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

Robust Speech Emotion Recognition under Different Encoding Conditions

Christopher Oates[1], *Andreas Triantafyllopoulos*[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

coates@audeering.com, atriant@audeering.com

Abstract

In an era where large speech corpora annotated for emotion are hard to come by, and especially ones where emotion is expressed freely instead of being acted, the importance of using free online sources for collecting such data cannot be overstated. Most of those sources, however, contain encoded audio due to storage and bandwidth constraints, often in very low bitrates. In addition, with the increased industry interest on voice-based applications, it is inevitable that speech emotion recognition (SER) algorithms will soon find their way into production environments, where the audio might be encoded in a different bitrate than the one available during training. Our contribution is threefold. First, we show that encoded audio still contains enough relevant information for robust SER. Next, we investigate the effects of mismatched encoding conditions in the training and test set both for traditional machine learning algorithms built on hand-crafted features and modern end-to-end methods. Finally, we investigate the robustness of those algorithms in the multi-condition scenario, where the training set is augmented with encoded audio, but still differs from the training set. Our results indicate that end-to-end methods are more robust even in the more challenging scenario of mismatched conditions.

Index Terms: speech emotion recognition, speech and audio compression acronym

1. Introduction

Data collection is an ongoing issue in the speech and audio machine learning community. We require large, clean and well-controlled data sets to perform rigorous experiments. However, we also need to test the robustness of our algorithms in real world conditions, and this involves 'in the wild' recordings with natural, mixed content and potentially poor quality signals. A vast amount of available online audio sources contain a virtually inexhaustible source of data for developing machine learning applications. However, due to storage and bandwidth constraints, this data is usually stored in a compressed form. Audio compression is known, in many cases, to degrade the perceptual quality of the audio [1, 2]. However, the consequences of using compressed audio for recognition models have not been sufficiently studied, since most of the available emotional corpora are recorded in very high quality conditions. With the rapid rise of speech emotion recognition (SER) research and its imminent advance into industry applications [3], we wish to investigate both the effects of using compressed audio to train SER models, and the potential pitfalls of using an existing model in a production environment under mismatched encoding conditions.

Previous studies can be found in the literature covering the topic of SER models trained with compressed speech. Albahri *et al.* [4] tested the low bitrate speech codecs AMR, AMR-WB, and AMR-WB+ and showed that accuracy does not always decrease with a decreasing bitrate. In fact, they showed how emotion classification accuracy can vary across different codecs and acoustic feature sets. On the other hand, Siegert *et al.* [5] showed that the psychoacoustic model within the OPUS codec can, as a by-product, remove emotion redundant information from a speech signal, resulting in improved model accuracy. In this case, the data rate of the speech signal is maintained but redundant and imperceivable audio information is removed. Siegert *et al.* [6] focused on emotion classification accuracy, obtained via support vector machines (SVMs) [7], when using MP3, AM-WB and SPX coded speech. The authors found that MP3 with a bitrate of 32 kbit/s or higher, was suitable for achieving satisfactory unweighted average recall (UAR) results. García *et al.* [8] investigated the call centre use case for SER and focused on band-limiting codecs like AMR-NB, SILK, and also downsampled original speech. The authors were able to show that there was little degradation in model accuracy when features were extracted from voiced segments only. On the other hand, model accuracy was severely degraded when features were extracted from unvoiced segments only. Frühholz *et al.* [9] investigated the narrow-band encoded and low-pass filtered cases for short-term speaker state and long-term speaker trait recognition. The study focused on narrow-band low-bitrate speech coders used in telecommunications and high dimensional feature exaction as input to an SVM classifier. The authors showed that, under the given conditions, the matched and multi-condition training methods showed only a slight performance degradation even at the lower bitrates for arousal and valence recognition. On the other hand, the mismatched training condition resulted in a performance degradation.

