

TUM

INSTITUT FÜR INFORMATIK

A New Data-Set for Research on Audio Detection and Modeling of Social Micro-Contexts

Georg Groh, Alexander Lehmann



TUM-I1011

Mai 10

TECHNISCHE UNIVERSITÄT MÜNCHEN

TUM-INFO-05-I1011-0/1.-FI

Alle Rechte vorbehalten

Nachdruck auch auszugsweise verboten

©2010

Druck: Institut für Informatik der
 Technischen Universität München

A New Data-Set for Research on Audio-Detection and Modeling of Social Micro-Contexts

Georg Groh , Alexander Lehmann
Fakultät für Informatik
Chair for Applied Informatics and Collaborative Systems
Technische Universität München, Germany
Email: {grohg,lehmann}@in.tum.de

May 15, 2009

Abstract—We introduce the concept of social micro-contexts and their relation to modeling social situations and discuss the problem of detecting and modeling social micro-contexts. With respect to the validity of the problem, we argue that an interesting class of applications for such models exist and thus the problem is indeed a valid research problem. We introduce a novel data-set intended to allow for the development and validation of distributed collaborative approaches for the detection and modeling of social micro-contexts and social situations on the basis of audio-signals. For both concepts we present suitable definitions. We describe the social experiment that the data was collected in as well as the data-collection methods. Furthermore the annotation application, annotation process and annotation policies that were used are described. We conclude with a road-map of planned further development activities.

I. INTRODUCTION

While social computing with all its various facettes has become an important topic of research, the rich social dynamics happening on small temporal and spatial scales has, to our knowledge, not been subject to quantitative modeling and has not been exploited for useful IT services on a large scale. A field where computer science made contributions in modeling aspects of interactions on such small scales is e.g. the field of meeting support [3], where a wide range of fixed audio-visual sensors in the room allowed for the identification of speakers, speech recognition, automatic assembly of meeting minutes etc.

Besides social computing, in the past few years, the fields of mobile computing and context-aware computing have also gained much attention by the scientific community [9]. Mobile social networking (often on a P2P-basis) is one newer sub-field [4]–[6], which usually aims at combining the context-awareness of mobile applications (usually location-awareness) with the communication approach of virtual community services. As an example consider awareness services like “where are my friends” [10] etc. A substantial effort at MIT-Media-Lab also aims at using mobile computing devices for the detection, modeling and usage of social contexts [1], [2].

The goal of the data-set introduced here is to provide a quantitative test-bed for research that also aims at combining

mobile and social computing in a special way. The basis of this research are sensors which can be integrated into a mobile computing device in an unobtrusive way and which also serve other purposes including (but not necessarily limited to) microphones, location- and orientation-sensors. On the basis of these sensors and with the help of suitable heuristics we aim at modeling social contexts on small temporal and spatial scales which we call social situations or social micro-contexts (in order to differentiate them from social macro-contexts which are typically modeled by virtual social networks (as in Facebook, MySpace etc.) on the Web today). These models can then be used in various ways:

- *Socio-contextual awareness services* like *Social Life-logging* applications that are an extension to the existing diary-like life-logging [7] In Social Life-logging, social situations (social contexts) that the user has been in are also recorded.
- *Communication services* that e.g. allow for *socio-context-casting* as an extension to geo-casting mechanisms, as e.g. in use in new Car-to-Car-Communication services [8]. A geo-cast will allow for transmitting a message over a mobile ad-hoc network in certain directions or with certain geometric primitives as target areas. These services are intended for e.g. hazard warning applications in Car-to-Car-Communication. Socio-context-casting allows for specifying social situations as targets of communication acts without necessarily exactly knowing the participating persons’ precise addresses.
- *Information services* that allow for information spaces attached to social situations by e.g. *automatically tagging* information items (photos, documents etc.) in the personal information spaces of the social situation participants that have been accessed, created or modified during a social situation.

