Robust Laughter Detection for Wearable Wellbeing Sensing

Gerhard Hagerer
Chair of Embedded Intelligence for Health Care and
Wellbeing, University of Augsburg
Augsburg, Germany
gerhard.hagerer@informatik.uni-augsburg.de

Nicholas Cummins
Chair of Embedded Intelligence for Health Care and
Wellbeing, University of Augsburg
Augsburg, Germany
nicholas.cummins@ieee.org

Florian Eyben
audEERING GmbH
Gilching, Germany
fe@audeering.com

Björn Schuller
Imperial College
London, United Kingdom
bjoern.schuller@imperial.ac.uk

ABSTRACT

To build a noise-robust online-capable laughter detector for behavioural monitoring on wearables, we incorporate context-sensitive Long Short-Term Memory Deep Neural Networks. We show our solution's improvements over a laughter detection baseline by integrating intelligent noise-robust voice activity detection (VAD) into the same model. To this end, we add extensive artificially mixed VAD data without any laughter targets to a small laughter training set. The resulting laughter detection enhancements are stable even when frames are dropped, which happen in low resource environments such as wearables. Thus, the outlined model generation potentially improves the detection of vocal cues when the amount of training data is small and robustness and efficiency are required.

CCS CONCEPTS

• **Computer systems organization** → **Neural networks**; • **Human-centered computing** → *Mobile devices*; • **Applied computing** → *Health informatics*;

KEYWORDS

Health Monitoring, Laughter Detection, Recurrent Neural Networks

ACM Reference Format:
Gerhard Hagerer, Nicholas Cummins, Florian Eyben, and Björn Schuller. 2018. Robust Laughter Detection for Wearable Wellbeing Sensing. In *DH'18: 2018 International Digital Health Conference, April 23–26, 2018, Lyon, France.* ACM, April 23-26, 2018, Lyon, France, 2 pages. https://doi.org/10.1145/3194658.3194693

1 INTRODUCTION

Over the last years, the market of mobile health software on wearables shows a constant and strong growth with respect to sales numbers of tracking apps analysing the behaviour and habits of customers in terms of health and wellbeing monitoring [1, 2]. A relevant use case therefore is laughter tracking on wearables as laughing affects health and wellbeing in a positive way [4, 5].

In terms of audio processing, automated laughter detection research so far has its primary focus on offline analysis of speech, e. g., [3, 6, 9]. However, for in-the-wild real-time monitoring tasks, there – to our best knowledge – appears to be no research considering robustness, e. g., frame drops and several noise types. Thus, we propose a data-driven method to improve laughter detection Recurrent Neural Networks (RNNs) on sparse laughter training targets while making it robust to difficult real-life scenarios.

In Section 2, we explain our modelling technique and data preparation. Section 3 describes the experimental design and results. These are concluded by Section 4.

2 LAUGHTER DATA & MODELLING

Our laughter detection system is based on the Long Short-Term Memory (LSTM) RNN methodology presented in [7, 8] with the main difference of containing a laughter detection output in addition to a voice activity detection (VAD) output.

Multitask Learning. For the present work, Multitask Learning is utilised, since the laughter data at hand is lacking many conditions which are relevant for our target use case scenario of laughter tracking on wearables. Some of the relevant factors are robustness to stationary and non-stationary background noises, signal loss, microphone types, and environmental impulse responses. By training a robust VAD [8] on a correspondingly prepared corpus, we explore the effects on integrated laughter detection when this data is combined with sparse laughter targets.

Laughter Corpus. As a basis to train our presented laughter detection models, we utilise the labelled laughter data from the SSPNet Vocalization Corpus (SVC) from the 2013 Interspeech Computational Paralinguistics Challenge Social Signals subtask [9].

VAD Corpus. The training data created for VAD is described in [8]. We mixed conversational and emotional speech with multiple background and convolutive noises to ensure robustness to difficult acoustic conditions. This corpus is called the *noisy VAD corpus*, the one without background noises is called *VAD corpus.*

Combined Laughter and VAD data. Since there is no laughter data in the VAD corpus, the annotated laughter recordings from the SVC dataset are added to the VAD training and development set. On this combination, both the VAD and laughter output of the net are trained. The sparse laughter annotations from the VAD

DH'18, April 23–26, 2018, Lyon, France

ACM ISBN 978-1-4503-6493-5/18/04.
https://doi.org/10.1145/3194658.3194693

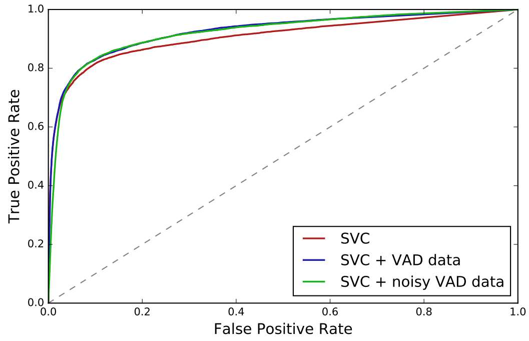

Figure 1: Receiver operator characteristics of frame-wise laughter detection. The proposed models trained on SVC and our VAD data improved the SVC-only baseline.

