34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Speech Activity Detection using Accelerometer*

Aleksandar Matic, Venet Osmani, Oscar Mayora

Abstract— **The level of social activity is linked to the overall wellbeing and to various disorders, including stress. In this regard, a myriad of automatic solutions for monitoring social interactions have been proposed, usually including audio data analysis. Such approaches often face legal and ethical issues and they may also raise privacy concerns in monitored subjects thus affecting their natural behaviour. In this paper we present an accelerometer-based speech detection which does not require capturing sensitive data while being an easily applicable and a cost-effective solution.**

I. INTRODUCTION

The association between the overall wellbeing and behavioral patterns of individuals has been long established. Everyday routines, which include sleep, nutrition, exercise, socializing are factors that affect various health outcomes [1]. In particular, the correlation between social activity and health has been the subject of scientific investigation for more than a century [2]. It was demonstrated that subjects with a low quantity of social relationships are less healthy, psychologically and physically, while manifesting higher risks for tuberculosis, accidents, and psychiatric disorders such as schizophrenia [3]. On the other hand, recent studies showed that an increased amount of social interactions can improve depressive symptoms [4] [5]. Individuals who maintain a certain level of social engagements are shown to be more successful in coping with stress, and in the case of the elderly, they are highly functional and independent [1]. However, while people demonstrate awareness of the general recommendations regarding physical activity and diet, they typically neglect other factors that impact wellbeing, such as social activities [1]. In this regard, the participation in social interactions constitutes an important aspect that should be monitored and assessed.

The standard methods for monitoring social interactions in health sciences rely on self-reports and recall surveys that suffer from several limitations including: 1) difficulties in recalling activities that occurred in the past, 2) a high effort for continuous long-term monitoring, 3) self-reports are subjective and may be affected by the current mood [6]. An alternative approach of engaging human observers to record communications in groups is inefficient particularly if the interactions occur in various physical locations or if the size of the group is large [7]. In this paper, we refer to social interaction as co-located, face-to-face interactions, excluding electronically mediated interactions such as chat, social networking and other kinds of electronic communication.

The last dozen years have brought automatic sensing methods for recognizing human behavior, based on improvements in computational power and miniaturization of sensing technology. Sensor-based monitoring paradigm is considered to be a breakthrough in the evolution of social behavior analysis due to its potential to overcome limitations of self reporting and observational methods [8]. Recognizing the occurrence of social interaction 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 [9] (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 [10] or specialized device such as Sociometer [7]). The limitations of these approaches include sensitivity to false positives since nearby conversations can be unintentionally picked up and activating microphone typically raises privacy concerns. Even though privacy sensitive recording techniques can be applied, the fact that microphone is activated still may raise concerns with the subjects, thus affecting their natural behavior. Furthermore, 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 [11]. A few alternative methods aimed to infer speech activity based on mouth movement, fidgeting, or gestures [11] [12] detected using video machine system. However, this restricts application scenarios to limited areas that are covered with the camera system while, like in the case of microphone, such approach may also raise privacy and ethical concerns. These reasons prompted us to investigate different ways for detecting speech activity in a mobile manner (not limiting application only to certain areas) aiming to provide an alternative to microphone-based methods commonly used by the systems for sensing social interactions.

Our method for detecting speech activity is based on identifying another manifestation of speech different than voice, namely the vibration of vocal chords. The phonation-caused vibrations spread from the area of larynx to the chest level, representing the exhibition of speech activity which can be automatically detected through the use of accelerometer. In the area of speech analysis, non-acoustic sensors were used so far to investigate speech attributes [13] [14], speech encoding [14], to augment communication possibilities in patients with special needs [15], however to the best of our knowledge there is no work which used accelerometers to detect the status of speech in social interactions. Relying on an off-the-shelf accelerometer attached at the chest level, we developed an easily applicable and cost-effective solution to recognize speech activity, considering that accelerometers are widely available and

*Research supported by INTERSTRESS EU project (http://interstress.eu). The authors are with CREATE-NET, via alla Cascata 56/D, Trento, 38123 Italy, (+39) 0461 408400; fax: (+39) 0461 421 157; e-mail: name.surname@create-net.org.

accepted sensors which have found their usage both in research and everyday life. Our approach does not require recording of sensitive data thus it is not expected to raise ethical issues and privacy concerns in comparison to microphone or video based approaches.

