representation of the proposed system in case scenario 1 (known number of speakers and roles): the clustering stops when the known number of clusters is obtained; Speaker decoding is done combining the acoustic information with prior turn-taking information induced by participants role

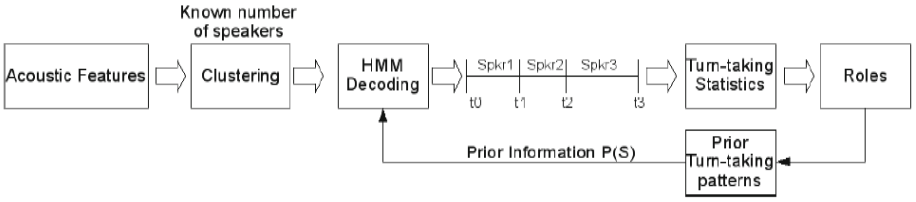


Fig. 2. Schematic representation of the proposed system in case scenario 2 (known number of speakers and unknown roles): the clustering stops when the known number of clusters is obtained; turn-taking statistics obtained from the diarization output are used to recognize speaker roles. Roles are then used to compute the prior probability of a speaker sequences $P(S)$ which is used then in the diarization system.

introduce a factor λ tuned on the development data set to scale $P(S)$ at the same order of magnitude of $p(X|S)$ and an insertion penalty:

$$\mathbf{S}^* = \arg\max_{\mathbf{S}} [\log p(X|S)\, p(\varphi(S))^{\lambda}] \tag{7}$$

Eq (7) can be solved using a Viterbi decoder that includes the prior probability of different speaker sequences. The development data set is used to estimate the probabilities $p(\varphi(S)$ and the scaling factor λ as well as the decoder insertion penalty. Performances are reported on the evaluation data set. In the most general case, the speaker roles are unknown. To incrementally study the integration of prior information $p(S)$, two different case scenarios are proposed.

5.1 Case 1

The number of participants K (thus speakers) in the debate is known as well as the mapping speakers-role $\varphi(.)$. The entire process is schematically depicted in Figure 1.

Those assumptions significantly simplify the problem. The clustering stops whenever the number of clusters is equal to the actual number of participants in the recording and the mapping speaker-role is obtained from the manual

reference thus the prior $P(S)$ can be directly estimated from Eq. (3). Table 4 reports the speaker error obtained with conventional decoding and with role-based decoding. The inclusion of the prior information reduces the speaker error from 6.2% to 4.6% i. e. a relative improvement of 25% for debates recordings and from 14.4% to 11.5% for meeting recordings, i.e., a 19% relative improvement. The improvements are verified on all the recordings from the data set. The largest reduction in the error rate is obtained using a bigram model, i.e., conditioning the turn to the role of the previous speaker. The use of trigram models only marginally improve over the bigram. It is interesting to notice that the approach appears effective on different type of acoustic conditions (far-field and close talk audio) and on different type of data, political debates and professional meetings. This suggest that the method could be applied to any type of multi-party conversation once a mapping from speakers to role is known.

Table 4. Speaker Error obtained using unigrams, bigrams and trigrams in case scenario 1. In brackets the relative improvement is reported w. r. t. the baseline where no prior information is available.

Prior	$P(\varphi(s_n)$	$P(\varphi(s_n)\|\varphi(s_{n-1}))$	$P(\varphi(s_n)\|\varphi(s_{n-1},\varphi(s_{n-2}))$
Debates - Sp. Err.	5.8 (+6%)	4.6 (+25%)	4.6 (+25%)
Meetings - Sp. Err.	13.8% (+4%)	11.8% (+18%)	11.5% (+19%)

5.2 Case 2

In this case we assume that the number of participants K in the debate is known but the mapping speakers-role $\varphi^*(.)$ is estimated from the segmentation T^*. The entire process is schematically depicted in Figure 2. As before, the clustering stops whenever the number of clusters is equal to the actual number of participants in the recording producing an initial solution T^*. The mapping speakers-role $\varphi^*()$ is estimated from the segmentation T^* using the following maximization:

$$\varphi^* = \arg\max_{\varphi} p(\varphi(s_0^*)) \prod_{n=1}^{N} p(\varphi(s_n^*)|\varphi(s_{n-1}^*)). \tag{8}$$

The optimization (8) is performed exhaustively searching the space of possible mappings speakers-roles, i. e., $\varphi(\{s_k\}) \rightarrow \{R\}$ and selecting the mapping that maximize the probability of the speaker sequence s^*, i. e., Eq. (8). Table 5 reports the speaker error obtained with conventional decoding and with role-based decoding. The inclusion of the prior information reduces the speaker error from 6.2% to 4.9% i. e. a relative improvement of 20% for debates recordings and from 14.4% to 11.9% for meeting recordings, i.e., a 17% relative improvement. Again the largest reduction in the error rate is obtained using a bigram model, i.e., conditioning the turn to the role of the previous speaker. The use

Table 5. Speaker Error obtained unigrams, bigrams and trigrams in case scenario 2. In brackets the relative improvement is reported w. r. t. the baseline where no prior information is available.

Prior	$P(\varphi(s_n))$	$P(\varphi(s_n)\|\varphi(s_{n-1}))$	$P(\varphi(s_n)\|\varphi(s_{n-1},\varphi(s_{n-2}))$
Debates - Sp. Err.	5.9 (+6%)	4.9 (+20%)	4.9 (+20%)
Meetings - Sp. Err.	14.4% (+3%)	12.0% (+16%)	11.9% (+17%)

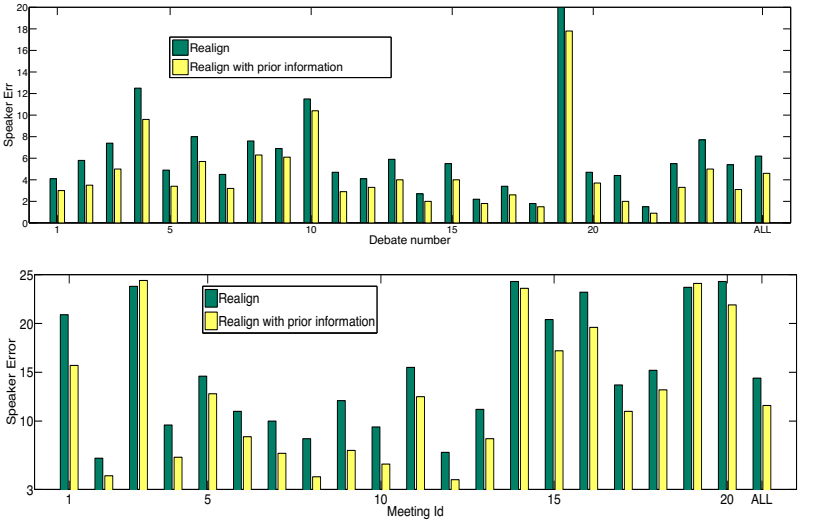


Fig. 3. Speaker error obtained using realignment with and without prior information for the 25 recordings that compose the debates test data set (top figure) and for the 20 recordings that compose the meeting test data set. The speaker error is reduced on all the debates as well as on 18 meetings out of 20.

of trigram models only marginally improve over the bigram. Improvements are slightly smaller compared to those obtained in Case 1 because of errors that occurs when roles are estimated using Eq. 5.

Figure 3 plots the speaker error with and without prior information for the 25 recordings that compose the test data set in Case 2. The proposed approach reduces the speaker error on 23 out of 25 debates in Case 2. The error does not decrease in two recordings with high speaker error. In Case 1 and Case 2 (not plotted), the improvements are verified on all the 25 recordings. We do not verify a degradation in performance in any recording.

Let us now investigate the differences between the systems outputs. Figures 4 plots the relative amount of total speaker time correctly attributed to each of the four roles by the baseline diarization and the proposed technique. Those statistics are averaged over the entire test set and normalized dividing by the total speaker time. The largest improvement in performance comes from the time

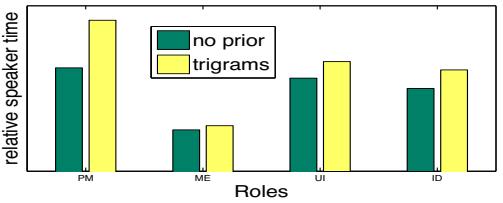


Fig. 4. Relative amount of speaker time correctly attributed to each of the four speakers labeled according to their roles by the baseline diarization and the proposed technique in case 2 in case of meeting recordings. Statistics are averaged over the entire test set.

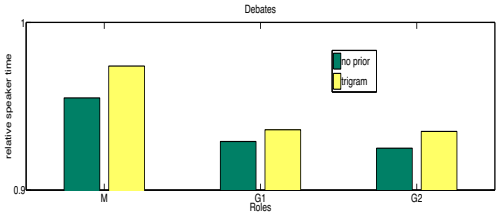


Fig. 5. Relative amount of speaker time correctly attributed to each of the four speakers labeled according to their roles by the baseline diarization and the proposed technique in case 2 in case of meeting recordings. Statistics are averaged over the entire test set.

correctly attributed to the speakers labeled as PM in meetings (see figure 4 (a)) and as moderator in debates (see figure 4 (b)). In the psychology literature, those roles (moderator and project manager) can be associated with the *gatekeeper* (see [1]), i.e., the speaker that encorages and regulates the discussion. In other words, most of the improvements comes from the speech attributed to the *gatekeeper* of the discussion rather then from speech attributed to the other roles.

Further analysis shows that the proposed method outperforms the baseline especially on short turns where the acoustic score may not provide enough information to assign the segment to a given speaker.

6 Discussions

A large body of recent works has focused on the recognition of roles in multi-party discussions. Turn-taking patterns, i.e. the tendency of participants to interact or to react to certain persons rather then others, represents a powerful cue for inferring the role that each speaker has in a discussion [3,15]. Speaker diarization represents a key techonolgy for automatic turns extraction.

This work discusses the use of turn-taking patterns as a priori information in diarization systems. In contrary to related works [5], the patterns are explicitly put in relation with the roles that each speaker has in the discussions and they are estimated on an independent development data set. Experiments are carried out on political debates and professional meeting recordings. Those two datasets

have different acoustic conditions (close talk speech for the first and far-field speech for the second) and represent different type of conversations (competitive debate in the first case versus professional collaborative meeting in the latter).

Results show that whenever the number of participants in the discussion as well as their roles are known the speaker error is reduced by 25% in case of debates and by 20% in case of meetings; whenever the second one is not available the improvements are 20% in case of debates and 17% in case of meetings. In summary the proposed method seem to reduce consistently the speaker error across different types of conversations and different acoustic conditions. The largest error reduction is obtained when bigram of roles are used; the use of trigrams marginally reduces the total error respect to the bigrams.

The largest part of the improvements come from speech attributed to the debate moderator or the meeting program manager; those roles can be associated with the *gatekeeper* (according to the social role coding scheme [1]), i.e., the speaker that encorages and regulates the discussion.

References

1. Bales, R.F.: Interaction Process Analysis: A Method for the Study of Small Groups. Addison-Wesley (1950)
2. Banerjee, S., Rudnicky, A.I.: Using simple speech based features to detect the state of a meeting and the roles of the meeting participants. In: Proceedings of International Conference on Spoken Language Processing, vol. (2-3), pp. 221–231 (2004)
3. Dong, W., Lepri, B., Cappelletti, A., Pentland, A., Pianesi, F., Zancanaro, M.: Using the influence model to recognize functional roles in meetings. In: Proceedings of the International Conference on Multimodal Interfaces (ICMI), pp. 271–278 (2007)
4. Garg, N., Favre, S., Salamin, H., Hakkani-Tür, D., Vinciarelli, A.: Role recognition for meeting participants: an approach based on lexical information and Social Network Analysis. In: Proceedings of the ACM International Conference on Multimedia, pp. 693–696 (2008)
5. Han, K.J., Narayanan, S.S.: Improved speaker diarization of meeting speech with recurrent selection of representative speech segments and participant interaction pattern modeling. In: Proceedings of Interspeech, pp. 1067–1070 (2009)
6. Laskowski, K., Ostendorf, M., Schultz, T.: Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In: Proceedings of the 9th ISCA/ACL SIGdial Workshop on Discourse and Dialogue, pp. 148–155 (June 2008)
7. McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V.: The AMI meeting corpus. In: Proceedings of the International Conference on Methods and Techniques in Behavioral Research, vol. 88 (2005)
8. Oreström, B.: Turn-taking in English conversation. Krieger Pub. Co. (1983)
9. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language, 696–735 (1974)
10. Tischler, H.: Introduction to sociology. Harcourt Barce College Publishers (1990)

11. Tranter, S.E.E., Reynolds, D.A.: An overview of automatic speaker diarization systems. IEEE Transactions on Audio, Speech, and Language Processing 14(5), 1557–1565 (2006)
12. Vijayasenan, D., Valente, F., Bourlard, H.: An information theoretic approach to speaker diarization of meeting data. IEEE Transactions on Audio, Speech, and Language Processing 17(7), 1382–1393 (2009)
13. Vinciarelli, A.: Capturing order in social interactions. IEEE Signal Processing Magazine 26(5), 133–152 (2009)
14. Vinciarelli, A., Dielmann, A., Favre, S., Salamin, H.: Canal9: A database of political debates for analysis of social interactions. In: Proceedings of International Workshop on Social Signal Processing, pp. 1–4 (2009)
15. Zancanaro, M., Lepri, B., Pianesi, F.: Automatic detection of group functional roles in face to face interactions. In: Proceedings of International Conference on Mutlimodal Interfaces, pp. 47–54 (2006)

Invisible, Passive, Continuous and Multimodal Authentication

Karen Renaud and Heather Crawford

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
Glasgow UK, G12 8RZ
karen@dcs.gla.ac.uk, hcrawford@fit.edu

Abstract. Authentication, as traditionally achieved by means of a shared secret, is effortful and deliberate. Frequent and repeated authentication easily becomes a hurdle, an annoyance and a burden. This state of affairs needs to be addressed, and one of the ways of doing this is by moving towards automating the process as much as possible, and reducing the associated effort — ie. reducing its visibility. A shared secret clearly does not have the flexibility to support this, and we need therefore to consider using biometrics. Biometrics are a well-established authentication method. Physiological biometrics require a biometric reader and explicit action by the user. Furthermore, there are always a minority of users who cannot have a particular biometric measured. For example elderly women often lose their fingerprints, and iris biometrics don't work for people with particular eye conditions.Behavioural biometrics, however, can be collected without the user having to take deliberate action. Hence there is a strong possibility that these biometrics could deliver the invisible and automatic authentication we are striving towards. One big advantage of these biometrics is that, since there is no reader, it is simple to utilise a number of different biometrics, and to combine these to authenticate the user. If one biometric fails the others can still perform authentication.

Here we propose using patterns such as keystroke dynamics, use patterns, and voice analysis techniques to create a multimodal biometric authentication mechanism. These *behavioural* biometrics take advantage of tasks that the user already performs thereby reducing the need for explicit authentication by more traditional means. In this way, the user is relieved of the burdens of constantly authenticating to multiple applications and devices.

1 Introduction

It is clear that the username-password "identity" combination, while perfectly satisfactory from a purely technical security perspective, is inherently flawed when used by fallible humans. Passwords are forgotten, shared and reused on multiple devices and applications. The policies implemented to strengthen passwords, such as requiring sufficient length or strength and changing the password

R. Murray-Smith (Ed.): MSSP 2010, LNCS 8045, pp. 34–41, 2014.

frequently, may not increase the security level since users simply find other ways of coping, such as writing the passwords down. Furthermore, the password, as a concept, does not authenticate the user; it authenticates an identifier. Someone else might be holding that identity, and the system errs in assuming that the verified identity authenticates the legal owner thereof. Clearly this mechanism is too weak to control access to many systems but, in the absence of a viable alternative, the flawed password prevails.

The type of security today's computers require, with their almost unlimited access to personal and private information, must go beyond secret knowledge and uniquely authenticate a particular *person*. We must be able to prove, with far more confidence than the password affords, that a person is who they claim to be, before granting them access to a restricted resource.

Authentication is traditionally achieved by using one of three classes of authenticator: something you *have*, something you *know*, or something you *are* [20, p. 29].

- *something you have*: the user is in possession of a physical device or token that aids in identification. The debit card associated with a particular bank account is an example of a token.
- *something you know*: requires a user to prove knowledge of a particular secret. Secret knowledge techniques such as knowledge of a password or PIN are examples of something you know.
- *something you are*: concerned with measuring a person's physical attributes as a unique identifier, and referred to as biometrics. Examples include fingerprints, retinal scans, and facial recognition. Biometric is "the science of recognizing an individual based on her physiological or behavioral traits." [18]. Interestingly, this definition also includes *behavioral* traits, which include typing style, device use patterns, and gait analysis, to name just a few. Unlike physical biometrics, which require the user to submit to their capture, behavioral biometrics can be captured while the user goes about their everyday tasks. This reduces reliance on the user to authenticate correctly and also allows authentication to take place invisibly. According to a 2004 study, users prefer biometrics to passwords since they believe biometrics would provide an increased level of security [16].

2 Motivation

In addition to increasing the need for more reliable access control and authentication mechanisms, the new generation of mobile computing devices has also increased the user's memory and cognitive load due to different usage paradigms imposed by these devices. Great care needs to be taken when deciding on an authentication mechanism for such devices. The user's main goal is to perform some task, such as checking their bank balance or calling a friend. Authenticating themlselves is extraneous, and it makes no sense to impose a complicated authentication onto them.

Authentication could benefit from being a "black box" – the user is aware of its operation and has confidence in the fact that they are accessing a secured device or action, but has little idea how the minitiae of the authentication procedure is being achieved. In order for authentication to be achieved in a black box fashion, it has to be designed with fault tolerance in mind. This needs to be achoeved by means of redundancy. The user should never be prevented from accessing needed resources if there is a failure in one component of the authentication mechanism. The mechanism should be able to function recover from partial failures so as to maintain its rationale of being as invisible as possible. Users should be freed to concentrate on their primary tasks, without being required to explicitly prove their identity from time to time.

The remainder of this paper examines three behavioral biometrics: keystroke dynamics, voice analysis, and use patterns. It is envisaged that these will be combined to achieve the invisible multimodal authentication that will facilitate a black box approach.

3 Multimodal Biometrics

A biometric identification system is called *multimodal* if it combines two or more biometric identifiers in order to authenticate a user. For example, physiological biometric systems can use a combination of, say, fingerprints and retinal scans to improve the probability of correctly identifying a user. The purpose of combining more than one biometric is to reduce the possibility of errors (either False Accept, where an unauthorized user is granted access, or False Reject, where an authorized user is denied access) and to reduce the dependence on a single identifier. Consideration must be given to how the biometrics are combined. The two possibilities are to combine each of the patterns into a single pattern, and make a decision based on that pattern, or to make a decision based on each pattern collected, and then combine the individual decisions into a final determination. Behavioral biometrics can also be combined into multimodal biometrics. This section examines three possible behavioral biometrics that are candidates for combination: keystroke dynamics, voice analysis, and use patterns.

3.1 Keystroke Dynamics

Keystroke dynamics attempts to uniquely identify a device user based on their typing patterns, either by requiring them to type a specific phrase or by simply sampling the user's typing patterns while they use applications that require keyboard input. Interest in keystroke dynamics as a potential distinguishing characteristic has a long history. It was applied to Morse code operators - clever listeners could distinguish one operator from another by the operator's *fist*, which is the distinctive pattern and speed of the dots and dashes transmitted. Keystroke dynamics on computers was first suggested as a behavioral biometric by Spillane in 1975 [21]. Since then, an extensive amount of research has been performed in this area. Studies have examined its viability as an authenticator on both mobile

devices [5,8,19] and desktop computers [1,10,15]. Early attempts used statistical classifiers to determine whether a person's keystroke patterns matched a previously stored pattern, but the state of the art is to now use neural networks, since their use has been shown to reduce the number of characters required to identify a user [6].

On its own, keystroke dynamics is not expected to be enough to uniquely identify an individual, although there is sufficient information to allow identity verification [13]. Its strength is that typing is something that most users do when interacting with a computing device, and therefore collecting typing characteristics can be undertaken by taking advantage of the users' current tasks. Although keystroke dynamics is not discriminatory by itself, it lowers the likelihood of accepting unauthenticated users. When combined with other similar biometrics, the data presented with keystroke dynamics is expected to provide enough information to uniquely identify a particular user.

3.2 Voice Analysis

Voice analysis compares samples of a person's voice to a pattern from the known authorized user. While voice recognition is a heavily researched field, the amount and quality of available research has declined since 2001. It has been found to be an area of limited potential because a person's voice alone is not considered unique enough to be the basis of an authentication mechanism [3]. If the main goal of the voice analysis is to *identify* a given person, the research supports this method, although there are still limitations. The quality of the microphone as well as distortions due to background noise can negatively affect the standard of the voice patterns gathered.

