Towards robust speech emotion recognition using deep residual networks for speech enhancement

Andreas Triantafyllopoulos, Gil Keren, Johannes Wagner, Ingmar Steiner, Björn Schuller

Angaben zur Veröffentlichung / Publication details:

Triantafyllopoulos, Andreas, Gil Keren, Johannes Wagner, Ingmar Steiner, and Björn Schuller. 2019. "Towards robust speech emotion recognition using deep residual networks for speech enhancement." In *Crossroads of speech and language: 20th Annual Conference of the International Speech Communication Association (INTERSPEECH 2019), Graz, Austria, 15-19 September 2019; Volume 3*, 1691–95. ISCA. https://doi.org/10.21437/interspeech.2019-1811.

INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Towards Robust Speech Emotion Recognition using Deep Residual Networks for Speech Enhancement

Andreas Triantafyllopoulos[1], *Gil Keren*[2], *Johannes Wagner*[1], *Ingmar Steiner*[1], *Björn Schuller*[1,2,3]

[1]audEERING GmbH, Gilching, Germany
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]GLAM – Group on Language, Audio & Music, Imperial College London, UK
atriant@audeering.com

Abstract

The use of deep learning (DL) architectures for speech enhancement has recently improved the robustness of voice applications under diverse noise conditions. These improvements are usually evaluated based on the perceptual quality of the enhanced audio or on the performance of automatic speech recognition (ASR) systems. We are interested instead in the usefulness of these algorithms in the field of speech emotion recognition (SER), and specifically in whether an enhancement architecture can effectively remove noise while preserving enough information for an SER algorithm to accurately identify emotion in speech. We first show how a scalable DL architecture can be trained to enhance audio signals in a large number of unseen environments, and go on to show how that can benefit common SER pipelines in terms of noise robustness. Our results show that incorporating a speech enhancement architecture is beneficial, especially for low signal-to-noise ratio (SNR) conditions.

Index Terms: speech enhancement, speech emotion recognition

1. Introduction

Recently, deep neural networks (DNNs) have provided state of the art results in a large number of applications. They are mostly known for their widespread usage in the image domain, reinforcement learning, and natural language processing (NLP), and their most successful application in the audio domain has been in the field of ASR [1]. However, these methods have also been gaining traction in the fields of source separation and speech enhancement [2].

Speech enhancement architectures are usually evaluated using some speech quality metric like PESQ [3] or STOI [4], or focus on improving metrics like word error rate (WER) for ASR [5, 6, 7]. In contrast, we consider the application of these architectures in the field of SER, where additive noise and reverberation have been shown to severely degrade the performance of algorithms [8].

Previous work has focused on the evaluation of SER algorithms based on acoustic features under white noise or for a limited number of noise environments. Schuller *et al.* [9] investigated the negative effects of white noise on two originally clean data sets and tried to mitigate them using feature selection. Schuller *et al.* [8] considered the effects of reverberation, both artificially added and originally present in the recordings, as well as additive white noise. Tawari and Trivedi [10] introduced speech enhancement as a preprocessing step, utilizing adaptive thresholding in the wavelet domain to deal with car and white noise. Eyben *et al.* [11] tried to counter the problem by augmenting the training set with noisy audio and also extended the problem to five different types of additive noise. Weninger *et al.* [12] used non-negative matrix factorisation (NMF) on mel spectra to augment a set acoustic features and reduce the effects of additive noise and reverberation, but limited their investigation to only two kinds of additive noise. Zhao *et al.* [13] used a sparse representation for robust SER under white noise. Avila *et al.* [14] evaluated the benefits of using a speech enhancement algorithm as a preprocessing step for an SER pipeline based on acoustic features for a single noise type, and how performance correlates with traditional speech quality measures.

While previous work in this area has illustrated the problems that noise can cause for SER applications, their scope has been limited to a small amount of environments and noise types. They also concentrate on algorithms based on acoustic features, and do not take recent DL approaches into consideration. Finally, the integration of a speech enhancement pipeline has only been considered as a preprocessing step.