Much of the previous work on the topic of SER with encoded speech has focused on the telecommunications use case. In contrast, we see the large amount of 'in the wild' data on platforms like YouTube as an invaluable resource of data that can be used both for training and testing a SER model. The most interesting case is that of testing an already trained classifier. Since the presence of coding artefacts are largely inconsequential to a human's ability to recognise an emotion [10], one would expect the same from a SER model which has truly built an internal representation of human emotion. Moreover, access to large amounts of speech data makes modern end-to-end deep neural networks (DNNs) which have demonstrated their superior performance over traditional machine learning algorithms in other domains, a viable option for classification.

This paper will focus on the following three applications of SER:

1. SER models trained and tested on compressed speech

http://dx.doi.org/10.21437/Interspeech.2019-1658

signals from the same codec (matched conditions)
2. SER models trained on uncompressed speech signals and tested on compressed speech signals (mismatched conditions)
3. SER models augmented with compressed speech signals and tested on different compressed speech signals (multi-condition)

To the best of the authors' knowledge, evaluating an SER model in the *mismatched* and *multi-condition* setting for wideband encoding has not been thoroughly investigated before. Moreover, utilising an end-to-end algorithm for classification in such a scenario has also not been thoroughly investigated before. In addition, our study encompasses a broad range of experimental conditions relevant to developing a production SER model, namely: (1) codec, (2) bitrate (3) classification algorithm, (4) training method, (5) data set and (6) feature set. We will focus our attention on three ubiquitous codecs, namely, MP3, AAC and OPUS [11] and test a wide range of bitrates common to all 3 codecs. We will evaluate the consistency of two different kinds of algorithms for 3 training methods over four data sets and four feature sets. This study aims to extend the body of knowledge on the topic with the unique and broad test cases presented.

The remainder of the paper is organised as follows: In Section 2 we provide an overview of the data sets, codecs, and algorithms used in our evaluation. In Section 3 the results are presented. Finally, in Section 4 we present our conclusions.

2. Experiment design

Speech Emotion Recognition is usually formulated as a simple classification of emotional labels or regression on emotional dimensions for single utterances, where it is assumed that the emotion is held constant within a relatively short temporal window. One of the most common approaches to modelling is to use a set of acoustic features over this temporal window, and then a machine learning algorithm to classify the utterance as belonging to one of the classes in the set. In contrast, modern deep learning methods attempt to fuse the feature extraction and classification steps, and train an algorithm that jointly learns rich representations and solves the downstream task.

In order to investigate the effects of encoding on both kinds of algorithms, we will first use a standard approach that utilises the openSMILE [12] open-source feature extraction toolkit and SVMs [7] for classification. Then, we also will use an end-to-end approach that makes use of convolutional and long short-term memory (LSTM) [13] layers to predict the arousal dimension.

2.1. Data sets

In this study, four standard emotion data sets were used which cover a range of acted and spontaneous emotions [14, 15, 16, 17]. The relevant information for each data set can be found in Table 1. EMO-DB, eNTERFACE, and Polish-Emo, contain acted emotional speech, with each utterance limited to expressing a single emotion, whereas RECOLA contains spontaneous, long-term interactions where the emotion of the speaker evolves naturally over time. In order to compare SER results across data sets, a common sampling rate must be used. By using the same sample rate we ensure that the number of bits available per sample of audio is the same for all speech samples. All data sets have therefore been resampled to 16 kHz.

Table 1: *Emotion data set information*

Data set	**EMO-DB**	**eNTERFACE**	**Polish-Emo**	**RECOLA**
Subjects (m/f)	5/5	34/8	12/12	19/27
Sample Rate	16 kHz	48 kHz	44.1 kHz	44.1 kHz
Language	German	English	Polish	French
Emotions	anger, boredom, fear, joy, sadness, neutral	anger, boredom, fear, joy, sadness, neutral, disgust	anger, disgust, fear, happiness, sadness, surprise	arousal, valence

