Automatic Sensing of Speech Activity and Correlation with Mood Changes

Aleksandar Matic, Venet Osmani, and Oscar Mayora

Ubiquitous Interaction Group, CREATE-NET,
Trento, Italy
{aleksandar.matic,venet.osmani,oscar.mayora}@create-net.org

Abstract. The association between social relationships and psychological health has been established fairly recently, in the last 30-40 years, relying on survey-based methods to record past activities and the psychological responses in individuals. However, using the self-reporting methods for capturing social behavior exhibits a number of shortcomings including recall bias, memory dependence, and a high end user effort for a continuous long-term monitoring. In contrast, automated sensing techniques for monitoring social activity, and in general, human behavior, has a potential to provide more objective measurements thus to overcome the shortcomings of self-reporting methods. In this paper, we present a privacy preserving approach to detect one component of social interactions - the speech activity, through the use of off-the-shelf accelerometers. Furthermore, we used the accelerometer based speech detection method to investigate the correlation between the amount of speech (which is an aspect that reflects the participation in verbal social interactions) and mood changes. Our pilot study suggested that verbal interactions are an important factor that has an impact on individuals' mood, while the study also demonstrated the potential of automated capturing social activity comparable to the use of gold standard surveys.

Keywords: speech activity detection, wearable computing, emotional response, mood changes.

1 Introduction

Throughout the history, the relationship between social interactions and emotions was an aspect of humanity analyzed by intellectuals from nearly all disciplines, from psychologists and philosophers to artists and poets. For instance, one of the most famous Shakespeare's plays, Othello, portrays characters from different backgrounds whose happiness depends mostly on social interactions. However, scientific evidence on the association between social relationships and psycho-physical health has been established fairly recently – the late 1970s and the 1980s [1]. The methods for collecting interaction data, which were used at that time and that are still prevalent in social and health related sciences, include surveys or human observers who continuously take notes on social behaviour of the

S.C. Mukhopadhyay et al. (Eds.): Pervasive & Mob. Sens. & Comput. for Healthcare, SSMI 2, pp. 195–205.
springerlink.com

monitored subjects. Survey-based methods exhibit a number of drawbacks including recall bias, difficulties in recalling activities that have occurred in the past, influence of the current mood [2], and a high end user effort for continuous long-term monitoring. In addition, self-reports correspond poorly to communication behaviour recorded by independent observers [3]. On the other hand, engaging human observers to record communications in groups is inefficient if the size of the group is large or if the interactions occur in various physical locations [4].

The use of automated sensing techniques for capturing social activity has been explored in the past decade, in order to address the limitations of self-reporting and observational methods. Recognizing the occurrence of social interactions in an automatic way is typically based on sensing proximity of subjects and/or on detecting speech activity. Since solely physical proximity does not always provide enough evidence for inferring social interaction [5] (for example, two colleagues sitting across from each other in the office and not interacting), methods for detecting social interactions usually include audio data analysis. This requires the activation of microphone that is either mounted in a monitored area or embedded in a mobile device (the mobile phone [6] or specialized device such as Sociometer [4]). However, in a number of situations (for example, in public spaces or in the case of monitoring patients) audio data cannot be obtained due to legal or ethical issues [7]. Moreover, activating microphone typically raises privacy concerns in subjects, even with methods that do not require continuous voice recording, thus affecting their natural behaviour.

In this work, we present the approach of detecting speech activity using an off-the-shelf accelerometer intended to identify another manifestation of speech different than voice, namely the vibration of vocal chords. The vibrations, caused by phonation, spread from the area of larynx to the chest level, which is a convenient place for attaching a sensor with an elastic band (similarly to attaching respiratory or cardio sensors) minimizing the interference with typical daily routines. We envision that the accelerometer-based speech detection can complement methods for automatic recording of social interactions, as an alternative to audio data analysis through microphones that may raise privacy concerns in subjects thus affecting their natural behaviour.

We used the accelerometer based speech detection method to investigate the correlation between the amount of speech, which is one aspect that reflects the participation in verbal social interactions, and mood changes. The current literature in social psychology reports several studies that examined how the social activity impacts the mood during the day [8] [9] [10] [11] [12], however none relied on the automated methods for collecting data. Despite being a pilot study with 10 subjects, the experimental setting was fully unconstrained yielding results that are consistent with the previous research [10][11][12] that reported positive association between the mood dimension of PA (positive affect) and the amount of speech activity.