Despite this negative result, some research has been done on using speech as an authentication mechanism, often in conjunction with another biometric to form a multimodal biometric system. Iwano *et al.* combined speech analysis with ear images to create a multimodal biometric system, but the speech analysis had significant Equal Error Rate (EER) values (around 40% with a low Signal to Noise ratio [12]. The results of combining speech analysis with ear images was more promising; the error rates dropped by about 75%, although were still far to high to be used as a biometric. Voice patterns were matched with facial recognition to create a multimodal system designed by Brunelli and Falavigna in 1995 [4], although the error rates for voice analysis alone were quite high at 14%. The BIOMET system developed by Garcia-Salicetti *et al.* uses five identifiers including voice patterns to distinguish one person from another [9]. The purpose for using such a large number of identifiers was to offset the failings of each identifiers with the strengths of the others. None of the systems studied so far has a low enough error rate for the voice analysis section alone to uniquely identify individuals.

Voice analysis as a possible authenticator need not be discarded completely, however. Research has been performed in the area of *conversational* voice analysis, where the patterns of a person's natural way of speaking (i.e., speed, pronounciation, word repetition) are used to identify a person [17]. In some cases,

it is the role the person plays in a conversation (say, boss and employee or interviewer and interviewee) that was studied rather than identifying the actual person [23]. This part of the voice analysis module is promising, and is likely to provide a suitable level of certainty that a person is who they claim to be. When combined with other behavioral biometrics, conversational analysis becomes a promising possibility for passive authentication.

3.3 Use Patterns

Use patterns involves collecting information on how the user interacts with a computing device (i.e., a laptop, desktop computer, or mobile device) and using the uniqueness in these patterns as a biometric identifier. Examples include who the user calls or sends text messages to, and how often, web sites visited, applications that are loaded, and what type of music is played and with what frequency. These uses of a device provide a rich source of information about who is using the device since it is unlikely that any two people use a device in exactly the same manner.

Use patterns have generated some interest in the area of behavioral biometrics. Clarke *et al.* mentioned "service utilization" as a possible behavioral biometric in their study of users' attitudes regarding authentication on mobile devices, but did not attempt to use it in a working system [7]. While it is clear that there is some identifying information to be found in tracking a person's device use, it has not been a well-researched area, particularly in the mobile device field. When combined with other authenticators such as voice analysis and keystroke dynamics, it is hoped that it will provide a method of further reducing error rates.

4 Pattern Classification

Biometric systems, both physiological and behavioral, use pattern classification methods to compare a known sample to a gathered sample. The two major fields of pattern classification that are used in biometrics research are statistical classifiers and neural networks. In practice, neural networks are often considered a sub-type of statistical classifiers [2, p. 8]. Statistical pattern recognition algorithms use statistical information about each biometric sample in order to classify them into groups. The patterns are feature sets that group together defining points (i.e., measurements) in the original signal in order to create a symbolic representation of that signal. Examples of statistical pattern classifiers are Bayesian filters, naive Bayes classifiers, and the k-Nearest Neighbor algorithm. Neural networks are an extension of statistical pattern recognition since they follow the same sorts of rules, but have improved upon them with the improvement in computing power and resources. The use of neural networks has reduced the length of the string required in order to authenticate a user via keystroke dynamics. [6], which makes authenticating using a short character string viable. Despite the long research history of statistical methods, they

have been found to produce higher error rates in keystroke dynamics research, when compared to neural networks [5]. However, there are always tradeoffs when selecting an algorithm to use in a computing environment; in this case, neural networks require a large training set in order to correctly classify patterns, they must be re-trained if a new user is added to the network, and they are computationally and memory-intensive pattern classification methods [5,14].

In addition to the practical concerns of what type of pattern classifier to use, consideration must also be given to the results of such comparisons. As with all biometric systems, samples of each person's voice, keystroke patterns, and use patterns must be made available for comparison and testing purposes, but there must also exist a large set of "world view" patterns for each of the three biometrics. The reason for this is twofold: first, the non-authenticated user patterns can be used to test whether the module in question reduces the confidence level in the presence of a non-matching pattern. The second reason is somewhat more important. In order to show that a particular pattern type is a good candidate for a behavioral biometric, it must be shown that the user's pattern is distinct enough from a representation of other users' patterns – the so-called world view. Therefore, not only must the chosen biometric identifier be unique enough for comparisons, it must also have a large corpus of non-authenticated patterns to make up the world view. Such corpora are widely available for voice patterns [11,22], but are not necessarily available for use patterns or keystroke information on a mobile device. Such corpora must be created in order to provide proof that the chosen biometric provides the distinctiveness required for authentication purposes.

5 Conclusion

Behavioral biometrics have strong potential as a passive authentication mechanism. Careful thought must be given to how uniquely identifying each biometric is, and whether the biometric pattern can meaningfully be combined with others to constitute a multimodal identifier. Such identifiers improve the system by providing additional certainty that the user is who they claim to be. Such redundancy can be seen in many mature and workable systems, and provides a measure of fault tolerance that is essential in biometric authentication systems. Consideration must also be given to showing that a particular biometric can be uniquely identified from a world view of other such patterns, in order to show that the biometric will match its owner's pattern, and no other, with a reasonable degree of certainty. These considerations will serve to guide the research process and their existence provides strong support for future research in the area of behavioural biometrics as an authentication method.

This mechanism is not without its concerns. There are privacy concerns related to the use of behavioral biometrics since the type of data gathered is from user's private emails, text messages, telephone calls, and physical location. This research will use anonymising techniques so that as much personal detail as possible is removed, although removal of all data will be impossible.

References

1. Ahmed, A.A.E., Traore, I., Almulhem, A.: Digital Fingerprinting Based on Keystroke Dynamics. In: Proceedings of the Second International Symposium on Human Aspects of Information Security & Assurance (HAISA 2008), Plymouth, UK, pp. 94–104 (July 2008)
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press (1995)
3. Bonastre, J.-F., Bimbot, F., Boe, L.-J., Cambell, J.P., Reynolds, D.A., Magrin-Chagnolleau, I.: Person Authentication by Voice: A Need for Caution. In: Proceedings of Eurospeech 2003 (2003)
4. Brunelli, R., Falavigna, D.: Person Identification Using Multiple Cues. IEEE Transactions on Pattern Analysis and Machine Intelligence 17(10), 955–966 (1995)
5. Buchoux, A., Clarke, N.L.: Deployment of Keystroke Analysis on a Smartphone. In: Proceedings of the 6th Australian Information Security Management Conference, Perth, Western Australia, pp. 40–47. SECAU - Security Research Centre (2008)
6. Cho, S., Han, C., Han, D.H., Kim, H.-I.: Web based Keystroke Dynamics Identity Verification Using Neural Network. Journal of Organizational Computing and Electronic Commerce 10(4), 295–307 (2000)
7. Clarke, N.L., Furnell, S.M., Reynolds, P.L.: Biometric Authentication for Mobile Devices. In: Proceedings of the 3rd Australian Information Warfare and Security Conference 2002, pp. 61–69 (2002)
8. Clarke, N.L., Furnell, S.M.: Authenticating Mobile Phone Users Using Keystroke Analysis. International Journal of Information Security 6(1), 1–14 (2007)
9. Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., Leroux les Jardins, J., Lunter, J., Ni, Y., Petrovska-Delacretaz, D.: BIOMET: A Multimodal Person Authentication Database Including Face, Voice, Fingerprint, Hand and Signature Modalities. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 845–853. Springer, Heidelberg (2003)
10. Gunetti, D., Picardi, C.: Keystroke Analysis of Free Text. ACM Transactions on Information and System Security 8(3), 312–347 (2005)
11. Hirschman, L.: Multi-Site Data Collection for a Spoken Language Corpus. In: Proceedings of the Workshop on Speech and Natural Language, pp. 7–14. ACM (1992)
12. Iwano, K., Hirose, T., Kamibayashi, E., Furui, S.: Audio-Visual Person Authentication Using Speech and Ear Images. In: Proceedings of Workshop on Multimodal User Authentication, pp. 85–90 (2003)
13. Jain, A.K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition. IEEE Transactions on Circuits and Systems for Video Technology 14(1), 4–20 (2004)
14. Karatzouni, S., Clarke, N.: Keystroke Analysis for Thumb-based Keyboards on Mobile Devices. In: Venter, H., Eloff, M., Labuschagne, L., Eloff, J., von Solms, R. (eds.) New Approaches for Security, Privacy and Trust in Complex Environments. IFIP, vol. 232, pp. 253–263. Springer, Boston (2007)
15. Obaidat, M.S., Sadoun, B.: Verification of Computer Users Using Keystroke Dynamics. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics 27(2), 261–269 (1997)
16. Price Waterhouse Coopers. Information security breaches survey 2004. Technical report, Department of Trade and Industry (2004)
17. Psathas, G.: Conversation Analysis: The Study of Talk-in-Interaction. Sage Publications (1995)

18. Ross, A., Jain, A.K.: Multimodal Biometrics: An Overview. In: Proceedings of the 12th European Signal Processing Conference (EUSIPCO), pp. 1221–1224 (September 2004)
19. Saevanee, H., Bhattarakosol, P.: Authenticating User Using Keystroke Dynamics and Finger Pressure. In: Proceedings of the 6th IEEE Consumer Communications and Networking Conference, Las Vegas, NV, USA, pp. 1–2. IEEE (2009)
20. Smith, R.E.: Authentication: From Passwords to Public Keys. Addison-Wesley (2002)
21. Spillane, R.: Keyboard Apparatus for Personal Identification. Technical Report 17, IBM Technical Disclosure Bulletin (1975)
22. Taussig, K., Bernstein, J.: Macrophone: An American English Telephone Speech Corpus. In: Proceedings of the Workshop on Human Language Technology, Plainsboro, NJ, USA, pp. 27–30. ACM (1994)
23. Vinciarelli, A.: Speakers Role Recognition in Multiparty Audio Recordings Using Social Network Analysis and Duration Distribution Meeting. IEEE Transactions on Multimedia 9(6), 1215–1226 (2007)

The Metaphysics of Communications Overload

Richard H.R. Harper

Microsoft Research Cambridge

Abstract. This paper enquires into the nature of the act of communication between two or more persons. It proposes that such acts are best conceived of as moral, as related to the performative consequences of the acts in question. Given this, the paper then asks what applicability phrases like 'overload' might have, and whether quantitative techniques have a role other than as a heuristic in understanding and designing tools for the control of communication overload between people.

Keywords: Communications, human activity, overload, moral, common sense reasoning, scientific concepts.

1 Introduction

In Microsoft, each employee sends and receives about 120 emails every day; many also receive alerts from RSS feeds; and, most, if not all, run *Link*, its own Instant Messaging client. Now of course the staff at Microsoft might like to think that they are busy, efficient and effective people, and that they are knowledgeable enough about the communications technologies of the 21st Century to leverage them for our own benefit. After all, Microsoft helped invent some of them and if not, then it certainly has a business interest in most. Consequently, Microsoft staff should know about these things. Yet any visit to a Microsoft office will find the staff complaining: they say that they are constantly interrupted; that they can't keep up with all the email; that they find it difficult to say Goodbye when IM'ing. The result, they say, is that there is not enough time to get their work done. Somehow the balance of things seems to have gone wrong, they will explain; the tools designed to let them work better seem to have had the opposite effect. It is not only at work that this malaise seems to be appearing. For these individuals will also go on to say that when they leave work their personal mobiles start bleeping as SMS's arrive; 'There are voice messages too!' they complain. And, worse, when they get home, there are traditional letters–not many to be sure, but always some–and these also have to be dealt with. So they say that if 'at work there is no time for work', so at home there is no time for 'being at home'. The point of their complaints is that their world–which is of course the world most readers of this short chapter occupy–seems to be getting harder to live in, busier than ever, fraught with more things said and communicated than ever before. It is no surprise, then, that each morning, over coffee, Microsoft staff can be heard to assert, 'Surely, a threshold is being reached! Enough, already! No more communication!'

R. Murray-Smith (Ed.): MSSP 2010, LNCS 8045, pp. 42–50, 2014.

Within research, this issue, the idea that some kind of tipping point beyond which the balance between what is practical and what is excessive has been or is about to be reached, is well known: the phrase *communications overload* is commonly heard. Many researchers are devising tools and techniques that can reduce this 'problem'. Some are devising machine learning applications that assess whether a change in the content of a website is sufficiently interesting to alert (via RSS feeds) a 'user' for example; others are devising filtering mechanisms that can let users 'triage' their in-trays more effectively. Yet others are designing ways of integrating messaging channels so as to reduce the burden of dealing with them all. Some of these solutions are, even by their author's own admission, forms of fire fighting. Assessing the degree of change in an RSS feed seems to be a case in point: all this does is put off to the future the moment when a user says, 'That's it! No more feeds!' Similarly, new ways of filtering and triaging only delay the day when the limits of time press down: 'When does one deal with the less urgent if all one ever has time for is that which is urgent?', and 'What about the simply important if not urgent?' one can hear a future user grumble.

Curiously, many of the researchers who are undertaking projects into these and other 'solutions' are doing something else, something that seems, at first glance, perplexing. These attempts at solving the communications overload are not by any means the primary focus of their research endeavors. Indeed, one might say quite the opposite: for in-between their continuous emailing and IM-ing, many of these researchers spend much of their time adopting new ways as they arise: keeping up with their newly acquired *Facebook* accounts for example or creating short messages via *Twitter*, on their mobiles. In other words, they seem to enjoy and indeed indulge in ever more forms of communication. And, even more curiously, these same people also put a great deal of effort in to devising new ways of communicating. They seek ways of conveying tactile experiences, as a case in point, to supplement audio-visual messaging; they devise new social communications systems that let people vote and comment and express *en masse*. In other words, they delight in the very thing that they seem so often to complain about: they gleefully produce the content that at other times they say weighs them down. At work and at play they fill their lives up with the thing that they say stops them working and playing. They communicate yet complain about communication; they express themselves in new ways yet berate the fact that there is not enough time to listen to others' expression.

Presented this way, this doing of one set of things and saying of another, might seem an amusing albeit lamentable fact of modern lives. Sure, we are all too busy these days, but what more can one usefully say? I think one can say something, something about where we have come from, how we got here, and where we might go in the future. I think one can also say something about how we have come to think about ourselves, what we think 'we are'—as a species who suffer from communications overload. I think all of this has partly to do with our desire to communicate and express, and partly the relationship between this and our ability to devise and exploit new technologies that foster and enable that same expression. Beyond this it also has to do with a philosophy about what a human is in this day and age. This philosophy constitutes a vision, a view about what the human who does all this communicating might be.

In my book *Texture* (MIT Press, 2010) I argue that why we communicate (and how and in what form), and, how, in turn, this communication keeps making more communication, is a measure of our age – for it ends up being a measure of us, of what we do, it seems to me. We are people who are communicants. But I also argue that this predilection for communication has also led us to create a new set of measures to apply to ourselves. Unfortunately, I do not think these measures are good or accurate. On the contrary, I argue that the measures conjure up a view of the human that is distant from how humans ought to understand themselves when it comes to the question of overload. These measures are derived from a sort of corrupt scientific vision of what the human communicator is and this vision is largely opposed to the vision of the human that people themselves use in everyday life when they think about and judge their own (and their friends and colleagues) acts of communication. I argue that if you look carefully at these every day or common sense techniques — the ones deployed in practical action – you will see that the value of communication is central, and that this value is constituted only in very small part quantitatively. A much more important set of elements concern the moral value that an act of communication delivers. Thus, for someone to say 'I love you' means a great deal when it is said once. This value may alter if it is said many times. But this value is moral, above all else, and this value has to do with the consequences the act has on the relationship between the participants. The quantitative aspect of this value, how often something is said, is not the central part to it, though it might create inflections to the moral consequences in question. Yet, it seems to me that the techniques derived from the purportedly scientific approaches used to judge questions like communications overload more or less willfully ignore this delicate but fundamental fact: that value, that moral consequences of communication, are the metric that ought to be applied when thinking about communication and communications overload.

I propose that many of those researchers who are looking at the problem of communications overload have been tempted by various concepts that derive from what I call the metaphysics of computer science – ideas deriving from Turing, for example, and more latterly from Bayes and the current manifestation of his ideas in computer science, namely machine learning, which take them away from asking questions about what values are delivered when people message to one another. These concepts (there are a bundle, nested with one another in numerous ways, combining as they do aspects of signal processing theory, cybernetics, theories of inference, as well as machine learning, statistics and much else beside) encourage a disregard of these values. Doing so, it seems to me, can lead to profound misunderstandings about what communications between people is all about and can prohibit sensible attempts to answer whether we do in fact suffer from communications overload; of greater salience to this book it can also scupper creative ways of using technology to address the problem of controlling communication.

2 An Example

In this chapter I do not want to explore every aspect of what those values might be or how they might leverage better answers, hoping instead that the reader might turn to my book for discussion of that in detail. But what I do want to note is how this temptation to overlook the values in what humans do when they communicate is so powerful and pervasive that it affects people from many disciplines, and not just those in, say, machine learning and signal processing, constitutive of the readership of this book. If one looks at some of these other instances one will find illustrations of just the confusion and misunderstanding that can result.

Take, as one such case, the view from what has come to be called *communication science* (or sometimes *media studies*). Central to this discipline is exploring the relationship between the human user (or recipient) of media content, especially broadcast content, and the content itself. This discipline looks at how content affects the recipient. When the discipline first emerged some twenty or thirty years ago, defining the media (and hence the message that affected the user in one way or another) was easy to do. But today, there are various sources of media, not just newspapers, radio, and television. The Internet has altered the landscape so much that a plurality of channels now mediate content to (and from) the user. Hence not only is it more difficult to ascertain the relationship between message and action, between content and the human, but in some instances media has no effect on the human. This is because people are becoming overloaded–and hence they cannot be subject to the consequences of some media, some message, since the content in question is likely to have disappeared in a chaos of media—TV, radio, YouTube, e-newspapers.

This is the conclusion of W. Russell Neuman and colleagues' report *Tracking the Flow of Information into the Home* (2007), a study of media consumption in the United States from 1960 to 2005. In this case, Neuman and his colleagues argue that a human can be treated as an information processor, a processor of *words*. Taking their cue from Itheil Pool's research in the 1980s (see Pool's 1983 article *Tracking the Flow of Information*), they argue that adults read 240 words per minute. With this base line, they analyze the time that the user has to consume the words sent to the home via the many channels or media that are "sent" or "pulled" into that setting. They conclude that there are too many words for the user to read or consume in the time available. Automated or intelligent systems will be necessary to select content on behalf of the user in the home of the future. Thus what Neuman *et al.* do is disregard the purpose of words, the 'reason behind the act of communication' and focus instead on simply counting the words.

This sounds like a kind of science but it comes at a cost. It is an odd thing to change a heterogeneous activity such as reading and distil into a simple metric like 240 words per minute. In this view, reading the back of a cornflake box is the same as reading a newspaper, a novel, a blog, a manual for a new washing machine—or a love letter. This view also makes the human choosing to do these different acts the same too. It makes reading a singular, mechanical act and makes the human equally mechanical. This approach can be appealing because it allows a simple quantification, but it offers a rather feeble vision of the human that reads, it seems to me. As I noted

with co-author Abi Sellen in *Myth of the Paperless Office* (2003), reading is an activity that is easy to oversimplify, and reading is a catch-all phrase for a number of activities that reflect something of the human in question—who they are and what they are seeking to do when they read. As it happens, only some activities labelled "reading in the workplace" can sensibly be understood in terms of speed. Indeed, speed is not the important dimension to be applied when thinking about reading technologies for work, for example. This is also likely to be the case in the home setting, the one that Neuman et al. concern themselves with. As Alex Taylor and I noted in 2003 (115–126) (in a study about television consumption), when people go home and pick up a newspaper or switch on the TV, they are not approaching that action as merely an information processing task. They might be doing so simply to turn themselves off. Reading the paper and watching TV here are ways to end the day's work and begin the day's leisure. These activities are not to be understood as being done on the basis of a choice between content formats or types or in terms of speed. However many words are read or news items watched, this type of activity is concerned with using twenty minutes to make a transition between work and home. And this, in turn, says something about the kind of person who chooses to break up their day in this fashion (not all people will do so, after all).

I do not want to suggest that in offering quantitative measures of an activity such as reading (and media consumption more generally), Neuman *et al.* are being disingenuous; they are not *intending* to lose sight of the phenomena they are seeking to analyse; nor am I suggesting that they are merely a bit lax in their science. It is rather that, in their desire to turn to this rendering of the phenomena (this quantification of media usage), they end up losing sight of what people are doing when they consume. Their approach prohibits understanding why people listen, watch, or read; it stops Neuman *et al.* understanding that reading is not always about consumption; it is sometimes about passing the time of day.