While an exact and all-embracing detection and modeling of social micro-contexts taking into consideration every social aspect can be considered to be an AI-hard problem, it is not the goal to cover all fine grained aspects of a social micro-

context but rather to provide a reliable basis for detection and modeling of social micro-contexts that is a basis for useful services. We will now investigate how social micro-contexts can be defined precisely.

II. SOCIAL MICRO-CONTEXTS

If a person A can perceive from her current point in time and space a set of persons $\{B_1, B_2, \dots, B_n\}$, we can define some notions in order to clarify what is meant by social micro-context:

An **individual, current real-world social micro-context frame** α_A of that person A is the “set” $\{s_1, s_2, \dots, s_n\}$ of the entireties s_i of all sensual perceptions by A of these persons B_1, B_2, \dots, B_n plus all additional knowledge about these persons that is *in principle* perceivable and knowable by A from her current position in time and space.

An **individual real world social micro-context** $\beta_A \subset \alpha_A$ of a person A is the “subset” $\{s'_1, s'_2, \dots, s'_n\}$ of the entireties $s'_i \subset s_i$ of all sensual perceptions by A of the persons B_1, B_2, \dots, B_n plus all additional knowledge about these persons that is *actually* perceived and known by A from her current position in time and space.

A **common real world social micro-context** β for a set of persons A_1, A_2, \dots, A_m is constituted by the *common* perceptions and additional knowledge by and of these persons: $\beta = \beta_{A_1} \cap \beta_{A_2} \cap \dots \cap \beta_{A_m}$.

A **social micro-context-model** λ is an algorithmically processable model for the sets s .

An **individual social micro-context representation** $\lambda(\beta_A)$ for a person A is a representation $\{\lambda(s'_1), \lambda(s'_2), \dots, \lambda(s'_n)\}$ of the entireties s'_i of all sensual perceptions by A of the persons B_1, B_2, \dots, B_n that can be measured by sensors and represented in an instance of λ plus all additional knowledge about these persons that can be represented in an instance of λ from A's current position in time and space.

A **common social micro-context representation** Λ for a set of persons $\{A_1, A_2, \dots, A_m\}$ is constituted by the *common* social micro-context representation by and of these persons: $\Lambda = \lambda(\beta_{A_1}) \cap \lambda(\beta_{A_2}) \cap \dots \cap \lambda(\beta_{A_m})$.

From these definitions a common social micro-context representation Λ is non-empty if and only if the associated persons are in mutual perception range. (Regard that the common additional knowledge about each other is considered to be “bound” to perceiving each other via the first assumption above). This illustrates the adjective “micro” with respect to time and space.

Of course, common social micro-context representations can overlap or be nested, which implies that Λ_X of a set of persons X can be overlapping with or be nested in Λ_Y of a set of persons Y if $X \cap Y \neq \emptyset$.

In order for a common social micro-context representation $\Lambda(X, t)$ to represent a *social situation* involving a set of persons $\{A_1, A_2, \dots, A_m\}$, it is necessary for $\Lambda(X, t)$ which is computed according to the upper definition with respect to a certain point in time t and implicitly through the perception

ranges of the set of persons $\{A_1, A_2, \dots, A_m\}$ a domain $X \subset \mathbb{R}^3$ to be persistent over some time-interval.

Social situations can also be nested or overlapping.

Social situations may act as instantiations of longer lasting n -ary social ties, if a subset of the involved persons is engaged in multiple social situations. The precise mechanisms of how a longer lasting social tie and social situations involving the people incident to that tie are related is subject to social psychology research.

A simple example for a social situation representation $\Lambda(T, X)$ involves unique ids for the involved persons, representations for the time-duration T the spatial extension X (possibly continuously dependent on $t \in T$), representations of the locations and relative orientations of the persons involved, and tags characterizing the semantic aspects of the social situation.

A. Services Using Social Micro-Contexts

The services using social situation models and corresponding representations involve awareness services (social life-logging as opposed to individual life-logging, blogging or micro-blogging (such as Twitter), indication of current social situations involving friends as a variant of the “where are my friends” service etc.), information services (attaching contextual information spaces to social situations or ongoing social situations etc.) or communication services (“socio-context-cast as means to address a social situation of the past without precisely knowing the addresses of all participants etc.)

