

MobileSSI: asynchronous fusion for social signal interpretation in the wild

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

Angaben zur Veröffentlichung / Publication details:

Flutura, Simon, Johannes Wagner, Florian Lingenfeller, Andreas Seiderer, and Elisabeth André. 2016. "MobileSSI: asynchronous fusion for social signal interpretation in the wild." In Proceedings of the 18th ACM International Conference on Multimodal Interaction - ICMI 2016: Tokyo, Japan — November 12 - 16, 2016, edited by Yukiko I. Nakano, Elisabeth André, and Toyoaki Nishida, 266-73. New York, NY: ACM Press. <https://doi.org/10.1145/2993148.2993164>.

MobileSSI: Asynchronous Fusion for Social Signal Interpretation in the Wild

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

ABSTRACT

Over the last years, mobile devices have become an integral part of people's everyday life. At the same time, they provide more and more computational power and memory capacity to perform complex calculations that formerly could only be accomplished with bulky desktop machines. These capabilities combined with the willingness of people to permanently carry them around open up completely new perspectives to the area of Social Signal Processing. To allow for an immediate analysis and interaction, real-time assessment is necessary. To exploit the benefits of multiple sensors, fusion algorithms are required that are able to cope with data loss in asynchronous data streams. In this paper we present MobileSSI, a port of the Social Signal Interpretation (SSI) framework to Android and embedded Linux platforms. We will test to what extent it is possible to run sophisticated synchronization and fusion mechanisms in an everyday mobile setting and compare the results with similar tasks in a laboratory environment.

CCS Concepts

•Human-centered computing → Open source software; Ubiquitous and mobile computing systems and tools;

Keywords

Social Signal Processing; Affective Computing; "In the Wild" Studies

1. INTRODUCTION

The contemporary relevance of Social Signal Processing (SSP) and emotion recognition can be seen in Microsoft's service for facial emotion recognition or Apple's acquisition of Emotient. However these services run on computer clusters in the cloud and therefore depend on a high bandwidth internet connection, rely on video as the main modality and require private data to be transferred. The target of SSP is



This work is licensed under a Creative Commons Attribution International 4.0 License.

Copyright is held by the owner/author(s).

ICMI'16, November 12–16, 2016, Tokyo, Japan
ACM. 978-1-4503-4556-9/16/11...\$15.00
<http://dx.doi.org/10.1145/2993148.2993164>

to bring social intelligence to computers. It is based on models of human-human interaction. Such models are difficult to develop just within laboratories, therefore SSP benefits from corpora collected outside the lab, "in the wild". Mobile devices allow us to collect data in a natural and unobtrusive manner and thus offer new perspectives and opportunities to SSP [31]. In particular, the following characteristics of mobile devices may be of benefit to SSP:

- 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 behavior and derive context information.
- Mobile devices are small and lightweight. They can be carried around for an extended period of time, which suits long-term and in-situ recording. 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 behavior. Mobile devices are worn in pockets most of the time and motion sensing using accelerometers is the modality of choice. But the amount of integrated sensors grows as well as computational power that enables online processing of social signals when they occur. As a consequence, time has come to exploit fusion techniques on mobile devices that combine the benefits of multiple modalities to compensate for the deficiencies of others. Fusion techniques are of particular relevance to mobile settings which typically have to face issues with data loss that might be mitigated by drawing on the modalities that show the best performance in a particular situation. The objective of this paper is to bring novel fusion techniques to mobile SSP. To this end, we developed MobileSSI, a port of the Social Signal Interpretation (SSI) framework [33] to Android and embedded Linux platforms. MobileSSI provides developers with tools to record and analyze human behavior in real-time on mobile devices. In addition, it offers an event-based fusion mechanism that is able to handle gaps in asynchronous data stream and therefore shows particular promise for mobile settings.

After a discussion of related work in Section 2, the paper will describe the technical realization of MobileSSI in Section 3 focusing on the engineering process of porting the existing SSI framework to mobile devices. To validate the approach, we tested MobileSSI in a real-life setting focusing on multimodal laughter recognition for groups. The study enables us a direct comparison with a previous laboratory study addressing the same topic. As a study in the wild, it bears advantages for laughter induction compared to laboratory studies, where careful design and laugh induction techniques have to be applied. As the range of available sensors on mobile devices is very limited in comparison to a laboratory setting, we discuss how to replace classical input modalities, such as video capturing and facial analysis, with modalities that can easily and permanently be obtained from smartphones (e. g. accelerometer data) and do not limit the users in their mobility. In Section 5, the results of our experiments are discussed focusing on challenges that we identified for mobile SSP. The paper ends with a conclusion in Section 6 and future work in Section 7.