We move beyond previous work in a number of ways. First, we examine the effects of noise on an end-to-end DL based architecture which operates on raw audio input, in addition to traditional, acoustic feature based algorithms. Secondly, we scale up the number of noise environments taken into consideration, essentially moving towards a production ready speech enhancement algorithm that can work reliably for different SER applications under very diverse conditions. In addition, we consider potential trade-offs of speech enhancement algorithms with respect to SER. We hypothesize that enhancement algorithms must not only remove the environment noise, but also preserve those qualities that characterise speech from an emotional perspective, and shed more light on this with our experiments. Finally, we explore the potential benefits of integrating a speech enhancement architecture in different stages of an SER pipeline compared to simply augmenting the training set with noisy audio or using an enhancement algorithm only in the preprocessing stage.

2. Speech enhancement

2.1. Data sets

We choose the Mozilla Common Voice database[1] as a source of speech signals because it offers a large amount of very diverse speech segments coming from a variety of speakers and recording conditions.

We use Audio Set [15] as the source of our noise signals. Audio Set is a very large corpus of audio segments extracted

[1]https://voice.mozilla.org/

http://dx.doi.org/10.21437/Interspeech.2019-1811

Table 1: *Emotion database information*

	RECOLA	EMO-DB	eNTERFACE
Subjects (m/f)	19/27	5/5	34/8
Sample Rate	44.1 kHz	16 kHz	48 kHz
Language	French	German	English
Utterances	N/A	465	967
Duration	2 h 15 min	21 min	45 min
Emotion	spontaneous	acted	acted
Labels	arousal	anger, boredom, fear, joy, sadness, neutral	anger, boredom, fear, joy, sadness, neutral, disgust

from YouTube videos and manually annotated according to a hierarchical ontology of 632 audio categories. From those, we exclude the ones that belong to the *Human Source* and *Music* categories, as they are likely to contain human speech.

2.2. Architecture

Our approach is based on stacked residual blocks [16] of 2D convolution layers, that have been previously shown to efficiently learn rich representations of input signals. We use an architecture similar to that of Keren *et al.* [5], with our main difference being that we do not use an embedding subnetwork to condition the network on the noise environment, which makes our algorithm easier to use in practice, since recordings of the background noise may be hard to acquire.

We use in total 8 residual blocks, the first 4 containing convolution layers with a 4×4 kernel, and the last 4 containing convolution layers with a 3×3 kernel. We start with 64 feature maps for each convolution layer in the first 2 blocks, then double the number of feature maps with each successive group of 2 residual blocks. We also apply a 2×2 stride in blocks 3, 5, and 7. The output of the last residual block is first processed by a 2D convolution layer to reduce its dimensionality before being fed to a fully-connected layer that maps it to the appropriate size. We use the output directly as our enhancement mask without applying any kind of activation.

The input to our network is log magnitude spectrograms, where the speech signal is initially resampled to 16 kHz and then mixed on the fly with a noisy signal randomly sampled from our noise data set. The two signals are mixed at a SNR selected randomly in the range of 0 dB to 25 dB with a step of 5 dB. We then compute the short-time Fourier transform (STFT) of the clean and noisy signals using a window of size 25 ms and a stride of 10 ms.

The network enhances a single frame on each forward pass. It takes as input a segment of n frames of the noisy signal (where we found the best value for n to be 35 using the validation set), and outputs an enhancement mask that is added to the original central frame of our input to compute the enhanced frame. We compute the mean square error (MSE) between the enhanced and the clean frame and use that as the loss function to train the network. The network was trained using stochastic gradient descent (SGD) with a learning rate of 0.01 and a batch size of 64 examples.

3. Speech emotion recognition

3.1. Data sets

We use three standard emotion corpora to test the performance of our speech enhancement architecture for SER, namely the REmote COLlaborative and Affective interactions corpus (RECOLA) [17], the Berlin Emotional Speech Database (EMO-DB) [18] and eNTERFACE [19]. For our experiments on RECOLA, we predict a continuous arousal value every 40 ms, matching its official annotation scheme, while for EMO-DB and eNTERFACE we predict a single emotion label on the utterance level.

