VADLite: An Open-Source Lightweight System for Real-Time Voice Activity Detection on Smartwatches

George Boateng
ETH Zürich
Zürich, Switzerland
gboateng@ethz.ch

Prabhakaran Santhanam
ETH Zürich
Zürich, Switzerland
psanthanam@ethz.ch

Janina Lüscher
University of Zürich
Zürich, Switzerland
janina.luescher@psychologie.uzh.ch

Urte Scholz
University of Zürich
Zürich, Switzerland
urte.scholz@psychologie.uzh.ch

Tobias Kowatsch
University of St. Gallen
St. Gallen, Switzerland
ETH Zürich
Zürich, Switzerland
tkowatsch@ethz.ch

ABSTRACT

Smartwatches provide a unique opportunity to collect more speech data because they are always with the user and also have a more exposed microphone compared to smartphones. Speech data could be used to infer various indicators of mental well being such as emotions, stress and social activity. Hence, real-time voice activity detection (VAD) on smartwatches could enable the development of applications for mental health monitoring. In this work, we present VADLite, an open-source, lightweight, system that performs real-time VAD on smartwatches. It extracts mel-frequency cepstral coefficients and classifies speech versus non-speech audio samples using a linear Support Vector Machine. The real-time implementation is done on the Wear OS Polar M600 smartwatch. An offline and online evaluation of VADLite using real-world data showed better performance than WebRTC's open-source VAD system. VADLite can be easily integrated into Wear OS projects that need a lightweight VAD module running on a smartwatch.

ACM Classification Keywords

H.5.m. Human-centered computing: Ubiquitous and mobile computing systems and tools

Author Keywords

Voice Activity Detection; Support Vector Machine; Wearable Computing; Smartwatch

UbiComp/ISWC '19 Adjunct, September 9–13, 2019, London, United Kingdom

ACM ISBN 978-1-4503-6869-8/19/09.
https://doi.org/10.1145/3341162.3346274

INTRODUCTION

Smartwatches as a platform provide a unique opportunity to assesses the mental health of individuals because of their sensors which have high proximity to the human body. Multimodal sensing of gestures (from the gyroscope and accelerometer), heart rate, physical activity, ambient light, Bluetooth signal strength, among others could be used to infer the mental state of people [1].

On the specific topic of audio data, smartwatches provide a unique opportunity to collect more speech data in the everyday lives of individuals since they are more likely to always be with the user given they are worn on the wrist. Additionally, the microphone is also prone to be more exposed as compared to a smartphone, which might be in a bag or a pocket.

Speech data could be used to infer various indicators of mental well being such as emotions [26], stress [17] and social activity [28]. Hence, real-time voice activity detection (VAD) on smartwatches could enable the development of applications for mental health monitoring. Researchers and developers that need real-time VAD on smartwatches, which use the Wear OS operating system have to build their own custom module since an API is not provided.

Prior work have developed VAD systems but they have not focused on real-time implementation of the developed algorithms [35, 29, 9, 23]. Important aspects such as computational efficiency, latency and accuracy in a naturalistic context were not addressed. Hence, it is not clear how well they will perform if they are implemented to run in real-time on smartwatches.

On the other hand, there are VAD systems that have been implemented to run in real-time as a smartphone app [34, 33, 18]. Unfortunately, the machine learning models that were used are not easily available for others who want to simply use those pre-trained models in their work. It is also not clear how well the models will work when they run on smartwatches with reference to computational efficiency and latency.

There was another VAD system developed to run on a smartwatch as a component of context recognizer [11]. The authors use 13 mel-frequency cepstral coefficients (MFCC) features and a convolutional neural network. Using three seconds of data, it takes 1.9 seconds to give a classification. The performance evaluation (eg. accuracy) is not provided and most importantly, the software component is not available for others to easily use in their smartwatch-based sensing work.

Then there are also open-source VAD systems such as WebRTC's VAD [12]. It is, however, a computer-based system, which does not have a module for smartwatches. Additionally, it has been reported that WebRTC's VAD performs poorly on real-world data collected with smartwatches [16].

Given the gap in obtaining an easy-to-use smartwatch-based VAD system, we developed VADLite, an open-source lightweight software system that performs real-time voice activity detection on smartwatches. VADLite extracts MFCC as features and classifies speech versus non-speech audio samples using a linear Support Vector Machine (SVM). We designed VADLite to meet our specific requirements; it is lightweight (i.e. runs efficiently on a constrained system such as a smartwatch), and also performs well in a real-world context in which we will be deploying the devices. In this work, we describe the process of developing and evaluating VADLite, and comparing its performance to WebRTC's VAD system since it is a widely used open-source VAD system.