2. RELATED WORK

Sensor-equipped mobile phones have promoted new forms of social sensing since they allow us to collect data of human-human interactions over extended periods of time in the wild. Such data can provide useful insights on social behavior patterns as well as psychological user states, such as mood and emotion. In the literature, a wide range of features has been investigated including the amount of conversation recorded by the smartphone’s microphone [5], communication data [14], postures [10] or proximity behaviors detected by Bluetooth patterns [2]. 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 indicator [30, 14, 24, 18]. Furthermore, attempts have been made to detect stress from the user’s voice in natural environments using the microphones on smartphones.

Devices that can be worn around the wrist today feature more computing power than high end desktop machines two decades ago. Nevertheless many approaches still focus on offline analysis. That is they employ the sensors integrated in the mobile phones to record data of human-human interactions in mobile settings. However, the acquired data are analyzed offline after transferring the data to servers. Typically, data conveying behavior, such as acceleration, skin conductance or voice, is obtained from mobile phones or wrist sensors and uploaded to a server for conducting statistical analysis. Typical applications include life logging systems, such as the smile and laughter detector presented by Fukumoto et al. [11]. Also a number of health care systems follow this approach. For example, Moturu et al. [23] transfer mobile data to a server to compute correlations between mood and human behaviors including their physical activity, social interactions and sleep.

Only recently, attempts have been made to analyze social cues in online mode in order to provide users with real-time feedback. MoodScope [14] trains predictive mood models in the cloud using data collected with mobile phones. The trained models are then transferred to the mobile phones to enable inferences on the user’s mood in online mode. Damian et al. [7] present a portable online system called Logue that provides real-time feedback about the quality of

a presenter’s performance in public speaking. Recommendations are automatically derived by analyzing openness, body energy and speech rate. Then recommendations are presented through a wearable display, such as Google Glass. Rachuri et al. [27] present a wearable system called SocialSense that monitors the users’ social interactions in an office environment in order to provide them with feedback to improve their sociability. In order to balance the trade-offs between energy consumption, performance and data traffic, the system offers a mechanism for adaptive sampling and distributed computation. In particular, the system decides in a context-sensitive manner whether to conduct computation tasks on the mobile device or on a remote server.

In the past, mobile social sensing was based on lightweight algorithms. With increasing computing and storage capabilities of mobile devices, more and more complex tasks can be run locally on these devices. As a consequence, a number of platforms have been developed that facilitate the development of mobile social sensing applications. Examples include StressSense [17] and AMMON [5], two platforms that offer feature extraction functionalities for vocal emotion recognition running on mobile devices. Encouraged by the recent success of deep neural networks in the area of audio and video processing, Lane et al. [13] conducted a number of experiments to explore the potential of deep learning in the mobile context. Their experiments revealed that deep learning techniques have the potential to increase robustness of various mobile sensing tasks, such as activity, emotion and speaker recognition, without requiring an unrealistic amount of resources. However, they relied on previously acquired mobile sensing data and did not yet employ their deep learning framework in the wild.

Despite of the increasing variety of sensors, fusion mechanisms running on mobile phones employ rather simple rules and treat multiple modalities usually in a strict complementary manner. For example, the above mentioned Logue system analyzes speech and body movements of a speaker. However, the two modalities are not fused in a synergistic manner, but handled separately to analyze the speaker’s performance. This observation is also reported in a recent survey paper by Palaghias et al. [26] who give a comprehensive overview of frameworks and systems that analyze human social data in the mobile context. The current paper aims to close this gap by providing tools to filter, transform, classify and fuse data streams in real-time, locally on mobile devices. We will also test to what extent it is possible to run sophisticated synchronization and fusion mechanisms in an everyday mobile setting and compare the results with similar tasks in a laboratory environment.

3. MOBILE SSI

Providing developers with tools to record, analyze and recognize human behavior in real-time on mobile devices has been our driving force to port the Social Signal Interpretation (SSI) framework [33] to run on mobile platforms. The SSI framework aims at closing the gap between offline analysis and the development of online systems. Therefore it provides an architecture that not only provides tools for data recording, feature extraction 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 un-

obtrusively monitor and analyze user behavior in the wild, it seems natural to extend the framework into the mobile world. Since mobile and desktop systems are a benefit to 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. These units can for example be sensors, transformers computing features on the signal or consumers for output or classification. An example pipeline is shown in Figure 1. Features are extracted from