2 Detecting Speech Activity with an Accelerometer

2.1 Privacy Issues in Interaction Data Collection

When detecting speech activity and, in general, sensing social interactions, an important issue is regard for subjects' privacy [13]. Capturing natural behavior pertains to recording people as they freely go about their lives, however there is a typically trade-off between the quality of collected data and the level of respecting subjects' privacy [6].

Privacy issues relate to an array of ethical norms that needs to be addressed. All subjects in the study should always know that they are being monitored, moreover they must have the right to authorize the use and the diffusion of the collected data [13]. If monitoring involves audio archives, they can be partially or totally deleted by subjects while recording uninvolved parties without their consent is considered unethical and illegal [6]. Despite addressing all the ethical norms, people are prone to change their behaviour if they have concerns about the way of monitoring, which negatively affects the reliability of the collected data. In the case of old methods such concerns can be raised due to a human observer, while for sensor-based approaches the presence of audio data analysis is critical. When automatically recording social behaviour, protecting privacy often implies discarding useful information [6] which is not always acceptable trade-off thus in some studies even the raw audio data was analyzed [14]. In response to this, several experimental designs included privacy-sensitive recording techniques, however the fact that microphone is activated may still raise concerns thus affecting subjects' natural behaviour. This often depends on the technical education and cultural background of monitored subjects, which can affect the perception of privacy [15] [16].

2.2 Our Approach

As an alternative to microphone-based methods for detecting speech, a few studies aimed to infer speech activity based on mouth movement, fidgeting, or gestures [7] [17] detected using machine vision. However, this limits application scenarios to areas that are covered with the camera system. Therefore, we attempted to avoid audio or visual cues for detecting speech activity, which led us to investigate physiological effects of phonation – vocal chords vibrations.

After the age of 20 the predicted fundamental frequency of vocal chords remains approximately 100Hz for male and 200Hz for female adults [18]. Therefore, identifying vibrations of these fundamental frequencies produced by vocal chords during phonation pertains to speech activity detection. Instead of a purpose-built accelerometer (with an appropriate shape, targeted frequency range and sensitivity), we investigated the use of an off-the-shelf accelerometer thus aiming for an easily applicable and cost effective solution. Since mounting sensors on the neck (close to the larynx area) may be too obtrusive, we selected the chest surface, in particular the central part of the sternum which is the area with the highest displacement amplitude of vocal chords vibrations [19]. This position is also convenient for attaching a sensor with an elastic bend (similarly to attaching respiratory or cardio sensors) minimizing the interference with typical daily routines.

2.3 Data Analysis

In our experiments we used Shimmer accelerometer [20] attached at the chest level to analyze frequency spectrum. The sensor specifications are as follows: the range of ±1.5 and ±6g, sensitivity of 800mV/g at 1.5g and a maximal sampling rate of 512Hz. According to the Nyquist-Shannon sampling theorem, the ceiling boundary frequency component that can be detected using this accelerometer is 256 Hz, which fulfils the requirements for the intended application (since the fundamental frequencies of vocal chords are approximately 100Hz for males and 200Hz for females). To analyze the frequency domain of acceleration time series (square roots of the sum of the values of each axis x, y and z squared), the method relied on Discrete Fourier Transform (DFT) defined for a given sequence xk, k = 0, 1, … N-1 as the sequence Xr, r = 0, 1, … , N-1 [21]:

$$X_r = \sum_{k=0}^{N-1} x_k \, e^{-j2\pi rk/N}$$

Frequency spectrum was analyzed in Matlab applying the Fast Fourier Transform (FFT) to calculate the DTF and then the power spectral density was computed.