However interesting the possibilities of *using* social situation representations are, robust and reliable techniques for detecting, measuring and characterizing social situations in order to build meaningful social situation representations must be developed first.

In order to create a suitable data-set which can be used to develop, test and compare approaches towards this goal, we conducted an experiment and annotated the resulting data in an extensive way.

III. EXPERIMENT BACKGROUND AND GOALS

We aim at using sensors for the detection, measurement and characterization of social micro-contexts and thus for social situations that can easily be integrated in a mobile computing device or wearable computing infrastructure, that do not require stationary support by snug equipment such as beacons or cameras, that are cheap and mass-compatible and that can be used for other applications as well.

One such piece of equipment is an audio-sensor, that is integrated in every mobile phone or PDA anyway.

The goal of the experiment is to create a data-set for the heuristic detection, measurement and characterization of social micro-contexts and social situations on the basis of audio signals.

We created a representative, medium-size social (super-)situation (a barbecue) and individually recorded all audio of all participants. We additionally recorded the event with several video cameras *only* to assist the later manual annotation of the data for later evaluation of the developed algorithms.

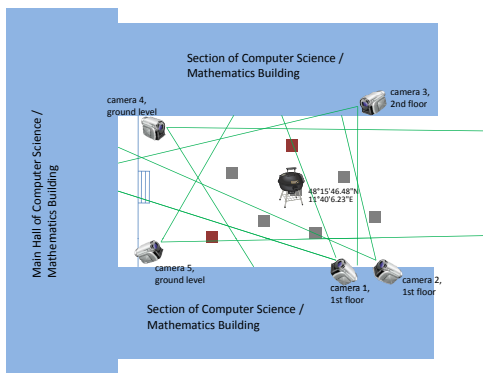


Figure 1. Experiment location with camera positions and rough angular views of the camera

IV. EXPERIMENT SETUP

The experiment took place in June 2008 at the TUM campus in Garching near Munich, Germany. A total of 25 study participants (students and researchers) plus 6 experiment supervisors took part in a 2 and a half hour afternoon barbecue. About two hours of the barbecue were recorded by a total of 5 fixed location cameras from different positions and one mobile camera operated in turn by the experiment supervisors. Each of the 25 participants was equipped with a low-fi microphone integrated into a cheap MP3-player.

A. Audio

Before commencement of the actual experiment, each participant provided a recording of their voice by speaking a sentence covering most of the vocals of German language:

The data were recorded using portable flash mp3 players (TEAC MP-114 Flash MP3 Player) at a sampling rate of 11 KHz and stored as WAVE files with IMA ADPCM encoding.

Furthermore, a central microphone was used to record sounds not related to any specific persons.

In order to be able to synchronize the multitude of audio- and video-recordings, a loud synchronization signal was sounded at the experiment's beginning and in certain intervals during the experiment.

B. Video

The locations and approximate perspectives of the 5 stationary cameras are depicted in figure 1. In order to ease the identification of the participants during later annotation the additional mobile camera was utilized for getting miscellaneous close-up shots of the participants as well as small groups of people at different times.

C. Social Network of Affection

In order to allow for research on how parameters of social situations (e.g. frequency of verbal interaction) and subjective estimation of social relations to a person interact, we also asked every participant to judge the social relations with respect to valence and intensity to as many other participants as desired. On average, each person estimated the relations to 5.68 (+/- 1.93) other persons. We will present partial results

for this affection network below after we have discussed the annotation process.

V. DATA-PROCESSING AND ANNOTATION

For the creation of the reference data-set it was the goal for each participant to annotate for each point in time, who exactly this participant was talking to.

A. Data-Preprocessing

In order to keep the annotation manageable, 30 minutes of the whole experiment were selected, starting from the highest amplitude of the first acoustic synchronization horn signal. The WAVE files of 30 minutes audio recordings were manually cut for each participant and cross checked at several points in these 30 minutes to rule out possible errors in the cutting process.

