Attentive Convolutional Neural Network based Speech Emotion Recognition: A Study on the Impact of Input Features, Signal Length, and Acted Speech

Michael Neumann, Ngoc Thang Vu

University of Stuttgart, Germany

{michael.neumann|thang.vu}@ims.uni-stuttgart.de

Abstract

Speech emotion recognition is an important and challenging task in the realm of human-computer interaction. Prior work proposed a variety of models and feature sets for training a system. In this work, we conduct extensive experiments using an attentive convolutional neural network with multi-view learning objective function. We compare system performance using different lengths of the input signal, different types of acoustic features and different types of emotion speech (improvised/scripted). Our experimental results on the Interactive Emotional Motion Capture (IEMOCAP) database reveal that the recognition performance strongly depends on the type of speech data independent of the choice of input features. Furthermore, we achieved state-of-the-art results on the improvised speech data of IEMOCAP.

Index Terms: Speech Emotion Recognition, Convolutional Neural Networks

1. Introduction

Speech emotion recognition has been attracting increasing attention recently. It is a challenging task due to the complexity of emotional expressions (affected by many factors such as age [1] and gender [2]) and the lack of a large dataset.

Deep learning (DL) has become a state-of-the-art method for many tasks such as speech recognition, computer vision and natural language processing (NLP). Convolutional neural networks (CNN) proposed in [3, 4] are a special kind of neural networks that have been successfully used not only for computer vision but also for speech [5, 6, 7]. For speech recognition, CNN proved to be robust against noise compared to other DL models [8]. Furthermore, [9] showed that CNNs are suitable for small memory footprint keyword spotting due to the parameter sharing mechanism.

More recently, attention based recurrent neural networks have been successfully applied to a wide range of tasks such as handwriting generation [10], machine translation [11], image caption generation [12] and speech recognition [13]. Researchers have also started to use attention mechanisms for CNNs in NLP tasks [14, 15, 16]. This seems to be helpful when the input signal is rather long or complex.

DL has been shown to significantly boost emotion recognition performance [17, 18, 19, 20, 21, 22]. Recently, several papers [23, 24] presented CNNs in combination with Long Short-Term Memory models (LSTM) to improve speech emotion recognition based on log Mel filter-banks (logMel) or raw signal. [24] demonstrated an end-to-end training from raw signal. This model, however, overfits easily due to the small amount of training data. Well known features, like MFCCs and logMel are fairly simple to extract and have a small number of dimensions which might be more suitable to a low-resource setting than raw signal.

In this paper, we propose an attentive convolutional neural network (ACNN) for emotion recognition which combines the strengths of CNNs and attention mechanisms. We focus on the comparison between different feature types. Furthermore, while previous models employed the complete signal to make predictions which costs recognition delays, we are interested in the robustness of the system against the signal length, i.e. finding the answer to the question: how long does the system need to wait to make an accurate prediction? Moreover, we analyze extensively performance differences between *improvised* and *scripted* speech. Finally, we report state-of-the-art results on the *improvised* subset of the IEMOCAP database.

2. Model

The model we apply to predict emotional categories from speech is depicted in Figure 1. It consists of two main parts: a CNN with one convolutional layer and one pooling layer and an attention layer. The CNN learns the representation of the audio signal, while the attention layer computes the weighted sum of all the information extracted from different parts of the input. The output from the pooling layer and the attention vector are then fed into a fully connected softmax layer.

2.1. Convolutional neural network

The input to the CNN is an audio signal divided into s overlapping segments represented by a d-dimensional feature vector. Thus, for each utterance, we form a matrix $W \in R^{d \times s}$ as input. For the convolution operation we use 2D kernels K (with width $|K|$) spanning all d features. The following equation expresses this operation:

$$(W * K)(x,y) = \sum_{i=1}^{d} \sum_{j=1}^{|K|} W(i,j) \cdot K(x-i, y-j) \quad (1)$$

After the convolution, we use max pooling to find the most salient features. Then, all feature maps are concatenated to one feature vector which is the input to the softmax layer.