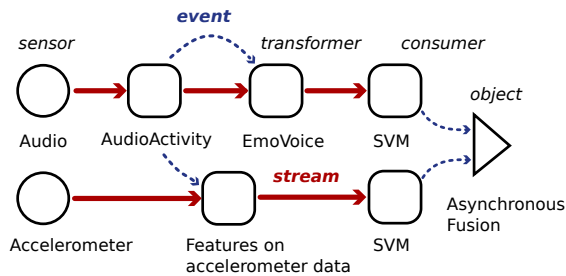


Figure 1: Components in an SSI pipeline used for laughter recognition in Section 4. Data flows in streams and events from one component to another.

the raw streams (see Section 3.2) when voice activity is detected in the audio channel. Support Vector Machine (SVM) classifiers recognize laughter events in the two channels that are combined using an asynchronous fusion scheme (see Section 3.3). SSI implements a plug-in system to dynamically load pipeline components at run-time. The structure of the 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.

3.1 Porting

Since SSI is written in C++ and was originally developed to run on Windows, porting the core system to Linux was a necessary intermediate step towards supporting Android. 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.

3.2 Features

One inherent property of SSI is its strict synchronization between various processing channels to allow 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 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. processing signals only when they convey something meaningful [28]. Hence, representing information in form of events becomes more important on a mobile plat-

form. Here, SSI's event handling system already provides a suited mechanism, though some extensions had to be made, e.g. serialization of events back into continuous streams.

Integrated sensors are a key feature of mobile devices as they allow us to constantly monitor the users' behavior without the requirement for extra wiring [20]. 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. By communicating with the Android Java API we are able to integrate Bluetooth devices. To handle Java sensors adequately, we integrate SSJ – a Java reimplement of SSI [6]. SSJ handles Java threads and components and again functions e.g. as a sensor for a native MobileSSI pipeline. For external sensors, we use a messaging protocol (XMPP) with a publish-subscribe model to be able to add sensors dynamically and aim at a more opportunistic approach [29]. Following raw data acquisition is feature extraction. Plugins for filtering and feature extraction are organized as transformers. These range from generic filters such as mean, derivation and the Butterworth filter to specific collections of features such as OpenSMILE [9] and EmoVoice [32] for audio processing.

Higher levels of an application depend on conclusive results rather than signal features. MobileSSI therefore integrates machine learning. Using model abstraction, different algorithms can be tested transparently. We mainly rely upon Support Vector Machines [4], but also have more advanced approaches, such as hierarchical classification [34]. The machine learning plugin supports automated structuring of data as well as automatic evaluation of trained models. These mechanisms serve as a basis and are tightly integrated with the asynchronous fusion approach described next.

3.3 Asynchronous Fusion

Originally, research on multimodal interfaces has focused on fusion mechanisms that integrate the meaning of multiple modalities, such as speech and pointing gestures, into a uniform representation (see [12] for a survey). To analyze human social behavior, fusion mechanisms are required not only at the semantic level, but also at the level of low-level social cues and high-level classes representing typical patterns of social behavior. A discussion of fusion approaches that employ different levels of abstraction is provided in [1]. Here, we focus on fusion at an intermediate level and propose events as meaningful interpretation units that lie between low-level features and high-level classes. Events are interpreted as short-termed cues that point to a searched target class. An ensemble of trained machine learning models is used to recognize these events in the available modalities.

To cope efficiently with changing sources of information that depend on sensors available and information emitted by the surrounding, we adopt an asynchronous fusion scheme. This fusion approach does not force decisions from all available channels for every time frame, but instead correlates occurrences of small windows of relevant information over time. Other ways of fusing modalities without steadily forcing decisions have been successfully investigated in previous research. Zeng et al. [36] apply Multi-stream Fused Hidden Markov Models, in which state transitions of different components of Hidden Markov Models are allowed to occur at differing times across multiple streams. Dupont et al. [8] model the asynchronous nature of audio and video streams using temporal topologies with multi-stream Hidden

Markov Models for continuous speech recognition. Methods pursuing a hand-modeled approach for the asynchronous fusion of streams using Petri-nets are applied by Navarre et al. [25]. Long Short-Term Memory Neural Networks have shown great success in paralinguistic tasks (see [3, 35]). They were used to replace traditional nodes with memory cells that allow the network to learn when to store or relate to bimodal information over long periods of time.