As expected, low amplitudes of the chest wall vibration were similar to the noise level thus making it difficult to distinguish accelerometer readings that contained speech from those that contained noise, only by analyzing the frequency spectra. In order to tackle the problem of noise, a simple noise cancelling strategy [22] was applied which consists of summing frequency spectra in time. This strategy is based on the assumption that the signal components are always focused in the same frequency range in contrast to noise that is, in this case, more random. Considering time frames for performing power spectral density analysis, the best accuracy was achieved by analyzing a sum of power spectral densities computed separately for five consecutive 2-second long time series (corresponding to 1024 samples in this case). Hence, each 10-seconds frame was represented with the power spectral density that was a sum of spectral densities computed for each 2 seconds. Therefore, our goal was to recognize the presence of spectral components that correspond to speech with the resolution of 10 seconds. Processing data in 10-second time frames resulted in the highest accuracy regardless of the duration of the speech i.e. whether there was only one word spoken or a continuous talk of 10 seconds. Decreasing the resolution corresponded to lower ratio between speech amplitudes and noise levels while processing data in longer time units was more likely to fail in detecting shorter durations of speech.

We investigated various classification algorithms (namely SVM, Naïve Bayes, Naïve Bayes with kernel density estimation and k-NN) and parameters for characterizing the spectral density (namely mean, maximal, minimal, and integral values regarding different frequency ranges). It turned out that Naïve Bayes with kernel density estimator applied on the two parameters – integral and mean values of the components between 80 Hz and 256 Hz, provided the highest classification accuracy. Note that the classification selection, a choice of signal parameters, frame size for calculating power spectral density and the resolution cannot be generalized since they strongly depend on the accelerometer's characteristics. In the following, we report the accuracy of our approach.

2.4 Results

We created a set of accelerometer data that contained speech activity of 19 subjects, 10 males and 9 females (overall, 2 minutes each subject, that is 38 minutes, divided in 10-second time frames) and accelerometer readings that contained physical movements without voice (approx. 2 hours of accelerometer readings that included sitting, standing and normal speed walking in 10-second data resolution). The voice recognition accuracy was estimated through leave-one-out method of sequentially selecting accelerometer readings that corresponded to one subject/one activity as a test unit while using the rest of the set for building the model (training set for Naïve Bayes with KDE classification). The voice was correctly recognized in 93% of cases while mild physical activities without voice induced false positives in 19% (Table 1a). The same model was used to test accelerometer readings acquired in more intensive activities such as fast walking or running which resulted in 29% rate of false positives (Table 1b).

Table 1. Speech Detection Accuracy

a)	Voice Detected	No Voice Detected
Voice	93%	7%
Mild Activities	19%	81%

b)	Fast Walking or Running
No voice detected (true negatives)	71%
Voice detected (false positives)	29%

Our approach demonstrates that the speech activity can be reliably detected in typical daily situations that include mild activities. More intense activities such as running may result in a higher rate of false positives. However, using different type of accelerometer may mitigate this.

The accelerometer-based approach does not require capturing privacy sensitive information. However, on the other hand, it imposes a sensor worn at the chest level, which may be perceived by subjects as obtrusive, consequently stigmatizing them. This issue, while currently a concern, may be mitigated, since accelerometers are increasingly becoming widely adopted both in research and everyday life. The shape and size of already accepted commercial accelerometer-based solutions can suit also the speech recognition purpose (such as Fitbit [28] – an accelerometer device for tracking wellbeing aspects of individuals' behavior), while the chest area is convenient for attaching a sensor with an elastic band (similarly to attaching respiratory or cardio sensors) minimizing the interference with typical daily routines. Therefore, imposing an accelerometer as an alternative to the use of microphone was a compromise for preventing privacy concerns in subjects while providing a mobile solution for continuous monitoring of speech activity.

In the following section we apply this approach to investigate the correlations between the amount of speech and the mood states.

3 Speech Activity and Mood Changes

The current literature reports several studies that examined how the social activity impacts the mood states during the day [8] [9] [10] [11] [12]. Vittengl et al. [8] and Robbins et al [9] demonstrated that different types of social encounters provoke diverse emotional effects, while there is also an association between the overall amount of social interactions and responses in positive affect [10][11][12]. All the studies were consistent in revealing the positive relation between social events and the mood dimension of positive affect (PA), while negative affect (NA) factors were shown to be correlated either with only certain types of conversations or not associated with social activity at all.