As I have remarked, their countings of media input and media consumption are typical not just of their discipline but of the ways that others, in quite different disciplines, also tend to think about humans and their acts of communication. There is nothing wrong with using quantitative tools; but one has to be careful: when one turns to them one has to ask, what does one gain and what does one loose? Is counting appropriate for the questions one is asking? Sometimes the answer will be yes but not always. Think about the chapters in this book, and the various questions that motivate the research reported in each: is quantification the right technique for all? Most? Just a few? What criteria would one use to judge? Besides, when I say counting, what I am suggesting is counted?: merely the volumes of messages or some property of the message? Or is it, for a third option, the sender or the recipient that is being counted? Beyond this, there is the question of how the counting is being done, what it entails: the example above of media consumption uses a kind of arithmetic, but when people use probalistic techniques to research aspects of human communication they are doing something different, something that might be more subtle. They might be pointing towards an emphasis, a tenor, a likelihood; not something strictly or even literally numerical, even though numbers are used in the calculation of this likelihood.

There are subtleties, here, some quite consequential. Nevertheless, my point remains the same: one still has to be careful: is a message sent after a prior message on the same topic 'probably too much'? How would one know? I have suggested above that one criteria that one might use to make such distinctions has to do with what might be the consequences of some act; hence what the act 'is'—that it seems to be the same as some prior message say—is not sufficient to analyse the thing 'itself'. I am applying this to the question of communications between people and suggested that it's not just what they say and thus how long, how quick, how often, or even whether they repeat themselves that matters, but what results when the act of doing the communication is considered too. So, one might ask why someone keeps repeating themselves: are they disregarding the possibility that they might overload the recipient of that message? Or are they deliberately trying to overload them, as a way of getting them to attend to something else, a prior message perhaps? Or are they being playful, seeking to annoy the person they are messaging to?

I would be the first to admit that treating the issue in this sort of way does not mean that answers are more easily come by. Asking what an act achieves extends the topic and the evidence that needs to be brought to bear. At least with a simple counting of, say, the words in an act of communication one limits the data; but how would one know when one has defined the consequence of an act? It is tempting to take the easy route, all the more so if we can say it is in the name of science. It is not just scientists and scholars who are so tempted. At the current time, many laymen tend to think of themselves in quantifiable ways. What I have suggested are the more apposite every day or common sense ways of understanding communication are being infected by what I think are infelicitous understandings. By laymen, I am thinking of all of the readers of this chapter, of ourselves in other words, but not as we are now: with our professional guises on. I am thinking of us when we take off our professional hats, go home, and orient to our lives in ordinary common-sense ways. It seems to me that then, however we might have thought about communications in the past, we often do look at the infinite number of channels on our TVs and wonder how we might consume them all. We do look at the news on the Web and wonder how much time we could allocate to reading it all. We do look at all our emails in our domestic accounts and the postings on our Facebooks and think, 'how can we deal with it?' We do, beyond this, start looking at ourselves in terms of inputs and outputs and start treating our communicative habits, all of them, the mediated communications as well as personal face to face ones, as visible measures of overload. Hence, we notice these acts of communications and start counting. We look at the numbers of messages received and wonder how we can balance the delight we get from their receipt against the labour we need to put in to reply. As we do so, we naturally turn to measures of our time and the pressures on it since this seems the most precious resource of all. We start from the *assumption* that quantitatively demonstrable overload is the measure of our age, and so we look at ourselves and our activities with that in mind and *make it so*.

If we don't start from this point, we soon learn that we ought to by the narratives produced by the experts—the media specialists like Neuman *et al.*, and by our computer science and HCI brethren offering us solutions to our computer mediated

overload. We thus find ourselves ignoring the fact that when we read the back of a cornflake box at breakfast, our eyes are simply caressing the words and not consuming them; and similarly we forget or ignore the fact that when we switch on our home computers and gaze at the evening news on our Web feeds, we aren't digesting what we see but are waiting for our minds to unravel the news in *our own* affairs, not in the world at large.

Our bodies might consume words then but not in the sense that Neuman *et al.* mean it or indeed those who offer various quantitatively based techniques to judge, parse and weigh our communications traffic. The goal of those who deploy these techniques is often to reduce communication. My concern is that in looking at communication as they do they can entirely miss the point of communication. Sometimes one will want to reduce communication to be sure, but if one starts from the assumption that communication is to be judged in terms of moral value, then what is or is not too much becomes a very sticky question to deal with altogether. No amount of inference, quantification or statistics will help with that unless one starts with understanding of human affairs. These affairs are often obtuse in their purpose and meaning, even though they are common, natural, 'common sense'.

3 Conclusion

Perhaps I am being too sensitive to what is popular at the current time. Some years ago Marta Banta noted in her book *Taylored Lives* (1993), that society had already become transfixed by numerical ways of thinking about our endeavors. Banta's analysis was written well before the onset of any concern with communications overload (it was about the desire to measure and monitor every activity in the home, at work, all with the expectation of managing ourselves better). Thinking of her draws attention to how questions about why people communicate and who the communicating human might be are as old as philosophy itself—perhaps even as old as language itself—and thus certainly older than computer science and the other disciplines that dominate our own time. In my view the best history of thinking about the relationship between how we think of the human as a communicating agent and the technologies we devise to enable that communication is John Durham Peters' *Speaking into the Air: A History of the Idea of Communication* (1999). Peters is particularly good at exploring the conceptual implications that various technologies have on the structure or hopes that are embedded in what he calls the "metaphysics of the idea" of communication. New technologies alter this metaphysics, he shows.

For example, the invention of recording devices in the nineteenth century that could 'copy' and 'replay' human voices helped cultivate the idea that people had a 'speaking soul' that was 'trapped inside a body'. This might sound odd to us today, but it is hard to capture just how startling people found the recorded voice at that time. The hearer of these early recordings thought that they were not hearing the same thing as they might when a person spoke in ordinary circumstances; somehow the recordings conveyed something ethereal, ghostlike; something transcendental. This led people at start thinking about "innerness"—of a thing, a spirit, perhaps a soul

trying to get out and transcend the body and its "skin" through words. Hence the title of Peter's book: 'Speaking into the air'. Even new words such as 'solipsism' were constructed coined as a result of the shock that people felt on hearing the recorded voice for the first time. Peters goes on to say that there is a contemporary metaphysics, too, though the one he focuses on is different to the one I have highlighted – for reasons we need not go in to now. He says that attempts (in the late twentieth century) to devise ways of seeing each other via video, for example, and relatedly attempts to offer more sensual aspects to communication to augment sight (like touch), draw attention to what he calls the erotic aspect in the act of communication. His view is not that people have always communicated for erotic reasons but that the late twentieth century and early twenty-first century have led us to think and act as if being in touch means just that—something erotic. Our technologies of communication have helped create what we think we are and hence give motive to our acts of communication. My view is that, as we enter the second decade of the twenty-first century, a somewhat different kind of metaphysics is coming to dominate ideas about communication—in this case one that says that people are processing machines of various kinds, and that problems like overload can be solved by determining what is the threshold beyond which these machines can no longer process. This view is held quite commonly as it has gradually suffused everyday reasoning. My discussion of how communications and media studies treat the user of media content illustrated this.

My case is that one ought to recognize that this view is somewhat arbitrary, and hence a kind of metaphysics. Other views could have come to dominate; Peter's erotic is another. But to say such views are arbitrary is not to deny they have causes, that they have emerged for good reason. Nor should it prevent us from being sensitive to the value a view might have. All views will have advantages; doubtless they will have disadvantages too, as I suggested with regard to the quantitative view. Though a view may be better or worse than others, one should nevertheless treat it according to whether it is useful or not. Sometimes it will be, sometimes it won't. My purpose in presenting these arguments has been to help the reader make such judgments about the views they use or read about in the chapters that follow. They will be better able to understand what the claims assume and posit, what is the metaphysics in each case. To be sure, some views will be grounded in appropriate understandings of what might be occurring when communication occurs. But there are many types of communication–what I have been remarking on is that peculiar type that occurs when people communicate with each other. It doesn't matter whether it is mediated or not, face to face in real time or conveyed asynchronously via, let us say, text. All of this is moral and is to be understood as such. Other types of communication, between a person and a machine, for example, can hardly be called moral. Information exchange might be a better phrase. But the point is that one needs to be aware of such distinctions. Otherwise we will misunderstand what is being argued and said, what is being communicated, and when for example, too much has been conveyed. It is only then can we facilitate new ways of communicating through technology more effectively.

References

1. Banta, M.: Taylored Lives: Narrative Productions in the Age of Taylor, Veblen and Ford. University of Chicago Press, Chicago (1993)
2. Harper, R.H.R.: Texture: human expression in the age of communications overload. MIT Press, Boston (2010)
3. Neuman, W., Park, Y.J., Panek, E.L.: Tracking the Flow of Information into the Home: An Empirical Assesment of the Digital Revolution in the US from 1960-2005 (2007), http://www.wrneuman.com/Flow_of_Information.pdf
4. Peters, J.D.: Speaking into the Air: a history of the idea of communication. University of Chicago Press, Chicago (1999)
5. De Sola Pool, I.: Tracking the Flow of Information. Science 211, 609–613 (1983)
6. Sellen, A., Harper, R.: Myth of the Paperless Office. MIT Press, Boston (2003)
7. Taylor, A., Harper, R.: Switching on to Switch Off. In: Harper, R. (ed.) Inside the Smart Home, pp. 115–126. Springer, Godalming (2003)

Capturing Performative Actions for Interaction and Social Awareness

Julie R. Williamson and Stephen Brewster

University of Glasgow, Glasgow UK

Abstract. Capturing and making use of observable actions and behaviours presents compelling opportunities for allowing end-users to interact with such data and eachother. For example, simple visualisations based on on detected behaviour or context allow users to interpret this data based on their existing knowledge and awarness of social cues. This paper presents one such "remote awareness" application where users can interpret a visualization based on simple behaviours to gain a sense of awareness of other users' current context or actions. Using a prop embedded with sensors, users could control the visualisation using gesture and voice-based input. The results of this work describe the kinds of performances users generated during the trial, how they imagined the actions of their fellow participants based on the visualisation, and how the props containing sensors were used to support, or in some cases hinder, successful performance and interaction.

1 Introduction

Capturing and using actions and behaviours for interaction has seen a wide variety of applications, from replacing traditional buttons with gestures (Crossan *et al.*, 2008) (Strachan *et al.*, 2007), to supporting self expression through performance in public places (Perry *et al.*, 2010) (Sheridan *et al.*, 2011), to creating remote awareness of friends and family through ambient interfaces (Dey and de Guzman, 2006). In the area of social signal processing, previous research has focused on creating a foundation of work aimed at sensing and detecting social signals, where effectively using or applying those signals for interaction remains an open challenge (Vinciarelli *et al.*, 2009). This paper presents a possible application area for these signals where simple actions and behaviours are sensed and visualized for intepretation by the users themselves in a remote awareness scenario.

An important aspect of this remote awareness scenario is that the actions and behaviours sensed by the system can be understood as *performances*. Indeed, nearly any action completed in a public place can be considered a performance of some kind. Goffman (1990) describes a wide variety of "performances" that people produce everyday, ranging from unconciously performed actions to specifically designed and directed personas and impression management. Goffman (1966) also describes social contexts as either focused or unfocused, where focused interaction are those with a single point of attention and

R. Murray-Smith (Ed.): MSSP 2010, LNCS 8045, pp. 51–63, 2014.

involve cooperation as opposed to unfocused interaction where people might be in the same place but not actively cooperating or interacting together. This performative perspective helps organize behaviour in public places into relevant categories and highlight behaviours of interest to social signal processing.

This paper presents an application that makes use of basic social signals to allow users to interpret these signals through a simple visualization. In the application, called MuMo, each user is represented by a fish in a virtual fish tank. Users' actions are displayed in the visualisation through their fish, where gestures or movements cause the fish to swim faster and audio or speech cause the fish to blow bubbles. Each user can view the fish tank visualisation as the wallpaper on a mobile phone. Thus, each user can gain some idea of other users' current context by looking at the visualisation. Ambiguity in the visualisation means that users can make a wide variety of interpretations based on what they see. Users can also perform intentionally for the interface knowing that others may be watching, and must balance the concerns of both their physically co-located and remote spectators. The MuMo system was evaluated in an "on-the-street" user study where pairs of users interacted with the system in both public and private spaces for two sessions spaced on week apart. The results of this study show what kind of actions users developed *in situ*, how they considered the influence of spectators, and how they interpreted the visualisation.

2 Background

Using a performative perspective on interaction, human actions can be viewed as a performance of some kind where people are constantly adjusting their own behaviours based on the presence (real or imagined) of spectators. Understanding action in this way has interesting implications for designers of interactive systems, where users can be viewed as actors, interaction spaces as stages, and spectators as the audience. Such a performative perspective can be used to leverage this perspective in design and how such performances can be captured.

2.1 Action and Performance

Goffman (1990) describes how everyday life can be understood from a performative perspective, a view that has seen growing popularity in human computer interaction. Goffman describes peoples' everyday behaviours as a performance, where people are constantly adjusting their actions based on feedback from spectators, using places as stages and making use of their appearance and props to support their intended impressions. Goffman describes a wide range of performances, from implicit performances of everyday action and impression management to explicit performances such as giving a formal presentation to an audience. These concepts can be further refined as impressions or performances *given* and *given off*, where impressions *given* relate to those explicit performances and impressions *given off* relate to implicit performances (Goffman, 1990). Implicit performances might be actions that are performed without being explicitly aware of them, but which are

unconsciously adjusted constantly throughout the day as feedback is gathered from spectators. More explicit performances carry with them significantly more intention from the performer and more clearly defined performer/spectator roles. Both impressions *given* and *given off* are interesting from a social signal processing point of view, where this performative perspective gives a sociologically motivated approach to organising these behaviours.

2.2 Performative Perspectives in HCI

The concept of interaction as a performance (Jacucci, 2004) provides a way of understanding interaction as the presentation of self and the experience of interacting in front of others. In interactive systems research, this means that performative concepts can be leveraged in design, such as the influence of spectators, users' percieved images of themselves, and narratives within interaction. Reeves *et al.* (2005) describes how the presence of spectators changes how people interact with systems in public places based on the size of their manipulations and the resulting effects. Dalsgaard and Hansen (2008) add to the performative perspective by developing the concept of "performing perception," describing in great detail the experience of performing with respect to the roles users must adopt throughout an interactive experience. Benford *et al.* (2012) describes how traditional narrative structures from theatre can be used to design uncomfortable but rewarding or fulfilling interactions. Each of these examples demonstrate how a performative perspective can be leveraged to inform the design of interactive systems.

2.3 Capturing Performances for Interactive Systems

There are a wide variety of sensors that have been used in activity recognition and social signal processing. Although accelerometers have been widely used in interactive systems, they have seen less action in social signal processing (Vinciarelli *et al.*, 2009). However, there are a several important signals that can be sensed through accelerometers and present interesting opportunities for visualisation and interaction. Crossan *et al.* (2005) demonstrate that accelerometers can be used to sense gait phase during mobile interaction for increased understanding of users' mobile context. This approach has also been used as a means to "instrument" users during evaluations to gain additional data about interaction context in the wild. Microphones have also been used as a mobile form of input in the instrumented usability scenario. Hoggan and Brewster (2010) used a phone's built in microphone to gather data about ambient noise levels to better understand users' current context during an in-the-wild study. However, accelerometers and microphones are not only used for such passive input. Jones *et al.* (2010) showed the possibilities of accelerometers for sophisticated input in a gesture-based text entry system. Scheible *et al.* (2008) created a system where throwing gestures performed on a mobile phone could "toss" content from that phone onto a large public display. Mobile phone sensors like these could be used to capture both actions "given" and "given off" to bring sophisticated social signal processing to a mobile context (Vinciarelli *et al.*, 2010).

3 Exploring Performance for Social Awareness

The study presented in the paper explored how simple behaviours and actions in a mobile context could be used in a remote awareness application. The evaluation explored not only how users interpretated this data but also how they experienced performing and generating this data, particularly when extravagent and exagerated actions were encouraged. During the study, participants were required to generate simple gesture and voice input *in situ* in public and private locations using a mobile remote awareness application with a partner over repeated trials. This application was designed to support divergent multimodal inputs with a high level of flexibility, create the experience of performing in different settings and participate as a distant audience member for a familiar other's performances.

This application, called MuMo, included a visualisation of a virtual fish tank where each user was represented by a fish in the tank that could be controlled using multimodal input. Users generated input by interacting with a small prop embedded with sensors. MuMo was designed to explore the issues of performance and the usage of props when the user was performing for two different audiences: one audience was the fellow participant watching the performance through the fish tank visualisation and the other was the immediately co-located spectators watching the live performance without necessarily being aware of its purpose or the interface itself. This application used highly flexible input methods, where participants were required to create their own performance style in real world locations using gesture and voice. Using this application, users were free to create a variety of performances to suit their current context and could participate as an audience member by watching the visualisation, where divergent imagined interpretations of the visualisation were possible. The possibility of this kind of extravagant performance Jones (2011) creates the opportunity for expression and imagination in real world contexts.

3.1 The MuMo Application

In the MuMo application, participants were each represented by a fish in a virtual fish tank, as shown in Figure 1. This visualisation could be seen as an animated background on each users' mobile phone and controlled using multimodal input. The application used a server/client architecture where each client updated the server with its current input values and pulled updates from every other user from the server roughly once per second. Thus, users could see the effects of their own actions in the visualisation alongside those of their fellow participants. Participants were told they could use gestures or motions to make their fish swim faster or use audio and voice input to make their fish blow more bubbles. In each case, the fish behaviour was based solely on the magnitude of input, although this was not explained to the participants. For audio input, the louder the sound level the more bubbles the corresponding fish would create. Thus, participants could perform any kind of speech or sound-based action and see the result in the fish tank. Similarly, changes in swimming movements were based

Fig. 1. Screenshots of MuMo application as an active wallpaper. Left shows fish tank visualisation as wallpaper, right shows visualisation with phone widgets.

on the magnitude of acceleration of the gesture performed. This type of sensing was designed specifically to support both extravagant and subtle input, meaningful and abstract input, or simply environmental input that could be reflected in the fish tank visualisation in real time. This flexible style of input afforded unconstrained interaction in order to encourage participants to generate creative methods of controlling the visualisation. This also allowed for imaginative interpretations for those watching the visualisation since the observed output in the visualisation could be generated in a variety of ways.

The interface was controlled using the SHAKE sensor pack[1] to collect accelerometer data with an added microphone as shown in Figure 2. This was then embedded into the various objects or props shown in Figure 3. These props were chosen to provide a variety of objects that could facilitate performance or demonstrate interaction in different ways. These included playful objects, an abstract object, an everyday object, and an object that displayed the bare electronics of the sensors. The playful objects included two plush toys and one solid toy in order to allow for enjoyable and playful interactions. The abstract object was a hollow red mould that would simply act to conceal the sensors. The everyday object was a book with a space hollowed out to conceal the sensors in order to disguise the interactive prop. The final prop was a clear glass jar that exposed the bare electronics of the system as a method for demonstrating the interactive purpose of the prop. These props were selected to provide different visual or cognitive clues for spectators about the performance in order to give performers different methods of exaggerating, disguising, or explaining their performance.

[1] More information: `http://code.google.com/p/shake-drivers/`

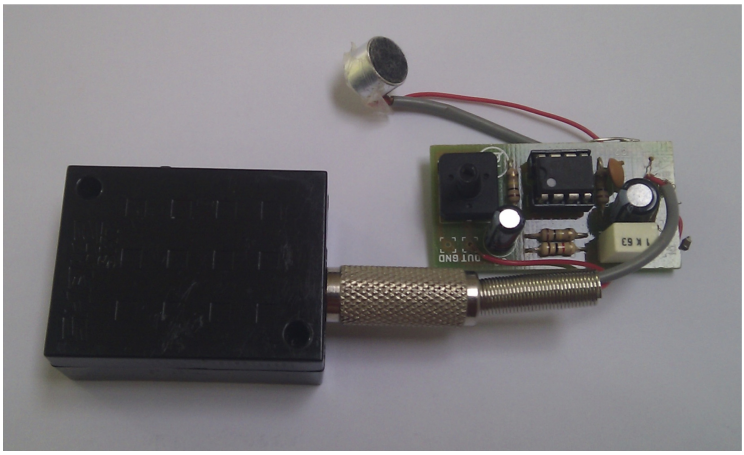


Fig. 2. The SHAKE sensing device with an added external microphone

3.2 The Study

Participants were recruited in pairs, where each pair completed two usage sessions spaced about one week apart. These sessions were repeated to give participants multiple chances to interact with the system and develop performance preferences based on multiple experiences. Before each session, participants were told only that they could control their fish's swimming behaviour using gestures and the bubbling behaviour using sound and were given a chance to briefly experiment with the system. Then, the session began with the first participant being taken to a public location, a busy pavement, while the second remained in a private indoor location. Once both were ready to begin, the first participant was asked to complete three performance tasks, such as creating more bubbles, while the second was asked to interpret the first participant's actions by watching the visualisation on the phone. After these performance tasks were complete, the first participant was then asked to interpret the other's actions while the second participant completed three performance tasks. The participants would then switch locations and the tasks were repeated. Each task lasted two minutes. This study design allowed participants to perform actions in both the public and the private setting as well imagine how their partner would perform actions in both settings. Once both participants had completed their tasks in each location, they were interviewed together about their experiences.

