INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

A System for Real-time Privacy Preserving Data Collection for Ambient Assisted Living

Fasih Haider, Saturnino Luz

Usher Institute of Population Health Sciences & Informatics
Edinburgh Medical School, the University of Edinburgh, UK
{Fasih.Haider, S.Luz}@ed.ac.uk

Abstract

Ambient Assisted Living (AAL) technologies are being developed which could assist elderly people to live healthy and active lives. These technologies have been used to monitor people's daily exercises, consumption of calories and sleeping patterns, and to provide coaching interventions to foster positive behaviour. Speech and audio processing can be used to complement such AAL technologies to inform interventions for healthy ageing by analyzing acoustic data captured in the user's home. However, collection of data in home settings present a number of challenges. One of the most pressing challenges concerns how to manage privacy and data protection. To address this issue, we have developed a low-cost system which can extract audio features while protecting the actual spoken content upon detection of voice activity, and store audio features for further processing which offer privacy guarantees. These privacy preserving features are being tested in the context of a larger project which includes health and well-being monitoring and coaching.

Index Terms: Ambient Assisted Living, Cognitive Health, Human Behaviour Analysis, Social Signal Processing

1. Introduction

Health and wellbeing monitoring using AAL technologies involves developing systems for automatically detecting and tracking a number of events that might require attention or coaching. In the SAAM project [1], we are employing AAL technologies to analyse activities and health status of elderly people living on their own or in assisted care settings, and to provide them with personalised multimodal coaching. Such activities and status include mobility, sleep, social activity, air quality, cardiovascular health, diet [2], emotions [3] and cognitive status [4]. While most of these signals are tracked through specialised hardware, audio and speech are ubiquitous sources of data which could also be explored in these contexts. Speech quality and activity, in particular, closely reflect health and well-being. We have explored the potential of speech analysis for automatically recognizing emotions [3], cognitive difficulties [4] and eating-related events [2] in the SAAM AAL environment. AAL technologies and coaching systems such as SAAM, which focus on monitoring of everyday activities, can benefit from recognition of these audio events in characterising contextual information against which other monitoring signals can be interpreted. One of the major challenges in collecting audio data in home environments for the development of health monitoring technology is user privacy. To address user privacy concerns, we developed a low-cost system which records content-free, anonymised audio features for automatic analysis. In particular, we extract features such as the *eGeMAPS* set [5] which we have used to detect specifiec behaviours in the above mentiond applications [2, 3, 4]. These features are computed using different statistical functionals over at the utterance level rather than at frame level, which makes it impossible to extract content information through, for instance, synthesis of speech from the extracted features or automatic transcription [6].

2. System for Data Collection

This section describes hardware and software components of the system which is used to collect data while preserving user's spoken content privacy.

2.1. Hardware Components

The hardware consists of a Matrix Creator board, consisting of a microphone array, an inertial measurement unit, and several other sensors, mounted on a Raspberry Pi 3 B+, as shown in Figure 1. This setup is meant to be installed in room where social activity and dialogue interaction occurs most frequently, such as a dinning room or a sitting room.

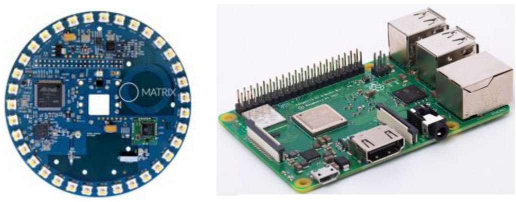

Figure 1: *Matrix Creator and Raspberry Pi 3 B+*

2.2. Software Components

For voice activity detection, we employed the Auditok[1] Python binding. Based on watchdog[2] input, the OpenSMILE [7] toolkit and a user recognition module are used to process the audio file and save the speech features in the attribute-relation file format (ARFF).

2.3. Feature Sets

Three features sets are extracted which have been widely used for emotion recognition [3, 5], eating conditions recognition [2] and cognitive state detection [8]. These sets are: a) *emobase*: an acoustic feature set containing Mel-Frequency Cepstral Coefficients (MFCCs), voice quality, fundamental frequency (F0), F0 envelope, LSP and intensity features along with their first and second order derivatives, along many statistical functionals, resulting in a total of 988 features for every speech utterance; b) *ComParE* [7], which comprises energy, spectral,

[1]https://pypi.org/project/auditok/ – last verified April 2019
[2]https://pythonhosted.org/watchdog/ – last verified April 2019

