Deep Recurrent Neural Networks for Emotion Recognition in Speech

Maximilian Schmitt[1], Björn Schuller[1,2]

[1] *ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany*

[2] *GLAM – Group on Language, Audio & Music, Imperial College London, U. K.*

Email: maximilian.schmitt@informatik.uni-augsburg.de

Abstract

Emotion recognition in speech (ERS) is a hot research topic in the field of affective computing. Giving computers the ability to know the emotions from a subject is an important aspect in naturalistic human-computer interaction or user profiling. Recent methods to tackle the complex task of ERS employ deep neural networks. As human emotion is an affective state, which changes over time, the neural network has the task of sequence-to-sequence modelling. This is usually implemented as a recurrent neural network (RNN), such as a long short-term memory RNN. Even though in recent work, the audio signal is directly fed into the network, the most common approach is to feed the RNN with acoustic low-level features, describing prosodic, spectral, or cepstral characteristics of the speech. Instead of using the raw low-level features, they can also be encoded in terms of the bag-of-audio-words approach, where the feature vectors are quantised using a previously learnt codebook of templates and the occurrence of each template is encoded in a sparse histogram vector. In this contribution, we propose a deep learning framework for ERS and compare different feature representations. Results are presented using a state-of-the-art benchmark database from the domain of affective computing.

Introduction

The automatic recognition of human emotions from speech or other vocalisations, such as laughter, is a prevailing topic in the field of *affective computing*. Applications range from naturalistic human-machine interaction to automated market research. Instead of categorising emotions into discrete classes such as *anger*, *happiness*, or *boredom*, the two-dimensional continuous model of *arousal* and *valence* is quite common [1, 2]. While arousal specifies the level of alertness of a person, valence (or *pleasure*) specifies whether the emotion is positive or negative. In contrast to personality traits, human emotion is a *state* that may vary rapidly over time [3], e. g., within successive speaker turns in a conversation. This is why recent multimodal affect databases, such as *RECOLA* [3] or *SEWA* [4] provide time-continuous annotations of emotional dimensions.

Most of the methods to tackle the complex problem of automatic emotion recognition proposed during the last two decades are based on *machine learning*, i. e., they are approaches that learn models from sample data. From the methodological point of view, a large variety of methods has been investigated, such as *hidden Markov models* [5], or acoustic feature brute-forcing [6] and *bag-of-audio-words* [7] in combination with a static machine learning model such as *support vector machine (SVM)*. The latter approaches involve the problem that they take into account only a limited amount of temporal context. Specific architectures of *neural networks*, the so-called *recurrent neural networks (RNNs)*, are capable of modelling time-series. A milestone in this context have been the *long short-term memory-RNNs (LSTM-RNNs)*, proposed by Hochreiter and Schmidhuber [8], making it possible to consider and memorise context over an (in theory) unlimited time. Eyben et al. have shown that the utilisation of LSTM-RNN architectures to dynamically model emotion in speech is superior to static modelling with an SVM [9]. Moreover, they have proposed to train the neural network simultaneously on multiple targets ('multi-task learning'), i. e., multiple emotional dimensions, in order to increase the accuracy of the prediction by providing the training process with side information and to exploit dependencies. They used an architecture consisting of two hidden LSTM layers, however, optimisation of all model hyperparameters is time-consuming. The winners of the 2017 *Audio-Visual Emotion Challenge (AVEC)*, a yearly research challenge in the field, also found that LSTM-RNNs outperform static classifiers and that multi-task learning outperforms single-task learning [10]. As acoustic features, they fused both neural network-learnt representations and hand-crafted bruteforced feature sets.