It is important to note that a) none of these data sets were seen during training of the speech enhancement architecture, and that b) two of the data sets are recorded in a different language than the data we used for training the enhancement network, indicating that the enhancement network indeed learns to remove noise from human speech and is not overfitting to the training data set. Details about the data sets can be found in Table 1.

3.2. Architectures

3.2.1. Acoustic features

We use openSMILE [20], our open-source feature extraction toolkit that is widely used in the field of SER, to extract features, because it comes prepackaged with numerous standardised feature sets that have been successfully used to build emotion recognition architectures, predominantly under the assumption that every utterance is characterised by a single label.

Out of the numerous available feature sets, we limited our analysis to the two most commonly used ones, namely ComParE [21] and extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [22]. In both cases, we used a support vector machine (SVM) with a radial basis function (RBF) kernel [23] as the classifier. We followed this approach for the EMO-DB and eNTERFACE data sets.

3.2.2. End-to-end

We base our implementation on that of Trigeorgis *et al.* [24], where the authors successfully employ a DNN architecture for modelling arousal on the RECOLA database using raw audio as input. In this context, end-to-end refers to training an SER model to predict arousal using raw audio as input, and does not also encompass our enhancement architecture, which is only used as a separate pre-processing step. The input to the network is a 6 s audio sample sampled at 16 kHz. It consists of two 1D convolution layers containing 20 and 40 filter banks, which are followed by two max pooling layers with a stride of 2 and 10, respectively. The output of the second pooling layer is fed to two uni-directional long short-term memory (LSTM) layers [25], each having 256 units. The output of the last LSTM is mapped to the arousal prediction through a fully-connected layer followed by a tanh activation. We also used a dropout of 0.5 after each convolution layer to prevent overfitting.

We train the model for 50 epochs with a batch size of 25 examples and a learning rate of 0.0001, use concordance correlation coefficient (ρ_c) as the loss function, and choose the model that performed best on the validation set based on ρ_c. We also performed all of the post-processing steps reported by Trigeorgis *et al.* [24], namely *median filtering*, *centring*, *scaling* and *time shifting*.

Table 2: *UAR (in %) results under matched conditions using openSMILE features and SVMs for EMO-DB and eNTERFACE*

Data set	Features	Clean	Noisy	Enhanced
EMO-DB	**ComParE**	72.34	52.23	58.06
	eGeMAPS	61.04	46.42	52.63
eNTERFACE	**ComParE**	65.60	46.19	44.22
	eGeMAPS	47.87	34.64	37.19

4. Results

We are interested in whether integrating a speech enhancement algorithm in an SER pipeline can be beneficial. There are three ways to do that, namely:

Matched conditions where we integrate the enhancement both in the training and testing phase when the SNR conditions are fixed. We compare the performance of SER algorithms trained using enhanced audio and tested on enhanced audio, with the performance of SER algorithms trained on noisy audio and tested on noisy audio.

Mismatched conditions where we integrate the enhancement in the testing phase only. We train our SER algorithms on the clean audio only and test on both the noisy and enhanced test sets.

Multi-SNR conditions where we integrate the enhancement both in the training and testing phase but with unknown SNR conditions.

In all cases, we investigated the performance of our algorithm in SNR levels ranging from −5 dB to 20 dB, with a step of 5 dB. We note that the enhancement network was trained with SNR levels ranging from 0 dB to 25 dB, so the −5 dB condition is much lower than that to which it was exposed during training.

We report ρ_c results on the arousal dimension for RECOLA, and unweighted average recall (UAR) of the emotion classes for EMO-DB and eNTERFACE. For RECOLA, we used the official training, validation and test sets of the AVEC 2016 challenge [26]. For eNTERFACE and EMO-DB we performed leave-one-speaker-out (LOSO) cross-validation (CV) to make our results easily reproducible.