The real-time implementation is done on a Wear OS smartwatch and our project files with the source code are available for use by others [1]. VADLite can be used by including the Java source code files in a Wear OS project. Also, the parameters of the trained model are in the Java files and hence, users can use our model parameters to build their own VAD pipeline for their Wear OS projects.

The rest of this paper is organized as follows. First, we give the motivation and use case of VADLite. Next, we give an overview and describe the development of VADLite. Then, we describe the real-time implementation of VADLite. After,we detail our experiments and results including a comparison of VADLite with WebRTC's VAD system. We address ethical implications and privacy concerns of our work. Finally, we summarize and conclude.

MOTIVATION

Our primary motivation for developing VADLite was our previous work DyMand, an open-source mobile and wearable system for assessment of couples' dyadic management of chronic diseases in everyday life [7]. The DyMand system collects self-report data about health behavior, and emotions, and sensor data about couples' dyadic management of chronic diseases. DyMand first determines if the partners are close using the Bluetooth signal strength between their watches. Then, VADLite is used as the next optimizing step to trigger collection of multimodal sensor data (heart rate, accelerometer, gyroscope,and ambient light) ensuring that data is collected only when the partners are speaking. This two-step process in which VADLite plays a key role ensures that DyMand collects more relevant speech data which is an improvement over how social psychologists currently collect ambulatory audio data for analysis: triggering data collection at random times of the day. [20, 21, 30, 31, 36].

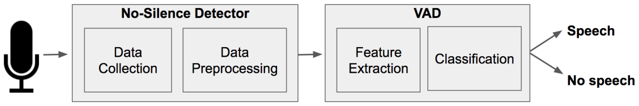

Figure 1. Overview of VADLite System

Our secondary motivation for developing VADLite is for it to be used in the development of a real-time smartwatch-based app for recognizing emotions of couples using speech prosody and the semantics of speech [26]. Emotion recognition from speech will then be used in combination with other sensor modalities from the smartwatch to perform real-time multimodal emotion recognition among couples [6]. Prior work has shown that social support among couples results in better health behavior when one partner has diabetes and also affect the emotions of the couple [22, 27, 14]. Real-time emotion recognition among couples would give an assessment of a key outcome of social support which could be used to develop just-in-time adaptive interventions [24] to enable couples better manage chronic diseases. In order to accomplish that goal, speech episodes in everyday life need to be recognized accurately and efficiently. VADLite fills that gap.

Beyond these specific use cases, real-time VAD in combination with other sensors could be used to infer social isolation or a lack of social activity, which is a predictor of mental health issues such as depression or suicidal ideation [37, 28]. An accurate measure of speech data could enable better prediction of social isolation. Using VADLite which runs on smartwatch will adequately enable the accomplishment of this goal.

OVERVIEW AND DEVELOPMENT OF VADLITE

VADLite is a 2-stage system consisting of a no-silence detector as the first part, and a voice activity detector as the second part (Figure 1). In developing VADLite, we used the pipeline of data collection and preprocessing, feature extraction and classification. We used a linear SVM. An SVM is a classifier that constructs a high-dimensional hyperplane to separate data of different classes [13]. SVM selects a hyperplane that maximizes the distance to the nearest data points on either side of the hyperplane in the case of binary classification. Previous work have used SVM for VAD successfully [9, 29, 2, 8, 15].

We use a linear SVM because it is memory and computationally efficient when incoming data is classified. For example, in comparison with a linear SVM, a radial basis function (RBF) SVM though only slightly outperformed a linear SVM for VAD took twice as much time for classification [15]. Prior work have used an implementation of linear SVM for real-time prediction on smartwatches for stress detection [5] and activity detection [3, 4].

[1] https://bitbucket.org/Jojo29/vadlite/

Data Collection

We collected real-world data using a protocol that was approved by the ethics commission of ETH Zurich. We collected data using a Polar M600 smartwatch where subjects (1) in the lab read a written text as a smartwatch recorded audio data for approximately 1-2 mins and (2) in the everyday life wore a smartwatch during waking hours as it continuously collected audio data. We used 16-PCM mono audio data and a sampling frequency of 8KHz. The data was annotated as speech or non-speech data. The speech data contained mostly conversations among several people (with at least 10 distinct speakers) at varying distances from the smartwatch's microphone. The non-speech data contained sounds from cars, trams, buses, wind, and music. The overall duration of the recorded sound data was 3.5 hours.