In principle, LSTM-RNNs can handle the raw acoustic short-term features (*low-level descriptors, LLDs*), extracted on audio frame-level, without the need to summarise them over a larger segment as it is needed for static classifiers. However, it is quite common to apply functionals, i. e., statistics over the LLDs of a short segment in time. In this work, we investigate the influence of different acoustic feature representations, *functionals* and *bag-of-audio-words (BoAW)* to be used for LSTM-RNN architectures with the AVEC 2017 emotion database. In the following section, we explain the corpus used in our experiments before the employed feature representations and *deep learning* models are presented. Afterwards, results are shown and discussed, followed by the conclusion and an outlook on future research.

Corpus

We used the German subset of the *SEWA* corpus[1], consisting of audio-visual recordings of 64 subjects in dyadic

[1] https://db.sewaproject.eu/

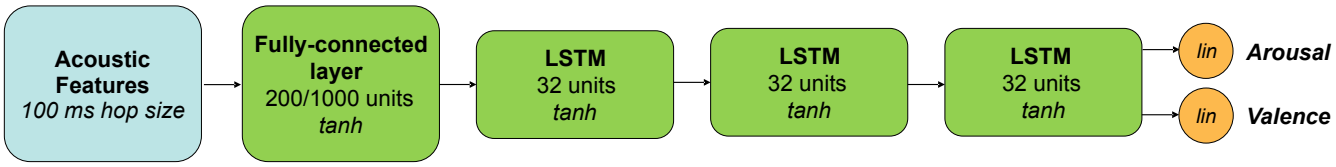

Figure 1: Structure of the investigated emotion recognition model.

interaction through a video chat platform. Each pair of subjects was recorded for up to 3 minutes while discussing a commercial watched beforehand, totalling up to 178 minutes of audio/video. The emotion of each subject was annotated continuously over time in terms of *arousal* and *valence* by six different annotators. The contours of each emotional dimension, annotated using a joystick, were smoothed using a median filter and combined considering inter-rater agreement to form a unique *gold standard*, with a step size of 100 ms. The SEWA corpus was used as the benchmark database in the *Affect* sub-challenge of AVEC 2017 [4]. The same partitioning (34/14/16 subjects for training/development/test) was used in our experiments to ensure comparability and reproducibility of the proposed approach. However, we used only the acoustic domain of the data, neither the video nor the manual text transcriptions that were provided for the challenge. The third dimension annotated in the dataset, *liking*, describing how much the subjects liked the product advertised in the commercial or the commercial itself, is not considered in our experiments as it is mainly reflected in the linguistic content of the speech.

Methodology

Although recent deep learning approaches to emotion recognition in speech (ERS) model the raw waveform of the speech signal directly, referred to as *end-to-end learning* [11, 12], most systems still rely on hand-crafted acoustic features. For ERS, an optimised acoustic feature set exists with the *Geneva Mimimalistic Acoustic Parameter Set* [13], whose *extended* version (eGeMAPS) was used in our experiments. eGeMAPS defines 23 acoustic low-level descriptors (LLDs) that are extracted sequentially from short frames (20 ms - 60 ms) of the signal, where the audio is considered quasi-stationary, with a shift (hop size) of 10 ms. The set mainly consists of voice related parameters, such as *Mel-frequency cepstral coefficients (MFCCs) 1 to 4*, the frequency and amplitudes of the *formants* 1 to 3 (and the bandwidth of the 1^{st} formant), *F0* (pitch), *jitter*, *shimmer*, and *harmonics-to-noise ratio*. In addition to that, *loudness* and 7 spectral balance descriptors (e. g., *spectral flux*) are included. The full list of LLDs is found in the article by Eyben at al. [13].

As the employed RNN architectures are *dynamic* machine learning models, the LLD series can be fed directly into them as an input. However, the target dimensions are given with a frequency of 10 Hz and would need to be upsampled to match with the input series, which increases the time required for training the network. Furthermore, it was observed that training converges better if the LLDs are summarised over a certain block (window) in time, to match with the targets. We compare three different ways to represent the LLDs:

1. The 88 functionals defined in eGeMAPS: *mean*, *standard deviation*, *percentiles*, and *slope* of the F0 and loudness contours and only *mean* and *standard deviation* for the other LLDs (MFCC, spectral descriptors, etc.). In addition to that, some rhythm-related features are computed, namely, the rate of loudness peaks, the mean length and the standard deviation of continuously voiced and unvoiced regions, and the number of continuous voiced regions per second.