II. Accelerometer to Detect Speech Activity

A. Methodology

Vocal chords (also known as vocal folds) are muscles within larynx that vibrate when air from lungs passes through thus producing voice [16]. The fundamental frequency of vocal chords vibrations depends on a variety of factors including age, gender and individual differences [17]. After the age of 20 the predicted fundamental frequency remains approximately 100Hz for male and 200Hz for female adults [17]. Therefore, identifying vibrations of these fundamental frequencies produced by vocal chords during phonation pertains to speech activity detection and in this paper we assess whether the characteristics of commercial accelerometer can suit this purpose. Since placing sensors on the neck (close to the larynx area) may be obtrusive, the chest surface was selected as a suitable body position. Chest area is already being used to place various sensors including cardio, respiratory and kinematic sensors. Sundeberg [13] identified a number of factors that contribute to the chest vibrations during phonation and examined the distribution of displacement amplitude over the chest wall surface, demonstrating that the vibrations can be detected all over the chest with the highest displacement amplitude located in the central part of the sternum, which is the area we chose to place the sensor on Figure 1. This position is also convenient for attaching a sensor with an elastic band (similar to attaching respiratory or cardio sensors) minimizing the interference with typical daily routines.

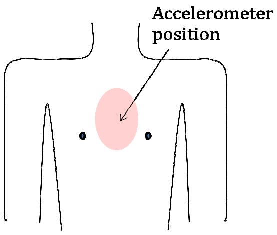

Figure 1. Area on the chest for placing accelerometer

B. Our Approach

The concept of using an accelerometer for recognizing speech activity is based on detecting phonation-caused vibrations at the chest level, targeting frequency range approximately between 100Hz and 200Hz. On the other hand, it is important to examine if there are potential sources in everyday life that produce components in the same range of frequencies which can be confused with speech activity. One may note that daily physical activities are not expected to overlap with vocal chords vibrations in the frequency domain since they typically occupy frequency ranges lower than 20 Hz [18]. However, this investigation is focused on the following two aspects. Firstly, it will be evaluated whether the characteristics of off-the-shelf accelerometers (i.e. not specifically designed for detecting small vibrations) are sufficient for recognizing speech activity and discriminate it from other components in the frequency spectrum. This concern refers mostly to low amplitudes of the chest wall vibrations [13] that may be similar to noise level, imperfect contact between the sensor and the chest, and physiological and acoustic differences between genders [17] and also across all individuals. Secondly, it will be investigated whether other sources of vibrations encountered in everyday life including elevator, car, bus, train or airplane, whose engines provide components in higher frequency ranges that may result in false positives for speech detection.

To evaluate the approach of detecting speech activity based on analyzing frequency spectrum of data acquired from an off-the-shelf accelerometer attached to the chest (Figure 1), the accelerometer produced by Shimmer was used in our experiments (coming as a part of an ECG device [19] thus it was not specifically adapted to detect small vibrations).The specifications are the following: 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 x_k, k = 0, 1, … N-1 as the sequence X_r, r = 0, 1, … , N-1 [20]:

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

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 [21] 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.

III. Evaluation results

In total, 21 subjects participated in the speech activity detection experiment (11 males, 10 females; 31.8±7.6 years old). Each subject was asked to read out loud the article from the latest newspapers for at least two minutes, while having the accelerometer attached to the chest with an elastic band. We evaluated the performance of our approach separately for each subject through cross-validation of two sets, one including the frequency spectra of 10-second frames containing subject's voice and the other including only spectra of accelerometer data samples recorded during mild physical activities without voice. 10 out of 11 male and 9 out of 10 female voices were successfully recognized, demonstrating that in large majority of the cases the accelerometer was sufficient to distinguish the frequency spectra of readings with and without voice despite the imperfect skin-sensor contact and individual subjects' characteristics.

