Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks

Egor Lakomkin[1], Cornelius Weber[1], Sven Magg[1] and Stefan Wermter[1]

I. RESEARCH MOTIVATION

In the near future, the presence of robots in home environments will become more common, helping humans with daily tasks, for instance assisting elderly people. One important area where robots can be very helpful is identifying possible dangerous situations in home environments. Robots can observe the situation at the moment and try to evaluate if there is a potential threat. A robot can be taught to do this with machine learning methods. As an example, given a speech segment, it would be interesting to predict if a person is excited or neutral. In this paper, firstly I outline main questions and directions in my PhD research, then I present current achieved results followed by the related and future work sections.

II. RESEARCH DIRECTIONS AND ACHIEVED RESULTS

I identify three main directions in my research: 1) features and signal representations learning for speech emotion recognition (SER) task. 2) investigation of neural architectures which allow robust to an internal robot's and an environmental noise emotion recognition 3) research on the methods and approaches to incorporate information contained in modalities other than auditory to improve speech emotion recognition. For example, linguistic analysis of a spoken text or facial expression recognitions can help in difficult situations when analyzing only acoustic signal is not enough to infer an affective state of the speaker.

A. FEATURES FOR SPEECH EMOTION RECOGNITION

I evaluate several dual architectures which integrate representations of the automatic speech recognition (ASR) neural network: a fine-tuning and a progressive network. The fine-tuning architecture reuses features learnt by the recurrent layers of a speech recognition network and can use them directly for emotion classification by feeding them to a softmax classifier or can add additional hidden SER layers to tune ASR representations. Additionally, the ASR layers can be static for the whole training process or can be updated as well by allowing to backpropagate through them. The progressive architecture complements information from the ASR network with SER representations trained end-to-end. Our experiments on the IEMOCAP dataset show 10% relative improvements in the accuracy and F1-score over the baseline recurrent neural network which is trained end-to-end for emotion recognition. Results were published and presented at the IJCNLP 2017 conference [1].

[1]University of Hamburg, Department of Informatics, Knowledge Technology Institute. Vogt-Koelln-Strasse 30, 22527 Hamburg, Germany lakomkin @informatik.uni-hamburg.de

B. LEARNING EARLY EMOTION RECOGNITION

Acoustically expressed emotions can make communication with a robot more efficient. Detecting emotions like anger could provide a clue for the robot indicating unsafe/undesired situations. Recently, several deep neural network-based models have been proposed which establish new state-of-the-art results in affective state evaluation. These models typically start processing at the end of each utterance, which not only requires a mechanism to detect the end of an utterance but also makes it difficult to use them in a real-time communication scenario, e.g. human-robot interaction. We propose the EmoRL model that triggers an emotion classification as soon as it gains enough confidence while listening to a person speaking. As a result, we minimize the need for segmenting the audio signal for classification and achieve around 50% latency reduction as the audio signal is processed incrementally. The method is competitive with the accuracy of a strong baseline model, while allowing much earlier prediction. The results will be presented at the ICRA 2018 conference.

C. ROBUST ACOUSTIC EMOTION RECOGNITION

Many neural network-based architectures were proposed recently and pushed the performance to a new level. However, the applicability of such neural SER models trained only on in-domain data to noisy conditions is currently under-researched. In this work, we evaluate the robustness of state-of-the-art neural acoustic emotion recognition models in human-robot interaction scenarios. We hypothesize that a robot's ego noise, room conditions, and various acoustic events that can occur in a home environment can significantly affect the performance of a model. We conduct several experiments on the iCub robot platform and propose several novel ways to reduce the gap between the model's performance during training and testing in real-world conditions. Furthermore, we observe large improvements in the model performance on the robot and demonstrate the necessity of introducing several data augmentation techniques like overlaying background noise and loudness variations to improve the robustness of the neural approaches. The results were published at the IROS 2018 conference[1].

[1]Video https://www.youtube.com/watch?v=js_TCxl_wF4

D. SEMI-SUPERVISED EMOTION RECOGNITION

One of the issues in the area of affective computation is that the amount of annotated data is very limited. On the other hand, the number of ways that the same emotion can be expressed verbally is enormous due to variability between speakers. This is one of the factors that limits performance and generalization. We propose a simple method that extracts audio samples from movies using textual sentiment analysis. As a result, it is possible to automatically construct a larger dataset of audio samples with positive, negative emotional and neutral speech. We show that pretraining recurrent neural network on such a dataset yields better results on the challenging EmotiW corpus. This experiment shows a potential benefit of combining textual sentiment analysis with vocal information. The results were published and presented at the EACL 2017 conference.