2. Only *mean* and *standard deviation* (Mean+Std) for all LLDs, resulting in 46 functionals.

3. BoAW representations of the LLDs. In this approach, each LLD vector in the audio is quantised according to a *codebook* of template 'audio words' that have been learnt from the training set using a random sampling. A histogram (BoAW) feature vector is then generated counting the occurrences of each template in a given audio segment or block [7, 14, 15]. While this vector is usually sparse, the sparseness can be reduced by assigning more than one template to each LLD vector. Besides this *number of assignments*, the *codebook size* (number of templates) is the most important parameter. To reduce the range of the histogram values, the term-frequencies are finally logarithmised. In this work, we considered codebook sizes of 100 or 1 000 and 1 or 10 assignments.

eGeMAPS features have been extracted using our feature extractor openSMILE [16], BoAW have been computed using our crossmodal bag-of-words toolkit openXBOW [17] with the default random seed. As in the SEWA corpus the audio signals belonging to either subject in a conversation are the same, an additional feature is added to the input feature sequence, derived from the speaker turn information, which was provided to all participants of AVEC 2017. For each timestamp, this feature is either 0 or 1, indicating whether the subject is audible or not. As the annotation of the SEWA corpus has been made in real-time, i. e., while watching the video and listening to the audio played back in normal speed, there is usually a temporal *delay* present between the emotion of a subject and the contour of the annotation [7]. Even though RNNs have the capability of

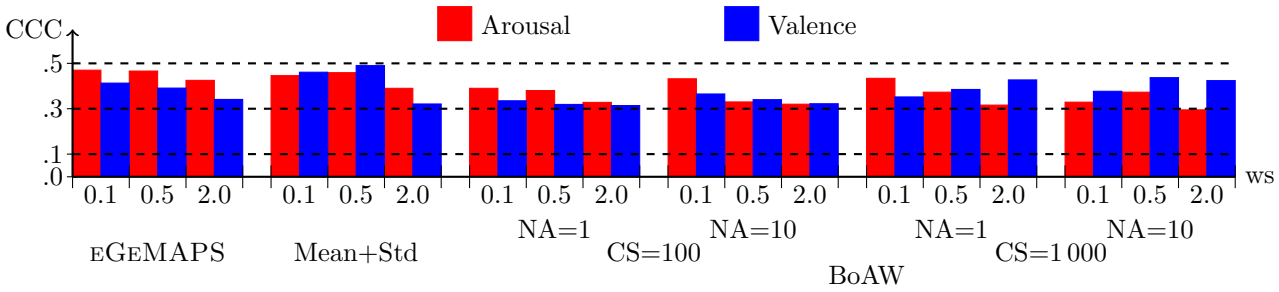

Figure 2: Results for the emotion recognition (in terms of the concordance correlation coefficient, CCC) for all investigated feature representations and three different window sizes (ws; 0.1 s, 0.5 s, 2.0 s).: eGeMAPS functionals, mean and standard deviation (Mean+Stddev), and bag-of-audio-words (BoAW, with codebook sizes, CS, 100 and 1 000 and numbers of assignments, NA, 1 and 10).

modelling this delay, we found that it improves the training when shifting the annotations to the front for half of the block size, i. e., the width of the window which summarises the LLDs, and shifting the predictions back for the same interval.

