Baby Cry Recognition in Vehicles

Jan Baumann, Timo Lohrenz, Jan Franzen, Tim Fingscheidt
Institute for Communications Technology, Technische Universität Braunschweig,
Schleinitzstr. 22, 38106 Braunschweig, Germany, Email: {jan.baumann, t.lohrenz, j.franzen, t.fingscheidt}@tu-bs.de

Abstract

In an automotive environment the driver's attention to traffic and the surrounding environment is a safety-critical aspect. Distracting situations and events around the driver can significantly decrease his/her focus and increase the risk of driving mistakes and accidents. Accordingly, it is desirable to prevent an unfocused state of the driver. Various systems have been designed to alert the driver in such situations and thereby prevent possible damage. While these systems rely on input from various sensors, it is also possible to obtain information from acoustic signals as they are acquired in the vehicle via the already existing hands-free microphone. Recent systems for automatic sound event detection can be used to detect and classify various types of sound events, even when high-level background noise is present. When it comes to sound events diverting the attention of the driver, a baby crying in the backseat is clearly a dangerous situation. For this use-case we propose a neural network-based detector for baby cries in vehicles. The method detects baby cries in the presence of typical vehicle noise, and in consequence triggers an increased support from driver assistance systems or autonomous driving functions. At moderate event-to-background power ratios, the method achieves an error rate below 10% and a strong F1 measure of above 95%.

Introduction

Humans in their role as drivers in an automotive environment, are perceiving numerous visual and audible events. They are usually able to detect and sort these events based on their importance for safely driving a vehicle. In this context, especially rare and/or high-power events, such as a baby crying in the back seat, causing the driver to direct his/her focus towards them, as they can implicit a situation with a need for reacting. An unfocused driver is more likely to make driving mistakes which can potentially endanger the vehicle and its surroundings of the vehicle, including the driver.

Various driver assistance systems for tracking the driver's focus have been proposed in the past, often relying on compute-intensive visual information, e.g., tracking the eyes and their blink frequency [1]. The automatic detection and/or classification of sound events (sound event detection, SED) has been a research topic for a long time. Earlier approaches were based on Gaussian mixture models and hidden Markov models for monophonic SED to model the probability distributions of sound events and their transition probabilities, e.g., [2, 3]. With the enormous evolution of neural networks, new classifiers were developed with the ability to predict multiple event classes at the same time, called polyphonic SED [4, 5]. These classifiers are able to outperform classical approaches [6], and sometimes even human listeners, in most classification tasks, where state-of-the-art performance can be achieved with large amounts of training data.

When it comes to audio classification tasks, very good or even the best performance can be achieved with a convolutional recurrent neural network (CRNN), not only for SED, but also ,e.g., for speech emotion recognition [7]. An example herefore, which our work is based on, is the SED-CRNN [8], which was successfully used in various applications and challenges, for example in [9], where it achieved a new state-of-the-art performance. This architecture is shown in Fig. 1 and is composed of a first, convolutional neural network (CNN) part, a second recurrent neural network (RNN) part and a last fully connected part. The CNN is capable of learning to extract neccessary information along the frequency axis for a short context over the time axis, thereby reducing the spectral resolution until it can be processed by the RNN, where a long-term context over the time axis is learned. The classifier output is processed with one or more fully connected layers after the RNN. On the last layer, one output node is deployed for each trained class, returning the event activity probability for each class.

Using the advancements in neural networks and sound event detection, we propose a neural network-based detector for baby cries in the presence of car noises. Thereby, we are laying the fundamentals for a support system, which is able to point out time windows of a possible unfocused driver and in consequence triggers increased support from driver assistance systems or autonomous driving functions.

Next, we present our novel classifier for baby cries and then, we discuss experiments for evaluating the performance of the proposed classifier.

Our Approach

As we design a single-event 'baby cry' classifier, we adapt the SED-CRNN [8] for our use-case, shown in Fig. 1. To extraxt features from an audio signal, we apply an FFT with a frame length of 40 ms and a frame shift of 20 ms to the monaural ($C = 1$ channels) input audio signal and extract $M = 40$ mel filterbank features $\boldsymbol{x}_\ell$ each frame. For frame with index ℓ, an input context of seven frames $[\boldsymbol{x}_{\ell-3}, \boldsymbol{x}_{\ell-2}, ..., \boldsymbol{x}_\ell, ..., \boldsymbol{x}_{\ell+3}]$ is needed to process the outputs.