2.2. Codecs

We used three ubiquitous codecs, MP3, AAC [18] and OPUS [2], because they find widespread use on many online platforms like YouTube, iTunes and many more. In addition, they are freely available, allowing for easy reproduction of the results presented. As reported in the *FFmpeg Codecs Documentation* [11], the libraries libmp3lame, libfdk aac, and libOPUS are the best performing implementations. All three codecs largely share the same available bitrates and were forced to use their constant bitrate mode to ensure a meaningful comparison between codec-bitrates. Variable bitrate modes are difficult or impossible to compare as they are based on subjective quality levels. The speech data was encoded at the following bitrates: 12, 16, 24, 32, 48, 64, 96, 128, 192 and 256 kbit/s. However, due to a limited frame buffer size in the AAC and MP3 codec, the upper-most bitrates available are 96 and 128 kbit/s, respectively. The OPUS codec was forced to use its internal SILK codec [2], since it is optimised for speech signals. MP3 and AAC, on the other hand, were primarily designed for music content, which may prove to be a disadvantage in the SER case.

2.3. Speech emotion recognition algorithms

2.3.1. openSMILE & SVM

In the traditional SER setting where the emotion is considered static within a single utterance, it is typical to extract a set of high level features in the pre-processing stage, which are then fed into a standard machine learning model for classification or regression.

Previous studies have predominantly utilised one feature set, usually the emobase [19] feature set from the openSMILE tool kit [5, 12, 6]. openSMILE has a range of example feature sets of various sizes and combinations of features. We have selected the feature sets, IS09_emotion with 384 features [20], ComParE with 6373 features [21], and emo_large with 6552 features [19], as they have been used as the baseline feature set in multiple Interspeech Computational Paralinguistic Challenges. These large feature sets aim to generate a rich feature space allowing the model to discover useful patterns for SER. In addition, we have also selected the eGeMAPS feature set with 88 features [22]. This feature set was built using physiological and signal processing expert knowledge of the voice production system. It aims to focus on known voice characteristics to distinguish between emotions. Each set provides a single feature vector per utterance. Each feature in the set is a statistical summary of a frame-wise low-level descriptor (LLD). One such example would be the mean fundamental frequency calculated over an utterance.

We chose a simple SVM with linear kernel as our algorithm

of choice, as it is commonly used as a baseline algorithm for SER. We also normalized our data using a mean and standard deviation normalisation which was computed on the training set and applied on the test set.

2.3.2. End-to-end

The RECOLA data set was successfully used to train an end-to-end architecture for predicting arousal by Trigeorgis *et al.* [23]. We consequently adopt this approach as well, to test the robustness of DNNs under different encodings. Our network consists of two 1D convolution layers containing 20 and 40 filter banks accordingly, which are respectively followed by two max pooling layers, with a stride of 2 and 10. The two max pooling layers effectively downsample the signal to allow for a more efficient implementation and help with learning higher level representations.

This first part of the network acts as a feature extractor. Its output is first flattened and then fed to two uni-directional LSTM layers [13], each of them having a size of 256. We also used a dropout of 0.5 after each convolutional layer to prevent overfitting. The output of the last LSTM is mapped to the arousal prediction through a fully-connected layer followed by a tanh activation.

We train the model to maximize the Concordance Correlation Coefficient (ρ_c) for 50 epochs with a batch size of 25 examples and a learning rate of 0.0001 and choose the model that performed best on the validation set. We also performed all of the post-processing steps outlined in Trigeorgis *et al.* [23], namely *median filtering*, *centring*, *scaling* and *time shifting*.

3. Experiments

We present the results of our experiments in Sections 3.1 and 3.2. As we shall see in Section 3.2, the end-to-end approach does not suffer from the effects of encoding. We therefore only include results for the most challenging scenario of mismatched conditions. For the acoustic features approach, we report the results from all scenarios.

3.1. openSMILE & SVM

Given the limited size of our data sets, we decided to report results on a leave one speaker out (LOSO) speaker independent cross validation (CV) for all our experiments, and report UAR. In Figures 1 to 3, we show the average UAR across all folds and an error bar denoting its standard deviation.

