MobileSSI - a multi-modal framework for social signal interpretation on mobile devices

Simon Flutura, Johannes Wagner, Florian Lingenfelser, Andreas Seiderer, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Flutura, Simon, Johannes Wagner, Florian Lingenfelser, Andreas Seiderer, and Elisabeth André. 2016. "MobileSSI - a multi-modal framework for social signal interpretation on mobile devices." In *Proceedings: 2016 12th International Conference on Intelligent Environments (IE) - IE 2016, 14–16 September 2016, London, United Kingdom*, edited by Randall Bilof, Juan Carlos Augusto, and Mehmet Karamanoglu, 210–13. Los Alamitos, Calif.: IEEE. https://doi.org/10.1109/ie.2016.47.

MobileSSI – A Multi-modal Framework for Social Signal Interpretation on Mobile Devices

Simon Flutura, Johannes Wagner, Florian Lingenfelser, Andreas Seiderer, Elisabeth André
Human Centered Multimedia,
Augsburg University, Germany
Email: {lastname}@hcm-lab.de

***Abstract*—Over the last years, new generations of mobile devices have found their way into our pockets. They provide more and more computational power and memory capacity to perform complex calculations that formerly could only be accomplished with bulky desktop machines. Moreover, mobile devices are equipped with a range of sensors to capture people's motion, environmental sound etc. These capabilities combined with the willingness of people to permanently carry them around open up completely new ways of observing human behaviour no longer in laboratories, but "in the wild". However, the detection and analysis of social cues is still a challenging task and requires adequate tools to synchronise, process and analyse relevant signals. This may be the reason why many studies and applications focus on offline analysis and typically collect data over long periods of time and analyse them afterwards. To allow for immediate feedback, real-time assessment is necessary. In this paper, we present MobileSSI, a port of the Social Signal Interpretation (SSI) framework to Android and embedded Linux platforms. The framework supports the joint development of processing pipelines for the analysis of social signals on a desktop computer and mobile devices. Throughout the paper we report on challenges we had to face when porting SSI to a mobile context. Furthermore, we summarise first experiences with a real-life setting in a pub where we focused on the analysis of multimodal social group dynamics investigating laughter as a sign of enjoyment.**

I. Introduction

With each passing season, a new generation of mobile devices finds their way into our pockets. They allow us to handle our daily home and work tasks without the help of bulky desktop computers. In fact, smart phones have become so omnipresent that people do no longer perceive them as computers any more. Meanwhile the computing power of those devices steadily increases and the amount of integrated sensors grows. This opens up completely new possibilities of combining personal data about an individual with context information that is autonomously acquired from the environment. Furthermore, it bears great potential for a research domain, which over the past years has gained increased attention: Social Signal Processing (SSP). SSP aims at making machines more human-like by equipping them with the ability to recognize, interpret and express nonverbal behavioural cues [20]. Since SSP targets phenomena observed in everyday communication it is difficult, if not impossible, to develop robust models of human-human interaction solely based on experiences gained in laboratory settings. Here, mobile devices provide completely new possibilities to collect data in a natural and unobtrusive way [19], [25]. The benefits of mobile devices for SSP can be summarized as follows:

- Mobile devices have become an integral part of people's everyday life. Therefore, they enable us to design experiments that balance realistic conditions and experimental control.
- Mobile devices are equipped with a wide array of sensors to monitor user behaviour and derive context information.
- Mobile devices are small and lightweight and can be carried around for an extended period of time, which suits long-term and in-situ recording. More spontaneous and natural interactions can therefore be expected.
- Mobile devices also allow us to go beyond short-term social and emotional cues and to create long-term user profiles. Battery power still comes in as a limitation which is mitigated by the fact that most people keep their phones charged routinely.

