INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

"Did you laugh enough today?" – Deep Neural Networks for Mobile and Wearable Laughter Trackers

Gerhard Hagerer[1,2], *Nicholas Cummins*[2], *Florian Eyben*[1], *Björn Schuller*[1,2]

[1]audEERING GmbH, Gilching, Germany
[2]Chair of Complex & Intelligent Systems, University of Passau, Germany
{ghagerer,feyben}@audeering.com, {Nicholas.Cummins,Bjoern.Schuller}@uni-passau.de

Abstract

In this paper we describe a mobile and wearable devices app that recognises laughter from speech in real-time. The laughter detection is based on a deep neural network architecture, which runs smoothly and robustly, even natively on a smartwatch. Further, this paper presents results demonstrating that our approach achieves state-of-the-art laughter detection performance on the SSPNet Vocalization Corpus (SVC) from the 2013 Interspeech Computational Paralinguistics Challenge Social Signals Sub-Challenge. As this technology is tailored for mobile and wearable devices, it enables and motivates many new use cases, for example, deployment in health care settings such as laughter tracking for psychological coaching, depression monitoring, and therapies.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics, laughter detection, neural networks, wearables

1. Introduction

With more than 100,000 new apps and 3 billion downloads each year, the market for mobile health software applications is increasingly becoming important for customers searching for ways to change their habits and every-day behaviours in such ways that they contribute positively to their overall state of health and wellbeing [1, 2]. This includes psychological health apps that monitor and give feedback to the customer on their mental state. As laughing has been shown to affect health overall in an overall positive way [3, 4] as well as contributing to social wellbeing [5], therefore software encouraging users to laugh more appears as a relevant use case in the health-app market. Whilst solutions in this regard have already been presented, e. g., [6], these neither show competitive results for noise robust laughter detection nor the ability to run in real-time directly on low resource technologies such as smartwatches. Furthermore, there is no similar affective technology based on deep neural networks which is executable on mobile and wearable devices currently being presented to a wider audience - at least not to the best of the authors' knowledge. Last but not least, our presented software shows the potential for actual context aware tracking of laughter during the day in terms of robustness, efficiency, and functionality.

2. Application Scenario

We realised an app running on a smartphone as well as on a wearable device, the recognition of laughter will be performed in realtime from the devices input audio microphone. For both scenarios, the app is capable of giving a visual indication when laughter is recognised from speech or not, as depicted in Figure 1. Additionally, general voice activity is detected and shown at the same time. The detection of both laughter and speech is performed in a noise robust way, which means that the detection works well also on inferior microphones, as imposed by smartphones and especially smartwatches, and despite background noise and reverberated rooms – the following methodology section explains the model training in detail.

Figure 1: *Runtime screenshots from a smartwatch: no speech, speech, and laughter as detected in real-time from the microphone signal.*

For the laughter detection, a counter is implemented showing the number of laughs that were detected since the start of the app. The counting performs such that it increments in the background, e. g., when the device is in standby mode. This means the app constantly logs statistics about the laughing activity during the day. This is a first step towards social and psychological context awareness. Although the app is not yet optimised for long continuous use, the current version demonstrates how such context awareness can be delivered, what it might look like, and which benefits it brings to potential customers.

3. Methodology

The underlying Machine Learning technology to detect laughter is built upon Long Short-Term Memory Recurrent Neural Networks (LSTM RNN), as these are especially suitable for context sensitive sequence analysis such as for audio streams. Regarding the details of the chosen neural network structure, the interested reader is referred to [7], which explains, not only the technology, but also most of the used data in full detail.

As an overview, our system consists of a LSTM RNN with two hidden layers (120 and 90 neurons), and two frame-wise regression outputs for general speech and laughter activity are used – see Figure 2. As input features, 50 Mel-spectra, 20 MFCCs, and the delta features for each of these are taken, i. e., $(50 + 20) \cdot 2 = 140$ features per 20 ms frame and 10 ms step size. The output frame and step size are identical.

Concerning the training data used, it is nearly identical to

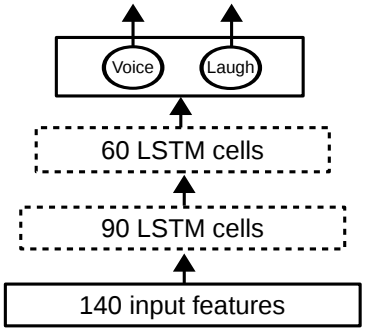

Figure 2: *Illustration of the LSTM structure: The inputs are audio feature vectors. Hidden layers are dashed.*

