An Automated Mood Diary for Older User's using Ambient Assisted Living Recorded Speech

Interspeech 2022
18-22 September 2022, Incheon, Korea

An Automated Mood Diary for Older User's using Ambient Assisted Living Recorded Speech

Fasih Haider and 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

In this paper, we describe a system for recording of mood diaries in the context of an ambient assisted living and intelligent coaching environment, which ensures privacy by design. The system performs affect recognition in speech features without recording speech content in any form. We demonstrate results of affect recognition models tested on data collected in care-home settings during the SAAM project (Supporting Active Ageing through Multimodal Coaching) using our custom-designed audio collection hardware. The proposed system was trained using Bulgarian speech augmented with training data obtained from comparable mood diaries recorded in Scottish English. A degree of transfer learning of Scottish English speech to Bulgarian speech was demonstrated.

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

1. Introduction

Health and wellbeing monitoring using Ambient and Assisted Living (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 wellbeing. 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 and processing audio data in home environments for the development and deployment 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 specific behaviours in the above-mentioned 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]. We further process those features for automatic mood detection and upon detection of mood the extracted features are deleted besides saved audio.

2. The MoodBox

The MoodBox's 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+. This setup is meant to be installed in the room where social activity and dialogue interaction occurs most frequently, such as a dinning room or a sitting room.

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 eGeMAPS features in the attribute-relation file format (ARFF). The *eGeMAPS* feature set is extracted which has been widely used for emotion recognition [3, 5], eating conditions recognition [2] and cognitive state detection [8]. *eGeMAPS* [5] 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. The process of feature extraction is shown in Figure 1. 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. The Python script used for this purpose is available through our git repository[3].

The proposed system also contains a MoodSlider through SAAM-APP [4]for self-reporting of the mood, as shown in Figure 2. The participants speak in-front of Moodbox while switching it On/Off through SAAM-APP and record 'mood diaries' where they talked about their days in free form. As a result, We gathered the *eGeMAPS* features, together with participants' self-assessment of mood in terms of arousal and valence.

2.1. Data Collection

The proposed system (i.e. MoodBox and MoodSlider) is used to collect privacy-based features from audio recording (Bulgarian

[1]https://pypi.org/project/auditok/ – last verified April 2022
[2]https://pythonhosted.org/watchdog/ – last verified April 2022
[3]git@git.ecdf.ed.ac.uk:fhaider/saam-av-capturing-system.git
[4]https://dev-saam-platform.eu/authentication/login?returnUrl=%2Fwall – last accessed April 2022

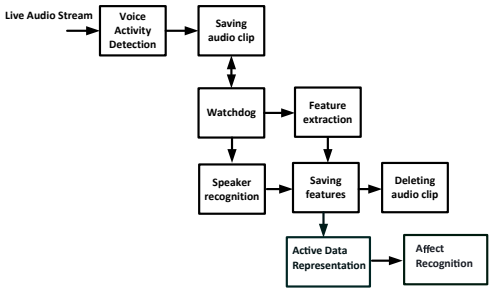

Figure 1: *MoodBox's architecture*

Figure 2: *MoodSlider: Arousal (top):'How "energised" do you feel right now and Pleasure (bottom): 'How pleased do your feel at the moment?'*

language) along with self-reported mood in the SAAM project. It is noted that, in this pilot study, the privacy-based features were extracted for 16 recordings from 4 participants. For regression analysis, we averaged the privacy-based features (extracted for each voice segment) over an audio recording. The regression analysis was performed using linear and random forest regression methods in leave one (recording) out cross-validation setting using MATLAB. We used the Concordance correlation Coefficient (CCC) and Pearson Correlation Coefficient (r) for evaluation purposes. It is noted that the linear regression (0.3376) provides better r than random forest (0.2555). However, random forest provides the best CCC (0.2371) for joy prediction. For predicting the arousal score, random forest provides better results than linear regression, as shown in Table 1.

Table 1: *Leave-one-recording out cross-validation results (i.e. Concordance correlation Coefficient (CCC) and Pearson Correlation Coefficient (r)) using Linear Regression (LR) and Random Forest (RF).*

Regression	Method	CCC	r
valence	LR	0.1129	0.3376
	RF	0.2371	0.2555
arousal	LR	-0.0457	-0.1231
	RF	0.0230	0.0256

2.2. Affect Recognition Models and Validation

As mentioned, the data collected in care-home settings in the context of the SAAM project contains only 16 recordings from 4 subjects. Due to the limited size of this pilot data set, we used data from one of our previous studies [9] where we collected the Scottish English data with self-reported mood for models training. We used models which are able to predict affect with a concordance correlation coefficient of 0.4230 (using Random Forest) and 0.3354 (using Decision Trees) for arousal and valence respectively using the active data representation method [9].

We also validated the models trained on Scottish English data on the Bulgarian speech data collected in the context of SAAM Project. We noticed that it provides CCC of -0.1043 and 0.1954 for arousal and valence respectively. A correlation coefficient of -0.1109 and 0.3445 is also observed for arousal and valence, respectively.

3. Conclusion

We described a system for recording of mood diaries in the context of an AAL and intelligent coaching environment. Our system ensures a level of privacy (content privacy) by design. We also trained machine learning models for automatic recognition of affect using the Bulgarian speech data which was collected in the scope of the SAAM project. Later, we evaluated the models trained using Scottish speech on Bulgarian speech, which resulted in almost the same results, hence demonstrating the transfer learning of acoustic features predictive of valence and arousal from Scottish English to Bulgarian speech. This was observed despite the fact that the Scottish speech data contains 108 audio recordings of 108 subjects, and the Bulgarian dataset contains only 16 audio recordings from 4 subjects. Future work may involve validating the proposed system in other languages.

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] S. de la Fuente Garcia, F. Haider, and S. Luz, "Covid-19: Affect recognition through voice analysis during the winter lockdown in scotland," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 2326–2329.