Hence, it is not surprising that there has been growing interest over the past few years in the development of mobile applications that monitor user behaviour. In order to help users improve their lifestyle, research has been conducted to identify correlations between mood and human behaviours including their physical activity, social interactions and sleep [10]. Typically, behavioural data, derived from sensor's data such as acceleration, skin conductance or voice, is obtained from mobile phones or wrist sensors. In addition to data provided by the mobile phone sensors, communication data, such as the number of text messages or missed calls, have been investigated as stress and mood indicators or predictors [16], [4], [11], [7]. Furthermore, attempts have been made to detect stress from the user's voice in natural environments using microphones on smartphones. Some approaches just make use of the microphones to collect data [12], while others [6], [1] developed a platform that offers feature extraction functionalities for vocal emotion recognition running on mobile devices. Renaud and Crawford [14] suggest employing behavioural biometrics, such as keystroke dynamics, use patterns and voice analysis techniques for passive authentication. Crossan et al. [2] present a multi-modal contact list to enhance remote communication by sharing selected context information, which is automatically derived by the system (e. g. the current user mood). Damian et al. [3] developed a portable system that provides real-time feedback about the quality of a presen-

ter's performance in public speaking. Recommendations are automatically derived by analysing openness, body energy and speech rate and presented through a wearable display, such as Google Glass.

However, most studies focus on offline analysis and typically data of people is collected over the day and analysed afterwards. To allow for immediate analysis and feedback, real-time assessment is necessary. Providing developers with tools to record, analyse and recognise human behaviour in real-time on mobile devices has been our driving force when porting the Social Signal Interpretation (SSI) framework [23] to run on Android platforms. In the following, we will briefly introduce SSI and describe particular challenges we had to face during porting. Finally, we will demonstrate the potential SSI offers on mobile phones by means of an application that detects user enjoyment in real-time in a multimodal context.

II. Mobile SSI

The SSI framework aims at closing the gap between offline analysis and the development of online systems. To this end, it provides an architecture that does not only provide tools for data recording and machine learning, but also supports the immediate implementation of a learned model in a real-time fashion. Originally, SSI was developed for desktop machines. However, given the mobile boom in the last years and the great potential mobile devices offer to unobtrusively monitor and analyse user behaviour in the wild, it seems natural to extend the framework into the mobile world. Since mobile and desktop systems benefit each other, we keep them as consistent as possible. In fact, with the current implementation it is possible to develop a system on a desktop machine and run it without (or only marginal) modification on a mobile device and vice versa.

The core idea of SSI is to accomplish complex signal processing pipelines from simple reusable units (see Figure 1). To this end, SSI implements a plug-in system to dynamically load pipeline components at run-time. The structure of a pipeline is described using plain XML. Both properties offer a sufficient level of abstraction to define a pipeline independent of the platform it will run on. Nevertheless, we had to face a number of challenges on the way.

Since SSI is written in C++ and was originally developed to run on Windows platforms, porting the core system to Linux was a necessary step to support embedded systems. CMake was chosen as a platform-independent build system. Wherever possible platform-dependent implementations were replaced by platform-independent solutions (e. g. switching threading to C++11 standard). The main challenges, however, arose from the limitations and peculiarities of mobile devices. Due to its wide distribution and open nature, we decided to primarily target Android as mobile operating system. One inherent property of SSI is its strict synchronisation between various processing channels to enable a proper integration of multi-modal information. On a desktop machine with a steady energy supply the primary way of processing information is in form of continuous streams at a fixed sample rate. On mobile

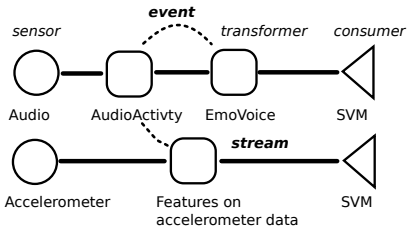

Fig. 1. Components in an SSI pipeline used for laughter recognition in Section III. Data flows in streams and events from one component to another.

devices, however, limited and heterogeneous computing power as well as inaccurate timers and battery usage have to be taken into account. Therefore, it often makes sense to handle signals in a "process-on-demand" fashion, i.e. to process signals only when they convey something meaningful [13]. Hence, representing information in form of events becomes more important on a mobile platform. Here, SSI's event handling system already provides a suited mechanism, though some extensions had to be made, e. g. serialisation of events back into continuous streams.

Integrated sensors are a key feature of mobile devices as they allow us to constantly monitor a user's behaviour without the requirement for extra wiring [8]. In addition, a mobile system can be extended with supplementary sensors worn by the user, as well as stationary ones placed in the surrounding environment. We use a messaging protocol (XMPP or MQTT) with a publish-subscribe model [18] to be able to add sensors dynamically and aim at a more opportunistic approach [15]. To combine XMPP events with information perceived by regular sensors we rely on an asynchronous vector-based fusion approach [5]. SSI also offers an appropriate way to cope with situations in which information is only partly available. The problem with missing data [22] is often ignored in laboratory settings, but represents a typical case in mobile applications where sensors may occasionally be out of range or deliver delayed input due to performance limitations.