4.1. Matched conditions

Our results for the matched conditions scenario on the eNTERFACE and EMO-DB data sets are presented in Table 2, where we average them across SNR levels. Our first observation is that the SER architectures achieve better performance when using enhanced audio compared to using noisy audio in almost all cases. The only exception is when training an SVM with ComParE features on eNTERFACE, but there the difference in UAR is small.

4.2. Mismatched conditions

4.2.1. Aggregated results

We present average results across SNRs in Table 3 for mismatched conditions. We can see that, on average, testing an SER architecture trained exclusively on clean audio on the enhanced audio improves ρ_c on RECOLA and UAR on EMO-DB, but decreases UAR on eNTERFACE compared to testing on the noisy audio.

An auditory inspection of the eNTERFACE data set revealed that the original audio recordings include a lot of reverberation noise, which degrades the quality of the data, whereas our enhancement network was only trained to remove additive noise and did not learn to deal with reverberation.

Table 3: ρ_c *(a) and UAR (b, in %) results under mismatched conditions.*

(a) *For RECOLA, we use the end-to-end architecture and evaluate on the arousal dimension.*

Data set	Features	Clean	Noisy	Enhanced
RECOLA	**raw signal**	0.4781	0.4066	0.4337

(b) *For EMO-DB and eNTERFACE we use openSMILE features and SVMs.*

Data set	Features	Clean	Noisy	Enhanced
EMO-DB	**ComParE**	72.34	24.56	32.99
	eGeMAPS	61.04	33.94	48.72
eNTERFACE	**ComParE**	65.60	35.91	30.88
	eGeMAPS	47.87	29.49	23.87

4.2.2. Detailed results

We present a more fine-grained analysis of our evaluation in Tables 4 and 5, where we report explicit results per SNR. We first examine the robustness of all SER architectures under additive noise. We observe that the algorithm based on openSMILE features and SVMs suffers from large drops in performance for all SNRs. In the lower SNRs in particular, the SVM architectures give results slightly above chance levels. We also note that eGeMAPS appears to be more robust to noise conditions than ComParE. This can be attributed to the much higher dimensionality of ComParE, which might lead to overfitting on the training conditions. The end-to-end architecture on the other hand is more robust to noise in the mid SNRs range, but still suffers a performance drop in the lower SNRs. We also note an increase in the performance of the end-to-end architecture trained on original audio when tested with noisy audio in the higher SNRs. We explain this unexpected observation through the presence of speaker pauses in the continuous recordings of RECOLA. In those instances the end-to-end model is lacking speech information and constantly outputs a low arousal score. The ground truth, however, does not suffer from the same lack of information as the raters based their judgment both on the audio and the video channel. In the presence of noise, it becomes more likely that the end-to-end model predicts a slightly higher arousal level, which may occasionally fit the ground truth and, thus, boost the performance of the model. This also explains why results drop again when speech enhancement is applied.

Next, we examine potential performance gains obtained by enhancing the noisy audio with our architecture. In general, we observe an improvement for all data sets and algorithms in the low SNRs. Results on EMO-DB are improved in all cases. Results on RECOLA are also better or equal to using noisy audio up to 10 dB and especially in the very low SNRs of -5 and 0 dB. In the higher SNRs, where the end-to-end architecture already performed better on the noisy audio, we actually observe a drop in performance, however, ρ_c is still close to the original. Finally, as mentioned, the eNTERFACE data set is more challenging because of the presence of reverberation noise in the original audio. Nevertheless, our enhancement architecture can still improve performance in the lower SNRs even for this data set.