Asynchronous fusion on event level has proven to be robust in affect recognition scenarios [16]. By monitoring occurrences of events over time, the higher level fusion plugin is able to decide what is going on at any point in time. This strategy provides an abstraction level that allows for easy adaption, as modalities that are able to provide events for the event based fusion algorithm can be easily added or removed. Therefore it is a good fit for in the wild signal processing, where it is not guaranteed to have all sensors available at all times (differing hardware, noisy data, energy consumption, etc.). Recognized events are initially weighted with regard to the confidence of the classification model and this weight is constantly decreased so that their influence on the final fusion result descends to zero over time. Currently active events give an appropriate overall picture that again is judged by the fusion model on the event level by calculating the center of mass based on currently active events and their updated weights (see Figure 2). This solves not only prob-

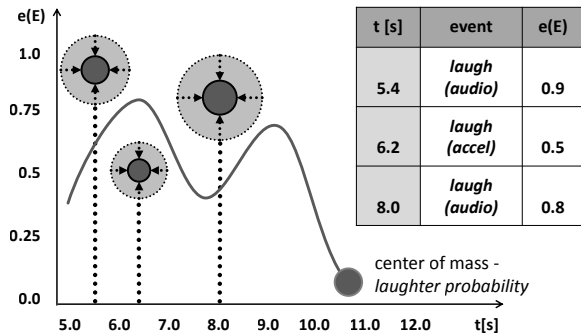


Figure 2: The fusion algorithm considers the temporal flow of laugh events. Their influence decreases over time. A framewise laughter probability is calculated as the center of mass of weighted events.

lems of different sample rates and segmentation windows on multiple streams. But it is especially of use when giving different relevance judgments to events in multiple modalities. From an engineering point of view, only transmitting events is leading to much lower network load than sending a raw data stream.

3.4 Networking

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 enables us to outsource heavy processing steps to a desktop computer and to immediately receive the result to continue processing. Beneath XMPP mentioned earlier, the MQTT protocol was integrated for low overhead communication that is especially of use for microcontrollers within the internet of things. In addition, MobileSSI features a webserver 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.

4. VALIDATION IN THE WILD

As a real world test we decided to conduct a laughter recognition study in an everyday setting. Laughter detection is a classic problem when it comes to social cues. Indeed, we have already built an enjoyment recognition system based on audiovisual laugh and smile detection [16]. Data acquisition, however, was done in a typical stationary lab setting in which up to four study participants were recorded while telling each other funny stories of their lives [22]. Now, our aim is to port the existing system to run on mobile phones to investigate the following questions:

- Can we replace the sensors of the previous system using solely sensor technology provided by mobile phones?
- Which parts of the SSP pipeline need adaption to work in a less predictable and changing environment?
- Can we expect a comparable recognition performance?

In principle, we could use cameras again for detecting visual laughter. However, we would have to either place them in the environment (which would limit the user’s mobility) or to attach them to the user. In the latter case, only visual laughter of the user’s interlocutors could be captured. Of course, the camera could also be attached in a way that it faces the user. However, this setup would result into a rather bulky device. Consequently, we had to find another solution. Accelerometers seem to be a promising option. Indeed there is evidence from previous work using visual markers [19] or a complete motion capture suit [21] that motion is a good indicator of laughter. Therefore we decided to replace the Kinect cameras that were used in the previous lab setting with accelerometers. For the audio modality no replacement was necessary since external microphones can be used to circumvent interferences from the pockets. The advantage of the new setup is that it uses only hardware that is available on smartphones or very easy to attach (microphones) and can be continuously assessed and analyzed.

4.1 Corpus

As a natural environment for our study we picked a pub, as it is a common place for people to meet and have enjoyable conversations. As described above, we decided to stick to audio and accelerometer sensors. The new setup is depicted in Figure 3 and shows three study participants, each of them equipped with a smartphone in his breast pocket connected to a clip microphone. The participants were acquired beforehand and given a brief introduction on how the setup worked. Apart from starting the session, no further interaction with the system was required from them. Throughout the session the participants were completely free in choosing the topic of their conversation, i.e. we did not give them any guidelines on the content to be discussed. Audio was recorded at 16 kHz, as it is the sample rate delivering the most reliable results on our target system vs. 48 kHz in the reference study. Accelerometer data were sampled at 100 Hz. For the study we used Samsung Galaxy S4 (GT-I9505) phones running Android 5.0.1 (latest official version).

First, we set up a pipeline to continuously record audio and accelerometer data and relied on SSI’s synchronization techniques [33] to ensure that captured signals are kept in



Figure 3: Our mobile setup: Three smartphones placed in breast pockets, clip-microphones.