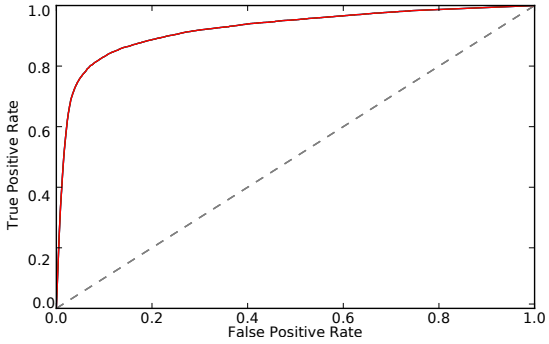

Figure 3: *Receiver operating characteristic for the LSTM RNN laughter regression outputs, evaluated on the SVC test set.*

that described in [7], which is conversational speech mixed with several kinds of stationary and non-stationary noises and convolved with various degrading impulse responses. However, due to the lack of laughter data therein, labelled laughter data from the SSPNet Vocalization Corpus (SVC) from the 2013 Interspeech Computational Paralinguistics Challenge Social Signals subtask [8] was added to the training and development set from the reference. The sparse laughter annotations from the referenced conversation and emotion corpora were suppressed if available at all. In return, the loss function was weighted during backpropagation for the laughter targets from the SVC set.

The performance measurements from the chosen approach are shown in Table 1 and Figure 3. Table 1 indicates that our approach performs comparable to key studies performed on the same laughter datasets; the mere differences in performance are most likely being due to the extra training data used to help ensure the noise robustness of our approach. This advantage is not apparent from the performance measurements (the test data does not contain noise), but from the real-life showcase.

4. Conclusions

In the present work, the technology of our laughter detection was outlined. Its outstanding features are real-time capability while natively running directly on both mobile and wearable devices, robustness for real-life usage, state-of-the-art performance, and laughter statistics during the day. These features and the concept are part of the show & tell demonstrations at the Interspeech 2017 conference.

The definite target of our implementation is to overcome issues like energy consumption such that battery usage does not suffer significantly when laughter activity is tracked through the whole day. This will make the app potentially market ready. It is our belief that future psychological and medical research will gain profit from such a tool which encourages subjects to focus on a behaviour for a healthier and more balanced lifestyle.

Table 1: *Measurements based on the SVC test set.*

Measurement	Percent
Best Accuracy	96.95
Best F1–Score	54.21
Equal Error Rate	13.93
Area Under the Curve	92.24
Area Under the Curve [9]	94.26
Area Under the Curve [10]	94.0
Area Under the Curve [11]	93.3

5. Acknowledgements

This work is supported by the EUs Horizon 2020 Programme through the Innovative Action No. 688835 (DE-ENIGMA).

6. References

[1] "mHealth App Developer Economics 2015," Research 2 Guidance, Tech. Rep. 5th, Nov 2015.

[2] "mHealth App Developer Economics 2016," Research 2 Guidance, Tech. Rep. 6th, Oct 2016.

[3] A. Clark, A. Seidler, and M. Miller, "Inverse association between sense of humor and coronary heart disease," *Int. J. Cardiol.*, vol. 80, no. 1, pp. 87–88, 2001.

[4] L. S. Berk, S. A. Tan, W. F. Fry, B. J. Napier, J. W. Lee, R. W. Hubbard, J. E. Lewis, and W. C. Eby, "Neuroendocrine and stress hormone changes during mirthful laughter," *Am. J. Med. Sci.*, vol. 298, no. 6, pp. 390–396, 1989.

[5] F. Lingenfelser, J. Wagner, E. André, G. McKeown, and W. Curran, "An event driven fusion approach for enjoyment recognition in real-time," in *Proc. of ACM MM*. Orlando, FL, USA: ACM, 2014, pp. 377–386.

[6] S. Flutura, J. Wagner, F. Lingenfelser, A. Seiderer, and E. André, "Laughter detection in the wild: demonstrating a tool for mobile social signal processing and visualization," in *Proc. ICMI*. Tokyo, Japan: ACM, 2016, pp. 406–407.

[7] G. Hagerer, V. Pandit, F. Eyben, and B. Schuller, "Enhancing lstm rnn-based speech overlap detection by artificially mixed data," in *Proc. ISCA*. Erlangen, Germany: AES, June 2017, pp. 1–8.

[8] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 148–152.

[9] G. Gosztolya, R. Busa-Fekete, and L. Tóth, "Detecting Autism, Emotions and Social Signals Using AdaBoost," in *Proc. INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 220–224.

[10] R. Brückner and B. Schuller, "Social Signal Classification Using Deep BLSTM Recurrent Neural Networks," in *Proc. ICASSP*. Florence, Italy: IEEE, 2014, pp. 4856–4860.

[11] R. Gupta, K. Audhkhasi, S. Lee, and S. Narayanan, "Paralinguistic event detection from speech using probabilistic time-series smoothing and masking." in *Proc. INTERSPEECH*. Lyon, France: ISCA, 2013, pp. 173–177.