3.1.1. Matched conditions

We initially investigate the effects of using compressed audio on each feature set's ability to accurately represent emotion. This is tested by comparing the performance of our models in the case where we train on compressed audio and test on compressed audio of the same kind as opposed to training and testing on the original audio itself.

As we see in Figure 1, for most combinations of feature sets, codecs and bitrates, the UAR does not differ from the uncompressed counterpart. With a bitrate greater than 64 kbit/s, there is little to no decrease in UAR across all codecs. This result implies that while there is a substantial perceptual loss in quality with a decreasing bitrate, there is still enough information in the compressed speech to maintain UAR results comparable to its uncompressed counterpart.

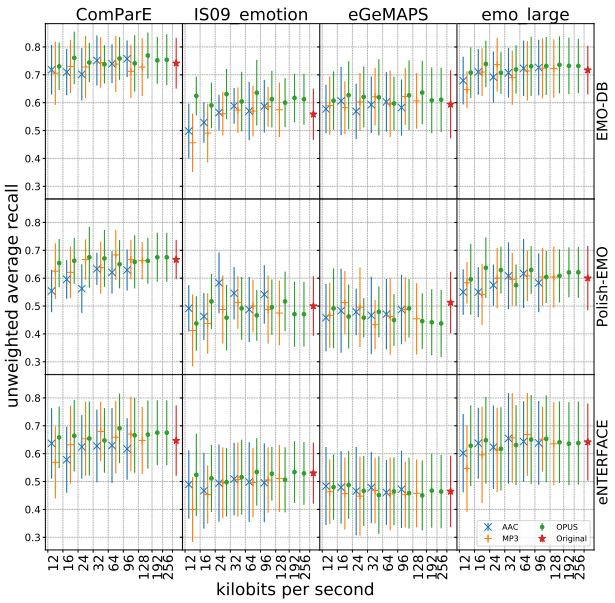

Figure 1: *UAR of SER models in the matched condition setting*

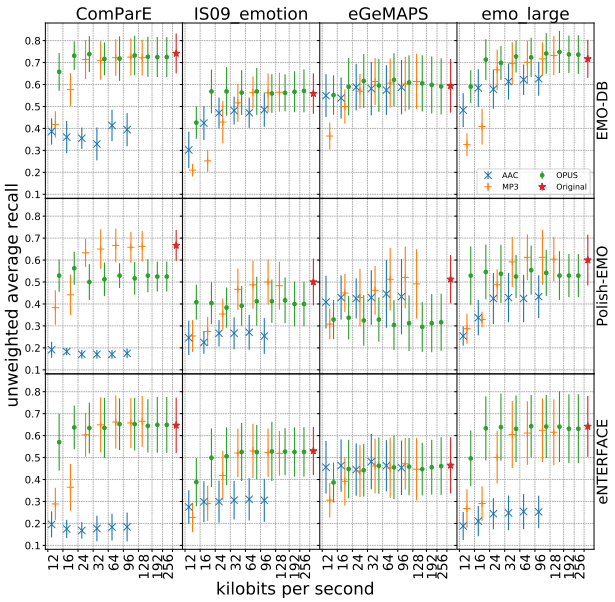

Figure 2: *UAR of SER models trained in the mismatched condition setting*

3.1.2. Mismatched conditions

After establishing that compressed audio still contains enough relevant information to identify emotion, we investigate whether an algorithm trained on uncompressed speech can perform satisfactorily in a production environment, where the incoming audio might have undergone any possible form of encoding.