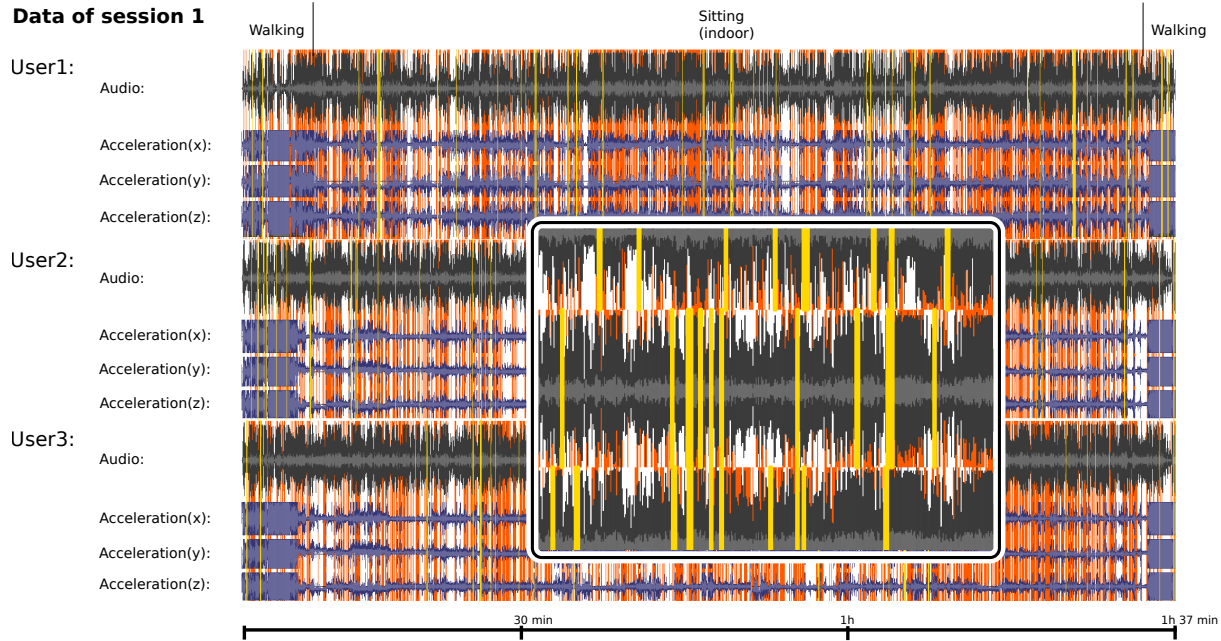


Figure 4: Overview of one session. Raw data containing audio and acceleration are plotted synchronized. Laugh (yellow) and talk (orange) events are marked. The zoom shows synchronized laughter between users.

sync. Figure 5 features a synchronized signal snippet showing speech followed by a laugh event. We ran two recording

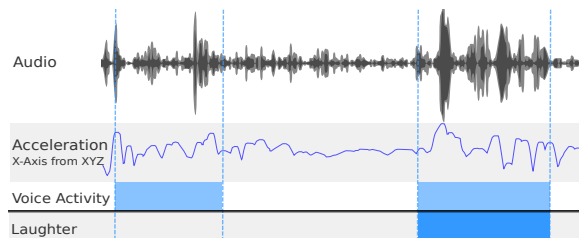


Figure 5: In addition to audio, we capture acceleration as an indicator of laughter.

sessions on different days and collected a total of about 3.5 hours of natural conversations per user. Our experiments showed that data can be reliably captured with the sensors provided by the smartphones for up to eight hours per charge. Feature processing of six hours and online recognition for seven hours is possible with one charge. In total we

extracted 21500 samples by using a sliding window of one second and 400 ms frame shift whereof 875 contain laughter. In comparison, the corpus acquired in the reference study [16] contains 27000 samples with the same window and shift. Audio was used as ground truth to annotate laughter on both modalities. Figure 4 shows that laughs are indeed infectious. Laughter of one person (indicated in yellow) is immediately followed by the others.

4.2 Features

In order to recognize cues for laughter in the observed channels, we need to extract relevant features from the segments of raw data. For audio we use the EmoVoice feature set (1451 in total) [32] - containing MFCCs, pitch, energy and more. For laughter recognition in audio data MFCCs have proven good indicators - not surprisingly as they are a useful tool in speech recognition and laughter has a lot in common with phonemes. For accelerometer data, we compute a series of nine features (listed in Table 1) for each of the three axis. We add the first and the second derivation

for each calculated feature, resulting in a feature vector of size 81 for the accelerometer modality.

Features on Accelerometer Data
Mean, Standard deviation
Minimum, Maximum, Range
Zero crossing rate
Peak count, Pulse rate
Energy

Table 1: Features used on accelerometer data, on each axis, additionally on first and second derivation of the signal.