Through a pilot study, we investigated the correlation between self-reported mood changes and the overall amount of speech within a certain interval that reflects participation in verbal social interactions. This study demonstrates the use of low cost sensing technologies for monitoring speech activity as one aspect of social behavior, which according to the previous studies, has an impact on emotional response of individuals.

While we automatically estimate the amount of speech activity, the mood in subjects was measured relying on the standard, questionnaire based method. Despite of an increasing attention that the field of automatic mood recognition has been receiving, the practical use of such methods, as a reliable alternative to standardized questionnaires, has not been demonstrated yet. Therefore, we opted for the method of assessing mood fluctuations during the day based on EMA (Ecological Momentary Assessment) approach in order to compare retrospective and momentary mood data [23]. The EMA approach, which involves asking participants to report their psychological state multiple times a day, reduces the critical issue of retrospective recall of extended time intervals. The retrospective recall is related to cognitive and emotive limitations that bias the recall of autobiographical memory influencing subject's report by most salient events during the recall interval. The questionnaire used in this study was derived from a well-established scale – the Profile of Mood States (POMS) that consists of 65 items in its standard version. However, long and repeated mood questionnaires become a burden on subjects; therefore we derived 8 adjectives from the POMS scale, namely cheerful, sad, tensed, fatigued, energetic, relaxed, annoyed and friendly that were rated on 5-point scale (1-not at all, 2- a little, 3- moderately, 4- quite a bit, 5- extremely). The points were summed across the items related to PA and NA dimensions while the difference in scores between two sequential questionnaires was taken as a measure of relative change of subject's mood states. The questionnaires were administered three times a day, scheduled to best fit with office workers' routines that participated our experiments. Typically, the questionnaires were answered in the morning, after lunch and at the end of working day.

3.1 Study Design

One's mood may depend on a number of different factors, such as circadian rhythms [24], type of environment [25], quality of sleep [26], state of health, private problems or some other factors incomprehensible not only through direct measurement but also difficult for an individual himself/herself to identify. Therefore, it may be impossible to consider all the factors that influence the mood and provide the ultimate conclusion about the exact cause of one's state of mood. For this reason, this study follows relative changes of the mood dimensions of PA/NA rather than focus on an absolute mood state, while assuming that interval between two mood assessments of a couple of hours (in the experimental design) is not sufficient for a significant change in "background" factors. It is hypothesized that these factors, such as private problems for example, are likely to be constantly present during relatively longer periods of time while, the activities within that period have pre-dominant influence on relative changes of mood. The goal of this research is to capture patterns of the amount of speech activity, in most cases, provoke similar responses in individuals' mood.

3.2 Experiments

In order to estimate the amount of speech activity within a certain period, the Shimmer accelerometer [20], attached on the chest, was continuously sampling and storing the data. Applying the model described in the previous section, each 10-second time frame of the acquired data was separately queried and classified according to the presence of speech. Afterwards, for each interval of interest we calculated the number of minutes in which at least one 10-second frame indicated speech status thus providing an aggregated number of minutes in which subjects were speaking. Overall 10 knowledge workers (7 males, 3 females) were recruited during one working week (5 working days). The characteristics of the sample are presented in Table 2.

The paper-based questionnaires were administered at 10:00, 14:00 and 18:00 (or with slight deviations when subjects were temporary unable to fill-out the questionnaire) thus dividing working day in two intervals of interest – one between 10:00 and 14:00 and another between 14:00 and 18:00. The amount of speech activity was expressed as the number of minutes in which speech status was identified, divided by the duration of the monitored interval. In total, 122 questionnaires were collected and the self-reported mood dimensions of PA and NA were analyzed with respect to the amount of speech activity detected in the previous time interval. Overall, 78 such intervals were analyzed, with the duration of 221±37 minutes, in which subjects spent 27.9±12.1% of time (minutes) in speech activity.

Table 2. Characteristics of the sample

Age (years)	33.3±9.4
Marital status	
Married	0%
Single, Divorced	100%
University/post diploma	90%
Work hours/week	39.2±1.7
Duration between two questionnaires (minutes)	221.3±37.0
Morning intervals (minutes)	250.3±37.5
Afternoon intervals (minutes)	192.3±41.2
Number of reported positive mood changes	4.9±1.5
Number of reported negative mood changes	5.4±2.0

Fig. 1 shows the distribution of Spearman correlation between the amount of speech activity as estimated from accelerometer readings and reported mood changes. The mean correlation between the amount of speech activity and PA and NA scores was 0.34±0.27 (min=-0.03, max=0.76) and -0.07±0.33 (min=-0.62, max=0.39) respectively.