2.2. Attention mechanism

For each vector x_i in a sequence of inputs x, the attention weights α_i can be computed as follows

$$\alpha_i = \frac{exp(f(x_i))}{\sum_j exp(f(x_j))} \quad (2)$$

where $f(x)$ is the scoring function. In this work, $f(x)$ is the linear function $f(x) = W^T x$, where W is a trainable parameter. The output of the attention layer, $attentive\_x$, is the weighted sum of the input sequence.

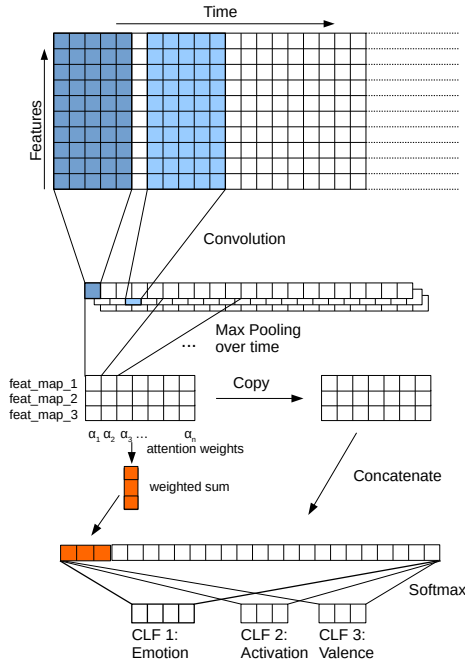$$attentive\_x = \sum_i \alpha_i x_i \quad (3)$$

Figure 1: *ACNN for speech emotion recognition, Classifier (CLF) 1 predicts emotional categories, CLF 2 and 3 predict activation and valence categories (multi-view learning).*

Our intuitions behind using an attention mechanism for emotion recognition are two-fold: a) speech emotion recognition is related to sentence classification with emotional content being differently distributed over the signal and b) the emotion of the whole signal is a composition of emotions from different parts of the signal. Therefore, attention mechanisms are suitable to first weight the information extracted from different pieces of the input and then combine them in a weighted sum. However, because the input signal is noisy, a max pooling layer is still helpful to only select the most salient features and filter noise. Therefore, we combine the CNN output vector after max pooling and the attention vector for the final softmax layer.

2.3. Multi-view learning

Emotions can be represented in two ways, either as categorical labels (e.g. *angry, happy*) or as continuous labels in the 2D activation/valence space. In [22], it is shown that multi-view (MV) learning with both categorical and continuous labels for training can improve prediction results. Similarly, we extend our model to incorporate activation and valence information.

3. Input Features

We use the following feature sets: (a) 26 logMel filter-banks, (b) 13 MFCCs, (c) a prosody feature set, and (d) the extended Geneva minimalistic acoustic parameter set (eGeMAPS). For all feature sets we apply mean and standard deviation normalization for each speaker independently.

We use the openSMILE toolkit [25] to extract all features. For logMel, MFCC, and prosody features, the audio signal is segmented into 25ms long frames with a 10ms shift. To extract logMel and MFCC features, a Hamming window is applied and the FFT with 512 points is computed. Then, we compute the logarithmic power of 26 Mel-frequency filter-banks over a

range from 0 to 6.5kHz. Finally, a discrete cosine transform (DCT) is applied to extract the first 13 MFCCs. The prosody feature set consists of the following features: PCM loudness, envelope of F0 contour, voicing probability, F0 contour, local jitter, differential jitter, and local shimmer.

The eGeMAPS is a hand-crafted feature set proposed for affective computing [26]. It consists of 25 low level descriptors (LLDs) containing frequency- and energy-related parameters and spectral parameters.

4. Data

We use the Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [27] for all experiments. It consists of about 12 hours of audiovisual data (speech, video, facial motion capture) from two recording scenarios: scripted play and improvised speech. Annotations are on turn level and consist of categorical labels (e.g. *happy, sad, angry*) and three continuous dimensions labeled with a discrete value from 1 to 5 each: activation, valence, dominance. For this study we use the same four categories as in [22, 28, 29, 30]: *angry, happy, sad,* and *neutral*. We merged *happy* and *excited* into one class: *happy*.[1] To be comparable with related work and to find out more about differences between *improvised* and *scripted* speech, we take three subsets from the data: only *improvised* (2,943 turns), only *scripted* (2,588 turns), and *all* sessions (5,531 turns).