4.3 Evaluation

Evaluation is carried out frame-wise over the two recorded sessions. As both sessions feature the same users, we decided to use two persons for training of recognition and fusion systems and keep the third for testing. This evaluation approach simulates the performance of an online system and allows us to draw a direct comparison to the reference system [16] evaluated the same way. To get a first impression of recognition performance, we train one SVM-model for each of the two modalities, audio and smartphone acceleration separately with two classes (laughter and no-laughter). Frame-wise recognition results are shown in Table 2. The tables present unweighted recognition results (average accuracy across classes), because the number of frames actually containing laughter is of course significantly lower than frames that show no hints of laughs. This prevents high detection rates by only favoring the dominant class (weighting the average with the classes’ sample count).

	Uni-Modal Classification	
	Accelerometer	Audio
Laughter	80.95 %	76.19 %
¬ Laughter	63.42 %	86.70 %
Average	72.19 %	81.45 %

Table 2: Results of classification per modality.

While the reference system reached an unweighted accuracy of up to 90 % for laughter recognition on audio frames, we can now observe a clear drop to 81 % in recognition accuracy. The detection rate for the accelerometer data was lower, too, yielding 72 % compared to 79 % obtained with the video modality in the laboratory study.

	Multi-Modal Classification	
	Decision Fusion	Event Fusion
Laughter	78.57 %	83.33 %
¬ Laughter	86.62 %	85.95 %
Average	82.59 %	84.64 %

Table 3: Results of classification fused using decision- and event-driven solutions

In order to compare the performance of the proposed event-based fusion approach we also applied a very basic decision-level fusion strategy (see Table 3). Decision-level

fusion using the product rule [15] improved the results by one percent point over uni-modal classification and scored 82.59 %. As a second method, asynchronous fusion on event-level features (Section 3.3) was conducted and improved the classification by three percent points to 84.64 %. Instead of fusing information over fixed time segments, recognized events are integrated frame by frame. To this end, the following parameters are taken into account. Each event is assigned a modality-specific weight to emphasize more reliable information sources. A decay parameter determines how fast the influence of events on the fusion result decreases (see Figure 2). If a particular threshold is achieved, a frame is classified as laughter. The optimal configuration of these parameters was learned on the training data by systematically testing parameter combinations following the grid search approach described in [16]. Within 12000 combinations of parameters, based on our previous research and additional adjustments for the new setting, 18 configurations were found that scored an average of 84 % detection rate. These configurations give events from the audio modality a higher influence (0.7 or 1.0) while accelerometer events are weighted lighter at 0.1 to 0.3. Audio and acceleration decay parameters are comparable and vary from 0.6 to 1.0 (audio) and 0.5 to 1.5 (acceleration). This is plausible as audio is the modality with better classification results in Table 2, therefore can be relied upon more and faster, while accelerometer events make a better contribution if they are weighted less.

5. DISCUSSION

Compared to the story-telling corpus, we found clear differences regarding the signal quality. For instance, the audio signals captured in the pub 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

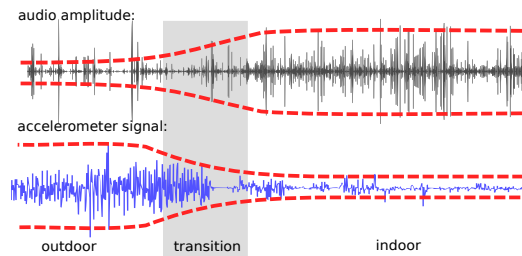


Figure 6: Change in audio amplitude and accelerometer energy before and after entering the pub.

to great changes, e.g. when the group initially enters / finally leaves the pub or is just temporarily leaving the pub for a smoke, noise cancellation schemes are required that are able to dynamically adapt to the current situation. Examples of such changes can be noticed at the start / end of session 1 visible in Figure 4 and in Figure 6 where a transition is shown in detail. On the other hand, the surrounding soundscape may contain 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 moving from a lab setting to a mobile environment. Our classification results are clearly lower than those obtained in the lab. However, techniques based on event fusion narrow the gap compared to uni-modal classification. The smile and laugh detector by Fukumoto et al. [11] obtained recognition rates of 89.2%. However, they had people watch videos of ten minutes only while we investigated social interactions over several hours in a mobile setting. Also they used a setup including glasses equipped with photo interrupters and relied on a PC for online processing whereas in our case all the computing was done on the phone. Since battery life of today’s smartphones 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.