In addition, we created a set of accelerometer data that contained speech activity of 19 subjects excluding 2 subjects that were not previously detected (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). This was done so that we can build a generic speech detection model. 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 I. a). 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 I. b). Furthermore, we investigated whether some sources that may be encountered in everyday life including elevator (5min of data), car (30min of data), bus (30 min of data), train (20min of data) or airplane (1 hour of data) whose engines provide components in higher frequency ranges result in false positives in speech detection. It turned out that elevator, train and airplane do not present an additional issue for the speech recognition, causing the same rate of false positives as physical movements performed in normal conditions (TABLE I. a and TABLE I. c) while travelling in a car or a bus increases the occurrence of false positives to the rate of 32%. In addition to phonation there are other causes of vocal chords vibrations, which can be incorrectly classified as speech activity such as coughing or mumbling; however, their occurrence is less frequent and typically negligible in comparison to speech.

Our approach demonstrates that the speech activity can be reliably detected in typical daily situations except in vehicles (such as car or bus) whose engine frequencies may result in a higher rate of false positives. However, this may be mitigated by using a different type of the accelerometer.

TABLE I. A) Voice/Mild Activities, B) Voice/Intensive Activities, C) Voice/sources of higher Frequencies

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

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

c)	Elevator	Bus/ Car	Train	Airpla ne
No voice detected (true negatives)	80%	68%	81%	79%
Voice detected (false positives)	20%	32%	19%	21%

IV. Conclusion

We presented the approach of detecting speech activity using an off-the-shelf accelerometer intended to identify another manifestation of speech which is different than voice, namely the vibrations of vocal chords. Our approach does not require recording of sensitive data thus it is not expected to raise privacy concerns in comparison to typical microphone-based methods. Such an approach represents an easily applicable and a cost effective mobile solution for recognizing speech activity, considering that accelerometers are widely used sensors (being already commonly used for detecting a number of activities). The shape of already accepted commercial accelerometer-based solutions can suit also the speech recognition purpose (such as Fitbit [22]).

The association between the level of social activity and the health status of individuals has been long established both on a theoretical and empirical basis [3]. It has been shown that social isolation can lead to a myriad of disorders, which even presents a major risk factor for mortality [3]. In this regard, monitoring social activity becomes an important aspect for the overall wellbeing (both physical and psychological) assessment. We envision that the accelerometer-based speech detection can complement methods for automatic recording of social interactions, being an alternative to audio data analysis which may raise privacy concerns in subjects thus affecting their natural behaviour. Furthermore, considering the fact that accelerometers were previously used for monitoring physical activity and sleep (both in research [6] and in commercial applications [22]), which are important aspects of wellbeing, our approach can also complement wellbeing applications by monitoring in the same time the amount of speech. Encouraging healthier lifestyle through such applications is based on the concept of providing people with self-monitoring tools aiming to increase the awareness of their own daily routines and consequently their wellbeing.

ACKNOWLEDGEMENT

This work was supported by EU FP7 INTERSTRESS project – http://interstress.eu

REFERENCES

[1] N. Lane, M. Mohammod, M. Lin, X. Yang, and H. Lu, “BeWell: A Smartphone Application to Monitor, Model and Promote Wellbeing,” in *5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth2011)*, 2011.

[2] E. Durkheim, *Suicide*. The Free Press, New York, 1897.

[3] J. S. House, K. R. Landis, and D. Umberson, “Social relationships and health.,” *Science (New York, N.Y.)*, vol. 241, no. 4865, pp. 540-5, Jul. 1988.