As has been said before, the video recordings had the sole purpose of aiding / enabling the annotation process. Thus the video recordings had to be split up into single frames allowing the human annotator to assess the respective current social situation. Assuming that during hot conversations speakers could change at a maximum of about every half of a second, it was decided that the timely video resolution should hence be reduced to 2 frames per second, resulting in 3600 frames for each camera for the 30 minute period which were then stored as JPEG images. For each camera the beginning of the 30 minute video sequence (the first frame of these 3600 frames) was manually determined according to the horn signal from the audio track of the video.

B. Annotation

A special annotation software was designed that implemented the following requirements:

- Ability to work with multiple audio- and video-sources in order to give as clear a picture of the social situation as desired.
- Support for distributed annotation and thus at best also platform-independence, since annotation can obviously become a relatively time-consuming task. The software must allow for each participant and each time frame of 0.5 seconds to specify who this person was talking to
- Data should be stored in a non-proprietary format that also allows for easy merging of possibly distributed annotation results and optional easy, yet extensive querying of the data.
- Reasonable adaptability to similar experiments and specifications.
- Easy installation, maintenance and usage of the software in conjunction with the accompanying source data as a whole.

Thus the result of the annotation process is for each user and each time frame a human estimation of the precise persons that the user is talking to. Figure 2 shows the GUI of the resulting C++ application. The cameras and the persons can be named as desired. In the figure the names of the original participants have been replaced by fantasy names. The main window is comprised of three major parts, namely video, audio