6. CONCLUSION

With MobileSSI, we presented a tool that brings SSP techniques to mobile and embedded devices. MobileSSI provides a flexible interface for interacting with multiple wearable sensing devices in a real-time and synchronized fashion while not constricting the user’s mobility. Our deployment in a real-life setting gave promising results and demonstrated its capability to run complex signal processing and machine learning tasks locally on mobile devices. Processing data captured in the wild is clearly more challenging compared to the analysis of 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 scheme that introduces events as an abstract intermediate layer and effectively decouples unimodal processing from the final decision making. Each modality serves as a client which individually decides when to add information. Signal processing components can be added or replaced without having to touch the actual fusion system, and missing input from one of the modalities does not cause the collapse of the whole fusion process. By exploiting multiple modalities, MobileSSI enables us to collect a large variety of behavioral cues from lower-level social cues to higher-level social group dynamics. In addition, it may be used to acquire a rich amount of context information. By correlating social cues with context information, a more holistic picture about social interactions is obtained that may provide social scientists with useful insights. For better transparency and as a contribution to the community MobileSSI is open source and available to the public¹.

7. FUTURE WORK

Several shortcomings of SSI on mobile platforms can be predicted as inevitable obstacles, such as energy efficiency, autonomous adaption to unforeseen events and annotation of masses of data in collaboration with the user. Smart filter algorithms are required that adapt to the current situation to handle noise sources in diverse situations. A classification of the surrounding environment (e.g. busy street vs.

quiet park) might help online recognition pipelines better fit the current situation. Apart from laughter, our corpus contains natural conversations and noises from the environment that invite to be used for data exploration. Speaking time per person and session as well as the proportion of laughter per person are information that can be easily extracted using our setup and application. In addition, we will conduct additional experiments with the analysis of social group dynamics to further validate and improve the MobileSSI framework. So far, the event-based fusion approach has been employed to integrate social cues of an individual. In the future, we will investigate to what extent social cues of a person can be predicted from the social cues of the surrounding interlocutors. Related studies can be found in the literature, but most of them have been conducted in stationary settings. In this paper, we presented experiments with a duration of a few hours. In the future, we plan to conduct more long-term studies over several weeks or even months. The amount of data that can be captured within these time frames will enable us to refine our assessment of human behaviors in the wild and test data-intensive technologies, such as deep learning.

8. ACKNOWLEDGMENTS

The work described in this paper is partially funded by the European Union under research grants H2020-RIA-645012 (KRISTINA) and H2020-RIA-645378 (ARIA-VALUSPA).

9. REFERENCES

- [1] E. André, J.-C. Martin, F. Lingenfelser, and J. Wagner. Multimodal fusion in human-agent dialogue. In M. Rocj and N. Campbell, editors, *Coverbal Synchrony in Human-Machine Interaction*, pages 387–410. CRC Press, 2014.
- [2] G. Bauer and P. Lukowicz. Can Smartphones Detect Stress-Related Changes in the Behaviour of Individuals? In *PerCOM ‘12 Workshops*, pages 423–426. IEEE, 2012.
- [3] R. Brueckner and B. Schuller. Social Signal Classification using Deep BLSTM Recurrent Neural Networks. In *ICASSP ‘16*, pages 4823–4827. IEEE, 2014.
- [4] C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.
- [5] 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 *BodyNets ‘11*, pages 71–77. ACM, 2011.
- [6] I. Damian, T. Baur, and E. André. Measuring the Impact of Multimodal Behavioural Feedback Loops on Social Interactions. In *ICMI ‘16*. ACM, 2016.
- [7] 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 *CHI ‘15*, pages 565–574. ACM, 2015.
- [8] S. Dupont and J. Luetttin. Audio-visual Speech Modeling for Continuous Speech Recognition. *IEEE Trans. Multimedia*, 2(3):141 – 151, 2000.