Data Preprocessing

We processed the data by first removing silence portions of the data using a one-second time window. Liaqat et al. found out that the real-world audio data they collected contained about 61.7% of silence and hence they implemented a silence detection algorithm to remove the silence part of their data [16]. Given that we use real-world audio data, we also removed silence segments of the data. We computed the root mean square (RMS) of each one-second time window of the whole data. We then check if the RMS value is below a certain threshold, in which case we mark that segment as silence and then remove it. To determine the threshold, we created a scatter plot of the RMS values of silence, speech and noise signals and then chose the value that separates silence from both speech and noise.

Feature Extraction

We extracted 13 MFCC features and use 12 of them (excluding the 1st coefficient, which is the DC component) over a time window of 25 ms. MFCC features have been widely used for VAD [11]. The parameters we used are as follows: 8KHz sample rate, window length of 25 ms, window step, 12 coefficients, 26 filters in the filterbank, FFT size of 512, 0 Hz as lowest band edge of mel filters, 4KHz as highest band edge of mel filter of (i.e., half the sampling rate), 22 lifters to apply to final cepstral coefficients and a Hamming windowing function. We used a Java implementation for the feature extraction.

Classification

Using the feature sets, we trained a linear SVM to classify speech or non-speech. We also perform grid search to pick the most optimal hyper-parameters of the linear SVM. We first normalized the features by subtracting the mean and dividing by the variance. This normalization is important for various algorithms such as SVM whose optimization assume that the features have a normal distribution [13]. We use the following metrics for evaluation: accuracy, speech hit rate (SHR), and false alarm rate (FAR). The SHR is the ratio of correctly detected speech frames to the total number of speech frames. By contrast, FAR is one minus the noise hit rate, where noise hit rate is the ratio of correctly detected noise frames to the total number of noise frames.

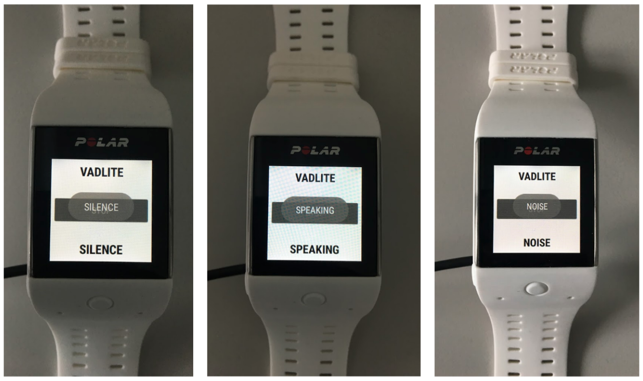

Figure 2. Real-time running of VADLite

REAL-TIME IMPLEMENTATION

We coded VADLite in Java for smartwatches that use the Wear OS operating system (formerly Android Wear). We used the Android Studio Integrated Development Environment (IDE) and the Android Software Development Environment (SDK). VADLite can potentially work on every Wear OS device. We used a Polar M600 smartwatch running Wear OS version 2.1 for testing which has the following specifications: Dual-Core 1.2GHz processor based on ARM Cortex-A7, 512MB RAM, 4GB flash storage, 500 mAh Battery.

Our implementation of VADLite is a Wear OS app, which collects 16-PCM mono audio data every second at a frequency of 8KHz (see Figure 2). We check if the one-second data is a non-silence segment by using an implementation of the non-silence detector from the previous section. We then process the data if it is non-silence. The one-second non-silence signal is then segmented into 25 ms frames. We then extract 12 MFCC features for each frame, which is then fed to a linear SVM for classification. We used the settings described in the previous section for the feature extraction. We used the Java implementation from the offline evaluation for extracting the MFCC features online.

We normalized the features using the stored normalization vectors before performing classification with a linear SVM. Our implementation of the linear SVM is a dot product of the stored coefficients with the features:

$$y = wx + b \tag{1}$$

where y is the result of the evaluation, w is the coefficient vector of length 12, and b is the intercept. We then assign y to be 1 (speech) if it is greater than zero, otherwise we assign it to be 0 (non-speech). We obtained w and b from the previously trained linear SVM. We output speech or non-speech classification for the whole one-second data. To accomplish this, we use majority voting of all the classified 25 ms samples within the one-second data.