Figure 1 shows the structure of the employed deep neural network (DNN). It was derived from preliminary experiments, in which we found that, a DNN consisting of *four* layers leads to competitive results on the SEWA corpus. The RNN layers consist of *LSTM cells*; it was also tried to replace layers with time-distributed fully-connected layers. For the first layer, the performance was not decreased by this, so we used a fully-connected layer to save time for training the DNN, optimising the number of units between 200 and 1 000, accounting for the varying dimensionalities of the input. However, the number of units in this layer did not have a big influence on the results. For all recurrent layers, a number of 32 units seemed to be an optimum. As there are dependencies and correlations in the emotional dimensions *arousal* and *valence* [10], we considered both *multi-task* learning, i. e., training a DNN with two outputs and two errors at the same time, and *single-task* learning. Usually, it is expected that a model considering both dimensions at the same time has certain advantages, exploiting cross-dependencies between the annotations [9]. Thus, the model in Figure 1 has two output neurons, with a *linear activation*, though a *hyperbolic tangent (tanh)*, which is employed for all other units, leads to similar results.

As in the AVEC 2017 challenge, the *concordance correlation coefficient (CCC)* is used as a metric to evaluate the results. It is taking into account both the correlation between the prediction and the target and the difference and thus, it is a good compromise between the linear correlation coefficient and the mean squared error. We use CCC also as an *objective function* when training the network, as superior results can be expected doing this [18].

In our preliminary experiments, we further found that *bidirectional* LSTM-RNNs, i. e., using one RNN in forward direction and a parallel one in backward direction, does not improve the performance. A similar outcome was also found in the work by Trigeorgis et al. [11]. Dropout did not have a meaningful influence on the performance, so we used a dropout of 10 % in all layers. We trained the DNN with the *full batch* in each epoch (maximum 300 epochs), training was stopped when there was no improvement in the predictions for either dimension on the *development set* (in terms of the CCC) for 10 epochs and the performance was evaluated for the model providing the largest CCC on the development set; we utilised *RMSprop* as an optimiser.

Results and Discussion

The results of the proposed ERS model on the test set are displayed in Figure 2. The *learning rate* as well as the number of neurons in the first layer have been optimised on the development set. Results are shown for all three types of representations, the four configurations of the BoAW, and three different window sizes (0.1 s, 0.5 s, 2.0 s). It can be seen that, in most cases, a short window size (0.1 s-0.5 s) works better, except for the BoAW representations with a codebook size of 1 000 for *valence*. A possible explanation for this is that a BoAW vector with a large codebook is very sparse and that summarising over a larger window softens this effect. For *arousal*, the eGeMAPS functionals provide the best results, while *valence* is better predicted using only Mean+Std as functionals.

The optimum results are shown in Table 1. Surprisingly, we also found that, single-task learning outperforms multi-task learning in some cases. Thus, the optimum result displayed for arousal has been obtained with single-task learning and Mean+Std features, while for valence, multi-task learning yielded a slightly better result for the same feature representation. Outperforming both the audio-only and the multi-modal (audio+video+linguistic) baselines of the AVEC 2017 challenge, we also outperform the winners' audio-only model for arousal while obtaining a similar performance for valence [10]. The winning team employed a combination of hand-crafted features and SoundNet features, learned from the raw audio waveform in a *transfer learning* approach on an audio-visual database [19].

Conclusion and Outlook

In this contribution, we proposed a model for time-continuous emotion recognition from the speech signal in terms of arousal and valence, using hand-crafted acoustic

Table 1: Results for automatic emotion recognition on the AVEC17 Affect Sub-Challenge corpus (SEWA-German) in terms of CCC; A: Arousal, V: Valence.

Method	Devel		Test	
	A	V	A	V
AVEC17 Baseline-Audio	.344	.351	.225	.244
AVEC17 Baseline	.373	.507	.375	.466
AVEC17 Winner (Audio)	—	—	.437	.494
AVEC17 Winner [10]	.823	.796	.672	.756
Proposed (Mean+Std)	.586	.516	.499	.489

features and an LSTM-RNN consisting of four layers. We showed that, the representation of the input feature has a meaningful influence on the performance of the model. Future work will include cross-cultural and multi-modal analysis of in-the-wild emotional corpora and further investigations of audio word embeddings for deep learning architectures.