¹<https://hcmllab.github.io/mobileSSI/>

- [9] F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Multimedia '13*, pages 835–838. ACM, 2013.
- [10] S. Feese, B. Arnrich, G. Tröster, B. Meyer, and K. Jonas. Detecting posture mirroring in social interactions with wearable sensors. In *ISWC '11*, pages 119–120. IEEE, 2011.
- [11] K. Fukumoto, T. Terada, and M. Tsukamoto. A smile/laughter recognition mechanism for smile-based life logging. In *AH '13*, pages 213–220. ACM, 2013.
- [12] D. Lalanne, L. Nigay, p. Palanque, P. Robinson, J. Vanderdonckt, and J.-F. Ladry. Fusion Engines for Multimodal Input: A Survey. In *ICMI '09*, pages 153–160. ACM, 2009.
- [13] N. D. Lane, P. Georgiev, and L. Qendro. DeepEar: Robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In *UbiCom '15*, pages 283–294. ACM, 2015.
- [14] R. LiKamWa, Y. Liu, N. D. Lane, and L. Zhong. MoodScope: Building a Mood Sensor from Smartphone Usage Patterns. In *MobiSys '13*, pages 389–402. ACM, 2013.
- [15] F. Lingenfelder, J. Wagner, and E. André. A systematic discussion of fusion techniques for multi-modal affect recognition tasks. In *ICMI '11*, pages 19–26. ACM, 2011.
- [16] F. Lingenfelder, J. Wagner, E. André, G. McKeown, and W. Curran. An Event Driven Fusion Approach for Enjoyment Recognition in Real-time. In *Multimedia '14*, pages 377–386. ACM, 2014.
- [17] 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 *UbiCom '12*, pages 351–360. ACM, 2012.
- [18] 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.
- [19] M. Mancini, G. Varni, D. Glowinski, and G. Volpe. Computing and evaluating the body laughter index. In *Human Behavior Understanding*, volume 7559, pages 90–98. Springer, 2012.
- [20] 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.
- [21] G. McKeown, W. Curran, C. McLoughlin, H. J. Griffin, and N. Bianchi-Berthouze. Laughter induction techniques suitable for generating motion capture data of laughter associated body movements. In *FG '13*, pages 1–5. IEEE, April 2013.
- [22] G. McKeown, W. Curran, J. Wagner, F. Lingenfelder, and E. André. The Belfast Storytelling Database – A Spontaneous Social Interaction Database with Laughter Focused Annotation. In *ACII '15*. IEEE, 2015.
- [23] 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 *SocialCom/PASSAT*, pages 208–214. IEEE, 2011.
- [24] 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.
- [25] D. Navarre, P. Palanque, R. Bastide, A. Schyn, M. Winckler, L. P. Nedel, and C. M. D. S. Freitas. A Formal Description of Multimodal Interaction Techniques for Immersive Virtual Reality Applications. In *INTERACT '05*, pages 170–183. Springer, 2005.
- [26] N. Palaghias, S. A. Hoseinitabatabaei, M. Nati, A. Gluhak, and K. Moessner. A Survey on Mobile Social Signal Processing. *ACM Comput. Surv.*, 48(4):57:1–57:52, Mar. 2016.
- [27] K. K. Rachuri, C. Mascolo, M. Musolesi, and P. J. Rentfrow. SociableSense: Exploring the trade-offs of adaptive sampling and computation offloading for social sensing. In *MobiCom '12*, pages 73–84. ACM, 2011.
- [28] A. Reiss, G. Hendeby, and D. Stricker. Towards Robust Activity Recognition for Everyday Life: Methods and Evaluation. In *PervasiveHealth '13*. IEEE, 2013.
- [29] 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 *2nd European Future Technologies Conference and Exhibition (FET)*, pages 173–174, 2011.
- [30] A. Sano and R. W. Picard. Stress Recognition Using Wearable Sensors and Mobile Phones. In *ACII '13*, pages 671–676. IEEE, 2013.
- [31] A. Vinciarelli, R. Murray-Smith, and H. Bourlard. Mobile Social Signal Processing: Vision and Research Issues. In *MobileHCI '10*, pages 513–516. ACM, 2010.
- [32] T. Vogt, E. André, and N. Bee. EmoVoice - A Framework for Online Recognition of Emotions from Voice. In E. André, L. Dybkjær, W. Minker, H. Neumann, R. Pieraccini, and M. Weber, editors, *Perception in Multimodal Dialogue Systems*, volume 5078, pages 188–199. Springer, 2008.
- [33] J. Wagner, F. Lingenfelder, T. Baur, I. Damian, F. Kistler, and E. André. The Social Signal Interpretation (SSI) Framework: Multimodal Signal Processing and Recognition in Real-Time. In *Multimedia '13*, pages 831–834. ACM, 2013.
- [34] J. Wagner, A. Seiderer, F. Lingenfelder, and E. André. Combining Hierarchical Classification with Frequency Weighting for the Recognition of Eating Conditions. In *INTERSPEECH '15*, pages 889–893. ISCA, 2015.
- [35] M. Wöllmer, B. Schuller, F. Eyben, and G. Rigoll. Combining Long Short-Term Memory and Dynamic Bayesian Networks for Incremental Emotion-Sensitive Artificial Listening. *J. Sel. Topics Signal Processing*, 4(5):867–881, 2010.
- [36] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang. Audio-Visual Affective Expression Recognition Through Multistream Fused HMM. *Trans. Multi.*, 10(4):570–577, 2008.