4 Results

The study involved eight participants recruited in pairs. The pairs included two couples and two pairs of friends, with four females and four males. The participants ranged in age from 20 to 28. The results focus on an in depth analysis of a relatively small user group in order to gain a highly personal qualitative

Fig. 3. Participants could select one of six objects containing an embedded sensor pack to control their fish in the tank

insight into the use and experience of this application. These results are based on the observation of the participants, recorded observations provided by the participants and transcripts of the interviews.

4.1 Creating Performances

Given that participants were allowed to create open ended performances using gesture and speech, it is not surprising there were a wide variety of styles and actions that resulted in the different locations where this study was completed. In each case, these actions can be analysed from a performative perspective to better understand how people generate actions and behaviours in this interactive context.

Performative Actions – Even though the sensors were contained solely within the prop, performances were not limited to interactions with that prop and often involved additional interactions purely as an enhancement to the experience and appearance of performing. For example, one participant chose to sing to the prop for voice input in the private indoor setting. Although this was an unnecessarily extravagant interaction, this participant found that performance enjoyable and amusing, especially when his partner imagined this performance. Another participant performed swimming motions with both hands while outdoors. Even though the prop would only sense the movement of one arm in this case, the participant still enjoyed performing with both arms. Perhaps this action better demonstrated the purpose of the participant's actions, where spectators might more easily understand the action of mimicking swimming with both arms. In these cases, the experience of performance was augmented with either playful

or meaningful actions to add to the functional aspects of interaction to make interaction more fun, more enjoyable, or more socially acceptable.

Hidden/Subtle Actions – Participants found ways of performing input that were subtle or hidden from passersby while still giving their fellow participant highly visible actions on the visualisation. Because the system was flexible enough to support both extravagant and subtle actions, participants could exploit this to balance their desire to perform for their partner while also considering the immediately co-located spectators around them. The hidden/subtle actions included input such as tapping the prop to make noise, fidgeting with the prop in hand, and using environmental noise to create input. For example, one participant chose to use the music of an outdoor performer as the input for their performance when audio was needed. These types of actions allowed participants to create meaningful input to the system without performing highly visual actions.

Functional Actions – In some cases, participants chose only to perform actions that completed the task without adding any additional performance or play. For example, participants would simply shake or wave the sensor to create gestures or say things like "I'm creating test speech for a system" or "I'm talking into the sensor now to see if something happens." In this case, participants did not try to actively hide or disguise their performance, but instead tried to demonstrate the purpose of the performance clearly by using "test speech" or rigid, purposeful actions. In this approach to impression management, participants aimed to make it clear they were interacting with a system by keeping the phone or prop visible and performing noticeably rigid actions in order to call attention to the action as purposeful input.

These different styles of performance were influenced as much by location as personality. For example, one participant performed purely functional actions while outside and highly performative actions inside. Another participant completed highly performative actions both inside and outside. Yet another participant completed hidden or subtle actions both inside and outside. Because the interface supported a variety of actions, participants were able to change their performance style as needed in order to continue participating and feel comfortable about interaction. These decisions varied between participants, depending on personal preferences and personality. These factors represent an interesting influence on social acceptability that needs further exploration.

4.2 Imagining Others

Because the MuMo application required participants to create their own input, fellow participants watching the interface could not be sure what kinds of actions their partner was performing given the current output. Participants had to imagine how they thought their partner might be performing based on what they could see in the visualisation and their knowledge about their partner's

current social context. This was both a positive and a negative aspect of this application, where some participants found it difficult to attach meaning to the interface while others enjoyed the process of imagining their fellow participant performing highly energetic, silly, or emotional behaviours. These imaginings not only contributed to the spectator experience of this application through the visualisation but also provided motivation for participants to generate performative input to the system.

For those participants that enjoyed imagining their partner performing through the interface, participants allowed and encouraged their partner to imagine highly divergent performances, even when this was not realistic or likely. For example, some participants imagined their partners singing or dancing as input for the visualisation even though their partner was in the outdoor setting and it was unlikely they would be singing or dancing there. Even though participants knew such energetic and performative actions where unlikely, participants were able to suspend their disbelief and enjoyed imagining these kinds of actions anyway. These creative imaginings occurred both when pairs of participants used highly visible, performative interactions and when pairs of participants used the most subtle and discreet methods of interaction. For example, one participant imagined her partner "singing a relaxing song" and "jumping with it [turtle] on one leg." These interpretations were recorded even though both participants used extremely subtle actions for input, such as microphone tapping. Participants enjoyed imagining these playful actions, even if they did not perform these kinds of actions themselves.

4.3 Props and Performance

During each of the two sessions, participants could select an object of their choice as their prop. The prop was an important part of the types of actions and behaviours participants would perform because the prop would be highly visible during interaction and could both support and hinder performative actions. For example, a playful prop like a toy might encourage fun interactions because toys are made to be played with while an everyday prop like a book may be more acceptable to carry around in public places but not typically be viewed as an "interactive" object. Of the props including in this study, the turtle object was chosen eight times, the dolphin was chosen five times, the book, jar, and owl were chosen once and the red mould was never chosen. When discussing their choices of these objects, participants described how the objects worked and failed as props.

Props as Toys – The most commonly picked objects were the turtle and dolphin plush toys. Participants favoured these props for their playful nature and their ability to relate to the lively and lighthearted application. These props were often used in a playful manner, even though participants knew that these kinds of actions would not provide any additional input to the application. For example, participants would move the fins of the turtle or cover it's eyes as part of their performance even though this did not generate additional effects.

Props as Pairs – Participants often chose their props based on their partner even though they knew the props would not be used together. Choosing props together allowed participants to better understand what kinds of performances their partner might complete and also provided a better connection between partners. For example, one participant stated that "first I wanted to pick the glass jar, but when I saw he picked a toy I wanted to pick a toy as well." Another participant stated that "I picked the dolphin because you picked a toy, so it's two soft toys. Otherwise, I would've picked the book."

Props as Everyday Objects – Although some objects, such as the book, represented common objects one might normally carry around, participants felt less comfortable using these objects when interacting with the application. While using the book as a prop, one participant stated that "when I was inside I sang a song, I just made it up. But when I was outside I tried to talk very quietly. It wasn't as normal as I thought it would be." When discussing other everyday objects that might be used as props, one participant stated that "you might put it [sensors] into an object that you walk around with, like a coffee cup, but you wouldn't talk into a coffee cup." Although these props might disguise or hide sensors effectively, they make poor interactive objects when it comes to performance.

Participants also discussed the benefits of different props with respect to physical attributes like size or texture. For example, when describing why the dolphin was a useful prop, one participant stated that "it's easier to hold than one of the hard objects, nicer to hold." When describing objects that would make the most desirable props, participants stressed the importance of using soft or flexible objects. The ability to manipulate the props and the comfort of holding a soft object made them easier to use. Participants also described the benefits of using different props to conceal the sensors. When describing why a prop would be better than simply holding the sensor pack, one participant stated that "it's bigger, so there's more you can do with it." Participants also described how the prop makes performance more comfortable. For example, one participant stated that "it was much easier to just wave around the turtle than it would've been to wave a bunch of sensors". Other participants would have preferred a more anonymous object. When discussing negative aspects of using props, one participant stated that "it made me more conscious of it, holding the object. If I just had the sensor in my hand people might not have noticed what I was doing." Because this application clearly had a playful nature, participants often chose props that encouraged this playfulness. However, props that are more abstract or anonymous were still desirable and in a different application area might have been more popular.

5 Discussion

This study provides some interesting insights into the ways in which these participants created performances in the wild, used props to enhance their interactions and demonstrated their intentions to co-located spectators. By performing

through the MuMo interface, participants were performing for the immediately co-located spectators as well as their fellow participant watching their actions through the visualisation. Thus, participants in this study were constantly performing for two audiences and had to balance the needs and expectations of these spectators simultaneously. For example, participants had to balance their desire to generate energetic or amusing input for their fellow participant with their desire to perform socially acceptable interactions in public places. In some cases, this meant that participants chose to limit their performance and the resulting output of the system, limiting the spectator experience for their fellow participant. In other cases, participants found ways of performing that were both comfortable for themselves and created ample output in the application for their fellow participant to enjoy.

Because this application required only basic actions but also supported extravagant ones, participants took full advantage of this flexibility and generated a wide variety of behaviours and actions through the system. The types of performances created were highly dependant on the location of the performance, with participants actively making decisions about their adoption of different performances based on their current location. In general, participants were more likely to perform highly visible or noticeable actions in the indoor location as compared to the outdoor location, which is in line with the results discussed in the previous chapter. Additionally, participants often adjusted their performances to match their fellow participant. Because the first session ended with an interview, participants learned what kind of actions their fellow participants had imagined them doing and what actions their fellow participant had actually performed during the first session. This was reflected in the second session where pairs of participants performed actions that were discussed during the first session. This included actions that might be amusing to their fellow participant or actions they thought the other participant might be performing as well. This demonstrates how social influence can affect adoption, even though this example is on a very small scale. For example, usage over time might allow constantly evolving practices and behaviours as the users of the application respond to each other and learn how to interpret the visualisation based on their knowledge of each other. Interpretations that come out of familiarity and extended use of an ambient display have been seen before Brewer *et al.* (2007), and certainly this emerging behaviour is an important aspect of these types of applications and how people might make use of sksocial signals in their everyday lives over time.

Participants' awareness of their partner watching the visualisation provided motivation for participants to perform amusing actions but also led participants to perform extremely subtle actions and simply allow or encourage their fellow participant to imagine more entertaining actions. Pairs of participants had varying degrees of enjoyment imagining the performance of their fellow participant, with the two couple pairs being the most imaginative. Even when both participants performed subtle actions in the outdoor settings, both participants enjoyed imagining amusing performances. Although these imaginings provided some motivation to perform amusing actions, these participants were still highly aware

of the co-located spectators, or passersby. In some cases, participants modified their performance when outside. For example, one participant used singing input while inside and conversational speech while outside. Both of these actions generated similar output in the visualisation, but participants used these different kinds of actions in order to maintain their comfort, experience, and enjoyment of the application. These adjustments show how considerations for both audiences must be balanced while using this application in public contexts.

6 Conclusion

The user study presented in this paper explored how participants generated and interpreted basic social actions and behaviour in real world settings. This involved using a remote awareness visualisation on a mobile device that could be controlled with gesture or voice based input. During the study, participants demonstrated three methods for generating multimodal output for the visualisation. Participants used highly performative actions, hidden or subtle actions, and simply functional actions when generating input for the system. Because the system supported both extravagant and subtle input equally, participants could perform a wide variety of actions as input and adjust these actions fluidly based on their current context. The variety of possible actions and the purposeful ambiguity in the visualisation also meant that participants could interpret the visualisation in many different ways, incorporating their knowledge of their fellow participant's personality, current context, and previous actions and inputs. These results demonstrate an interesting scenario for making use of basic social signals as part of a remote awareness application.

References

Benford, S., Greenhalgh, C., Giannachi, G., Walker, B., Marshall, J., Rodden, T.: Uncomfortable interactions. In: Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems, CHI 2012, pp. 2005–2014. ACM, New York (2012)

Brewer, J., Williams, A., Dourish, P.: A handle on what's going on: combining tangible interfaces and ambient displays for collaborative groups. In: Proceedings of the 1st International Conference on Tangible and Embedded Interaction, TEI 2007, pp. 3–10. ACM, New York (2007)

Crossan, A., Murray-Smith, R., Brewster, S., Kelly, J., Musizza, B.: Gait phase effects in mobile interaction. In: CHI 2005 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2005, pp. 1312–1315. ACM, New York (2005)

Crossan, A., Williamson, J., Brewster, S., Murray-Smith, R.: Wrist rotation for interaction in mobile contexts. In: Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services, MobileHCI 2008, pp. 435–438. ACM, New York (2008)

Dalsgaard, P., Hansen, L.K.: Performing perception—staging aesthetics of interaction. ACM Trans. Comput.-Hum. Interact. 15, 13:1–13:33 (2008)

Dey, A.K., de Guzman, E.: From awareness to connectedness: the design and deployment of presence displays. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2006, pp. 899–908. ACM, New York (2006)

Goffman, E.: Behavior in public places: notes on the social organization of gatherings. Free press paperback. Free Press (1966)

Goffman, E.: The presentation of self in everyday life. Penguin Psychology. Penguin (1990)

Hoggan, E., Brewster, S.A.: Crosstrainer: testing the use of multimodal interfaces in situ. In: Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, pp. 333–342. ACM, New York (2010)

Jacucci, G.: Interaction as Performance. Cases of configuring physical interfaces in mixed media. PhD thesis, University of Oulu (2004)

Jones, E., Alexander, J., Andreou, A., Irani, P., Subramanian, S.: Gestext: accelerometer-based gestural text-entry systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2010, pp. 2173–2182. ACM, New York (2010)

Jones, M.: Journeying toward extravagant, expressive, place-based computing. Interactions 18, 26–31 (2011)

Perry, M., Beckett, S., O'Hara, K., Subramanian, S.: Wavewindow: public, performative gestural interaction. In: ACM International Conference on Interactive Tabletops and Surfaces, ITS 2010, pp. 109–112. ACM, New York (2010)

Reeves, S., Benford, S., O'Malley, C., Fraser, M.: Designing the spectator experience. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI 2005, pp. 741–750. ACM, New York (2005)

Scheible, J., Ojala, T., Coulton, P.: Mobitoss: a novel gesture based interface for creating and sharing mobile multimedia art on large public displays. In: Proceedings of the 16th ACM International Conference on Multimedia, MM 2008, pp. 957–960. ACM, New York (2008)

Sheridan, J., Bryan-Kinns, N., Reeves, S., Marshall, J., Lane, G.: Graffito: crowd-based performative interaction at festivals. In: Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA 2011, pp. 1129–1134. ACM, New York (2011)

Strachan, S., Murray-Smith, R., O'Modhrain, S.: Bodyspace: inferring body pose for natural control of a music player. In: CHI 2007 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2007, pp. 2001–2006. ACM, New York (2007)

Vinciarelli, A., Murray-Smith, R., Bourlard, H.: Mobile social signal processing: vision and research issues. In: Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services, MobileHCI 2010, pp. 513–516. ACM, New York (2010)

Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image Vision Comput. 27(12), 1743–1759 (2009)

Negotiation Models for Mobile Tactile Interaction

Dari Trendafilov[1,2], Saija Lemmelä[1], and Roderick Murray-Smith[2]

[1] Nokia Research Center, Helsinki, Finland
[2] University of Glasgow, Scotland, UK
{Dari.Trendafilov,Saija.Lemmela}@nokia.com, rod@dcs.gla.ac.uk

Abstract. With the recent introduction of mass-market mobile phones with touch-sensitive displays, location, bearing and motion sensing, we are on the cusp of significant progress in a highly interactive mobile social networking. We propose that such systems must work in various contexts, levels of uncertainties and utilize different types of human senses. In order to explore the feasibility of such a system we describe an experiment with a multimodal implementation which allows users to engage in a continuous interaction with each other by using capacitive touch input, visual and/or vibro-tactile feedback and perform a goal-oriented collaborative task of target acquisition. Initial user study found the approach to be interesting and engaging despite the constraints imposed by the interaction method.

Keywords: Mobile social interaction, multi-modal, mobile, touch, tactile, human communication.

1 Introduction

With the recent introduction of mass-market mobile phones with touch-sensitive displays, location, bearing and motion sensing, we are on the cusp of significant progress in a highly interactive mobile social networking [5]. Strachan and Murray-Smith [13] described bearing-based interaction with content and services, and in linked work Robinson et al. [11] describe its use for GeoBlogging. It is also an obvious step to couple this with social networking applications, where users can probe and point at and engage with nearby friends [12]. The richness of the sensing, and the context-sensitivity and person-specific nature of such communications suggest that designers should beware of implementing overly prescriptive mechanisms for allowing individuals to interact in such systems. Currently capacitive touch displays are primarily used for human-computer interaction, i.e. for navigating through the user interface of the device, however they have the potential to be used for a much more exciting range of interaction styles including dynamic human-human interaction. We present in this paper such an interaction style and explore its potential also for eyes-free mobile context. We introduce and describe a system, with an initial user study, which examines the interaction between human users using embodied touch-based interaction, exploring whether it is possible for users to track each other and locate objects in the virtual environment with realistic sensing conditions.

R. Murray-Smith (Ed.): MSSP 2010, LNCS 8045, pp. 64–73, 2014.

2 Mobile Social Interaction

Mobile Social Interaction enables users to interact with each other via a hybrid physical/virtual environment using their mobile device. Users can remotely touch each other in a mediated environment and in an eyes-free manner and scan the space for objects using finger touch.

A fluid and unrestricted collaboration between two or more people connected remotely via a computer has long been a goal in the fields of virtual and augmented reality. Collaborative Virtual Environments [1] enable a sense of shared space and physical presence in the virtual world. The increasing power and ubiquity of continually connected and continuously sensing mobile devices enables us to generalise down to the mobile realm with the development of Mobile Collaborative Virtual Environment (MCVE). In this paper we present an approach enabling the creation of a hybrid 'eyes-free' physical/virtual world in which users can interact using their mobile device as a probe for objects or for other users, while receiving vibro-tactile feedback dependent on the nature of their probing. A key aspect to the success of this kind of interaction is the provision of a sense of embodiment or presence in the virtual environment. Greenhalgh and Benford [7] tackle this with the DIVE and MASSIVE systems by providing a number of graphical representations of embodied participants. The major advantage that these systems have is that they are highly visual and significant emphasis is placed on the provision of visual feedback to the user. Oakley et al. [10] presented a mechanism for haptic collaboration in synchronous shared editors and a study where haptic communication appeared to facilitate collaboration and improve usability.

One of the major functions of social cognition in humans is to allow the creation of a shared world in which interaction can take place. Communication between two or more people is greatly enhanced by the adoption of a shared vocabulary that enables us to share goals, so that we can engage in joint activity. For successful interactions to take place it is necessary that the interactors achieve the same perception of the world, referred to as 'common ground' [2]. This is slightly easier to achieve in a hybrid virtual/physical environment than in a completely abstracted space. Since an MCVE is located in the real world the user is not completely immersed in the virtual world, they have access to real-world visual cues and some of this natural intuition regarding the interaction with the physical world may be transferred into the virtual world. Espinoza et al. [3] describe their GeoNotes system that allows users to leave virtual messages linked to specific geographical positions. They strive to socially enhance digital space by blurring its boundary with the physical space. However little has been achieved in terms of active interaction or collaboration between two or more users in such environments and it still remains a challenge. The starting point for this kind of active and collaborative interaction is to align the focus of our attention in the environment, typically achieved in real life by pointing at an object or watching bodily movements that can give us some idea about the intentions of the other person [4]. Our bodies are used to provide continuous and fine-grained social cognitive signals about our psychological state, our presence, activity and our attention via gestures, facial expressions or general body posture. It is important that this kind of information is not lost completely in the virtual environment. Lenay et al. [9] show in their studies of perceptual crossing and reciprocal tactile perception how the feeling

of sharing a common space with another intentional being can emerge by switching between two kinds of perception: perceiving the other as part of environment, versus perceiving the activity of other perceiving me. They also present a theoretical framework and models for assisting the conception of tactile communication devices.

Most experimental research on human communication relies on methods that entail the use of pre-established natural or artificial languages, which tap into the processes leading to the emergence of communication systems only indirectly. Galantucci [6], however, described a method introducing the complexity of human behavior into a controlled experimental setting, in the absence of pre-established human communication systems. The scientific understanding of such complex processes would greatly benefit from experiments that elucidate how these systems emerge and develop in the context of joint human activities.