The mean length of all turns is 4.46s (max.: 34.1s, min.: 0.6s). Since the input length for a CNN has to be equal for all samples, we set the maximal length to 7.5s (mean duration plus standard deviation). Longer turns are cut at 7.5s and shorter ones are padded with zeros.

We group activation and valence labels into three categories each for the MV approach. The same range mapping as in [31] is used: low: [1,2]; medium: (2,4); high: [4,5].

5. Experimental Results

5.1. Setup

The IEMOCAP data consists of five sessions with one male and one female speaker each. To train the models in a speaker-independent manner, we use leave-one-session-out cross validation. We take data from 8 speakers to construct training and development sets and use the remaining two speakers as test set.

We conduct two sets of experiments: Firstly, we compare the performance of CNN and ACNN (both with single-view (SV) and MV learning) regarding different input features. We run each combination of model, dataset and feature set six times with different random seeds. In doing so, we are able to report result variations due to random parameter initialization. We consider the averaged results produced this way more reliable than only reporting the single best number.

Secondly, we intend to find out how much information in terms of length of an utterance is sufficient to predict the affective state. We train and test our model with decreasing utterance length (by cutting the speech signals at 7, 6, 5, 4, 3, 2, and 1 seconds respectively).

5.2. Hyper-parameters

Our CNN models are implemented with the Theano library [32, 33]. We use stochastic gradient descent with an adaptive

---

[1]Class distribution: angry: 1,103; happy: 1,636; sad: 1,084; neutral: 1,708

learning rate (Adam [34]). For regularization dropout is applied to the last hidden layer [35]. The system's hyper-parameters are: 100 kernels with two different widths each (a total of 200 feature maps); a batch size of 30 for logMel and eGeMAPS, and 50 for MFCC; a dropout rate of 0.8; a pool size of 30, and stride of 3 for all configurations.

5.3. Experiment 1: Different data and feature sets

For all experiments, we report weighted accuracy (WA, accuracy on the whole test set). All results are shown in Tables 1-3. The tables present averaged results across six runs and the respective minimum and maximum accuracy.

Improvised speech (Table 1). The best performance is reached with logMel filter-banks. The ACNN with MV learning performs best with 62.11% mean accuracy. The single best result of 63.85% – which outperforms the state-of-the-art result of 62.85% reported by [36] – is reached with ACNN and SV learning.

Scripted speech (Table 2). Prediction results are in general notably lower than for *improvised* speech. For this dataset, MFCC and eGeMAPS features lead to higher accuracies than logMel. The best performance of 53.19% is achieved with the ACNN (MFCC with SV and eGeMAPS with MV).

All data (Table 3). MFCC and logMel features produce similar results, the accuracy with eGeMAPS is slightly lower, whereas prosody features perform notably worse. The best mean accuracy of 56.10% is achieved with logMel features using ACNN and MV learning. This model outperforms related work on the same data reported in [28, 29]. However, our focus does not lie on competing with state-of-the-art results (60.8% and 60.6% WA published in [22, 30]). In this work, we focus on the comparison of different input features, as well as the

interpretation of our results and a thorough error analysis (cf. section 6).

Feature fusion. In addition to the results in Tables 1-3, we test early fusion of logMel and prosody features (only one run of each model configuration). These results show slight improvements for *scripted* data (53.69%, ACNN with MV), but decreasing results for the complete dataset and *improvised* speech. This suggests that the CNN model cannot learn more discriminatory features from this additional information. This might be due to the convolution kernels spanning all features.

All results show that prosody features alone perform worse than cepstral features like logMel and MFCC. In [37], the authors state that prosodic features are strongly speaker-dependent and that their use is debatable in speaker-independent emotion recognition. To confirm this with our results, a comparable speaker-dependent experiment would be necessary. We assume that the prosody feature set contains too little information (only seven features) to compete with the others. The performance differences between logMel, MFCC and eGeMAPS are in general small. This suggests that the CNN is able to learn high-level features equally from these different input features. To find out whether the same information is learnt by the model from different input, further investigation is needed. In general, MV learning improves the prediction only slightly, if at all. The attention mechanism brings slight improvements on the *improvised* and *scripted* data for most of the feature sets, but has almost no effect on the complete dataset. Further, we see that there is high variation between single runs of the same model/feature combination (up to 4.2% between min and max results).