III. RELATED WORK

Deep neural networks significantly boosted the performance of acoustic emotion recognition models. The majority of recent work focuses on learning to extract useful input representations and searching for neural architectures for emotion recognition, as neural approaches outperform traditional ones like support vector machines and decision trees [2].

Recurrent neural networks have an ability to model long-term context information and were successfully applied to emotion recognition [3], [4]. Convolutional neural networks can capture only a local context, but have an ability to model longer dependencies when their architecture was designed with a deep hierarchy [2]. Commonly, these methods train neural networks on pre-extracted features: MFCC coefficients, spectrograms and high-level information like formants, pitch, and voice probability. Alternatively, Trigeorgis et al. demonstrate a model that learns how to recognize the affective state of a person directly from the raw waveform [5]. Another explored direction is transfer learning: adapting audio representations trained initially for other auxiliary tasks, like gender and speaker identification [6] or speech recognition [1], [7].

Robustness to noise was a subject of several previous work. Attention mechanisms [3], [8] aim to identify useful regions for emotion classification automatically by assigning a low importance to irrelevant inputs, for example, non-speech or silence frames. Adding background noise during training improved the robustness of neural models in several acoustic classification tasks [9]. Different types of data augmentation methods were explored by Zhou et al. [10] to improve the performance of speech recognition. Supervised domain adaptation was proposed by Abdelwahab et al. [11] to mitigate the problem of training and testing mismatch conditions by tuning the model on the small set of test samples.

Our work on the robustness of the speech emotion recognition is close to Lane et al. [9] and our main difference is that our testing conditions are not synthetically constructed by overlaying clean samples with additive noise, but recorded on the iCub robot which adds a significant amount of ego-noise. We argue that distortions introduced by playing a sample through speakers, changing room conditions and distance from the speech source to the robot, reverberations, added external acoustic events and the robot's internal noise introduce non-linear deformations which are challenging for the neural network to deal with.

IV. FUTURE WORK

In future work, I plan to investigate further ways to enhance the data augmentation pipeline for a robust speech emotion recognition. For example, data-driven generative models, like generative adversarial networks, can produce realistic speech samples, which potentially can be useful during training. I plan to evaluate an option to enrich input representation with the information on the spoken text under noisy conditions as it appears to be difficult to analyze valence without it.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642667 (SECURE)

REFERENCES

[1] E. Lakomkin, C. Weber, S. Magg, and S. Wermter, "Reusing Neural Speech Representations for Auditory Emotion Recognition," *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, vol. 1, pp. 423–430, 2017. [Online]. Available: http://www.aclweb.org/anthology/I17-1043

[2] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for Speech Emotion Recognition," *Neural Networks*, vol. 92, pp. 60–68, 8 2017.

[3] C.-W. Huang and S. Narayanan, "Attention Assisted Discovery of Sub-Utterance Structure in Speech Emotion Recognition," *Proceedings of Interspeech*, pp. 1387–1391, 2016.

[4] J. Lee and I. Tashev, "High-level Feature Representation using Recurrent Neural Network for Speech Emotion Recognition," *Interspeech*, 2015.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3 2016, pp. 5200–5204.

[6] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, E. M. Provost, and A. Arbor, "Progressive Neural Networks for Transfer Learning in Emotion Recognition," *Interspeech*, pp. 1098–1102, 2017.

[7] H. M. Fayek, M. Lech, and L. Cavedon, "On the Correlation and Transferability of Features between Automatic Speech Recognition and Speech Emotion Recognition," *Interspeech*, pp. 3618–362, 2016.

[8] M. Neumann and N. T. Vu, "Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech," *Interspeech*, pp. 1263–1267, 2017.

[9] N. D. Lane, P. Georgiev, L. Qendro, and B. Labs, "DeepEar: Robust Smartphone Audio Sensing in Unconstrained Acoustic Environments using Deep Learning," *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 283–294, 2015.

[10] Y. Zhou, C. Xiong, and R. Socher, "Improved Regularization Techniques for End-to-End Speech Recognition," *CoRR*, vol. abs/1712.07108, 2017. [Online]. Available: http://arxiv.org/abs/1712.07108

[11] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4 2015, pp. 5058–5062. [Online]. Available: http://ieeexplore.ieee.org/document/7178934/