The motivation for our experiment was to explore the feasibility of the presented new interaction method and investigate questions related to performance, cognitive load, user experience and in particular to sense of engagement and togetherness. The future outlook of the proposed method includes 'in-pocket' interaction, where simple tasks could be performed in an eyes-free manner.

3 Experimental System

Imagine the following scenario. Andy is in a meeting room with other colleagues while his friend John is walking on a busy street. They certainly cannot have a phone conversation at the moment, but would like to agree on a specific time for a call. Since they are not aware of each other's schedule they would have to negotiate. Ringing each other up intrusively is not an option; texting extensively is not too convenient either. Instead they could work this out in a more dynamic and fluid manner by probing each other on their shared membrane and agreeing on the time. In this situation one important concern is privacy. They would not like the other one to have full visibility of their schedule, which makes it more complicated than if they would have shared calendars (Fig. 1a). Instead they will have to negotiate a common time slot suitable for both of them without revealing too much private information.

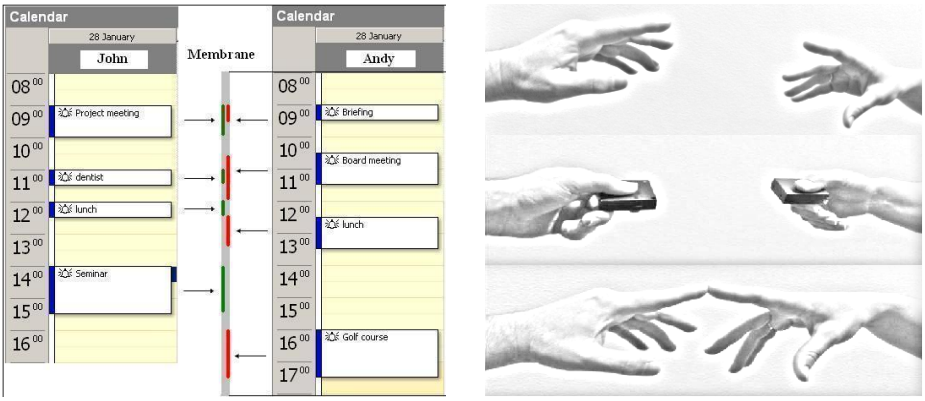


Fig. 1. (a) Membrane concept for diary alignment, (b) Remote touch using SHAKEs

Our interaction mechanism takes the concept of an abstract membrane as a medium to convey the sense of touch. The metaphor of the membrane dynamic system could facilitate and enrich interaction in scenarios as described above. It enables people to touch each other remotely and engage in a continuous dynamic interaction by sliding their fingers and pushing gently on both sides of the membrane. The feedback is visual and tactile, and provides rendering of the internal states of the simulated dynamic system. Allowing users to perceive the changing physical characteristics of the modeled system can be used to convey much richer information about the current state of the person they are interacting with, via continuous interaction and rich feedback, than a static event-based technique would.

The system is intended to illustrate an example of how shared environments can be created with low-latency multimodal feedback and capacitive sensing. It builds on the membrane concept of touch at a distance utilizing mobile tactile devices (Fig. 1b). In this case the membrane has certain number of holes on both sides. The interaction concept includes two users exploring simultaneously the membrane from their side respectively and trying to find a hole through which they could touch each other. In order to explore the effect different modalities have on this kind of interaction we have designed both visual and tactile feedback while trying to keep in mind results in crossmodal interaction research [8]. The main goal of the design was to present the same information in both modalities in a consistent and logical way. We designed the feedback displays so as to allow users to sense each other whenever their fingers meet on the shared membrane and to sense the holes in their side of the membrane whenever they locate one. We display the membrane in a section as a vertical gray-colored stripe (Fig. 2a). The visual representation of the finger is a bell-shaped pointer, while the tactile one is a fast and sharp vibration. The visual shape of a hole is a black square and the tactile one is a slow pulsing vibration.

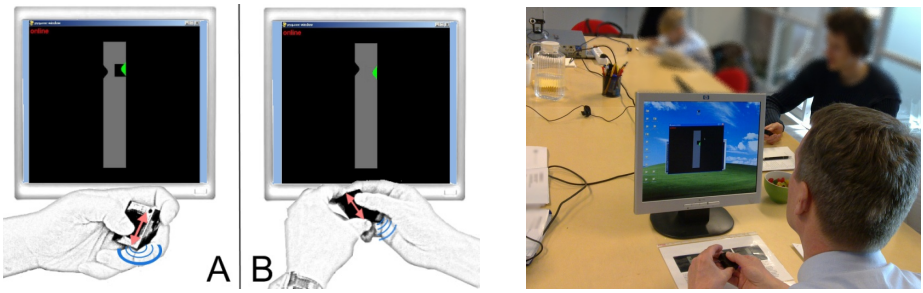


Fig. 2. (a) User A (in green on his display) has found a hole presented by black square and slow pulsing vibration. He can also see and feel his partner (in black) on the other side of the membrane. User B (in green on his display) feels and sees only his partner (in black), since there is no hole on his side of the membrane in this location, (b) Study participants familiarizing with the system.

The system uses capacitive sensing devices for finger touch input, a dual vibro-tactile engine and a visual display. With the input device the user can probe the membrane up and down, in the vertical direction and search for holes and for the remote partner. Holes and the remote partner can be sensed only when the user gets in their close proximity (Fig. 2a), otherwise they are hidden. The user can obtain information only by sensing for impact events, whenever their pointer collides with objects in the shared environment. The task requires users' active exploration of the membrane and locating a hole which is common for both sides. The hole is acquired only when both users locate it simultaneously and hold on still for 0.5 sec, which eliminates incidental acquisitions. Both sides have three holes each, sensible only from their respective side, of which only one is common.

The prototype system consisted of two laptops and two SHAKE SK7 sensor packs. The SHAKE provides 8-bit resolution capacitive sensing from 12 square pads in a 4x3 configuration at 100Hz. It also provides a dual vibro-tactile feedback display - a pager-style vibration motor and a pulsed resonant actuator. The former provides good low frequency while the latter provides excellent high frequency actuation. We use the former for representing the holes and the latter for representing the fingers. The system is implemented in Python and uses WiFi link between the laptops at 100Hz, which in turn are connected to the respective SK7 over Bluetooth.

4 User Study

In this exploratory study our aim was to examine the feasibility of the presented human collaborative task given the restricted modes for communication. We were interested in the limits of these unusual interaction methods in terms of performance and cognitive load and especially in how people cope with the tactile-only system. In our study participants performed a collaborative task in pairs via shared environment, while sitting in separate rooms. Performance depended on their cooperation, which in turn required some sort of communication. Since use of standard communication systems as speaking, writing and body language was prevented, in order to succeed participants were enforced to converge onto a way of using the available resources for interaction.

The experiment consisted of three phases, each 5 minutes long, corresponding to the three different conditions – Visual, Tactile, and Combined (visual&tactile). In each phase the pairs had to complete a set of tasks, where a new task started automatically 5 seconds after the previous task was completed successfully. The location of holes in different tasks was randomized and all pairs had to perform the same list of tasks in the same order, however the order of the three conditions varied for different pairs.

Twenty-six people, 18 males and 8 females participated in the study. Four pairs knew each other well (friends, couple), two pairs were colleagues, and five pairs did not know each other well previously.

Participants were first given an introduction to the system, before being allowed to practice the Combined version for up to 10 minutes, while still sitting in the same room (Fig. 2b) and being allowed and encouraged to discuss.

After completing each phase of the experiment, the participants filled out a section in the questionnaire including extended NASA-TLX, and a set of questions created especially for this experiment. In addition to this they defined 3-5 positive and negative words reflecting the positive and negative aspects of the interaction method they just tried out. At the end the participants answered a set of questions surveying the experience, preferences and game strategies used.

5 Results

The NASA-TLX results revealed that overall workload was significantly higher in Tactile condition ($p<0.001$). This was the trend in Mental ($p<0.001$) and Physical demand ($p<0.05$), Time pressure ($p<0.03$), Frustration and Effort ($p<0.001$). Combined had significantly higher performance level than Tactile ($p<0.03$). The majority of the subjective preferences ranked Combined the highest. Visual was considered unhurried and slow, while Combined was described as more responsive and active. Tactile was linked to words as 'togetherness', 'collaboration' and 'connection'. According to the questionnaires participants believed this technology could encourage communication with people.

The total number of holes acquired by all 13 pairs in different conditions shows a difference between Tactile (32) and the other two, Visual (135) and Combined (122). On the pair level differences between Visual and Combined seem to be due to the execution order, the first one being lower. The learning effect in the visual conditions shows an increase of 50% for most pairs in the subsequent phases. Top scoring pairs had a consistent strategy, executed relatively well in the Visual and Combined and less successfully in Tactile. Even when certain strategies failed to materialize some pairs tried and succeeded in creating new ones that eventually worked.

Some pairs performed surprisingly well in Tactile, even though the scores were lower than in the other conditions, which shows the potential in this approach. Tactile was considered challenging, novel, emphasizing the contact between the partners and the sense of togetherness and collaboration as suggested in [9].

One of the pairs who admitted having a working strategy described it as moving 'together from the top down' (Fig. 3). Fig. 4 shows parts of their time series in more detail. Fig. 3 reveals that this pair did not have clearly defined leader and follower roles. Instead they implemented a sort of turn-taking leadership, which resembles more to a smooth dance than a command-and-control behavior. After the experiment they commented that it was 'interesting', 'fun' and 'surprisingly social experience'.

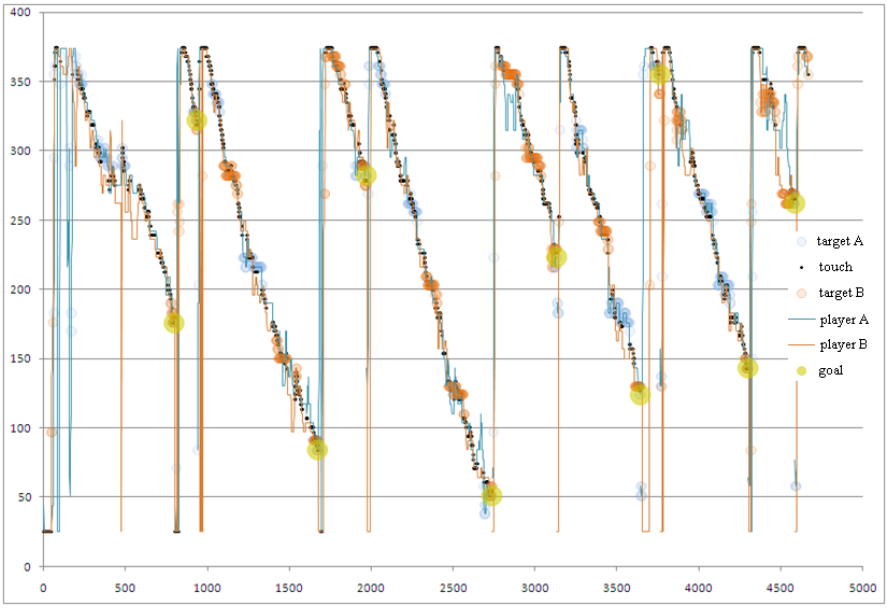


Fig. 3. Time series of a top performing pair

Another pair (Fig. 5), who admittedly did not have a strategy – or as one of the partners put it 'at least not a common one' – achieved only a fraction of the top performers' results. Although they claimed that it was easy to learn the technique and to find the holes and the partner, they found that the most difficult was 'to get the other to move to the same direction'. Since our system provided only limited means for exercising a command-and-control style behavior, this pair was unable to interact successfully (Fig. 6b). They could not literally drag or control, but only perceive each other and exactly this minimalistic kind of interaction was the purpose of our feasibility study.

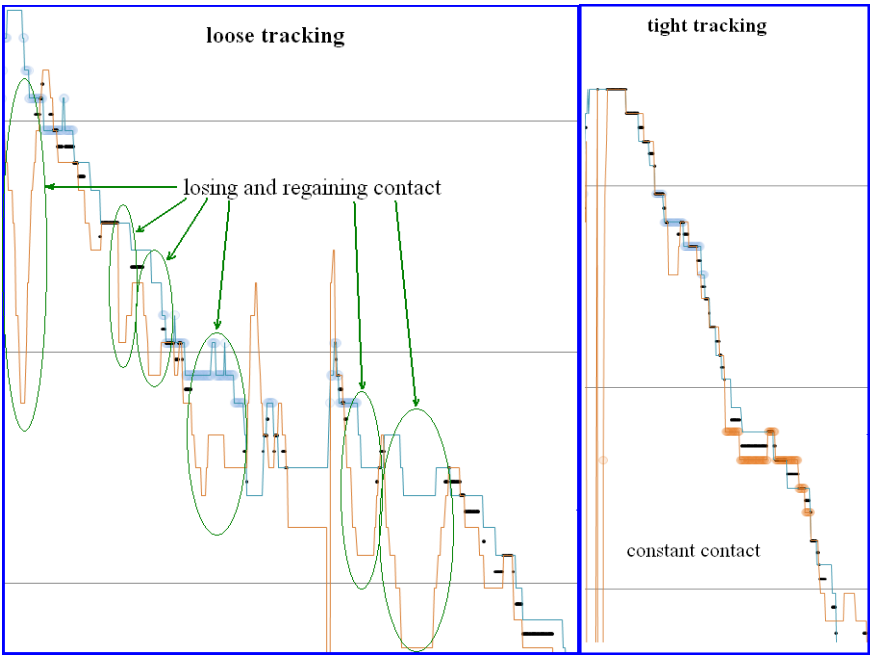


Fig. 4. Consistent strategies - (a) loose tracking, (b) tight tracking

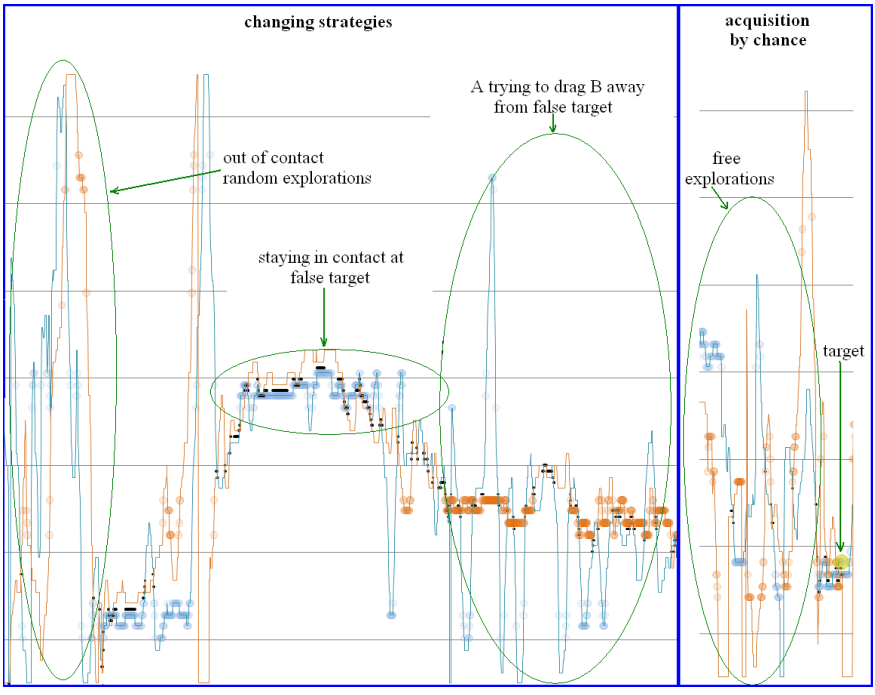


Fig. 5. Random strategies

The phase plot of a successful strategy (Fig. 6a) shows that after the pair managed to 'get in touch' they successfully executed their strategy while staying close (in touch) until they reached the target.

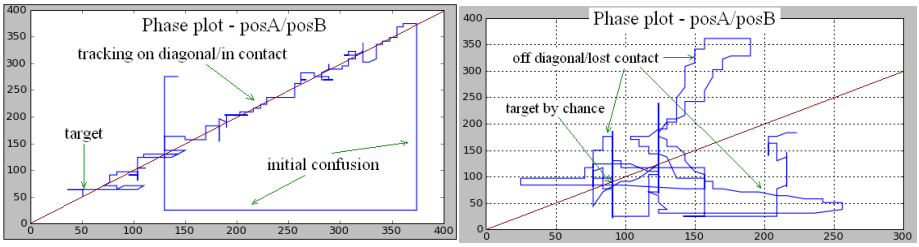


Fig. 6. Phase plots of (a) tight tracking and (b) random strategy.

Further successful strategies are shown in Fig. 7.

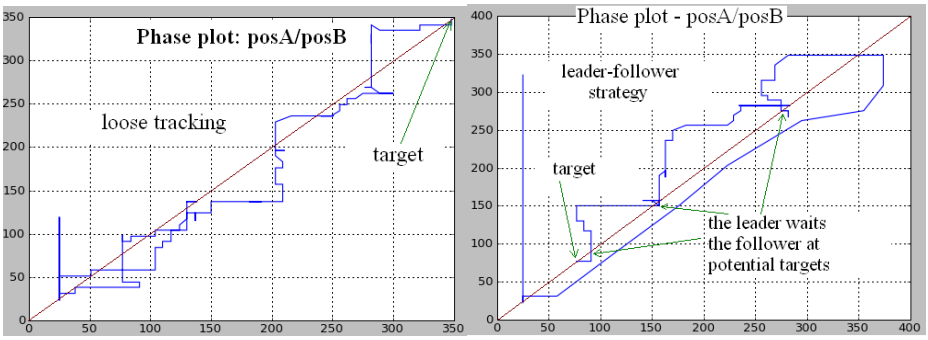


Fig. 7. Phase plots of (a) loose tracking and (b) leader-follower pair

6 Conclusion and Future Work

This paper introduced a new form of embodied social interaction using current technology and demonstrated its potential via an initial user study. It was found that pairs, with limited level of training, succeeded in their tasks to different degrees – the main discriminating factor being the existence or the lack of negotiating strategy. The participants found this new form of interaction interesting and engaging, and believed it could encourage communication with people, which opens new possibilities for the development of richer social interactions. The limited potential of the tactile-only version, shown by significant increase in overall workload and decrease in performance, however brings new research topics for future 'in pocket' interaction studies – namely by incorporating other eyes-free modalities like audio. The experiment in this paper involved human users interacting in pairs, from which an extensive amount of data was collected. One next step is to build models of human

behavior fitting the collected data. Future experiments using simulated agents and human users would give us more control of the activity levels and would improve repeatability. This would give us firmer ground for observing the detailed interactions that evolve as people engage and disengage from remote contact with each other.

Acknowledgments. Our thanks to Suuntaamo and Demola for facilitating the user studies. We are grateful for the support from Janne Bergman, Johan Kildal, Stephen Hughes and Andrew Ramsey.

References

1. Benford, S., Greenhalgh, C., Rodden, T., Pycock, J.: Collaborative virtual environments. Commun. ACM 44(7), 79–85 (2001)
2. Clark, H.: Using Language. Cambridge University Press (1996)
3. Espinoza, F., Persson, P., Sandin, A., Nyström, H., Cacciatore, E., Bylund, M.: *GeoNotes*: Social and navigational aspects of location-based information systems. In: Abowd, G.D., Brumitt, B., Shafer, S. (eds.) UbiComp 2001. LNCS, vol. 2201, pp. 2–17. Springer, Heidelberg (2001)
4. Frith, C.D., Frith, U.: How we predict what other people are going to do. Brain Research, Multiple Perspectives on the Psychological and Neural Bases of Understanding Other People's Behavior 1079, 36–46 (2006)
5. Fröhlich, P., Baillie, L., Simon, R.: Realizing the vision of mobile spatial interaction. Interactions 15(1), 15–18 (2008)
6. Galantucci, B.: An experimental study of the emergence of human communication systems. Cognitive Science 29(5), 737–767 (2005)
7. Greenhalgh, C., Benford, S.: Massive: a distributed virtual reality system incorporating spatial trading. In: Proceedings of the 15th International Conference on Distributed Computing Systems, pp. 27–34 (1995)
8. Hoggan, E., Brewster, S.: Designing audio and tactile crossmodal icons for mobile devices. In: Proceedings of 12th International Conference on Multimodal Interfaces, New York, USA, pp. 162–169 (2007)
9. Lenay, C., Thouvenin, I., Guénand, A., Gapenne, O., Stewart, J., Maillet, B.: Designing the ground for pleasurable experience. In: Conference on Designing Pleasurable Products and Interfaces, pp. 35–58 (2007)
10. Oakley, I., Brewster, S., Gray, P.: Can you feel the force? An investigation of haptic collaboration in shared editors. In: Proceedings of the Eurohaptics, Birmingham, UK (2001)
11. Robinson, S., Eslambolchilar, P., Jones, M.: Point-to-GeoBlog: gestures and sensors to support user generated content creation. In: 10th International Conference on Human Computer Interaction with Mobile Devices and Services, MobileHCI 2008, pp. 197–206. ACM, New York (2008)
12. Strachan, S., Murray-Smith, R.: Geopoke: Rotational mechanical systems metaphor for embodied geosocial interaction. In: Proceedings of the Fifth Nordic Conference on Human-computer Interaction, pp. 543–546. ACM, New York (2008)
13. Strachan, S., Murray-Smith, R.: Bearing-based selection in mobile spatial interaction. Personal and Ubiquitous Computing 13(4) (2009)

Direct Tactile Coupling of Mobile Phones with the FEELABUZZ System

René Tünnermann, Christian Leichsenring, and Thomas Hermann

Bielefeld University, CITEC, Bielefeld, Germany
{rtuenner,cmertes,thermann}@techfak.uni-bielefeld.de
http://feelabuzz.org/

Abstract. Touch can convey emotions on a very direct level. We propose FEELABUZZ, a system implementing a remote touch connection using standard mobile phone hardware. Accelerometer data is mapped to vibration strength on two smartphones connected via the Internet. This is done using direct mapping techniques, without any abstraction of the acceleration signal. By this, FEELABUZZ can be used for implicit context communication, i. e. the background monitoring of the natural movements of the users themselves or their environments, as well as for direct communication, i. e. voluntary and symbolic signalling through this new channel.