The results presented in Figure 2 show that in most cases there is a large drop in UAR on the compressed test sets. The AAC codec shows on average the worst performance, with none of the bitrates achieving consistent results across data sets or

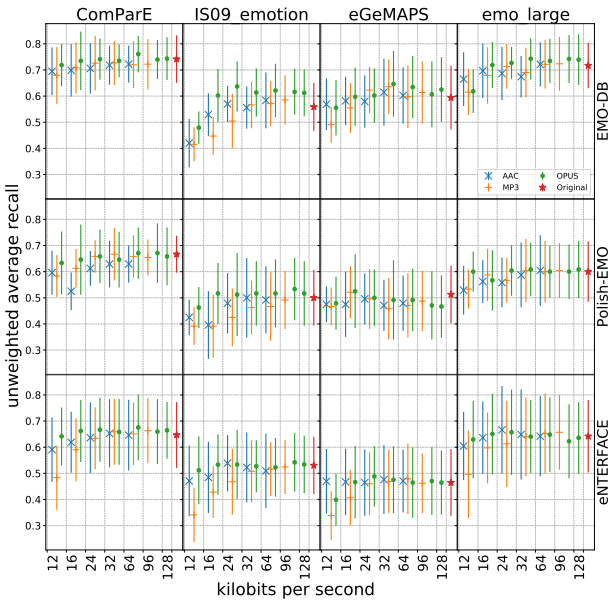

Figure 3: *UAR of SER models in the multi-condition setting*

feature sets. This results suggests that the AAC codec is not suitable in such a scenario. For the MP3 codec, only bitrates greater than 96 kbit/s perform well across data sets and return UAR scores on par with the uncompressed test sets.

A closer examination of the Polish-EMO data set, where OPUS suffers from the biggest decrease in UAR, −20 % on average, revealed that there is a DC offset in all of its samples. The OPUS codec applies a highpass filter to remove the DC offset, a pre-processing step that is not performed in the MP3 and AAC codecs, and this is the likely cause of this huge drop in performance, especially compared with OPUS' consistency across the other two data sets. It should however be noted that having a negligible DC offset is standard recording practice, as in the EMO-DB and eNTERFACE data sets, where the OPUS codec is able to maintain a UAR score comparable to its uncompressed counterpart for all but the lowest bitrates. Nevertheless this inconsistency across data sets means that in such a scenario OPUS is also not suitable for highly robust SER.

3.1.3. Multi-condition

Finally, we investigate the most realistic production setting, where our training set is augmented with compressed audio signals of specific bitrates, but our test set consists of audio signals compressed with a different bitrate. Ensuring different bitrates between the training and test sets simulates a production environment scenario when the actual bitrate is unknown or variable. We always use higher bitrates to augment our training set because the more interesting case is when the test set contains lower audio quality. Specifically, we use the two immediately higher available bitrates for augmenting our training set, e.g. for testing the 12 kbit/s audio, we augment our training set with 24 and 32 kbit/s samples.

The results presented in Figure 3 show that the worst case AAC tests have greatly improved after data augmentation. In particular, for the ComParE feature set, the AAC codec improves by 35 % on average across the data sets. A similar improvement can be seen for the MP3 codec at its lowest bitrates. As expected, the OPUS codec now achieves UAR results similar to the uncompressed results across all data sets, feature sets and all but the lowest bitrates. In many cases, by augmenting the uncompressed training set with just two compressed versions of itself we are able to build a model which is robust to the temporal and spectral artefacts introduced by the compression process.

3.2. End-to-end

As seen in Table 2, the end-to-end approach does not suffer from the effects of encoding. We see only marginal differences in the concordance correlation coefficient (CCC) results under all encodings and bitrates. This indicates that such architectures can potentially be more robust to encoding noise than traditional machine learning approaches.

Table 2: *RECOLA CCC results in mismatched conditions*

Original	Bitrate (kbit/s)	OPUS	MP3	AAC
0.4195	12	0.4068	0.4221	0.4189
	16	0.4089	0.4221	0.4205
	24	0.4089	0.4221	0.4216
	32	0.4091	0.4221	0.4121
	64	0.4162	0.4198	0.4130
	96	0.4151	0.4195	0.4135
	128	0.4147	0.4195	
	192	0.4153		
	256	0.4150		