To support distributed processing of sensor input over multiple machines and platforms, SSI offers a socket-based interface to start multiple pipelines in-sync and hosts a time server to keep timers from drifting apart. This feature also allows us to outsource heavy processing steps to a desktop computer and immediately receive the result to continue processing. In addition, MobileSSI features a web server for communication via web sockets, for instance, to visualize information in a browser either on the mobile device itself or an external machine in the network.

Like SSI, MobileSSI is open source and available for the public.[1]

III. Laughter Detection in a Daily Environment

To validate our approach, we decided to implement a laughter detection system with MobileSSI. We have already built an enjoyment recognition system based on audiovisual laughter and smile detection [5]. Data acquisition, however,

[1] https://hcm-lab.de/git/project/mobileSSI

Fig. 2. Our mobile setup: Three smart phones placed in breast pockets, clip-microphones.

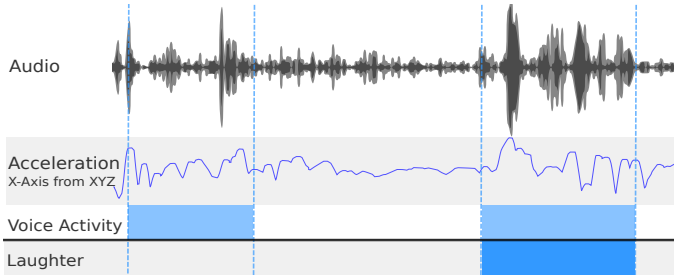

Fig. 3. In addition to audio analysis we capture acceleration, which is a useful indicator of body movement to differentiate laughters from spoken conversation.

was done in a typical static lab setting in which up to four subjects were recorded while telling each other funny stories of their lives [9].

The aim of our present research is to port the existing system to run on mobile phones in order to research the following topics:

- Can we replace the sensor devices of the previous system using solely sensor technology provided on mobile phones?
- Which parts of the signal processing pipeline needs adaptation to work in a less predictable and changing environment?
- Can we expect an acceptable recognition performance?

As a natural environment for our study we picked a pub, as it is a common place for people to meet and have funny conversations. Since we decided to use only hardware that is available on smart phones and continuously provides data for analysis, we could no longer rely on face recognition and depth image processing as in the previous study. Instead we decided to stick to audio and accelerometer sensors. The new setup is depicted in Figure 2 and shows three probands, each of them equipped with a smart phone in their breast pocket connected to a clip microphone. Figure 3 features a signal snippet showing a speech event followed by a laughter event. Throughout the whole session probands were completely free in choosing the topics of their conversations, i.e. we did not specify any guidelines regarding the discussed content. For the experiment, we used Galaxy S4 phones running Android 5.0.1.

First, we set up a pipeline to continuously record audio and accelerometer data. To ensure that captured signals are kept in sync, we relied on SSI's synchronisation techniques (see [23]). We ran two recording sessions on different days and collected a total of four hours of natural conversations per user. To

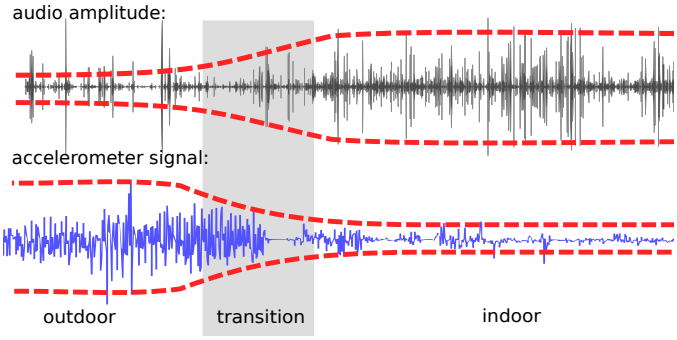

Fig. 4. Change in audio amplitude and accelerometer energy before and after entering the pub.

prevent data loss, we turned off the sleep mode of the mobile phones. Our experiments showed that data can be reliably captured with the sensors provided by the smart phones for up to five hours per charge. Audio was recorded with 16 kHz and acceleration data with 100 Hz. When reviewing the data, we also found a significant amount of laughter (about 50 events per session and user). In total, we extracted 21500 overall samples by using a sliding window of one second and 400 ms frame shift. This figure compares well to the story–telling corpus collected in the lab consisting of 27000 samples. For the annotation task we used the ELAN tool[2] [24].