We describe the system and its implementation, discuss its possible implications and verify the system's ability to recognizably transmit different actions in a preliminary user study.

Keywords: mobile devices, wearable computing, haptic display, tactile feedback, mediated communication.

1 Introduction

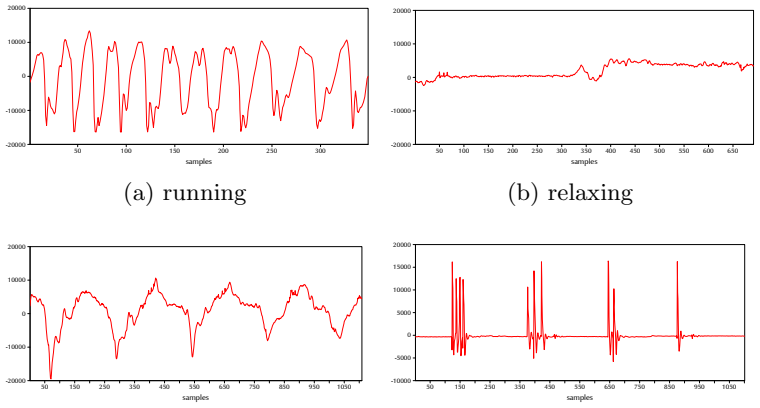
Touch is arguably the most immediate, the most affective, and – when it comes to media – one of the most neglected modalities used for human communication. It can convey emotions and feelings on a direct and primordial level [5,10,18].

We propose FEELABUZZ – a system to directly transform one user's motion into the vibrotactile output of another, typically remote device. Unlike previous work on tactile communication [3] we do so using only mobile phones without any additional gear. This is possible because mobile phones these days almost universally have accelerometers as well as vibration motors which can be used for the sensing of movement and vibrotactile actuation respectively. Mobile phones have the key advantages of not only being widespread to the point of omnipresence but also to usually be worn on the user's body. Furthermore, not having to buy and more importantly to carry around an extra piece of hardware is a property whose importance cannot be overstated. Using phones also makes it easy to integrate the new haptic channel with existing auditory, visual and maybe textual channels, thereby extending the phone's capabilities as a communication device. As we have our phones with us or nearby most of the time, they are well suited

R. Murray-Smith (Ed.): MSSP 2010, LNCS 8045, pp. 74–83, 2014.

not only for *direct communication* but also for *implicit context communication* (e. g. walking or riding the bus; cf. Section 3 for a more detailed discussion of these concepts). Being able to assess a contact's current context could equally be important when it comes to determining a good time to call.

The choice of vibration as an output modality not merely stems from its prevalence on the chosen platform and its availability and unobtrusiveness when carrying the phone in a pocket but also from the fact that movement such as impacts or strokes naturally transforms into tangible vibration in the real world (e. g. footsteps on the floor, multiple persons using one stair rail, someone stirring on a sofa or even the feedback to one's own hand when stroking something).



(a) running (b) relaxing

(c) swinging (d) hand tapping on the phone

Fig. 1. Accelerometer data of different movements recorded at 100 Hz

2 Related Work

Heikkinen et al. [10] provide insights on the expectations of users regarding haptic interaction with mobile devices. Their results underline our design considerations. The participants brought up poking and knocking metaphors as well as the idea of a constantly open "hotline" between two participants. Their participants even saw the possibility of the emergence of a haptic symbolism or primitive language, which have been developed during the evolution of the interaction.

O'Brien and Mueller [15] created special devices of various forms to examine the needs of couples when "holding hands over a distance". A main critique of their participants was concerned with the cumbersome and unfashionable design of their devices: "The participants stressed how they wanted a device that was more personal and easy to carry. They desired it to be small enough to fit it in their pocket. One participant noted that she wanted something she could relate to personally" [15]. Furthermore, their users disliked that the special device draw to much public attention.

Eichhorn et al. build a pair of stroking devices for separated couples. Each device has a sensor and a servo which expresses the stroke initiated by the

remote device. The device functions as a proxy object to stroke each other over a distance.

A lot of the work already conducted on vibrotactile interaction is focused either on the recognition of haptic gestures or on mapping different cues to haptic stimuli [14,16,2,4,6].

With FEELABUZZ we aim at creating a personal, lightweight and always ready-to-hand haptic communication channel. An earlier prototype of the system has already been presented [12]. In this work we will first discuss aspects of haptic communication, introduce the new FEELABUZZ system and then present and discuss the results of the informal user study.

3 Concepts

The information conveyed by FEELABUZZ can be split into two parts that we call *implicit context communication* and *direct communication*.

3.1 Implicit Context Communication

The most obvious kind of information that is conveyed by FEELABUZZ are the unintentional and implicit movements of the device. These can either originate from the users or from the environment, as already proposed by Murray-Smith et al. [14]. The time-series data in Figure 1 show that different kinds of activities by the users themselves lead to very different acceleration profiles. Likewise, sitting in a driving vehicle will lead to an acceleration pattern that is notably different from those caused by human movements. Note that none of this has to be detected by pattern recognition software. There are no predefined classes. Instead, the interpretation of many movement patterns is expected to come quite naturally and involve all the rich context information and world knowledge humans have. Additionally, the sophistication of the interpretations can fluently increase with the user experience. As there are rarely clear class boundaries in the real world, transitions between different types of movement can be perceived in all their ambiguity and fuzziness in a near-analogue fashion without the need to make clear distinctions. While regression models could do so as well, the subsequent mapping back to artificial vibrotactile stimuli in a way that allows direct access as well as in-depth learning of subtle features would be a major challenge to say the least. Actually one would have to know and reliably detect any such subtlety in advance before playing it back to a user in an alienated way. Relying on the human's long-evolved ability to interpret rich real-world data streams seems to be a more promising way in terms of effectiveness and a much more interesting way in terms of unintended uses and exploration by future users.

3.2 Direct Communication

Providing people with the possibility to intuitively induce tactile feedback in another person's mobile phone presents a new communication channel that can also

be used deliberately in a number of ways. The channel's possibilities for readily understood signals are limited though. Apart from *knocking* to do simple things such as requesting attention, synchronizing or timing pre-decided behavior, or giving short binary feedback, few intentional tactile communication events will be understood by the naïve user. Although there are sophisticated means of communication through such narrow channels, most notably Morse code, we expect that to be employed only by experts and not to become widespread. Instead, we rely on people's ability to develop their own adapted communication strategies using a mixture of implicit and explicit negotiation. Quite complex and effective communication systems can emerge via such mechanisms [7,9,8,17,11,1].

The general lack of interpretation and abstraction on the side of the system enables users to become creative in that they use the system in ways that were not intended by the system designer. It will be an interesting area of future research to see if and how people start to use FEELABUZZ in ways that fall under the definition of *direct communication*.

4 Implementation

4.1 Technology

The FEELABUZZ prototype hardware, which was used for the evaluation, consists basically of two *Palm Pre* mobile phones. On the phones we gather the accelerometer data which is then preprocessed, transmitted and mapped to the vibrotactile actuator of the other phone. The data is transmitted over two direct *Open Sound Control* (OSC) connections [19] between the paired devices. The OSC connection is run over a wireless network connection. OSC is a UDP-based simple push protocol which is widely available in common programming languages. On the device itself we are using the Python programming language to preprocess the sensor data, to connect the devices over the network and eventually to excite the vibration motor.

4.2 Signal Processing and Vibrotactile Mapping

To map the S accelerometer readings $\mathbf{s}(t)$ with $s_i(t) \in [0, s_{max}], 1 \leq i \leq S$ to the vibration module input value $y(t) \in [0, y_{max}]$ we perform a couple of steps.[1] First we compute the magnitude of the acceleration vector:

$$m(t) = \rho \|\mathbf{s}(t)\| = \rho \sqrt{\sum_{i=1}^{S} s_i(t)^2} \tag{1}$$

with ρ being a normalization factor:

$$\rho = \frac{y_{max}}{\sqrt{S s_{max}^2}} \tag{2}$$

[1] For the Palm Pre, our prototype hardware, the number of sensors S is 3, s_{max} is 2 and $y_{max} = 100$. The sensor sampling rate was set to 30 Hz.

Now an RC high-pass filter is applied to the sensor values with the decay constant $\alpha_h = 0.967$

$$b_h(t) = \alpha_h\Big(b_h(t-1) + \big(m(t) - m(t-1)\big)\Big) \tag{3}$$

which gets rid of the gravitational acceleration and other constant or long-term acceleration influences[2] without losing as much inertia as a simple derivation would. Subsequently, an exponential smoothing is applied with smoothing factor $\alpha_l = 0.157$:

$$b_l(t) = \alpha_l |b_h(t)| + (1-\alpha_l) b_l(t-1) \tag{4}$$

This is important to give more inertia to the system in a controlled way so that a lot of activity from the sender will add up to give an increasingly strong signal on the receiving end (cf. Figure 2). This turned out to be what best matched our intuitive a-priori expectations of how the system *should* behave.

It has the drawback of levelling out all of the more impulse-like parts of the signal which are a salient feature and also quite important for signalling. To preserve these impulse components as well, we add them back in with a simple kind of spike detection. This also has the benefit of making the system more responsive to quick accelerations as the then-detected spike will kick-start the acceleration motor.

For this we compute the moving average over the last n time steps, defined for any function $x(t)$ as

$$MA_n(x,t) = \frac{1}{n}\sum_{i=0}^{n-1} x(t-i) \tag{5}$$

and check if the high-pass-filtered signal $b_h(t)$ exceeds a certain threshold of $\beta_a = 5$ times the moving average. If this is the case we perform an exponential mapping of the spike signal and add it back to the low-pass-filtered signal with the adjusting coefficients $\beta_{b_h} = 2$ and $\beta_{b_l} = 3$:

$$k(t) = \begin{cases} y_{max}\left(\frac{\beta_{b_h} b_h(t)}{y_{max}}\right)^{\alpha_e} & \text{if } b_h(t) > \beta_a MA_n(b_h,t), \\ 0 & \text{else.} \end{cases} \tag{6}$$

$$y(t) = \min\Big(\eta\big(k(t) + \beta_{b_l} b_l(t)\big), y_{max}\Big) \tag{7}$$

with $n = 5$ and $\alpha_e = 0.4$. The normalization constant η is necessary on some platforms to linearly correct for sensor or actuator sensitivities that are too low.

[2] When using a sample rate of 30 Hz it is possible to shake the phone so hard that the accelerometers will register a constant acceleration. In an earlier prototype [12] the accelerometers were capable of 100 Hz which was enough to circumvent this phenomenon. To prevent the high-pass filter from eliminating the constant maximum acceleration on platforms that cannot read from the sensors fast enough, it turned out to be inelegant yet appallingly effective to artificially set the sensor value to 0 when a threshold number of successive near-maximum acceleration frames is exceeded.

For the Palm Pre we found a value of $\eta = 2.5$ to work well. Finally, the output is cropped to y_{max}.

Figure 2 shows the behaviour of these steps combined. A burst of delta pulses increasingly excites the system and this excitation takes a comparatively long time to wear off. At the same time, the pulses themselves are perfectly preserved and amplified.

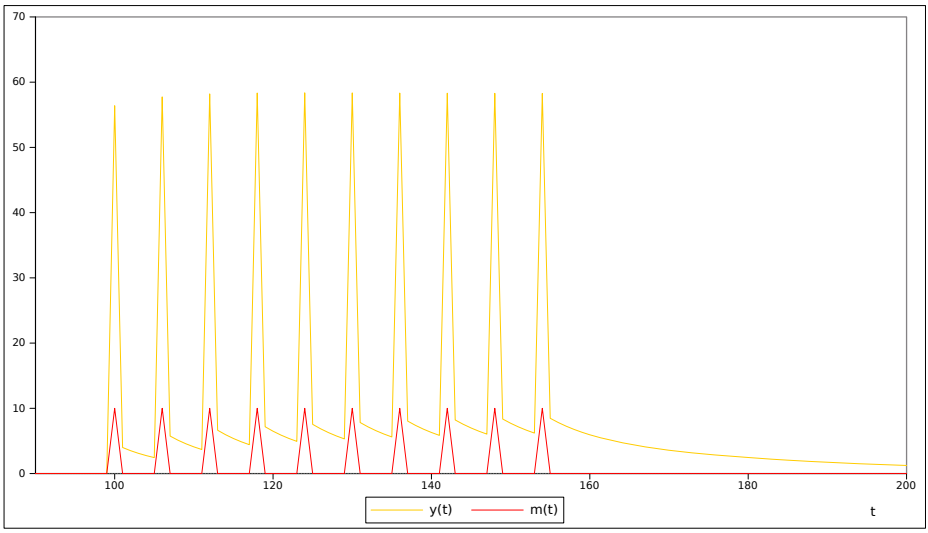


Fig. 2. Filter response $y(t)$ to a burst of delta pulses $m(t)$

5 Evaluation

5.1 Method

To verify that basic activity types can be distinguished with FEELABUZZ we did a study with 10 participants, 5 male and 5 female. The participants went through the study in pairs who were known to each other. Accordingly there were two phones running FEELABUZZ that were bidirectionally transmitting the acceleration data. As the first step of each trial, the general idea and basic properties of the acceleration-vibration mapping were explained to the participants. Each participant was then given the opportunity to familiarize him- or herself with both phones at the same time to get a better first impression of the mapping. When they both felt familiar with the system, they split up the phones so that both participants had one of them. They were again asked to explore the system until feeling familiar with it. They were then explained the following procedure.

The two participants were separated so that they could no longer see or hear each other. One of them was asked to perform one of three activities while wearing the telephone in their pocket: resting, walking or running. The other participant was instructed to guess which of these activities was being performed, holding the telephone in their hand. This step was repeated ten times before the roles were switched between the two participants. The schedule of activities each participant had to perform was randomly generated in advance and different for each participant.

Finally, the participants were asked to fill in a questionnaire. The questionnaire we used is based on the Computer System Usability Questionnaire (CSUQ) by Lewis [13]. We removed or adapted questions that did not make sense in our scenario and ended up with 12 multiple-choice questions using a 7-point Likert scale. We also added six free-response questions.

5.2 Results

The results of the activity classification can be seen in Table 1 as a confusion matrix. All four misclassifications occurred between the classes "running" and "walking" and only when a participant was first confronted with one of these activities.

Figure 3 shows the responses to four of the questions as histograms. The most favourably answered items were "It was simple to use this system." and "It was easy to learn to use this system.", both of which were "strongly agree"d upon by all participants (average 1.0). The items that scored worst were "I believe I would use this system on a regular basis." with an average of 3.7 (cf. Figure 3) and "This system has all the functions and capabilities I expect it to have." with an average of 3.714.

Table 1. Confusion matrix of the participants' activity recognition using FEELABUZZ

		actual activity		
		resting	walking	running
guess	resting	**35**	0	0
	walking	0	**29**	2
	running	0	2	**32**

6 Discussion

The classification results show that it is possible to distinguish different activities using only the FEELABUZZ system. Although it was a task that was fairly easy to solve, the practically perfect performance of all participants is very encouraging. In addition, most users liked using the system (cf. Figure 3). Future studies with more complex and more diverse activities will have to show whether the level of recognition of simple activities holds or if it gets degraded when the users move

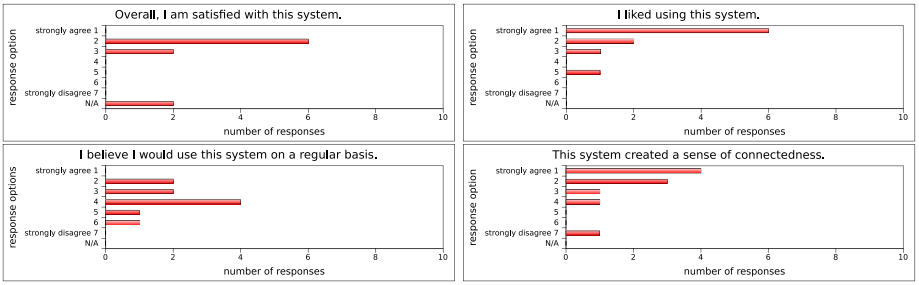


Fig. 3. Responses to four of the questions in the questionnaire. The average values for these questions from left to right and top to bottom are 2.25, 1.8, 3.7 and 2.4.

out of this narrow domain. Even more interesting, though, is the question of whether users will actually accept such a system, how they will use it and what they gain from it emotionally. Longitudinal studies in actual relationships will have to show this but there are some hints already that can be taken from this basic study. Figure 3 shows that the participants had a feeling of *connectedness* to a varying degree. They were more divided, though, on their assessment of whether they would use FEELABUZZ on a regular basis at all. In the free-response questions participants emphasized the aspects *ease of use* and *learnability* that also showed up clearly in the multiple-choice part of the questionnaire. They noted for example that the system was "easy to use", "uncomplicated" and "easy to understand". One participant noted to have "liked the buzzing, it's smooth". Another user mentioned that it was "possible to submit actions without actively operating the device". Some participants found it unlikely to constantly use the system at all time though. One user commented that observation: "I cannot imagine to use it all the time. But it could be handy for those 'what is XY doing right now?' moments."

This feedback to us suggests that there is potential for an emotionally significant connection of people with FEELABUZZ but the right mode of operation regarding the individual timing of the vibration output and the control thereof will be a delicate part of the application design and further investigations.

7 Conclusion

We presented the concept and a prototype of a near-analogue coupling of the accelerometers built into modern mobile phones to the likewise included vibration motors of a remote device to create a feeling of connectedness over a distance. We described a mapping to transmit such acceleration data and implemented it for a pair of *Palm Pre* phones. Furthermore we reported the results of an informal users study. The study showed that users are able to sense if the other person is resting, walking or running just by feeling the activation of the vibration motor.

Our future work is focused on how to run FEELABUZZ on many users' own phones by providing an improved application for download. This will not only

make it possible to put future evaluations of our method on a broad basis but also to collect experiences with haptic communication channels in general with a handy device to which the subjects can personally relate and which accompanies them in their daily life.

Acknowledgements. We thank the DFG and the Center of Excellence for Cognitive Interaction Technology who funded this work within the German Excellence Initiative. We also thank Sebastian Hammerl for porting our prototype system from the Openmoko to the Palm webOS platform.