For the CNN layers (Conv), we use $F = 96$ filter kernels

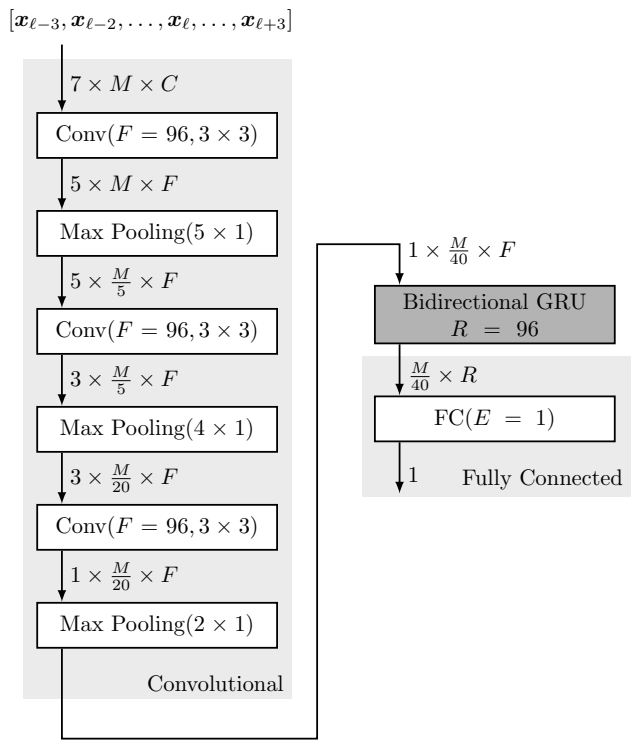

Figure 1: Overview of the proposed CRNN topology.

with the size of 3×3 each. Symmetrical zero padding is added to the feature maps before each convolution to retain the same feature map size for the convolution output. The size along the frequency axis is reduced step-by-step with max pooling of the size $\{5, 4, 2\} \times 1$, so that the resulting feature map size after three layers of convolution and max pooling is 1×1. These feature maps are concatenated per frame to an $1 \times F$ feature vector, which is processed by the RNN layer, in this case a bidirectional gated recurrent unit (GRU), allowing the processing of future context. This requires some further lookahead in a realtime application, and in exchange improves the classification accuracy. After the recurrent layer, a vector of $R = 96 \times 1$ features is processed by the fully connected (FC) layer with a single output node $E = 1$, which is applied to each output of the RNN layer with the same weights.

Experiments

In this section, we present experiments to investigate the practicability of our baby cry detector. First, we generate a data set suitable for this task. Afterwards, our CRNN is optimized so that its performance on the test set can be evaluated.

For evaluating our classifier, we use the segment-based error rate $\mathrm{ER}_{\mathrm{sg}}$ and F1-score $\mathrm{F1}_{\mathrm{sg}}$, as described in [10]. The annotations and outputs are evaluated in one second long segments. An annotation or output during one segment will count the whole segment as an annotated, i.e., detected, event. The segment-based error rate is calculated by

$$\mathrm{ER}_{\mathrm{sg}} = \frac{I + D}{N},$$

where I is the count of segments with insertions, D is the count of segments with deletions and N being the count of all annotated segments each in the entire used test set. As another measure for the classifier performance, we use the segment-based F1-score $\mathrm{F1}_{\mathrm{sg}}$. It is calculated with two times the true positives $2 \times TP$, divided by the sum of all event outputs $TP + FP$ and all positive annotations $TP + FN$:

$$\mathrm{F1}_{\mathrm{sg}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}.$$

For our experiments, we generate a data set consisting of 1000 sequences of 30 seconds duration with a sample rate of $f_s = 48$ kHz. The sequences are composed of a 30 second background excerpt from proprietary vehicle noise recordings and baby cry events taken from freesound [11] and annotated for use in SED in context with the DCASE 2017 challenge task on rare SED [12]. Generating the sequences, a random background excerpt is chosen from approximately five hours of total recordings and one baby cry recording is randomly added from a total of 148 different baby cries. Two separate data sets are generated with an event-to-background power ratio (EBR) set to either +0 dB or +10 dB. We split the generated data set into a training set of 600 sequences, a validation set of 200 sequences and a final test set consisting of 200 sequences as well. The used background sequences and baby cries are disjunct between all three subsets.

We train our CRNN classifier on the training set and minimize the segment-based error rate $\mathrm{ER}_{\mathrm{sg}}$ on the validation set. We use the Adam algorithm [13] for optimizing the CRNN weights. The training is performed for a maximum of 200 epochs with an early stopping patience of 50 epochs after the last improvement of the error rate on the validation set. The resulting network is then evaluated on the test set under matching EBR conditions, yielding the final event based error rates $\mathrm{ER}_{\mathrm{sg}}$ and F1-scores $\mathrm{F1}_{\mathrm{ev}}$.