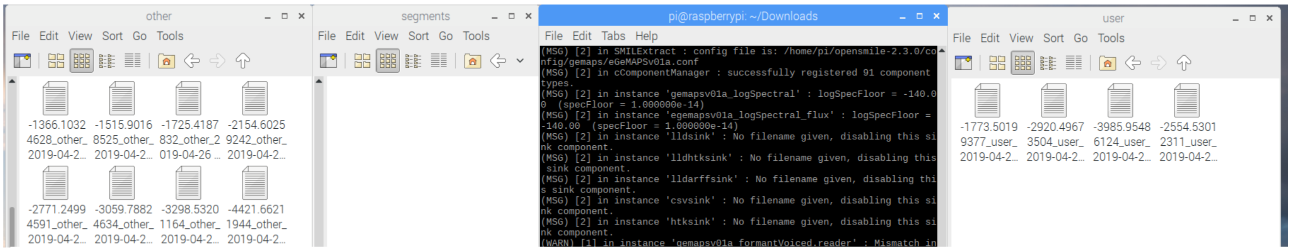

Figure 2: *A snapshot showing the process of feature extraction*

MFCC, and voicing related low-level descriptors (LLDs), including logarithmic harmonic-to-noise ratio, voice quality features, Viterbi smoothing for F0, spectral harmonicity and psychoacoustic spectral sharpness, for a total of 6373 acoustic features per utterance; and c) *eGeMAPS* [5] which contains the F0 semitone, loudness, spectral flux, MFCC, jitter, shimmer, F1, F2, F3, alpha ratio, hammarberg index and slope V0 features, and their functionals, for a total of 88 features per utterance.

2.4. Feature Extraction Process

The process of feature extraction is shown in Figure 3. Auditok is used to detect voice "chunks" using the energy of the audio signal in real time, and save them into pulse-code modulation (PCM) streams. A user recognition module and openSMILE take these streams as input. The user recognition module processes it to distinguish those streams which contain the target user's speech from other sounds (such as other speach or noise). The latter are ignored, to further protect the privacy of non-consented interlocutors. Upon extraction of the privacy preserving acoustic features, the PCM streams are immediately deleted. A snapshot showing feature extraction in progress is shown in Figure 2. The Python script used for this purpose is available through our git repository[3].

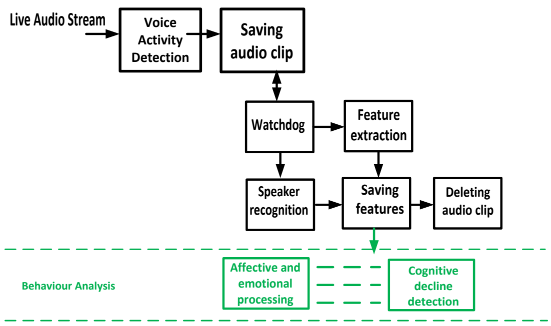

Figure 3: *System architecture*

3. Conclusion

AAL can benefit from unobtrusive, privacy-preserving systems for gathering and processing of speech at home. This paper describes a simple and inexpensive system for capturing of speech features which can be used in a number of automatic wellbeing monitoring tasks, in the context of an AAL-based coaching system for healthy ageing. Privacy protection and preservation in audio and speech can be regarded from different perspectives, including the protection of a person's identity, protection of the content spoken, and protection of inferences one may be able to draw from characteristics of a person's voice (such as cognitive or emotional status) [9]. A current limitation of our approach is that it only addresses the second of these aspects. In future work we aim to address identity and inference protection within the general framework presented here.

[3] git@git.ecdf.ed.ac.uk:fhaider/saam-av-capturing-system.git

4. Acknowledgements

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 769661, SAAM project.

5. References

[1] D. Yordan, G. Zlatka, Žnidaršič Martin, Ženko Bernard, V. Vera, and M. Nadejda, "Social activity modelling and multimodal coaching for active aging," in *Procs. of Personalized Coaching for the Wellbeing of an Ageing Society, COACH'2019*, 2019.

[2] F. Haider, S. Pollak, E. Zarogianni, and S. Luz, "SAAMEAT: Active feature transformation and selection methods for the recognition of user eating conditions," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. New York, NY, USA: ACM, 2018, pp. 564–568. [Online]. Available: http://doi.acm.org/10.1145/3242969.3243685

[3] F. Haider and S. Luz, "Attitude recognition using multi-resolution cochleagram features," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 3737–3741.

[4] S. Luz and S. D. la Fuente, "A method for analysis of patient speech in dialogue for dementia detection," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, D. Kokkinakis, Ed. Paris, France: European Language Resources Association (ELRA), may 2018.

[5] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva minimalistic acoustic parameter set GeMAPS for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[6] L. Lajmi, "An improved packet loss recovery of audio signals based on frequency tracking," *Journal of the Audio Engineering Society*, vol. 66, no. 9, pp. 680–689, 2018.

[7] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[8] H. Akira, F. Haider, L. Cerrato, N. Campbell, and S. Luz, "Detection of cognitive states and their correlation to speech recognition performance in speech-to-speech machine translation systems," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015, pp. 2539–2543.

[9] M. A. Pathak, B. Raj, S. D. Rane, and P. Smaragdis, "Privacy-preserving speech processing: cryptographic and string-matching frameworks show promise," *IEEE signal processing magazine*, vol. 30, no. 2, pp. 62–74, 2013.