References

1. Bickerton, D.: Roots of Language. Karoma (1981)
2. Brewster, S., Brown, L.M.: Tactons: Structured tactile messages for non-visual information display. In: AUIC 2004: Proceedings of the Fifth Conference on Australasian User Interface, pp. 15–23. Australian Computer Society, Inc., Darlinghurst (2004)
3. Chang, A., O'Modhrain, S., Jacob, R., Gunther, E., Ishii, H.: Comtouch: Design of a vibrotactile communication device. In: DIS 2002: Proceedings of the 4th Conference on Designing Interactive Systems, pp. 312–320. ACM, New York (2002)
4. Deepa, M.: vsmileys: Imaging emotions through vibration patterns (2005)
5. Eichhorn, E., Wettach, R., Hornecker, E.: A stroking device for spatially separated couples. In: MobileHCI 2008: Proceedings of the 10th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 303–306. ACM, New York (2008)
6. Enriquez, M.J., MacLean, K.E.: The hapticon editor: A tool in support of haptic communication research. In: International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, p. 356 (2003)
7. Galantucci, B.: An experimental study of the emergence of human communication systems. Cognitive Science: A Multidisciplinary Journal 29(5), 737–767 (2005)
8. Goldin-Meadow, S., Mylander, C.: Spontaneous sign systems created by deaf children in two cultures. Nature 391(6664), 279–280 (1998)
9. Healey, P., Swoboda, N., Umata, I., King, J.: Graphical language games: Interactional constraints on representational form. Cognitive Science: A Multidisciplinary Journal 31(2), 285–309 (2007)
10. Heikkinen, J., Olsson, T., Väänänen-Vainio-Mattila, K.: Expectations for user experience in haptic communication with mobile devices. In: MobileHCI 2009: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 1–10. ACM, New York (2009)
11. Kegl, J.: The nicaraguan sign language project: An overview. Signpost 7(1), 24–31 (1994)
12. Leichsenring, C., Tünnermann, R., Hermann, T.: feelabuzz – direct tactile communication with mobile phones. International Journal of Mobile Human Computer Interaction 3(1) (2011)
13. Lewis, J.: Ibm computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. International Journal of Human-Computer Interaction 7(1), 57–78 (1995)

14. Murray-Smith, R., Ramsay, A., Garrod, S., Jackson, M., Musizza, B.: Gait alignment in mobile phone conversations. In: MobileHCI 2007: Proceedings of the 9th International Conference on Human-Computer Interaction with Mobile Devices and Services, pp. 214–221. ACM, New York (2007)
15. O'Brien, S., Mueller, F.F.: Holding hands over a distance: Technology probes in an intimate, mobile context. In: OZCHI 2006: Proceedings of the 18th Australia Conference on Computer-Human Interaction, pp. 293–296. ACM, New York (2006)
16. Rovers, A., van Essen, H.: Him: A framework for haptic instant messaging. In: CHI 2004 Extended Abstracts on Human Factors in Computing Systems, pp. 1313–1316. ACM, New York (2004)
17. Senghas, A., Kita, S., Ozyurek, A.: Children creating core properties of language: Evidence from an emerging sign language in nicaragua. Science 305(5691), 1779 (2004)
18. Vetere, F., Gibbs, M.R., Kjeldskov, J., Howard, S., Mueller, F.F., Pedell, S., Mecoles, K., Bunyan, M.: Mediating intimacy: Designing technologies to support strong-tie relationships. In: CHI 2005: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 471–480. ACM, New York (2005)
19. Wright, M., Freed, A.: Open Sound Control: A new protocol for communicating with sound synthesizers (1997)

A Multimodal Contact List to Enhance Remote Communication

Andrew Crossan[1], Grégoire Lefebvre[2], Sophie Zijp-Rouzier[2], and Roderick Murray-Smith[3]

[1] School of Engineering and Built Environment, Glasgow Caledonian University, Glasgow, Scotland
andrew.crossan@gcal.ac.uk
[2] Orange Labs, Meylan, France
{gregoire.lefebvre,sophie.zijprouzier}@orange.com
[3] School of Computing Science University of Glasgow, Glasgow, Scotland
roderick.murray-smith@glasgow.ac.uk

Abstract. The Multimodal Contact List provides a mechanism to browse context information and communicate with friends in a contact list both visually and through touch and sound. Each contact can share with their friend group selected information on their current context such as mood and availability. Users are able to gain a quick overview of the context provided over all of their contacts, allowing them to close the loop with them via touch, audio, visual feedback or a combination of all three. A user can then progressively probe the contact for more detailed information, eventually allowing the user to open a real-time multimodal voice and tactile communication channel to the contact for verbal or discreet tactile communication.

Keywords: Multimodal, Mobile, Remote Communication, Vibrotactile Feedback.

1 Introduction

Mobile devices provide the ability to stay connected and engaged with our friends, family and work colleagues in a wide range of contexts. They can now be seen as small, general-purpose, portable computers, but the ability to communicate with contacts has remained one of their most important and compelling features. Some of this connectedness now occurs through online social networking such as Facebook and Twitter, as well as the more traditional voice and SMS communication. The communication offered by these devices through voice or text however is in some ways limited in expressiveness due to its inherently remote and distant nature. Messaging systems attempt to get around this lack of expressiveness through the introduction of emoticons. Similarly, in remote voice communication the non-verbal cues from the context as well as contextual cues are not present. Humans are extremely good at reading these non-verbal cues in face-to-face conversations that can

R. Murray-Smith (Ed.): MSSP 2010, LNCS 8045, pp. 84–100, 2014.

guide the tone and quantity of communication as well as affecting the meaning. King [10] argues that context is fundamental to communication. In mobile communications, we lose a lot of the context information that we take for granted in everyday conversations. Partially, this relates to the loss of non-verbal cues during a conversation. However, King's definition of context includes factors such as psychological, environmental, cultural and situational contexts that can play an equally important role during communication. For example, the quantity and quality of communication might vary dramatically between situations where one of the parties is in a home or work environment, or when external sources of noise might interfere with the communication, or even when the remote person is engaged in some other activity that requires concentration. While these factors might be easy to discern in a shared communication space, important cues are lacking when the two parties are in different locations.

Sellen [15] demonstrates, however, that even in situations where both good quality video and audio are present, remote communication mediated through technology can result in communication that has a tendency to be more formal and less spontaneous suggesting that even with some context information, the remote nature of the communication can have an effect.

1.1 Scenarios

Here we propose a system where a user can choose to display certain context information (such as happiness or availability) to their friends through a multimodal contact list application. The ability to determine whether someone would be available for a phone call without calling is one immediate benefit of sharing context, helping prevent a contact's phone ringing at inappropriate times such as during a meeting. There are also for more subtle situations where privacy or discreetness might be an issue. Similarly, there are many situations where voice communication is impossible (e.g. on a loud subway), inappropriate (e.g. in a meeting or at cinema) or simply not desired (e.g.in public [13]). In these cases, non-verbal communication (touch-based device for example) could provide a private channel for communication. If a user sees a contact is unavailable, he may choose to tap out a message on the screen that the remote contact will feel as vibrotactile message. The length and rhythms of taps being sent might be used to convey some information to the remote contact, allowing expressive, discreet communication with the potential for people to develop simple personal tactile languages.

Shared context also allows users to filter the contact list according to some parameters, for example to find out about people's availability and general mood. So assuming the scenario of a night out, a user may look for friends to go out by filtering for contacts that are happy and not busy. Contacts are then filtered according to these input parameters, thus showing people who are most likely to be interested in going for a night out.

There is also the potential to multiplex the different channels of communication to provide either context or emphasis at different points in the communication. During a phone call, feedback presented through the standard vibrotactile actuator in a phone

could be used to convey the emotional state of a caller or more contextual information for the remote contact in the communication displaying for example whether he or she is walking and how quickly, or whether they are communicating through a headset, or in a car, for example.

The multimodal nature of the communication also offers possibilities for providing an accessible means of communication. Touch-enhanced communication opens another potential remote communication channel for people with visual or hearing impairments or both. The addition of an expressive vibrotactile channel would provide obvious benefits over preset vibration patterns, allowing users to customise their messages and develop more complex communications.

2 Related Work

There are different mechanisms that have previously been used to attempt to convey more context information in communication (see [5]). One common mechanism is the social network 'status update' that allows a user to display a general purpose piece of information to a group of friends. Here, the user manually enters whatever information they choose into the message. With the availability of more sensors in phones, we are starting to see more systems that take advantage of automatically sensed context. Location aware systems would be the most common example where applications such as maps or navigation aids can use the user's location to present context sensitive information. Google's Latitude system now extends this location awareness to incorporate social networking features where friend groups share their locations. Many users are now willing to share more context information with selected friends or colleagues, leading to social features being incorporated into a growing number of applications, allowing users to maintain an awareness of their friends' and families' days, even if they do not get a chance to meet up with them.

The remote communication problem extends to issues of the availability of conversation partners. It is difficult to get an indication of whether someone is free to be contacted or whether current situational and environmental factors will make it difficult or socially inappropriate to respond. Researchers have started to address these issues in a number of ways. In instant messaging systems, they adopt the strategy of allowing users to explicitly state whether they are available or not. This concept has been extended to more general communication in systems such as the BuddyClock [9]. BuddyClock is an augmented alarm clock that allows people to know whether it is appropriate to make a call by alerting a friend group whether a person is awake or asleep. Using this friend group feature, it further incorporates social networking aspects to allow people to share sleeping behaviours with each other.

More recently, in a mobile setting, the Feelabuzz system [18] provides a mechanism for communicating continuous awareness signals between sensor-enabled mobile devices. An accelerometer is used to measure movements and these data are transmitted in real time to a receiver in a direct, non-abstract way, without the use of pattern recognition techniques, in order not to destroy the 'feel'. The goal is that over

time, a user would learn to recognise common vibration patterns such as walking or on a bus. This enables direct communication as well as implicit context communication with display to the receiver through vibrotactile actuators. The real-time aspect of the communication allows the context information to enrich and augment voice communication between remote communicators.

In related work, Murray-Smith et al. [12] examined how to augment voice communication using the vibrotactile channel to indicate to each participant the walking behaviour of the other participant. They were able to demonstrate how this additional channel could affect synchronisation of the step rate between participants with the results being different for spontaneous and scripted communication. Brown and Williamson [3] allow users to send audio/tactile messages to each other using gestures. These gestures can be used to represent common events such as 'home safely' and are translated into preset audio/tactile messages that the remote user can learn to interpret. There have been a number of systems based on asynchronous haptic messaging [1, 4, 6, 14]. Chang et al. [4] have examined enhancing communication through vibration, converting hand pressure on the sender's phone directly to the receiver, mapping pressure to vibrational intensity. Again this real-time tactile channel is used to allow the user to add expressiveness to the communication. They further allowed symbolic communicative function through a tactile interface that is attached to a mobile device and utilizes encoded haptic patterns in communication. The first claimed advantage of adding the tactile channel was redundancy between voice and haptic, for emphasis. The second was the ability to incorporate mimicry of haptic patterns to indicate attention and camaraderie. At the minimum, it represents an additional communication channel to provide a means to incorporate turn taking signals into the conversation. This area is of particular importance for interfaces looking to support intimate interactions.

There have been many examples of systems that aim to support the communication of emotion. As so much is communicated non-verbally when face-to-face, research has looked to replace these non-verbal cues to both communicate more accurately the intention of the speaker, but also to support couples in a relationship. Wang, and Quek describe Touch & Talk where squeezing a customised mobile device causes a remote armband to squeeze the arm of a loved one [19]. This example demonstrates the potential of tactile feedback for affective communications in a remote context.

Similarly, recent work by Hemmert et al. [7] uses unconventional interaction mechanisms to support couples in a long distance relationship. Users interact by grasping, kissing or whispering into a custom device. The sensations are conveyed to the remote user through tactile sensations with a band that contacts around the user's hand communicating a squeeze, air jets communicating a whisper and a semi-permeable membrane and water conveying a kiss. These sensations are designed to be as close a representation of the sensation as technology allows, and as such requires specialist hardware.

Kontaris et al. [11] convey emotion through spatially distributed vibration patterns designed to supplement video communication. The patterns were generated by gesturing on a custom touch surface, with the vibrations being displayed through a 4x4 array of tactile actuators. The authors suggest that custom gestures could eventually form part of a personal tactile language between the couple.

Kaye [8] examines a different way of supporting couples in remote relationships without the need for custom technology. His minimalist approach uses a 1-bit communication channel initiated by clicking on a button. Although minimal information was conveyed, Kaye's study suggests that a rich channel of communication can be built over time by incorporating context. Alternatively, Bales et al. [2] use explicit codes for building tactile messages that support a couple's awareness of their partner's location. By defining recognisable vibrotactile patterns for arrive and depart, they could combine these cues with locations such as home or work to build a tactile message in a language that can be learned by both partners in the relationship.

One common factor in all these systems is that they use touch as an output channel. Touch provides a very personal communication channel that naturally lends itself to supporting affective means of communication.

Another relevant are of research is automated sensing of context. For example, Williamson et al. [20] demonstrate how accelerometers could be used to determine context of a user in a usability study during the morning commute. They were able to determine whether the participants were stationary, walking or using public transport.

We may also choose to automatically sense the mood of the content of a communication. Shirazi and Schmidt [16] suggest how audio alerts shaped by the content of a message can be used to provide non-visual feedback to the user. Their work identified four different types of common messages, happy messages, sad messages, questions and answers and responses. Through identification of these different types of messages using emoticons, punctuation and common words, the system can give the user an audio preview of the message content or mood.

3 The Multimodal Contact List

Here, we augment both the representation of a user in a contact list, as well as communication between the users to provide a mechanism for sharing context information. The goal is to provide a system that allows a user to initially get a brief overview, or glimpse of the status of users within their contact list. Further, the system attempts to provide a way to probe for more detailed information about a contact, and eventually open a multimodal communication channel with that contact. We choose to augment the contact list to display this information as it is an application already used for more traditional forms of communication and the user's contacts are already collected in one place.

The Multimodal Contact List has been implemented for Android Phones (2.0 and above). Figure 1 shows the main interface. The header describes the current situation: the last event status (here "New Mood Vector" message from Gregoire) and the current tactile communication status (i.e. communication enabled/disabled with voice on/off). Each entry in the contact list represents a user and provides some context information about that user visually through a texture and, when explored on screen, through the phone's internal vibrotactile actuator. This context information is stored as a texture with parameters controlled by some aspect of the users' context. These

context parameters can be automatically inferred from an automatic classification of a combination of sensor readings, or could be controlled explicitly by the user. In this example we use the concept of a 'Mood Vector' associated with each user. This vector can be explicitly set by the user to indicate different aspects of their current mood with each parameter affecting the visual appearance and tactile representation of the contact. Currently the four parameters of the mood vector are 'Joy' (to indicate happiness or sadness), 'Aggression' (to indicate calm or stressed states), 'Mobility' (as an indication of their recent motion and mobility characteristics) and 'Busyness' (as an indication of their current workload). Similar parameterizations of mood have previously been used successfully in the widely used Moodagent music player that builds playlists based on the user's mood (moodagent.com). Each parameter can be set explicitly using the 10-point sliders shown in Figure 2. Eventually, the mood vector concept could be extended to inferred mood using sensor information such as phone location or acceleration activity, along with other contextual cues such as calendar entries, current music playlist mood content, or whether the contact is engaged in a call or not.

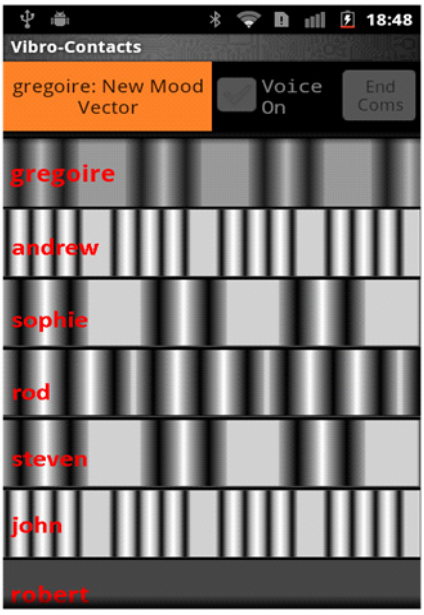


Fig. 1. The main Multimodal Contact List interface. The status bar at the top shows the most recent message received and the current communication status. Below is the list of contacts with mood vector indicated by their texture.

The parameters are then mapped to a visual and tactile texture displayed to the user for each contact when browsing the contact list. The visual texture initially provides a user with a low effort mechanism for gaining a fast overview of the general context information shared by their friends allowing the mood of each contact to be ascertained (for instance the contact named Robert is offline; consequently its visual

texture is black). For more detailed interactions with a particular contact, we use a 'modality scheduling' mechanism [6]. When the user is interested in more detail about one particular contact, they can use a low-attention approach, where they interact through the phone's touchscreen, with increasing engagement with a contact leading to increasingly detailed multimodal feedback, and eventually tightly-coupled direct communication. Interactions with the contact list can be separated into vertical swipes (to move up and down the contact list), horizontal swipes (to feel the texture of a contact's Mood vector), and horizontal exploration (moving along a contact to find out more detailed information and eventually open a communication channel).

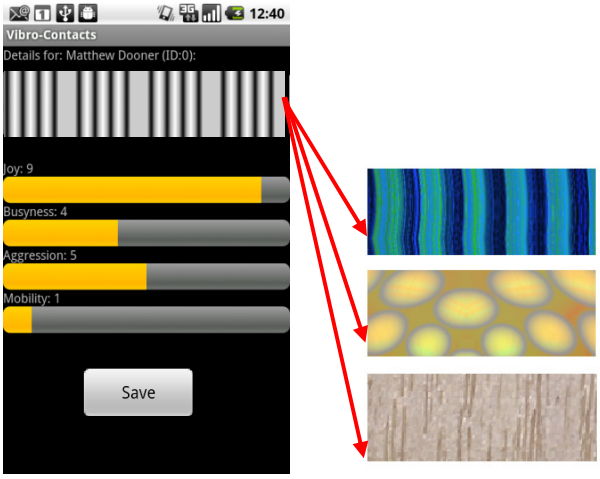


Fig. 2. Users set their Mood Vector parameters between 1 and 10 using the four sliders shown (Left). There are many ways this vector can be mapped to an arbitrary texture (right).

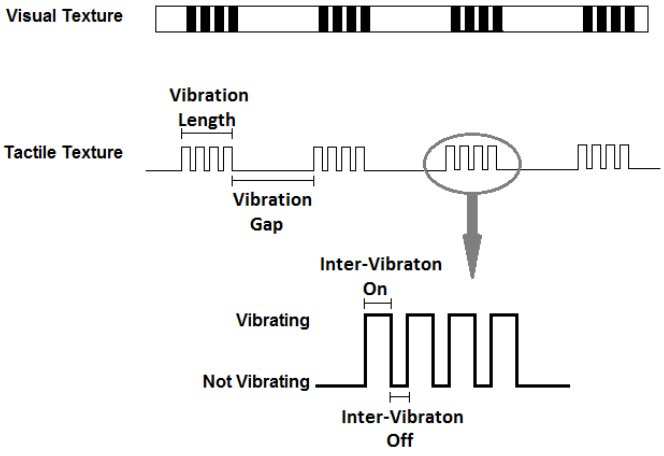


Fig. 3. An example of the visual and vibrotactile textures used. The four parameters of the texture are varied as the mood parameters change.

3.1 Mapping the Mood Vector to a Texture

There are many different ways we could map an arbitrary vector to visual and tactile texture. For example, using an approach such as physical modeling, rich representations of a vector can be built that vary with any number of parameters. The specific mapping should be tuned to best fit the technology available. In the current implementation, we have designed it for use with a phone's standard internal vibrotactile actuator.

Our current mood vector has four parameters that are mapped to a visual and tactile texture. Our texture is constructed through short regular pulses of vibration patterns. We define the four parameters of our texture as follows: the Vibration Length (VL), the Vibration Gap (VG), the Inter-Vibration On Length (IVOn), and the Inter-Vibration Off Length (IVOff). Figure 3 shows how each of these parameters are mapped to the vibrotactile signal. VL and VG are used to set the low frequency of the texture, with IVOn and IVOff contributing to the higher frequency components of the vibration that affect the perceived roughness of the vibration signal. Here we make simple mappings from these characteristics to the moods: IVOn length increases as 'Joy' increases, the VL increases as 'Aggression' increases, IVOff increases as 'Mobility' increases, and VG decreases as 'Busyness' increases. Consequently, when user is joyful and in urban transport, other contacts can feel a vibration intense with a high impulse rhythm. In opposition, when this user is sad and in his office, the vibration is smooth and slow. When the user aggression and busyness are high, pattern vibration are long and gaps between them are short revealing user is occupied and has few time for communication.

Likewise, the visual texture is built with ridges and grooves from a gradient pattern from the previous four parameters. The light grey corresponds to the vibration gap, and the grey levels reveal in black the grooves and in white the ridges.

3.2 Tactile Communication

The paper [11] proposes a method for creating tactile textures without force feedback by using a simple motion sensor and a single vibrotactile actuator. This proposal is based on wavetable synthesis driven by the user's hand movements. The results show envelope ridge length and spatial density were distinguishable design parameters and ridge length and spatial density influence perceived roughness and flatness similarly as with real textures.