The classifier is tested on the test set, where we achieve an error rate $\mathrm{ER}_{\mathrm{sg}}$ of 9.28% and an F1-score $\mathrm{F1}_{\mathrm{sg}}$ of 95.42% on the test set, as shown in Table 1. We can observe a significantly better performance on the validation set (dev set in Tab. 1) than on the actual test set for an EBR of +0 dB, which is likely caused by some model overfitting to the training data. Also, higher EBRs tend to allow for lower error rates as one would expect comparing it with speech intelligibility at high and low SNR conditions.

	Dev Set		Test Set	
Model / EBR test cond.	$\mathrm{ER}_{\mathrm{sg}}$[%]	$\mathrm{F1}_{\mathrm{sg}}$[%]	$\mathrm{ER}_{\mathrm{sg}}$[%]	$\mathrm{F1}_{\mathrm{sg}}$[%]
EBR= +10 dB	9.05	95.58	9.28	95.42
EBR= +0 dB	13.65	93.43	21.49	90.11

Table 1: Performance of our CRNN used for baby cry detection.

Conclusion

We proposed a neural network as a detector for baby cry in an automotive environment. The exprimental results show good performance under laboratory conditions, e.g., an F1-Score of over 90%, so that a deployment and testing under real-world conditions would be desirable. It can be used to enhance existing driver assistance systems, indicating a possible lack of attention, when a baby is crying in the back seat. Furthermore, our classifier could be expanded to a polyphonic detector, also being capable of detecting other sound events known to impair the driver's behaviour.

References

[1] F. Zhang, J. Su, L. Geng and Z. Xiao: "Driver Fatigue Detection Based on Eye State Recognition", in *Proc. of CMVIT*, Singapore, Feb. 2017, pp. 105–110

[2] P. Transfeld, S. Receveur, and T. Fingscheidt: "Towards Acoustic Event Detection for Surveillance in Cars", in *Proc. of 11th ITG Conference on Speech Communication*, Erlangen, Germany, Sept. 2014, pp. 1–4

[3] P. Transfeld, S. Receveur, and T. Fingscheidt: "An Acoustic Event Detection Framework and Evaluation Metric for Surveillance in Cars", in *Proc. of Interspeech*, Dresden, Germany, Sept. 2015, pp. 2927–2931

[4] G. Parascandolo, H. Huttunen and T. Virtanen: "Recurrent Neural Networks for Polyphonic Sound Event Detection in Real Life Recordings", in *Proc. of ICASSP*, Shanghai, China, Mar. 2016, pp. 6440–6444

[5] J. Baumann, T. Lohrenz, A. Roy and T. Fingscheidt: "Beyond the DCASE 2017 Challenge on Rare Sound Event Detection: A Proposal for a More Realistic Training and Test Framework", *accepted for publication on ICASSP 2020*, Barcelona, Spain, May 2020, pp.1–5

[6] O. Gencoglu, T. Virtanen and H. Huttunen: "Recognition of Acoustic Events Using Deep Neural Networks", in *Proc. of EUSIPCO* Lisbon, Portugal, Sept. 2014, pp. 506–510

[7] P. Meyer, Z. Xu and T. Fingscheidt: "Improving Convolutional Recurrent Neural Networks for Log-Mel Spectrogram-based Speech Emotion Recognition", *submitted to Interspeech*, Shanghai, China, Oct. 2020, pp. 1–5

[8] E. Cakir, G. Parascandolo, T. Heittola, H. Huttunen and T. Virtanen: "Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection", in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, Number 6, June 2017, pp. 1291–1303

[9] E. Cakir and T. Virtanen: "Convolutional Recurrent Neural Networks for Rare Sound Event Detection", in *DCASE2017 Challenge*, Munich, Germany, Sept. 2017, pp. 1–5

[10] A. Mesaros, T. Heittola and T. Virtanen: "Metrics for Polyphonic Sound Event Detection", in *Applied Sciences*, Vol. 6, Number 6, May 2016, pp. 1–17

[11] B. d. Jong, A. Porter and others: "Freesound", `https://freesound.org`, [Online; accessed 01.03.2020]

[12] A. Mesaros, T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj and T. Virtanen: "DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System", in *Proc. of DCASE2017*, Munich, Germany, Nov. 2017, pp. 85.92

[13] D. P. Kingma and J. Ba: "Adam: A Method for Stochastic Optimization", in *arXiv Preprint*, arXiv:1412.6980, Dec. 2014, pp. 1–15