Acknowledgements

The research leading to these results has received funding from the European Union's 7th Framework and Horizon 2020 Programmes under grant agreements No. 338164 (ERC StG iHEARu) and No. 645094 (IA SEWA).

References

[1] Russell, J. A.: A circumplex model of affect. Journal of Personality and Social Psychology, vol. 39, no. 6 (1980), 1161–1178

[2] Kuppens, P.; Tuerlinckx, F.; Russell, J. A. & Barrett, L. F.: The relation between valence and arousal in subjective experience psychological bulletin. American Psychological Association, 139 (2012), 917-940

[3] Ringeval, F.; Sonderegger, A.; Sauer, J & Lalanne, D.: Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. Proc. IEEE FG (2013), 1-8

[4] Ringeval, F.; Schuller, B.; Valstar, M.; Gratch, J.; Cowie, R.; Scherer, S.; Mozgai, S.; Cummins, N.; Schmitt, M. & Pantic, M.: AVEC 2017 – Real-life depression, and affect recognition workshop and challenge. Proc. AVEC (2017), 3-9

[5] Schuller, B.; Rigoll, G. & Lang, M.: Hidden Markov model-based speech emotion recognition. Proc. ICASSP (2003), 4 pages

[6] Schuller, B.; Steidl, S. & Batliner, A.: The Interspeech 2009 emotion challenge. Proc. INTERSPEECH (2009), 312-315

[7] Schmitt, M.; Ringeval, F. & Schuller, B.: At the border of acoustics and linguistics: bag-of-audio-words for the recognition of emotions in speech. Proc. INTERSPEECH (2016), 495-499

[8] Hochreiter, S. & Schmidhuber, J.: Long short-term memory. Neural computation, vol. 9, no.8 (1997), 1735-1780

[9] Eyben, F.; Wöllmer, M. & Schuller, B.: A multitask approach to continuous five-dimensional affect sensing in natural speech. ACM Transactions on Interactive Intelligent Systems (TiiS), vol. 2, no. 1 (2012), 29 pages

[10] Chen, S. ; Jin, Q.; Zhao, J. & Wang, S.: Multimodal multi-task learning for dimensional and continuous emotion recognition. Proc. AVEC (2017), 19-26

[11] Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M. A.; Schuller, B. & Zafeiriou, S.: Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. Proc. ICASSP (2016), 5200-5204

[12] Tzirakis, P.; Trigeorgis, G.; Nicolaou, M. A.; Schuller, B. W. & Zafeiriou, S.: End-to-end multimodal emotion recognition using deep neural networks. IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8 (2017), 1301-1309

[13] Eyben, F.; Scherer, K. R.; Schuller, B. W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L. Y.; Epps, J.; Laukka, P.; Narayanan, S. S. & Truong K. P.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, vol. 7, no. 2 (2016), 190-202

[14] Pancoast, S. & Akbacak, M.: Bag-of-audio-words approach for multimedia event classification. Proc. INTERSPEECH (2012), 2105-2108

[15] Pokorny, F.; Graf, F.; Pernkopf, F. & Schuller, B.: Detection of negative emotions in speech signals using bags-of-audio-words. Proc. WASA held in conj. with ACII (2015), 879-884

[16] Eyben, F.; Weninger, F.; Groß, F. & Schuller, B.: Recent developments in openSMILE - the Munich open-source multimedia feature extractor. Proc. ACM Multimedia (2013), 835-838

[17] Schmitt, M. & Schuller, B.: openXBOW - Introducing the Passau open-source crossmodal bag-of-words toolkit. The Journal of Machine Learning Research, vol. 18 (2017), 1-5

[18] Weninger, F.; Ringeval, F.; Marchi, E. & Schuller, B. W.: Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. Proc. IJCAI (2017), 2196-2202

[19] Aytar, Y; Vondrick, C.; Torralba, A.: Soundnet: Learning sound representations from unlabeled video. Advances in Neural Information Processing Systems (2016), 892-900