We use here three types of tactile feedback automatically generated by user finger movements. Firstly, the tactile channel is used for list manipulation: the user can feel a short buzz for each contact moved over when scrolling up or down the contact list (vertical touchscreen movement). Secondly, the tactile channel is used to display the remote contact's mood vector providing context awareness (horizontal touchscreen movement along the contact item): when the user glides his finger along a contact item, he can feel a vibrotactile texture representing the mood vector of this contact, sharing their current mood. Finally, we use the tactile channel for inter personal communication augmenting voice communication (moving back and forward along a

contact item): when the user strokes the contact item, the remote contact receives a (visual or tactile) notification. The remote contact can then open the tactile communication channel by stroking the user's contact list entry. Then they can start a tactile dialogue, playing with the duration of vibration and duration of silent: each time the user touches his screen, the device of the remote contact starts to vibrate. When the user stops touching his screen, the remote device stops vibrating.

The tactile communication can either take the form of preset patterns of vibration that might be customisable to a user, but have a specific meaning to that user (such as a notification for a particular contact being unique). Alternatively, real time communication allows us to provide a channel to support the more affective aspects of the communication. Initially, as one user explores the context of a contact, the remote contact may be made aware by feeling a tactile representation of the finger movements of the user across the touch screen. Mapping the vibration pattern to finger speed could provide a simple but potentially expressive method of interaction, with the potential to allow users to develop their own tactile languages.

3.3 Real Time Communication

With the advent of the *eXtensible Messaging and Presence Protocol* (XMPP – **www.xmpp.org**), the transfer of generic data in real time is now available on a wide range of devices. Traditionally, this protocol is used for instant messaging and VoIP applications, however we can take advantage of this infrastructure to start transmitting different types of data. Here we use XMPP to provide a mechanism for communication of generic data across a communication channel in real time. Currently, these messages are received either as Voice Over IP (VOIP) messages or in coded tactile form allowing simultaneous tactile and audio communication.

Figure 4 presents our real time communication architecture. The local phone uses sensor information as microphone or finger displacement on touchscreen to build a message. The message is processed and compressed to follow the XMMP recommendation. When receiving the message, the remote phone parses the input stream and decompresses the content to resituate the initial information through visual, audio or tactile feedback.

4 Evaluation

Prior to user testing, we evaluated the real time nature of the communication in different circumstances with particular focus on the latency of the communications. We do this over WiFi and 3G as well as in close proximity (both participants in the same city) and over a larger distance (where one participant is in the UK and one in the France). We also examine latency during low bandwidth (in this case tactile messages) and high bandwidth (when the audio channel is open) communication. Table 1 shows mean latencies in the different situations. Unsurprisingly, there is a larger latency when messages are being sent over longer distances and lower bandwidth communication channels. The round trip time of around 200ms in these

instances translates to a network latency of approximately 100ms for a communication. This value might be acceptable for one-way communications and manageable for standard voice communication, but may cause issue in any more complex interactions where users try to synchronise their movements. One example of this would be a more complex version of the tactile feedback where supporting a remote relationship with both users simultaneously sending messages simultaneously where the timing of the messages may become important to any personal tactile language developed.

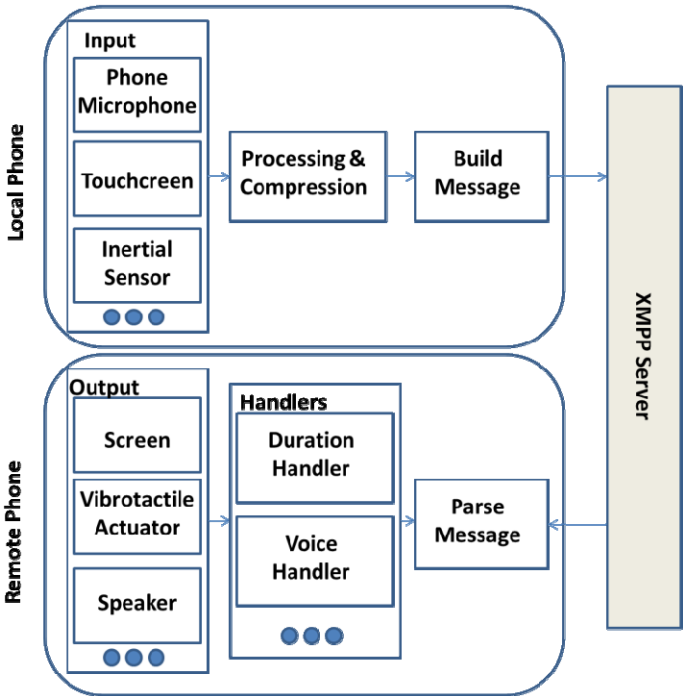


Fig. 4. The real time communication architecture

In the close proximity WiFi situations, a one-way latency of less than 20ms may lend itself better to more complex real time interactions.

We have completed an early user evaluation of the system. This took place in two stages; firstly to examine usability and user acceptance of the interactions and secondly to test the system with a small group of potential users.

4.1 Focus Groups

The first session took place as a focus group between 3 pairs of friends. The goal here was to test their acceptance and understanding of the application. The focus groups were used to generate discussion on potential usage and issues with the system. The two users were co-located but each had their own phone (Google Nexus One) with the

app installed. The interactions were first described and demonstrated to the participants and then they were given time to explore the interactions and communicate with each other. A discussion of each of the features of the system and their potential uses was then held. Each focus group session took around 30 minutes.

While the reception of the technology was extremely positive, these sessions raised some potential issues that users will face using the multimodal contact list. These finding we feel will generalise to any multimodal communication system that attempt to use technology to communicate in unusual ways. The four main findings are described below.

Table 1. Round Trip Time (ms) data for different situational contexts and communication channels. Close refers to two users communicating in the same local area, where as remote refers to communications between two users in the UK and France. Low data load refers to situations where tactile message were being sent and high data load refers to RTT during audio communication.

	Round Trip Time (ms) WiFi - WiFi	Round Trip Time (ms) 3G - 3G
Close (low data load)	37.4 (std. dev. 19.6)	153.8 (std. dev. 38.8)
Close (high data load)	33.1 (std dev 22.6)	223.7 (std. dev. 59.6)
Remote (low data load)	96.6 9 (std.dev. 19.2)	N.A.
Remote (high data load)	163.2 (std. dev 34.9)	N.A.

4.1.1 Appropriate Metaphors Are Required

When developing novel methods of communication, we must be aware that users are not use to communicating in this manner and the emphasis is on the designer to provide methods that allow users learn new interaction techniques. One way of supporting this learning is by using interface metaphors. In this application the users stroked the appropriate contact onscreen to open a tactile communication channel. The contact then lit up on the screen indicating to the user that communication could begin. This stroke interaction and lighting up effect proved confusing. We could potentially take cues from face to face communication where a tap on the shoulder or nudge is sometimes used to alert the other participant discretely that you want their attention. Stroking an onscreen contact may be more appropriate for more intimate communication (between a husband and wife for example) and less appropriate when chatting to friends or work colleagues.

Additional feedback could support the metaphor when the channel is opened. For example, showing the user's face to indicate that you now have their attention.

4.1.2 Be Wary of Using Touch as a One Way Channel

Touch is naturally a bi-directional channel. We tap someone on the shoulder to get his or her attention, but we also received feedback through our haptic sense that we have performed the action. Here we use touch communication as two separate one-way channels. We open a tactile channel and send tactile feedback when the user presses the flat touchscreen. There was a strong sense in the focus groups that this one-way communication was not appropriate. Even co-located, participants were sending a signal and then asking "Did you feel that". This lack of feedback adds a level of confusion to the communication that is rarely an issue with channels such as audio. One immediate change to the interface that was made was to include visual feedback to the user that their tactile message was being sent. A more complete system might also use phone sensors on the remote side (such as accelerometers or capacitive sensors) to determine whether the remote user is holding their device and communicate this back to the sender to close the loop and complete the communication channel.

4.1.3 Unusual Combinations of Technologies

Voice and vibration are rarely used together in phone communication. In this case the vibration interfered with the voice communication, particularly if the device was on a hard surface. An audible vibration could be heard through the audio communication channel. While this was an issue for this experiment, we do not believe this is a fundamental issue, as the problem of crossover and feedback has been dealt with in normal speakerphone design, and it should be possible to develop appropriate filters.

4.1.4 Overloading the Vibrotactile Channel

In the real world, the tactile channel is relatively high bandwidth and can detect subtle differences in the shape, texture, softness or temperature of objects. The standard vibrotactile device on phones severely limits this communication channel providing limited control and generally a single actuator. In this application we use the vibrotactile channel for different purposes such as list scroll events, receiving one or more tactile messages, and when exploring the mood vector. The key issue here is 'how does a user distinguish one vibration from another?' We can choose to use different vibration profiles for each, however the poor level of control for the actuator limits what can be achieved here. There is also the issue that two simultaneous events (such as two tactile messages sent by different users) will interfere. It is difficult to separate these events one tactile message from an ongoing communication with another from a new user without resorting to visual or auditory feedback. Care must be taken when rely on the tactile channel for multiple purposes that the user can distinguish between the separate events.

4.2 User Evaluation

We evaluated a modified version of the Multimodal Contact List app. The app was modified to better support feedback during communication, indicating more clearly

when a communication channel was open and when tactile messages were being sent. A group of seven users took part in the study over a period of five days. Six users were work colleagues and friends while the seventh was in a remote relationship with one of the other participants. Each user had the same contact list containing all seven users of the system including them so that they could see their own mood vector and test the system by sending messages to themselves.

For this user study, we used a simplified version of the mood vector. We map two parameters onto the texture: Joy and Busyness. Joy is increased by increasing the 'Inter-Vibration On' parameter and Busyness is increased by increasing the 'Vibration Gap' parameter. Examples textures generated from this mapping are shown in Figure 5.

Training was provided to each of the users taking part in the study with a user manual describing how to perform each of the interactions being provided along with a training session where the experimenter demonstrated the different features of the application and then asked the user to try each of the interactions until they were satisfied that they could set their mood vectors, and send audio and tactile messages successfully. Results were collected at the end of the study through a questionnaire, which probes user acceptance of each of the features and their attitudes on how these ideas could be used in a wider context.

Here we discuss the main findings from the study for each of the functionality and an overall view of the goals of the app separately.

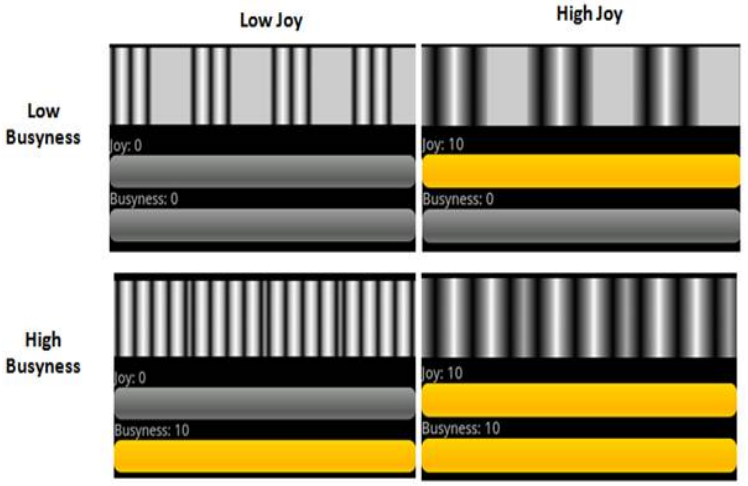


Fig. 5. Examples of the extremes of the mood vector used during the study. This application mapped 2 parameters onto the texture; Joy and Busyness.

4.2.1 The Mood Vector

The participants generally received the mood vector negatively. One participant "found the mood vector to be the least useful of the functions". The mapping chosen was described as confusing by more than one participant, which may have contributed

to it being updated infrequently. This infrequent use again did not help reinforce the mapping with the users leading to problems remembering the mapping once learnt. Further to this, users seemed to use the visual mood vector but had little need for the tactile version.

The abstract nature of the mapping forced users into an extra 'interpretation' stage that could potentially have been avoided through an iconic representation that had a semantic link between the image and the user's mood. The goal of this feature was to allow users to gain a lot of information about the general mood of their contacts with a visual glance. A better semantic link using visual properties quickly and easily identified with a glance may better support this goal.

4.2.2 Vibrotactile Messages

All users perceived Vibrotactile messages positively. There was a general feeling that it provided a discreet means of communication that would be appropriate in a number of different situations both in work contexts and when with family and friends. It was seen as a means to initiate a conversation in real time and potentially probe contacts about their availability to communicate.

There were reservations with the difficulty of communicating more complicated messages through vibration. With extended use, some "vibration code" may develop between participants but this was not evident in this short trial. Users drew attention to the fact that vibration was not suitable in all instances. Firstly, users felt that the vibrotactile feedback was a very personal way of communicating and not appropriate for unknown contacts.

Secondly, it was not considered appropriate for emergencies situations where it was important the other contact got the message. The discreet nature of tactile feedback is advantageous when communicating in situations where a lot of motion or noise may be inappropriate. However, this also leads to the fact that if the user is not touching the phone, the message is easily missed. Potentially, this could be resolved by using other sensors in the phone (such as accelerometers or capacitive sensors) to detect whether the user is holding the device and can therefore perceive the vibrotactile message.

4.2.3 Voice messages

Users were obviously more used to voice communication with phones. One innovation with this system however is that users could initiate a voice call without calling the contact first. Participants felt that some form of notification and acceptance was essential to avoid noise in inappropriate settings. The couple in a remote relationship particularly highlighted this. They chose to use the system in a playful manner "like two kids with walkie-talkies". They used it to supplement Skype communication combining voice and vibrotactile messaging with a video feed.

There were also concerns about the bandwidth required for VOIP. This was to some extent mitigated by allowing users to restrict network communication in situations where no WiFi connection was available, however as this was a global setting on the app this also did not allow communication using the other modalities. Allowing more control to the user over the connection to allow low bandwidth forms of communication while blocking high bandwidth data would resolve this issue.

4.2.4 General Perspectives

There were a number of changes and additions to the available functionality suggested by the users. The major suggestions for change revolved around improvements to the mood vector to provide useful and easy to interpret feedback on the current context of contacts. This could be supplemented by text to support learning of the mapping and provide more detailed information when required. Availability for communication was a key factor here. This is a key feature missing from standard phone communication that is available and plays a useful role in many other forms of remote communication such as Instant Messaging and VOIP systems.

The mood vector concept could easily be adapted to support the user in their choice to communicate with a contact just now or later. The vector could be extended to show different information to different groups of users (e.g. Joy might be appropriate for friends but not work colleagues), however appropriate representation is key to allowing the user to browse quickly and easily extract the important information. This could also be enhanced by allowing the user to filter contacts on different parameters (e.g. Find me contacts that are not currently busy).

Participants also felt there were other types of information that could potentially be used to communicate. Sketches and handwriting were suggested as useful forms of communication that could support both useful and playful communications. Images were not seen as something that would regularly be shared, as there are other more traditional channels that allow easier sharing to a wider group of friends.

The vibrotactile messaging was seen as a useful feature, however, key to acceptance is providing feedback on whether the message was received or not. Even with the additional feedback added after the focus groups, there was still no indication of whether the contact received the message unless they replied. It was generally used to notify contacts, however there is the potential to include more information. The vibrotactile messages generated were simple vibrations of variable length (controlled by the user). With more complex control, users could potentially build up more powerful forms of vibrotactile communication over a longer time period.

One oft-mentioned issue was that users were acutely aware of their data limits. In an app such as this which relies of an always on data connection, it is often difficult to track data usage which could lead to expensive overruns of data over the 3G network. This is particularly true when using a higher bandwidth channel such as audio. Without any sort of visibility of the data usage, users were choosing to restrict the app to WiFi-only usage. While this allowed participants to use the app without worrying about cost, it meant that they were restricted to interacting only in fixed locations; usually at home or at work. This situation could have been somewhat mitigated by providing an option for tiered access to the 3G network for low bandwidth channels allowing tactile messages and mood vector messages to be sent without access to WiFi. This would have allowed users to maintain more of a presence in the system and encouraged more interaction between users.

Other improvements will look to support a history of communications such that any messages sent and missed will be available to be experienced later.

5 Conclusions and Future Work

The Multimodal Context List allows users to share context information and communicate both verbally and, discreetly, through touch. Visual and tactile textures are used initially to provide a quick overview of the context information for contacts in a contact list. By interacting further with the contact on the screen, we can 'drill down' deeper into the contact's context information, eventually opening a multimodal audio and tactile channel of communication. This information will allow a user better understanding of the psychological, situational, and environmental context that the remote contact is in.

The context represented can be based on explicitly set parameters (as described here with the Mood Vector which provide psychological context) or through some context inferred through a fusion of sensor values that maybe used to provide environmental and situational contexts. The application had been demonstrated on small numbers of Android 2.0 phones, which provide the basis for a larger field trial that will investigate the benefits in a longer-term usage scenario over a larger number of users.

References

1. Ahmaniemi, T., Marila, J., Lantz, V.: Design of Dynamic Vibrotactile Textures. IEEE Transactions on Haptics 3(4), 245–256 (2010)
2. Bales, E., Li, K.A., Griwsold, W.: CoupleVIBE: mobile implicit communication to improve awareness for (long-distance) couples. In: Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (2011)
3. Brown, L.M., Williamson, J.: Shake-to-talk: Multimodal Messaging for Interpersonal Communication. In: Proceedings of the 2nd International Workshop on Haptic and Audio Interaction Design, Seoul, Korea (2007)
4. Chang, A., O'Modhrain, S., Jacob, R., Gunther, E., Ishii, H.: ComTouch: design of a vibrotactile communication device. In: Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, p. 320. ACM (2002)
5. Dey, A.K., Hakkila, J.: Context-Awareness and Mobile Computing. In: Lumdsen, J. (ed.) Handbook of Research on User Interface Design and Evaluation for Mobile Technology, ch. 13. Idea Group, Inc. (2008)
6. Heikkinen, J., Olsson, T., Vaananen-Vainio-Mattila, K.: Expectations for user experience in haptic communication with mobile devices. In: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI 2009). ACM, New York (2009)
7. Hemmert, F., Gollner, U., Löwe, M., Wohlauf, A., Joost, G.: Intimate mobiles: grasping, kissing and whispering as a means of telecommunication in mobile phones. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, Stockholm, Sweden, August 30-September 02 (2011)
8. Kaye, J.: I just clicked to say I love you: rich evaluations of minimal communication. In: CHI 2006 Extended Abstracts on Human Factors in Computing Systems (CHI EA 2006), pp. 363–368. ACM, New York (2006)

9. Kim, S., Kientz, J.A., Patel, S.N., Abowd, G.D.: Are you sleeping?: sharing portrayed sleeping status within a social network. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work, pp. 619–628. ACM (2008)
10. King, D.: Four principles of interpersonal communication, http://www.pstcc.edu/facstaff/dking/interpr.htm (accessed June 10, 2010)
11. Kontaris, D., Harrison, D., Patsoule, E., Zhuang, S., Slade, A.: Feelybean: communicating touch over distance. In: Proceedings of the 2012 ACM Annual Conference Extended Abstracts on Human Factors in Computing Systems Extended Abstracts, CHI EA 2012 (2012)
12. Murray-Smith, R., Ramsay, A., Garrod, S., Jackson, M., Musizza, B.: Gait alignment in mobile phone conversations. In: Proceedings ACM MobileHCI, pp. 214–221. ACM, New York (2007)
13. Rico, J., Brewster, S.: Gestures all around us: user differences in social acceptability perceptions of gesture based interfaces. In: Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2009, Bonn, Germany, September 15-18, pp. 1–2. ACM, New York (2009)
14. Rovers, A.F., van Essen, H.A.: HIM: A framework for haptic instant messaging. In: CHI 2004 Extended Abstracts on Human Factors in Computing Systems (CHI EA 2004), pp. 1313–1316. ACM, New York (2004)
15. Sellen, A.J.: Remote conversations: The effects of mediating talk with technology. Human-Computer Interaction 10, 401–444 (1995)
16. Shirazi, A.S., Schmidt, A.: Using Mobile Phones to Maintain Intimacy and Connectedness. In: Proceedings of MobileHCI (2011)
17. Strachan, S., Williamson, J., Murray-Smith, R.: Show me the way to Monte Carlo: density-based trajectory navigation. In: Proceedings of ACM SIG CHI Conference, San Jose, pp. 1245–1248 (2007)
18. Tünnermann, R., Mertes, C., Hermann, T.: Feelabuzz – Direct Tactile Communication with Mobile Phones. In: Proceedings of Workshop on Mobile Social Signal Processing (2012)
19. Wang, R., Quek, F.: Touch & talk: contextualizing remote touch for affective interaction. In: Proceedings of the Fourth International Conference on TEI, pp. 13–20. ACM (2010)
20. Williamson, J.R., Crossan, A., Brewster, S.: Multimodal Mobile Interactions: Usability Studies in Real World Settings. In: Proceedings of ICMI, Alicante, Spain, pp. 361–368. ACM Press (2011)

Author Index