IV. Results and Discussion

As in our previous study [5], we trained a Support Vector Machine (SVM) classifier for each modality and used leave-one-user-out for evaluation, calculating an average over the results of each users. We make use of the EmoVoice feature extraction algorithms that cover 1451 in total [21] for audio in addition to a series of 14 features (mean, variance, peaks, ...) for motion that we compute for each axis and their first and second derivation. Recognition results are shown in table I. For the audio channel, we found a clear drop from 90% to 67% in recognition accuracy compared to the result obtained in the controlled laboratory setting. The detection rate for the accelerometer data was lower as well yielding 67% compared to 79% obtained with the video modality in the reference study. When altering the window size, audio classification improved for larger frames (up to 78% for frames of 2.0 s) whereas accelerometer detection rates were more stable for smaller frames. An asynchronous fusion scheme using both modalities and an ensemble of features is under work.

	Classification Results in %	
	Audio	**Accelerometer**
Laughter	59.37	57.24
¬ Laughter	74.24	77.54
Average	**66.80**	**67.39**

TABLE I

Compared to the story–telling corpus we found clear differences regarding the quality of the signals. For instance,

[2]http://tla.mpi.nl/tools/tla-tools/elan/

in the pub the captured audio signals were overlaid with diverse sources of noise: music playing in the background, surrounding conversations of varying intensity, utterances of the waitress while taking orders, interferences with mobile network activity etc. These disturbances present great challenges to voice activity detection and audio classification and should be addressed, for instance, by applying noise reduction techniques. Since the environment in a mobile setting is subject to great changes, e. g. when the group is temporarily leaving the pub for a smoke (see Figure 4), noise cancelling schemes are required that are able to dynamically adapt to the current situation. It is important to note that the surrounding sound scape also contains relevant data that should be analyzed to gain further information about the environment and the user's activity. For instance, tailored classification models could be used for outdoor and indoor settings.

Overall, our experiment demonstrated the benefits of MobileSSI when porting existing lab settings into a mobile environment. Since battery life of today's smart phones is sufficient to record and process data in real-time for several hours, we are able to run real-life experiments, which provide better insights on the actual challenges we have to face when applying social signal processing in the wild.

V. Conclusion

With MobileSSI, we presented a tool that brings social sensing to mobile and embedded devices. Our deployment in a real-life setting gave promising results and demonstrated the capability of MobileSSI to run complex signal processing and machine learning tasks locally on mobile devices. Processing data captured "in the wild" is clearly more challenging than analyzing data recorded in laboratory settings. MobileSSI does not only help developers pinpoint these challenges, but also offers a flexible software framework to implement algorithms that are able to address them. In order to cope with partially missing, unreliable or noisy data, we provided an event-based fusion approach that tolerates gaps in data streams. In the CARE project [17], we currently employ MobileSSI to provide elderly people with personalized life style recommendations based on context information about their living environment.

References

[1] K.-h. Chang, D. Fisher, J. Canny, and B. Hartmann. How's my mood and stress?: An efficient speech analysis library for unobtrusive monitoring on mobile phones. In *Proceedings of the International Conference on Body Area Networks*, pages 71–77, 2011.

[2] A. Crossan, G. Lefebvre, S. Zijp-Rouzier, and R. Murray-Smith. A multimodal contact list to enhance remote communication. In R. Murray-Smith, editor, *Mobile Social Signal Processing*, volume 8045 of *LNCS*, pages 84–100. Springer, 2014.

[3] I. Damian, C. S. S. Tan, T. Baur, J. Schöning, K. Luyten, and E. André. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 565–574, 2015.

[4] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the International Conference on Mobile Systems, Applications, and Services*, pages 389–402, 2013.

[5] F. Lingenfelser, J. Wagner, E. André, G. McKeown, and W. Curran. An event driven fusion approach for enjoyment recognition in real-time. In *Proceedings of the International Conference on Multimedia*, pages 377–386, 2014.