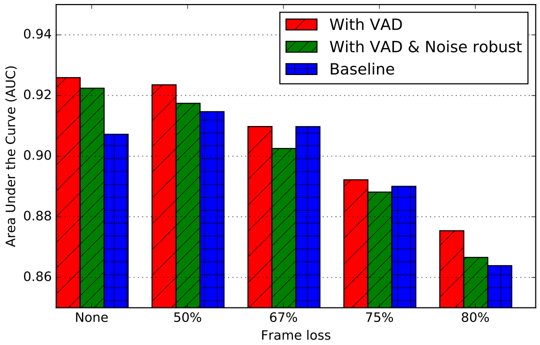

Figure 2: Area Under the Curve (AUC) measures decline the more frames are dropped. The combined VAD+SVC models outperform the baseline in most cases.

data are suppressed. As compensation, the SVC laughter targets are weighted during backpropagation by a factor of 10.

3 EXPERIMENTS & RESULTS

Experiments. Three laughter detection models are compared to each other; differing on the data they were trained on:

SVC: Only SSPNet Vocalization Corpus (SVC) with laughter targets. This is referred to as the *baseline model* or *dataset*.

SVC + VAD: This is the combination of SVC laughter *and* VAD data *without* background noise. The VAD targets on SVC are set to undefined. On the VAD set, the laughter targets are suppressed.

SVC + noisy VAD: Same as SVC + VAD, except of background noise being added to the VAD audio data, but not on SVC.

The assumption is that the laughter detection RNN improves when VAD is included in the model. Thus, the network distinguishes not only between laughter and non-laughter, but also between laughing, general, and absence of speech. We speculate this improves the laughter detection, since it has a more accurate understanding of what is *not* laughter, i. e., general speech activity and background noise. The influence of the latter is considered separately in our experiments by the SVC + noisy VAD train data.

Results. Regarding performance, we evaluate our models only on the laughter targets of the SVC test set. Figure 1 gives an impression in terms of receiver operator characteristics (ROC); the blue and green curves, coming from the two SVC + VAD models outperform the SVC only baseline. This is also reflected by greater *area under the curve* values – see [7]. From this it is apparent that training with the additional VAD data is beneficial for laughter detection.

As we found out when running our laughter detection system under real-life conditions [7], the runtime environment of wearables is randomly dropping audio samples and feature frames respectively due to processing overload. However, from Figure 2 it is apparent that our models consistently perform better than the baseline even when frames are dropped as we did it in our experiments. Moreover, a comparison with laughter detectors from related work given in [7] indicates that our approach performs comparable to key studies performed on the same dataset.

4 CONCLUSION

This paper shows an easily extensible way to increase accuracy, generalisation, and robustness of a vocal cue detector exemplified for laughter detection by adding an according output to a VAD neural network and respective vocal cue data to a VAD training corpus. Furthermore, we showed that these advantages hold even when frames are dropped, as we experienced it on smartwatches. Thus, our approach can be used for in-the-wild use cases, where real-time processing, efficiency, and robustness to signal loss or noise are required. In the future, we plan on training and evaluating integrated models for the detection of a variety of vocal cues.

5 ACKNOWLEDGEMENTS

This work was supported by the EU's 7th Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 688835 (RIA DE-ENIGMA).

REFERENCES

[1] 2015. *mHealth App Developer Economics 2015.* research2guidance. 5th annual study on mHealth app publishing based on 5,000 plus respondents.

[2] 2016. *mHealth App Developer Economics 2016.* research2guidance. 6th annual study on mHealth app publishing based on 2,600 plus respondents.

[3] Raymond Brueckner and Bjorn Schuller. 2014. Social signal classification using deep BLSTM recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on.* IEEE, 4823–4827.

[4] Adam Clark, Alexander Seidler, and Michael Miller. 2001. Inverse association between sense of humor and coronary heart disease. *International journal of cardiology* 80, 1 (2001), 87–88.

[5] Lee S. Berk et al. 1989. Neuroendocrine and stress hormone changes during mirthful laughter. *The American journal of the medical sciences* 298, 6 (1989).

[6] Rahul Gupta, Kartik Audhkhasi, Sungbok Lee, and Shrikanth Narayanan. 2016. Detecting paralinguistic events in audio stream using context in features and probabilistic decisions. *Computer Speech & Language* 36 (2016), 72–92.

[7] Gerhard Hagerer, Nicholas Cummins, Florian Eyben, and Björn Schuller. 2017. "Did you laugh enough today?"–Deep neural networks for mobile and wearable laughter trackers. *Proc. Interspeech 2017* (2017), 2044–2045.

[8] Gerhard Hagerer, Vedhas Pandit, Florian Eyben, and Björn Schuller. 2017. Enhancing LSTM RNN-based Speech Overlap Detection by Artificially Mixed Data. In *Audio Engineering Society Conference: 48nd International Conference: Semantic Audio.* Audio Engineering Society.

[9] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. (2013).