VADLite had an average processing time of 2 ms for each 25 ms frame and 76 ms for the total one-second duration. As a result, throughput was met since the frame processing

Table 1. Evaluation results from the offline evaluation of VADLite and WebRTC's VAD

Model	*Accuracy*	*SHR(%)*	*FAR(%)*
WebRTC(0)	66.7	91.6	60.7
WebRTC(1)	71.7	89.0	47.4
WebRTC(2)	71.4	79.5	37.5
WebRTC(3)	75.9	55.3	2.1
VADLite	83.7	83.4	16.0

time was less than the 25 ms segment duration. Likewise, the processing time for whole duration was less than one second.

EXPERIMENTS AND RESULTS

We evaluated VADLite offline and also online. Additionally, we compared the classification performance of VADLite with a popular open-source VAD system, WebRTC's VAD [12]. WebRTC's VAD uses frequency band features and a pre-trained Gaussian Mixture Model (GMM) classifier [32]. We used a Python implementation of the system [38]. It gives the option to set an aggressiveness mode using an integer from zero to three with zero being the least aggressive about filtering non-speech audio. It only accepts 16-bit PCM mono audio sampled at 8KHz, 16KHz, 32KHz or 48KHz. It also processes data in frames with a duration of either 10, 20 or 30 ms. Our implementation used 8 KHz sampling rate and 20 ms time window to match the settings of VADLite.

Offline Evaluation

We split the data into train and test using about 70%-30% split. The speech and noise train data were 73.9 and 71.8 minutes long respectively. The speech and noise test were 24.6 and 22.4 minutes long respectively. We performed 10-fold stratified cross-validation on the train data using VADLite's linear SVM model. We used the scikit-learn library for our experiments [25]. The model achieved 82.6% accuracy, 80.2% SHR and 14.9% FAR. We then trained the VADLite model on the whole train dataset and then we used the test data to evaluate both the VADLite's model and those of WebRTC's VAD. The results of the evaluation are shown in Table 1.

VADLite's model outperforms WebRTC's VAD when its aggressiveness mode is two and three. WebRTC's VAD with settings zero and one though have very high SHR, their FAR are high, which will result in a lot of noise being classified as speech, which is not acceptable. VADLite's model provides a good enough tradeoff between SHR and FAR. These results indicate that VADLite is better than WebRTC's VAD. These results support those by Liaqat el al. who found that WebRTC's VAD performed poorly on real-world smartwatch-based audio [16].

Online Evaluation

To evaluate the real-time performance of VADLite with real-world data, we recorded audio data from a naturalistic context. We then played the recorded audio through a loudspeaker as the VADLite app performed real-time classification of the audio just like was done by Feng et al [10]. The audio had a duration of 15 minutes each for speech and noise. We stored the classification and compared it with real labels of the audio. We report the classification results below. We also ran the audio data through WebRTC's VAD. VADLite had SHR and FAR of 91.6% and 5.5% respectively. WebRTC's VAD's best performing mode had SHR and FAR of 73% and 18% respectively. Consistent with the results from the offline evaluation, VADLite outperforms WebRTC's VAD. Once again, these results supports those by Liaqat el al. who found that WebRTC's VAD performed poorly on real-world smartwatch-based audio [16].

ETHICAL IMPLICATIONS AND PRIVACY CONCERNS

This work has ethical implications as the system could be used in a manner that violates the privacy of others. We envision that this system can be used in two main ways.

The first is that it could be used to collect raw speech data from subjects, which will be stored for processing later. For this approach, it is especially important that the study protocol is subjected to review and approval from the ethics committee of the overseeing institution, as is standard practice. And additional steps need to be taken such as giving subjects the option to listen to the recorded audio and to delete any as they wish without any explanation. This approach has been used in our studies [19] and those of others [30, 31]. Depending on the use case, the app may need to give subjects the option to completely disable audio recording as needed. Also, to protect the privacy of subjects not taking part in the study, it might be necessary to have subjects wear a tag indicating to others around them that they may be recorded.

The other way we envision this system being used is to derive important summary statistics. In this case, no raw audio will be stored. Rather, various inference such as conversation frequency and duration (total duration of speech per day, times of the day with most speech etc.) will be computed [28]. This use case is less invasive but again as usual, ethical approval needs to be obtained for such a study since summary data could be considered personal and private for some subjects.