Table 4: *RECOLA ρ_c results for the arousal dimension under mismatched conditions where the training set consists of samples of clean audio.*

Data set	Clean	SNR (dB)	Noisy	Enhanced
RECOLA	0.4781	−5	0.2167	0.4022
		0	0.3206	0.4165
		5	0.4067	0.4400
		10	0.4530	0.4504
		15	0.5097	0.4486
		20	0.5327	0.4444

Table 5: *UAR (in %) results under mismatched conditions using openSMILE features and SVMs on EMO-DB and eNTERFACE.*

		EMO-DB		eNTERFACE	
Features	**SNR (dB)**	**Noisy**	**Enhanced**	**Noisy**	**Enhanced**
ComParE	−5	14.73	20.75	19.81	25.73
	0	18.05	26.85	23.13	30.68
	5	20.30	33.95	34.59	32.15
	10	23.02	36.65	37.28	32.95
	15	31.00	39.34	47.03	30.66
	20	40.27	40.39	53.60	33.13
eGeMAPS	−5	16.29	36.14	16.20	22.94
	0	19.82	45.07	20.69	23.65
	5	28.35	50.99	27.69	25.09
	10	39.36	53.25	32.92	23.72
	15	46.34	52.84	37.22	24.17
	20	53.49	54.07	42.22	23.65

These results also confirm our original hypothesis that incorporating an enhancement architecture in an SER pipeline is not straightforward. In the lower SNRs, all SER models benefit from the enhancement. However, as the SNR increases, the performance gains become smaller. Even on EMO-DB, where the enhancement architecture helps in all cases, we see that the performance gains are smaller in the higher SNRs, and the performance gap between the enhanced and the clean audio is still large.

An auditory inspection of the results for these two data sets revealed the following. In the low SNR levels, there remains some residual noise even after the enhancement. This could be why the SER performance in the low SNRs is still low. In the higher SNRs, we observed very little residual noise. However, there were cases where the enhanced audio appeared distorted, and cases where the enhancement network introduced small artefacts in parts of the audio. These changes in the audio quality could explain why there are small or no gains in using the enhanced audio in the high SNR cases, and illustrate the importance of preserving the quality of the original speech signal for SER.

4.3. Multi-SNR conditions

Finally, we consider the case where a researcher or developer assumes that noise will be present during the testing phase, but does not know which SNR level to expect. In that case, he or she will likely try to make their SER pipeline more robust against noise, either by augmenting the training set with noisy audio, or by integrating an enhancement architecture. We investigate the following two scenarios:

- Training on *noisy* audio and testing on audio of the same kind but of potentially different SNR levels and noise environments without incorporating a speech enhancement architecture at all;
- Training on *enhanced* audio and testing on audio of the same kind but of potentially different SNR levels and noise environments, essentially integrating the speech enhancement architecture in both the training and testing phases.

Table 6: *RECOLA ρ_c results for the arousal dimension under multi-SNR conditions where the training set consists of samples of either noisy or enhanced audio, and the test set consists of samples of the same type but potentially different SNR levels.*

Data set	Train SNR (dB)	Test SNR (dB)	Noisy	Enhanced
RECOLA	5	5	0.4460	0.4439
	5	10	0.4419	0.4500
	5	15	0.4409	0.4411
	10	5	0.4489	0.4447
	10	10	0.4646	0.4526
	10	15	0.4650	0.4455
	15	5	0.4545	0.5040
	15	10	0.4678	0.5133
	15	15	0.4846	0.5074
Average			0.4571	0.4669

We present results for the RECOLA data set using the end-to-end architecture in Table 6 for three different SNR levels in the mid-SNR range where we saw the SER algorithm transition from performing better on the enhanced audio to performing better on the noisy audio in the mismatched case. Results show that integrating the speech enhancement architecture in the training phase still increases performance in the lower SNRs, but without suffering from a similar drop as in the mismatched condition in the higher SNRs. In particular, we observe similar performance when training on audio of lower SNR and testing on the same or higher SNR, between noisy and enhanced audio. However, there is an improvement when we train on a high SNR and test on same or lower SNRs. This indicates that we can benefit from including the speech enhancement architecture in both the training and the testing phase of an SER pipeline even when the test conditions are not known.