Overall, our model performs better on free speech (*improvised*) than on acted speech independent of the choice of fea-

Table 1: *CNN prediction results on improvised sessions (weighted accuracy).*

Features (dim.)	CNN						Attentive CNN					
	SV			MV			SV			MV		
	μ	min	max	μ	min	max	μ	min	max	μ	min	max
logMel (26)	**61.71**	60.40	62.66	**62.06**	61.08	62.86	**61.95**	61.19	*63.85*	***62.11***	61.41	63.34
MFCC (13)	61.31	60.85	61.94	61.35	60.85	62.28	60.85	60.10	61.41	61.35	60.68	62.12
eGeMAPS (25)	60.25	59.41	60.94	60.28	59.34	60.93	60.26	59.45	61.27	61.27	60.50	62.12
Prosody (7)	56.34	55.82	57.57	56.33	56.02	56.88	57.11	56.17	58.84	57.12	56.61	57.71

Table 2: *CNN prediction results on scripted sessions (weighted accuracy).*

Features (dim.)	CNN						Attentive CNN					
	SV			MV			SV			MV		
	μ	min	max	μ	min	max	μ	min	max	μ	min	max
logMel (26)	51.07	48.78	52.99	51.64	50.73	52.78	52.64	51.27	53.53	51.70	51.16	52.58
MFCC (13)	**52.35**	51.22	52.97	**53.01**	52.37	53.97	***53.19***	52.84	54.21	52.72	52.31	53.45
eGeMAPS (25)	51.84	50.93	53.98	52.82	52.15	54.25	52.31	51.16	54.16	***53.19***	52.57	*54.31*
Prosody (7)	49.17	48.46	50.06	48.76	48.16	49.65	48.69	47.71	49.70	49.02	48.16	50.25

Table 3: *CNN prediction results on the complete dataset (weighted accuracy).*

Features (dim.)	CNN						Attentive CNN					
	SV			MV			SV			MV		
	μ	min	max	μ	min	max	μ	min	max	μ	min	max
logMel (26)	**55.38**	54.58	56.52	**55.92**	55.24	56.85	54.86	54.14	55.57	***56.10***	55.24	56.85
MFCC (13)	55.33	54.70	55.82	55.74	54.76	*57.02*	**55.12**	54.02	55.55	55.40	54.46	56.64
eGeMAPS (25)	54.73	52.64	55.33	54.71	53.71	56.00	54.93	54.12	55.47	54.78	54.46	55.43
Prosody (7)	48.90	48.57	49.23	48.79	47.73	49.68	48.99	48.36	49.81	49.13	48.65	49.49

tures. These findings show that speech emotion recognition can be very sensitive to the type of speech data (in line with findings by [38]). Hence, it is important to carefully select suitable training data for a particular application.

5.4. Experiment 2: Signal length

In the second experiment, we use the ACNN with MV learning to perform emotion recognition on signals with decreasing length. We use logMel and MFCC features because these performed best previously. Results are presented in Figure 2. In
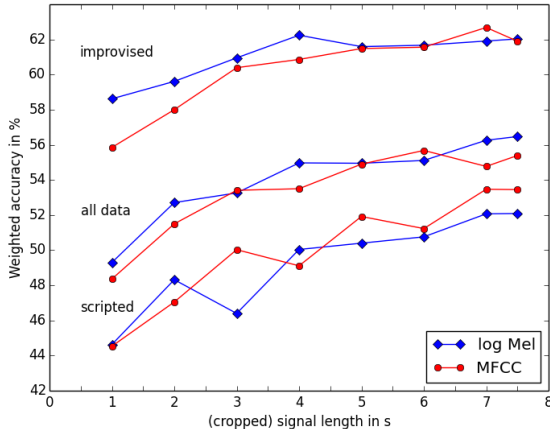

Figure 2: *System performance with decreasing signal length.*

general, accuracy decreases with shorter input. We observe a notable difference in the performance drop between *improvised* and *scripted* speech, especially with logMel features (3.4% and 7.5% drop). From these results we assume that in spontaneous speech, it is more likely that an utterance carries emotional content in the first seconds already, whereas in scripted speech it is more difficult to predict the emotion from only the first one or two seconds. In general, the results show that a relatively short snippet of a speech signal can be sufficient to perform emotion recognition with only a small accuracy loss. This is an important finding for the development of real-time applications which aim to make a prediction while the user is still speaking. Moreover, the training time of the system can be reduced.