[4] V. Isaac, R. Stewart, S. Artero, M.-L. Ancelin, and K. Ritchie, “Social activity and improvement in depressive symptoms in older people: a prospective community cohort study.,” *The American journal of geriatric psychiatry official journal of the American Association for Geriatric Psychiatry*, vol. 17, no. 8, pp. 688-696, 2009.

[5] H. B. Bosworth, J. C. Hays, L. K. George, and D. C. Steffens, “Psychosocial and clinical predictors of unipolar depression outcome in older adults.,” *International Journal of Geriatric Psychiatry*, vol. 17, no. 3, pp. 238-246, 2002.

[6] M. Rabbi, T. Choundhury, S. Ali, and E. Berke, “Passive and In-situ Assessment of Mental and Physical Well-being using Mobile Sensors,” in *13th International Conference on Ubiquitous Computing (UbiComp’11)*, 2011.

[7] T. Choudhury and a. Pentland, “Sensing and modeling human networks using the sociometer,” *Seventh IEEE International Symposium on Wearable Computers, 2003. Proceedings.*, no. 1997, pp. 216-222, 2004.

[8] N. N. Eagle, “Machine Perception and Learning of Complex Social Systems,” Massachusetts Institute of Technology, 2005.

[9] D. Wyatt, T. Choudhury, J. Keller, and J. Bilmes, “Inferring Colocation and Conversation Networks from Privacy-sensitive Audio with Implications for Computational Social Science,” in *ACM Transactions on Intelligent Systems ans Technology*, 2010.

[10] D. Wyatt, T. Choudhury, and J. Bilmes, “Inferring colocation and conversation networks from privacy-sensitive audio with implications for computational social science,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 1, 2011.

[11] M. Cristani, A. Pesarin, and A. Vinciarelli, “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.

[12] R. Rao and T. Chen, “Cross-modal prediction in audio-visual communication,” in *IEEE international Conference on Acoustics, Speech, and Signal Processing. ICASSP-96.*, 1996, pp. 2056–2059.

[13] J. Sundberg, “Chest wall vibrations in singers.,” *Journal Of Speech And Hearing Research*, vol. 26, no. 3, pp. 329-340, 1983.

[14] T. H. Falk, J. Chan, P. Duez, G. Teachman, and T. Chau, “Augmentative communication based on realtime vocal cord vibration detection.,” *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 18, no. 2, pp. 159-63, Apr. 2010.

[15] T. F. Quatieri et al., “Exploiting Nonacoustic Sensors for Speech Encoding,” *Language*, vol. 14, no. 2, pp. 533-544, 2006.

[16] “Medicine Net.” [Online]. Available: http://www.medterms.com/script/main/art.asp?articlekey=6224. [Accessed: 15-Nov-2011].

[17] I. Titze, “Physiologic and acoustic differences between male and female,” *J. Acoust. Soc. Am*, pp. 1699-1707, 1989.

[18] M. J. Mathie, A. C. F. Coster, N. H. Lovell, and B. G. Celler, “Accelerometry: providing an integrated, practical method for long-term, ambulatory monitoring of human movement,” *Physiological Measurement*, vol. 25, no. 2, p. R1-R20, Apr. 2004.

[19] “Shimmer - Wireless Sensor Platform for Wearable Applications.” [Online]. Available: http://www.shimmer-research.com/p/products/sensor-units-and-modules/wireless-ecg-sensor . [Accessed: 15-Nov-2011].

[20] “Linear Systems, S.M. Tan, The University of Auckland, Chapter 9 The Discrete Fourier transform,” pp. 1-8.

[21] B. Widrow, J. G. Jr, and J. McCool, “Adaptive noise cancelling: Principles and applications,” *Proceedings of the IEEE*, vol. 63, no. 12, pp. 105-112, 1975.

[22] “Fitbit.” [Online]. Available: http://www.fitbit.com/. [Accessed: 10-Mar-2012].