5. Conclusions

In this work, we investigated the impact of noise on two popular SER architectures and the potential benefits of integrating speech enhancement in SER applications. In general, our enhancement architecture performs favourably for the lower SNRs in the very challenging scenario of training an SER architecture only on original audio and simply integrating the enhancement as a pre-processing step in the testing phase. In the very low and negative SNRs in particular, the enhancement network was able to render the SER algorithms usable again.

We also discovered that using an enhancement architecture can potentially degrade the audio quality and introduce artefacts that make an SER algorithm perform worse compared to the noisy audio in high SNRs, which further supports our hypothesis that speech enhancement algorithms must be designed with an emphasis on preserving the emotional information in the signal for robust SER.

6. References

[1] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015. DOI: 10.1016/j.neunet.2014.09.003.

[2] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018. DOI: 10.1109/TASLP.2018.2842159.

[3] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 2, 2001, pp. 749–752. DOI: 10.1109/ICASSP.2001.941023.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011. DOI: 10.1109/TASL.2011.2114881.

[5] G. Keren, J. Han, and B. Schuller, "Scaling speech enhancement in unseen environments with noise embeddings," in *International Workshop on Speech Processing in Everyday Environments*, 2018, pp. 25–29. DOI: 10.21437/CHiME.2018-6.

[6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99. DOI: 10.1007/978-3-319-22482-4_11.

[7] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology*, vol. 9, no. 5, 2018. DOI: 10.1145/3178115.

[8] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. IV, 2007, pp. 941–944. DOI: 10.1109/ICASSP.2007.367226.

[9] B. Schuller, D. Arsic, F. Wallhoff, and G. Rigoll, "Emotion recognition in the noise applying large acoustic feature sets," in *Speech Prosody*, 2006. [Online]. Available: https://www.isca-speech.org/archive/sp2006/sp06_128.html.

[10] A. Tawari and M. M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *International Conference on Pattern Recognition*, 2010, pp. 4605–4608. DOI: 10.1109/ICPR.2010.1132.

[11] F. Eyben, B. Schuller, and G. Rigoll, "Improving generalisation and robustness of acoustic affect recognition," in *ACM International Conference on Multimodal Interaction*, 2012, pp. 517–522. DOI: 10.1145/2388676.2388785.

[12] F. Weninger, B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognition of nonprototypical emotions in reverberated and noisy speech by nonnegative matrix factorization," *EURASIP Journal on Advances in Signal Processing*, 2011. DOI: 10.1155/2011/838790.

[13] X. Zhao, S. Zhang, and B. Lei, "Robust emotion recognition in noisy speech via sparse representation," *Neural Computing and Applications*, vol. 24, no. 7–8, pp. 1539–1553, 2014. DOI: 10.1007/s00521-013-1377-z.

[14] A. R. Avila, J. Alam, D. O'Shaughnessy, and T. Falk, "Investigating speech enhancement and perceptual quality for speech emotion recognition," in *Interspeech*, 2018, pp. 3663–3667. DOI: 10.21437/Interspeech.2018-2350.

[15] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[17] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013. DOI: 10.1109/FG.2013.6553805.

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Interspeech*, 2005, pp. 1517–1520. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2005/i05_1517.html.

[19] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *International Conference on Data Engineering Workshops*, 2006. DOI: 10.1109/ICDEW.2006.145.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: The Munich versatile and fast open-source audio feature extractor," in *ACM International Conference on Multimedia*, 2010, pp. 1459–1462. DOI: 10.1145/1873951.1874246.

[21] B. W. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. C. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The Interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Interspeech*, 2016, pp. 2001–2005. DOI: 10.21437/Interspeech.2016-129.

[22] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "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. DOI: 10.1109/TAFFC.2015.2457417.

[23] B. Scholkopf and A. J. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT Press, 2001.

[24] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5200–5204. DOI: 10.1109/ICASSP.2016.7472669.

[25] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997. DOI: 10.1162/neco.1997.9.8.1735.

[26] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016 – depression, mood, and emotion recognition workshop and challenge," in *International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10. DOI: 10.1145/2988257.2988258.