6. Error analysis

We analyze the predictions of the ACNN (logMel features, MV learning). Figures 3a-3c show the confusion matrices.

For *improvised* speech (Fig. 3a) the most striking observation is that the model predicts *happy* for 49.12% *angry* samples. This counter-intuitive mistake becomes more plausible when looking at the activation information. Both *angry* and *happy* have a high activation level. Hence, the system's frequent confusion is due to the fact that valence is harder to predict than activation [39, 24, 26]. The category *sad* is predicted best (73.01%). This observation is in line with findings by [37, 27]. Further, the neutral class is frequently confused with other classes. This seems plausible because the neutral state is located in the center of the activation-valence space, what makes the discrimination from other classes difficult.

In contrast, for *scripted* sessions the accuracy for *angry* is surprisingly high, and relatively low for *sad* and *happy*. In general, there are more errors in almost all classes. One reason for the high discrepancy in the class *angry* is the different class distribution (many *angry* samples in scripted sessions). But this does not explain all other differences. The analysis suggests that *improvised* speech is in general more variable and therefore makes it easier to discriminate affective states. Investigation with more data would be helpful to confirm these findings. Note the high percentage of *sad* samples predicted as *happy* (23.08%). To find out the reason for this frequent confusion, further analysis is necessary. The error distribution on the complete dataset (Fig. 3c) lies between those seen in Figures 3a and 3b. There are similar patterns as for *improvised* data (the *angry/ happy* confusion is not as severe).

7. Conclusion

We presented a comparison of different features for speech emotion recognition using an attentive CNN. The results with logMel, MFCC, and eGeMAPS features are similar, but notably lower with prosodic features. A reason for that could be the smaller number of features in the latter. The similar results suggest that for a CNN the particular choice of features is not as important as the model architecture and the amount and kind of training data. We found strong differences between *improvised* and *scripted* speech, obtaining better results on the first. Experiments with decreasing signal length showed that the performance decreases slightly, but remains at a relatively high level even for short signals down to two seconds. Future work includes testing the presented ACNN on a different database.
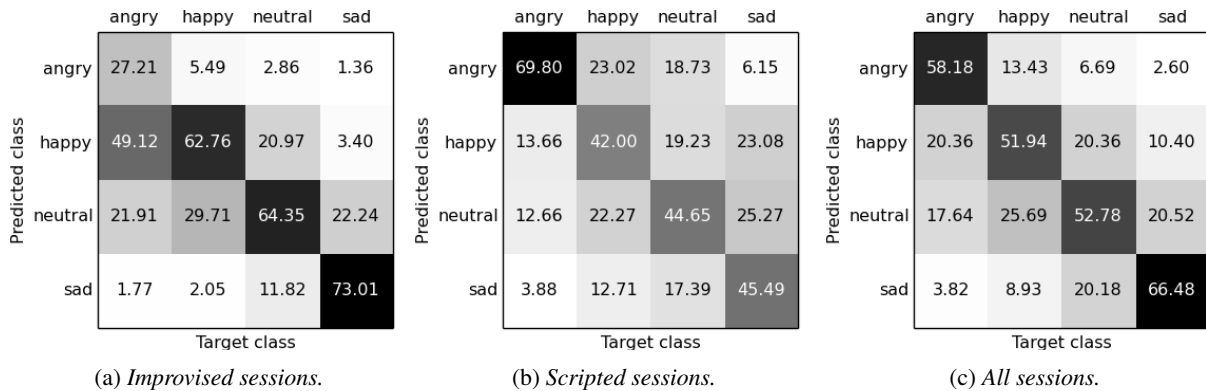

(a) *Improvised sessions.*

(b) *Scripted sessions.*

(c) *All sessions.*

Figure 3: *Error distribution of the ACNN model with MV learning.*

8. References

[1] A. Mill, J. Allik *et al.*, "Age-related differences in emotion recognition ability: a cross-sectional study." *Emotion*, vol. 9, no. 5, p. 619, 2009.