The distribution of the correlations between the amount of detected speech and PA scores were significantly greater than 0 (t=4.009, P<0.005) and not significantly skewed. The distribution related to NA scores was not significantly less than 0 (t=-0.721) and was significantly negatively skewed.

The results suggest that the time spent in speech activity (reflecting the participation in verbal social interactions) was positively correlated with changes in reported PA and was not related to the changes in NA scores. On the other hand, the mood score reported at the beginning of monitored interval and the amount of speech activity within that interval showed no significant correlations, 0.153 and 0.225 for PA and NA respectively, indicating that participation in verbal social interactions was not influenced by the initial subjects' mood. This may be due to the fact that working environment typically imposes conversations leaving no options for the one to choose the level of socialization depending on the current state of mood.

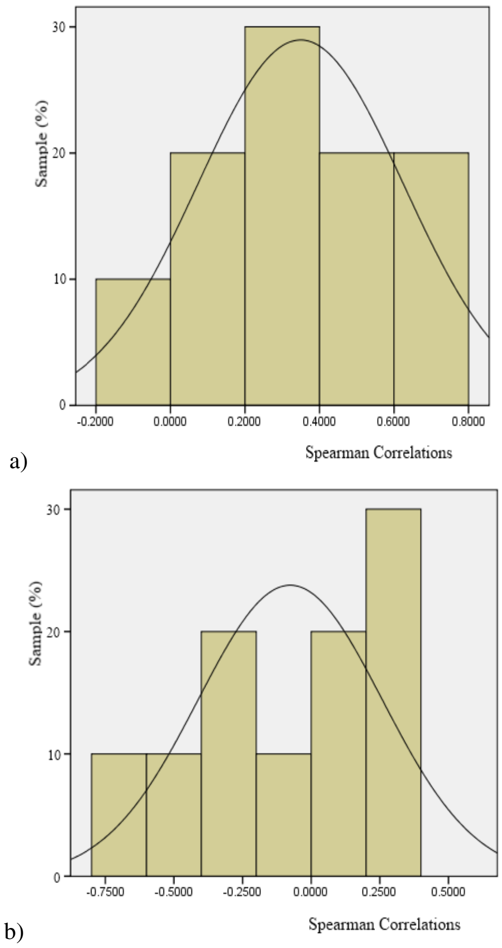

Fig. 1. Distributions of Spearman correlations between an amount of speech activity and a) PA and b) NA

4 Conclusion

In this paper we presented the concept of using an off-the-shelf accelerometer to infer speech activity by detecting vibrations at the chest level. This approach does not require capturing sensitive data, with a trade-off that includes a sensor worn at the chest level, which may be perceived by subjects as obtrusive. However, as an alternative to microphone-based method, the use of an accelerometer was a compromise for preventing privacy concerns in subjects while providing a mobile solution for continuous monitoring of speech activity. Such an approach allows for privacy-preserving collection of a large amount of speech activity data while being an easily applicable and a cost effective solution.

We investigated the correlation between the amount of speech and the mood changes through a pilot study relying on the accelerometer based approach to detect speech activity. The results of our study suggest that the amount of speech, which reflects the engagement in verbal communications, positively relate to the reported PA while no evidenced correlations were found for NA. These results show that verbal interactions are an important factor to be considered when taking into account the overall wellbeing of subjects in general and knowledge workers in particular.