FUTURE WORK

VADlite could be extended so that it additionally makes various inference such as conversation frequency and duration. These additions would make VADLite more useful in smartwatch-based applications that seek to improve the mental well-being of people.

CONCLUSION

In this work, we developed VADLite an open-source lightweight software system for real-time VAD on smartwatches. VADLite uses MFCC as features and classifies speech versus non-speech audio samples using linear SVM with a real-time implementation on a Wear OS smartwatch. Our evaluation of VADLite showed SHR and FAR of 83.4% and 16.0% respectively for offline, and 91.6% and 5.5% respectively for real-time classification. Benchmarking of our system against WebRTC's VAD showed better performance. Our open-source system, VADLite can be easily integrated into Wear OS projects that need a lightweight voice activity module running on a smartwatch. VADLite can be integrated into the development of various well-being specific apps.

ACKNOWLEDGEMENT

We are grateful to Kemeng Zhang who helped collect the audio data. We thank Filipe Barata and Chen-Hsuan Shih for their feedback during this work. This project was funded by a grant from the Swiss National Science Foundation with funding number CR12I1_166348/1.

REFERENCES

1. Saeed Abdullah and Tanzeem Choudhury. 2018. Sensing technologies for monitoring serious mental illnesses. *IEEE MultiMedia* 25, 1 (2018), 61–75.

2. M Baig, S Masud, and M Awais. 2006. Support vector machine based voice activity detection. In *2006 International Symposium on Intelligent Signal Processing and Communications*. IEEE, 319–322.

3. George Boateng, John A Batsis, Ryan Halter, and David Kotz. 2017. ActivityAware: an app for real-time daily activity level monitoring on the amulet wrist-worn device. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 431–435.

4. George Boateng, John A Batsis, Patrick Proctor, Ryan Halter, and David Kotz. 2018a. GeriActive: Wearable app for monitoring and encouraging physical activity among older adults. In *2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 46–49.

5. George Boateng and David Kotz. 2016. StressAware: An app for real-time stress monitoring on the amulet wearable platform. In *2016 IEEE MIT Undergraduate Research Technology Conference (URTC)*. IEEE, 1–4.

6. George Boateng, Prabhakaran Santhanam, Janina Lüscher, Urte Scholz, and Tobias Kowatsch. 2018b. Multimodal Affect Detection among Couples for Diabetes Management. In *Poster: 2nd Black in AI Workshop at the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*.

7. George Boateng, Prabhakaran Santhanam, Janina Lüscher, Urte Scholz, and Tobias Kowatsch. 2019. DyMand: An Open-Source Mobile and Wearable System for Assessing Couples' Dyadic Management of Chronic Diseases. In *14th International Conference on Design Science Research in Information System and Technology (DESRIST)*.

8. Cheng Dai, Linkai Luo, Hong Peng, and Qingyun Sun. 2018. A Method Based on Support Vector Machine for Voice Activity Detection on Isolated Words. In *2018 13th International Conference on Computer Science & Education (ICCSE)*. IEEE, 1–4.

9. Dong Enqing, Liu Guizhong, Zhou Yatong, and Zhang Xiaodi. 2002. Applying support vector machines to voice activity detection. In *6th International Conference on Signal Processing, 2002.*, Vol. 2. IEEE, 1124–1127.

10. Tiantian Feng, Amrutha Nadarajan, Colin Vaz, Brandon Booth, and Shrikanth Narayanan. 2018. TILES audio recorder: an unobtrusive wearable solution to track audio activity. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*. ACM, 33–38.

11. Claudio Forlivesi, Utku Günay Acer, Marc van den Broeck, and Fahim Kawsar. 2018. Mindful interruptions: a lightweight system for managing interruptibility on wearables. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*. ACM, 27–32.

12. Google. 2011. WebRTC. (2011). Retrieved April, 2019 from `https://webrtc.org/`

13. Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, and others. 2003. A practical guide to support vector classification. (2003).

14. Masumi Iida, Mary Ann Parris Stephens, Karen S Rook, Melissa M Franks, and James K Salem. 2010. When the going gets tough, does support get going? Determinants of spousal support provision to type 2 diabetic patients. *Personality and Social Psychology Bulletin* 36, 6 (2010), 780–791.

15. Tomi Kinnunen, Evgenia Chernenko, Marko Tuononen, Pasi Fränti, and Haizhou Li. 2007. Voice activity detection using MFCC features and support vector machine. In *Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, Vol. 2. 556–561.