[2] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*, 2006.

[3] A. Waibel, T. Hanazawa *et al.*, "Phoneme recognition using time-delay neural networks," *IEEE transactions on acoustics, speech, and signal processing*, vol. 37, no. 3, pp. 328–339, 1989.

[4] Y. Le Cun, B. Boser *et al.*, "Handwritten digit recognition with a back-propagation network," in *Advances in neural information processing systems*, 1990.

[5] O. Abdel-Hamid, A.-r. Mohamed *et al.*, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4277–4280.

[6] T. N. Sainath, A.-r. Mohamed *et al.*, "Deep convolutional neural networks for lvcsr," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 8614–8618.

[7] T. N. Sainath, B. Kingsbury *et al.*, "Deep convolutional neural networks for large-scale speech tasks," *Neural Networks*, vol. 64, pp. 39–48, 2015.

[8] D. Palaz, R. Collobert *et al.*, "Analysis of cnn-based speech recognition system using raw speech as input," in *Proceedings of Interspeech*, 2015.

[9] T. Sainath and C. Parada, "Convolutional neural networks for small-footprint keyword spotting," in *Proceedings of Interspeech*, 2015.

[10] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention." in *ICML*, vol. 14, 2015, pp. 77–81.

[13] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015, pp. 577–585.

[14] H. Adel and H. Schütze, "Exploring different dimensions of attention for uncertainty detection," *arXiv preprint arXiv:1612.06549*, 2016.

[15] F. Meng, Z. Lu, M. Wang, H. Li, W. Jiang, and Q. Liu, "Encoding source language with convolutional neural network for machine translation," *arXiv preprint arXiv:1503.01838*, 2015.

[16] W. Yin, H. Schütze, B. Xiang, and B. Zhou, "Abcnn: Attention-based convolutional neural network for modeling sentence pairs," *arXiv preprint arXiv:1512.05193*, 2015.

[17] A. Stuhlsatz, C. Meyer *et al.*, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5688–5691.

[18] N. E. Cibau *et al.*, "Speech emotion recognition using a deep autoencoder," *Proceedings of the XV Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2013), San Carlos de Bariloche*, 2013.

[19] L. Li, Y. Zhao *et al.*, "Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 312–317.

[20] C. Huang *et al.*, "A research of speech emotion recognition based on deep belief network and svm," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[21] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine." in *Interspeech*, 2014, pp. 223–227.

[22] R. Xia and Y. Liu, "A multi-task learning framework for emotion recognition using 2d continuous space," *IEEE Transactions on Affective Computing*, 2015.

[23] G. Keren and B. Schuller, "Convolutional rnn: an enhanced model for extracting features from sequential data," *arXiv preprint arXiv:1602.05875*, 2016.

[24] G. Trigeorgis, F. Ringeval *et al.*, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*, 2016.

[25] F. Eyben, F. Weninger *et al.*, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 835–838.

[26] F. Eyben, K. R. Scherer *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[27] C. Busso, M. Bulut *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[28] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4749–4753.

[29] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, "Representation learning for speech emotion recognition," *Interspeech 2016*, pp. 3603–3607, 2016.

[30] V. Rozgic, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, "Ensemble of svm trees for multimodal emotion recognition," in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*. IEEE, 2012, pp. 1–4.

[31] A. Metallinou, M. Wollmer *et al.*, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, 2012.

[32] J. Bergstra, O. Breuleux *et al.*, "Theano: A cpu and gpu math compiler in python," in *Proc. 9th Python in Science Conf*, 2010, pp. 1–7.

[33] F. Bastien, P. Lamblin *et al.*, "Theano: new features and speed improvements. nips workshop on deep learning and unsupervised feature learning," *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2012.

[34] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] N. Srivastava *et al.*, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[36] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition." in *INTERSPEECH*, 2015, pp. 1537–1540.

[37] V. Chernykh, G. Sterling, and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," *arXiv preprint arXiv:1701.08071*, 2017.

[38] L. Tian, J. D. Moore, and C. Lai, "Emotion recognition in spontaneous and acted dialogues," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 698–704.

[39] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 552–557.