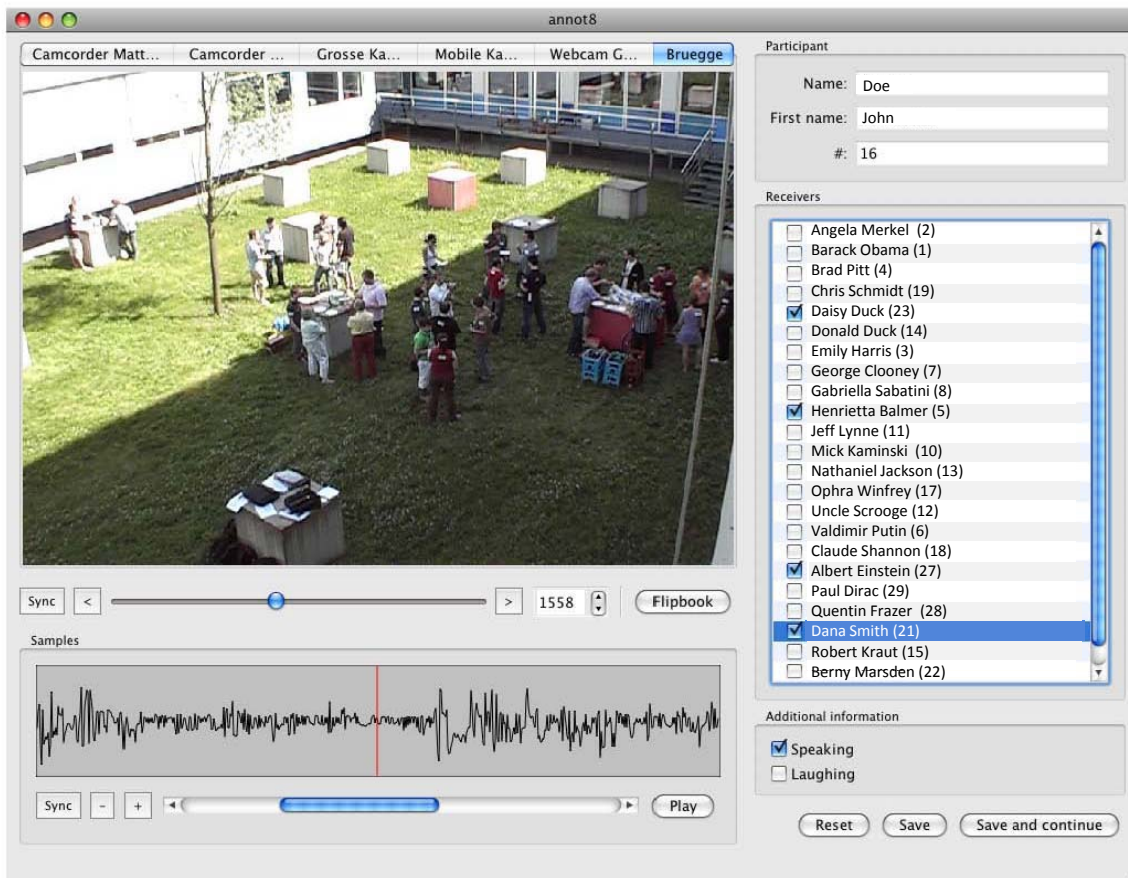


Figure 2. Annotation application

and annotations. While the video part can be used to select the active video source from the set of available video sources and current frame, start respectively stop the so-called flipbook mode or synchronize the current frame to the audio position, the audio part allows for previewing, navigating and zooming of the recorded samples, starting and stopping audio playback as well as synchronizing the audio position to the current frame. The right-hand side of the main window displays the annotations part which can be subdivided into three portions: The upper portion is called Participant and is only used to show information about the currently selected sender, i.e. the participant who was chosen for annotation. As in only show”, nothing can be edited or modified here. In order to select another sender, the application must be restarted and the desired sender then be selected. The middle portion is called Receivers and displays a list of all participants excluding the sender. A checkbox next to each participant shows whether he or she has been selected as receiver for the current frame. Also, these checkboxes can be used to toggle the selection state. The lower portion is called Additional information and shows the attributes that were specified for the given frame. Again, the checkboxes can be used to toggle the attributes selection state. Finally, the push buttons Reset, Save and Save and continue can be used to reload or save the current frames annotations.

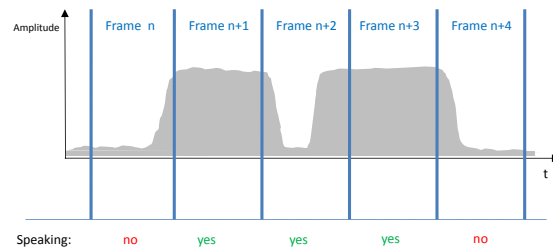


Figure 3. Annotation of frames

The latter additionally advances to the next frame. Also, in order to ease quickly changing annotations, the applications menu provides Cut, Copy and Paste accordingly. The most important thing to note, however, is that annotations are always made with respect to only the current at a time. Hence, again, no matter what the current audio position is, only the current frame number is used to determine the timestamp for possible annotations. Keyboard shortcuts have been assigned to the relevant (if not all) elements of the user interface and some special functions (e.g. changing the video source from the active source to another source and back).

Together with our volunteer annotators, best practices for using the application were designed and rules for all situations

Pearson	Valence	0.395
	Intensity	0.501
Spearman	Valence	0.583
	Intensity	0.476

Table I

CORRELATIONS OF THE NUMBER OF FRAMES A ADDRESSED B AND THE VALENCE AND INTENSITY OF $A \rightarrow B$. (THE VALUES ARE $\alpha = 0.05$ -SIGNIFICANT).

where the audience of a speech act in a frame was unclear were set up, in order to ensure that the annotations were all compatible and consistent:

- We generally annotate who is *addressed* by a speaking person B and *not* who is actually *listening* to B which would be much harder if not impossible to determine visually or by listening to the audio-signal.
- We annotate on a per frame basis (0.5 seconds). An ongoing social interaction may thus well contain frames of silence (silence means no annotations).
- Each person A within a 2 meter radius around the person B speaking and facing B is considered to be addressed by the speaking person if not explicitly perceivable otherwise.
- Each person A (no matter how far away or no matter how oriented) that is perceivably addressed by a speaking person B is annotated as being addressed by B.
- Sub-chatting: If n Persons A_1, A_2, \dots, A_n are perceivably part of a social situation with m participants $A_1, A_2, \dots, A_n, A_{n+1}, A_{n+2}, \dots, A_m$, and the n persons have a communication of their own with one of them, let's say A_i $i \leq n$ addressing the other $n - 1$ while at the same time someone from the larger social situation let's say A_j $n < j \leq m$ perceivably addresses all $m - 1$ other participants (including the n persons from the sub-situation) we annotate this as such: A_j addresses all $m - 1$ other participants whereas A_i only addresses her sub-chat encompassing its $n - 1$ members.
- We only annotate someone as speaking (to whom ever) in a frame of 0.5 seconds if for the majority of the frame duration the person is speaking. Figure 3 depicts this policy.

We cross checked the annotations systematically for plausibility and manually re-edited sequences where inconsistencies were detected.

C. Social Network of Affection: Partial Results

As a very simple example of how the data can be used, we computed the correlations between the number of frames, a person A addressed a person B and the valence and intensity of the social relation of the relation from A to B for all social relations that the users estimated in the questionnaire accompanying the experiment. The results are shown in table I. Although the linear correlations are not perfect, we can clearly see that the frequency of contacts (deduced from the

frequency of verbal contacts) is an indicator for an intense and positively valued social relation. While this result may not be too surprising by itself, it can be seen as an indicator that audio signals may indeed be used to allow for meaningfully deducing social structures, regarding that here only the most simple means of analysis (simply counting) was used.

VI. FUTURE WORK

The first part of future work on the data-set will be to review the existing approaches on signal processing for speaker diarization and adapt them for our purposes. This will allow for individual social micro-context representations on the basis of audio to be computed. We will then have to develop a distributed approach to

- 1) Improve the individual social micro-context representations through mutual adjustments (calibration, synchronization, cross-validation etc.)
- 2) Agree on common social micro-context representations if possible.
- 3) Agree on common social situation representations if possible.

We also intend to investigate in how far the structure and sequence of detected social situations match the subjective estimations of valence and intensity of social ties.

VII. CONCLUSION

We presented a new extensively annotated data-set for the development and evaluation of methods for collaborative distributed detection and modeling of social micro-contexts and social situations. Both concepts were suitably defined. We argued that solutions for this problem have interesting applications and we thus consider the problem as a valid open problem worth investigating. In conjunction with the data-set we introduced our view on social situations and a principal road-map for the development of the respective methods for collaborative distributed detection and modeling of social micro-contexts and social situations. We intend to present first results on these approaches in the very near future.

REFERENCES

- [1] N. Eagle *Behavioral Inference Across Cultures: Using Telephones as a Cultural Lens*, IEEE Intelligent Systems 23:4, 62-64.(2008)
- [2] N. Eagle and A. Pentland *Reality Mining: Sensing Complex Social Systems*, Personal and Ubiquitous Computing, Vol 10, 4, 255-268.(2006)
- [3] *State of the Art Report: Automatic Dialogue Act Recognition November 6, 2007* AMI Consortium <http://www.amiproject.org> (URL, May 2009)
- [4] G. Groh *Groups and Group Instantiations in Mobile Communities—Detection, Modeling and Applications Proc. ICWSM07*, Boulder, Co, USA, Mar 2007
- [5] *Sociallight* www.sociallight.com (URL, May 2009)
- [6] *Noserub* www.noserub.com (URL, May 2009)
- [7] Mann, S. *Wearable computing:toward humanistic intelligence* Intelligent Systems 16 (3): 1015. 2001
- [8] C. Maihöfer *A Survey of Geocast Routing Protocols* IEEE Communications Surveys & Tutorials, vol. 6, no. 2, pp. 32-42, 2004
- [9] A. Dey *Understanding and Using Context* Personal and Ubiquitous Computing Journal, Volume 5 (1), 2001, pp. 4-7.
- [10] *Where are my friends?* Google Android Application <http://code.google.com/p/wherearemyfriends/> (URL, May 2009)