[6] H. Lu, D. Frauendorfer, M. Rabbi, M. S. Mast, G. T. Chittaranjan, A. T. Campbell, D. Gatica-Perez, and T. Choudhury. StressSense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the Conference on Ubiquitous Computing*, pages 351–360, 2012.

[7] Y. Ma, B. Xu, Y. Bai, G. Sun, and R. Zhu. Infer daily mood using mobile phone sensing. *Ad Hoc & Sensor Wireless Networks*, 20(1-2):133–152, 2014.

[8] H. Martín, A. M. Bernardos, J. Iglesias, and J. R. Casar. Activity logging using lightweight classification techniques in mobile devices. *Personal and Ubiquitous Computing*, 17(4):675–695, 2013.

[9] G. McKeown, W. Curran, J. Wagner, F. Lingenfelser, and E. André. The belfast storytelling database – a spontaneous social interaction database with laughter focused annotation. In *International Conference on Affective Computing and Intelligent Interaction*, Xi'an, China, Sept. 2015.

[10] S. T. Moturu, I. Khayal, N. Aharony, W. Pan, and A. Pentland. Using social sensing to understand the links between sleep, mood, and sociability. In *Proceedings of the International Conference on SocialCom/PASSAT*, pages 208–214. IEEE, 2011.

[11] A. Muaremi, B. Arnrich, and G. Tröster. Towards measuring stress with smartphones and wearable devices during workday and sleep. *BioNanoScience*, 3(2):172–183, 2013.

[12] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas. Emotionsense: A mobile phones based adaptive platform for experimental social psychology research. In *Proceedings of the International Conference on Ubiquitous Computing*, pages 281–290, 2010.

[13] A. Reiss, G. Hendeby, and D. Stricker. Towards robust activity recognition for everyday life: Methods and evaluation. IEEE, 5 2013.

[14] K. Renaud and H. Crawford. Invisible, passive, continuous and multimodal authentication. In *Mobile Social Signal Processing*, volume 8045, pages 34–41. Springer, 2014.

[15] D. Roggen, A. Calatroni, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, A. Ferscha, M. Kurz, G. Hölzl, H. Sagha, H. Bayati, J. del R. Millán, and R. Chavarriaga. Activity recognition in opportunistic sensor environments. In *Proceedings of the European Future Technologies Conference and Exhibition*, pages 173–174, 2011.

[16] A. Sano and R. W. Picard. Stress recognition using wearable sensors and mobile phones. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, pages 671–676, 2013.

[17] A. Seiderer, S. Hammer, E. Andre, M. Mayr, and T. Rist. Exploring digital image frames for lifestyle intervention to improve well-being of older adults. In *Proceedings of the 5th International Conference on Digital Health 2015*, DH '15, pages 71–78, New York, NY, USA, 2015. ACM.

[18] C.-F. Sørensen, M. Wu, T. Sivaharan, G. S. Blair, P. Okanda, A. Friday, and H. Duran-Limon. A context-aware middleware for applications in mobile ad hoc environments. In *Proceedings of the 2Nd Workshop on Middleware for Pervasive and Ad-hoc Computing*, MPAC '04, pages 107–110, New York, NY, USA, 2004. ACM.

[19] A. Vinciarelli, R. Murray-Smith, and H. Bourlard. Mobile social signal processing: Vision and research issues. In *Proceedings of the International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 513–516, 2010.

[20] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image Vision Computing*, 27(12):1743–1759, 2009.

[21] T. Vogt, E. André, and N. Bee. EmoVoice - a framework for online recognition of emotions from voice. In *Perception in Multimodal Dialogue Systems*, volume 5078, pages 188–199. Springer, 2008.

[22] J. Wagner, F. Lingenfelser, E. André, and J. Kim. Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Transactions on Affective Computing*, 99(PrePrints), 2011.

[23] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André. The Social Signal Interpretation (SSI) framework: multimodal signal processing and recognition in real-time. In *International Conference on Multimedia*, pages 831–834, 2013.

[24] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes. Elan: a professional framework for multimodality research. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, 2006.

[25] S. Zhang and P. Hui. A survey on mobile affective computing. *CoRR*, abs/1410.1648, 2014.