References

[1] House, J.S., Landis, K.R., Umberson, D.: Social relationships and health. Science 241(4865), 540–545 (1998)

[2] Rabbi, M., Choundhury, T., Ali, S., Berke, E.: Passive and In-situ As-sessment of Mental and Physical Well-being using Mobile Sensors. In: 13th International Conference on Ubiquitous Computing, UbiComp 2011 (2011)

[3] Bernard, H.R., Killworth, P., Kronenfeld, D., Sailer, L.: The Problem of Informant Accuracy: The Validity of Retrospective Data. Annual Review of Anthropology 13(1), 495–517 (1984)

[4] Choudhury, T., Pentland, A.: Sensing and modeling human networks using the sociometer. In: Proceedings of Seventh IEEE International Symposium on Wearable Computers, 2003, vol. (1997), pp. 216–222 (2004)

[5] Wyatt, D., Choudhury, T., Keller, J., Bilmes, J.: Inferring Colocation and Conversation Networks from Privacy-sensitive Audio with Implications for Computational Social Science. ACM Transactions on Intelligent Systems and Technology (2010)

[6] Wyatt, D., Choudhury, T., Bilmes, J.: Inferring colocation and conversation networks from privacy-sensitive audio with implications for computa-tional social science. ACM Transactions on Intelligent Systems and Technology (TIST) 2(1) (2011)

[7] Cristani, M., Pesarin, A., Vinciarelli, A.: Look at who's talking: Voice activity detection by automated gesture analysis. In: Workshop on Interactive Human Behavior Analysis in Open or Public Spaces (2011)

[8] Vittengl, J.R., Holt, C.S.: A Time-Series Diary Study of Mood and Social Interaction. Motivation and Emotion 22(3), 255–275 (1998)

[9] Robbins, P.R., Tanck, R.H.: A study of diurnal patterns of depressed mood. Motivation and Emotion 11(1), 37–49 (1987)

[10] Berry, D.S., Hansen, J.S.: Positive affect, negative affect, and social interaction. Journal of Personality and Social Psychology 71(4), 796–809 (1996)

[11] Clark, L.A., Watson, D.: Mood and the mundane: relations between daily life events and self-reported mood. Journal of Personality and Social Psychology 54(2), 296–308 (1988)

[12] Watson, D., Clark, L.A., McIntyre, C.W., Hamaker, S.: Affect, personality, and social activity. Journal of Personality and Social Psychology 63(6), 1011–1025 (1992)

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

[14] L.A.F. E., Nie, E.: Computational modeling of face-to-face social interaction using nonverbal behavioral cues (2011)

[15] Lee, Y., Kwon, O.: Information Privacy Concern in Context-Aware Personalized Services: Results of a Delphi Study *. Journal of Information Systems 20(2) (2010)

[16] Nguyen, D.H., Kobsa, A., Hayes, G.R.: An empirical investigation of concerns of everyday tracking and recording technologies. In: Proceedings of the 10th international conference on Ubiquitous computing - UbiComp 2008, p. 182 (2008)

[17] Rao, R., Chen, T.: Cross-modal prediction in audio-visual communication. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1996, pp. 2056–2059 (1996)

[18] Titze, I.: Physiologic and acoustic differences between male and female. J. Acoust. Soc. Am., 1699–1707 (1989)

[19] Sundberg, J.: Chest wall vibrations in singers. Journal Of Speech and Hearing Research 26(3), 329–340 (1983)

[20] Shimmer - Wireless Sensor Platform for Wearable Applications, http://www.shimmer-research.com/p/products/sensor-units-and-modules/wireless-ecg-sensor (Accessed November 15, 2011)

[21] Tan, S.M.: Linear Systems, The Discrete Fourier transform, ch. 9, pp. 1–8. The University of Auckland

[22] Widrow, B., Glover Jr., J.R., McCool, J.: Adaptive noise cancelling: Principles and applications. Proceedings of the IEEE 63(12), 105–112 (1975)

[23] Smyth, J.M.: Ecological Momentary Assessment Research in Behavioral medicine. Journal of Happiness Studies 4(1), 35–52 (2003)

[24] Clark, L.A., Watson, D., Leeka, J.: Diurnal variation in the possitive affect. Motivation and Emotion 13(3), 205–234 (1999)

[25] Adan, A., Guàrdia, J.: Circadian variations of self-reported activation: a multidimensional approach. Chronobiologia 20(3-4), 233–244

[26] Volkers, A.C., Tulen, J.H.M., Broek, W.W.V.D.: Relationships between sleep quality and diurnal variations in mood in healthy subjects (1998), http://www.nswo.nl/userfiles/files/publications/jaarboek-1998/volkers.pdf