16. Daniyal Liaqat, Robert Wu, Andrea Gershon, Hisham Alshaer, Frank Rudzicz, and Eyal de Lara. 2018. Challenges with real-world smartwatch based audio monitoring. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*. ACM, 54–59.

17. Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 351–360.

18. Hong Lu, Wei Pan, Nicholas D Lane, Tanzeem Choudhury, and Andrew T Campbell. 2009. SoundSense: scalable sound sensing for people-centric applications on mobile phones. In *Proceedings of the 7th international conference on Mobile systems, applications, and services*. ACM, 165–178.

19. Janina Lüscher, Tobias Kowatsch, George Boateng, Prabhakaran Santhanam, and Urte Scholz. 2019. Social Support and Common Dyadic Coping in Couple's Dyadic Management of Type II Diabetes: Study Protocol for an Ambulatory Assessment Application. In *JMIR Protocols*. JMIR.

20. Matthias R Mehl, James W Pennebaker, D Michael Crow, James Dabbs, and John H Price. 2001. The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers* 33, 4 (2001), 517–523.

21. Matthias R Mehl, Megan L Robbins, and Fenne große Deters. 2012. Naturalistic observation of health-relevant social processes: The Electronically Activated Recorder (EAR) methodology in psychosomatics. *Psychosomatic Medicine* 74, 4 (2012), 410.

22. Daisy Miller and J Lynne Brown. 2005. Marital interactions in the process of dietary change for type 2 diabetes. *Journal of Nutrition Education and Behavior* 37, 5 (2005), 226–234.

23. Mohammad H Moattar, Mohammad M Homayounpour, and Nima Khademi Kalantari. 2010. A new approach for robust realtime voice activity detection using spectral pattern. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4478–4481.

24. Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2017. Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine* 52, 6 (2017), 446–462.

25. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

26. Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

27. Gabriele Prati and Luca Pietrantoni. 2010. The relation of perceived and received social support to mental health among first responders: a meta-analytic review. *Journal of Community Psychology* 38, 3 (2010), 403–417.

28. Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 385–394.

29. J Ramírez, P Yélamos, JM Górriz, and JC Segura. 2006. SVM-based speech endpoint detection using contextual speech features. *Electronics letters* 42, 7 (2006), 426–428.

30. Megan L Robbins, Elizabeth S Focella, Shelley Kasle, Ana María López, Karen L Weihs, and Matthias R Mehl. 2011. Naturalistically observed swearing, emotional support, and depressive symptoms in women coping with illness. *Health Psychology* 30, 6 (2011), 789.

31. Megan L Robbins, Ana María López, Karen L Weihs, and Matthias R Mehl. 2014. Cancer conversations in context: Naturalistic observation of couples coping with breast cancer. *Journal of Family Psychology* 28, 3 (2014), 380.

32. Sergey Salishev, Andrey Barabanov, Daniil Kocharov, Pavel Skrelin, and Mikhail Moiseev. 2016. Voice Activity Detector (VAD) Based on Long-Term Mel Frequency Band Features. In *International Conference on Text, Speech, and Dialogue*. Springer, 352–358.

33. Abhishek Sehgal and Nasser Kehtarnavaz. 2018. A convolutional neural network smartphone app for real-time voice activity detection. *IEEE Access* 6 (2018), 9017–9026.

34. Abhishek Sehgal, Fatemeh Saki, and Nasser Kehtarnavaz. 2017. Real-time implementation of voice activity detector on ARM embedded processor of smartphones. In *2017 IEEE 26th International Symposium on Industrial Electronics (ISIE)*. IEEE, 1285–1290.

35. Jongseo Sohn, Nam Soo Kim, and Wonyong Sung. 1999. A statistical model-based voice activity detection. *IEEE signal processing letters* 6, 1 (1999), 1–3.

36. Adela C Timmons, Theodora Chaspari, Sohyun C Han, Laura Perrone, Shrikanth S Narayanan, and Gayla Margolin. 2017. Using multimodal wearable technology to detect conflict among couples. *Computer* 50, 3 (2017), 50–59.

37. Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. 2016. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth* 4, 3 (2016), e111.

38. John Wiseman. 2016. Python interface to the WebRTC Voice Activity Detector. (2016). Retrieved April, 2019 from `https://github.com/wiseman/py-webrtcvad`