4. Conclusion

In this study, we established that enough emotion relevant information survives the encoding process to produce UAR scores similar to the uncompressed counterpart. This effect was demonstrated across data sets and feature sets. However, in the mismatched condition scenario, the changes imposed by the encoding process can result in a large loss in SER UAR. In many cases, a model trained on high-quality uncompressed audio will not work with low bitrate encoded input. In addition, even at the higher bitrates, large inconsistencies in SER UAR were found across data sets, most notably and consistently for the AAC codec but also for the OPUS codec with the Polish-EMO data set. These negative effects can be mitigated by augmenting the training set with encoded audio. In fact, we have shown that in the majority of cases, augmenting with only two encoded versions is required to regain the lost UAR. Indeed, if, in the multi-condition training, a representative subset of a codec's bitrate range is used for augmentation, the resulting model is likely to be robust to the artefacts of the codec. For the conditions investigated in the multi-condition case withstanding, this result means that speech data can be procured from an online source without concern for the encoding process affecting the SER result. End-to-end architectures, on the other hand, seem not to suffer from this effect at all and can be reliably used independent of the encoding conditions.

5. Acknowledgements

We would like to thank Uwe Reichel and Hagen Wierstorf for their helpful and constructive comments on the manuscript.

6. References

[1] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram and N. Harte, 'Perceived audio quality for streaming stereo music', in *ACM International Conference on Multimedia*, 2014, pp. 1173–1176. DOI: 10.1145/2647868.2655025.

[2] C. Hoene, J. Valin, K. Vos and J. Skoglund, 'Summary of Opus listening test results', IETF, 2013. [Online]. Available: https://tools.ietf.org/html/draft-ietf-codec-results-03.

[3] B. W. Schuller, 'Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends', *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018. DOI: 10.1145/3129340.

[4] A. Albahri, M. Lech and E. Cheng, 'Effect of speech compression on the automatic recognition of emotions', *International Journal of Signal Processing Systems*, vol. 4, no. 1, pp. 55–61, 2015. DOI: 10.12720/ijsps.4.1.55-61.

[5] I. Siegert, A. Requardt, O. Egorow and S. Wolff, 'Utilizing psychoacoustic modeling to improve speech-based emotion recognition', in *International Conference on Speech and Computer*, 2018, pp. 625–635. DOI: 10.1007/978-3-319-99579-3_64.

[6] I. Siegert, A. Requardt, L. Linda Duong and A. Wendemuth, 'Measuring the impact of audio compression on the spectral quality of speech data', in *Elektronische Sprachsignalverarbeitung*, O. Jokisch, Ed., TUD Press, 2016, pp. 229–236.

[7] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[8] N. García, J. Vasquez, J. D. Arias-Londoño, J. Vargas-Bonilla and J. R. Orozco, 'Automatic emotion recognition in compressed speech using acoustic and non-linear features', in *Symposium on Signal Processing, Images and Computer Vision*, 2015, pp. 1–7. DOI: 10.1109/STSIVA.2015.7330399.

[9] S. Frühholz, E. Marchi and B. Schuller, 'The effect of narrowband transmission on recognition of paralinguistic information from human vocalizations', *IEEE Access*, vol. 4, Sep. 2016. DOI: 10.1109/ACCESS.2016.2604038.

[10] A. Requardt, I. Siegert, M. Maruschke and A. Wendemuth, 'Audio compression and its impact on emotion recognition in affective computing', Mar. 2017.

[11] (2019). FFmpeg codecs documentation, [Online]. Available: https://www.ffmpeg.org/ffmpeg-codecs.html.

[12] 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.

[13] 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.

[14] F. Burkhardt, A. Paeschke, M. A. 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.

[15] 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.

[16] P. Staroniewicz and W. Majewski, 'Polish emotional speech database – recording and preliminary validation', in *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*, A. Esposito and R. Vích, Eds., Springer, 2009, pp. 42–49. DOI: 10.1007/978-3-642-03320-9_5.

[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, pp. 1–8. DOI: 10.1109/FG.2013.6553805.

[18] K. Brandenburg, 'MP3 and AAC explained', in *International Conference on High-Quality Audio Coding*, 1999. [Online]. Available: http://www.aes.org/e-lib/browse.cfm?elib=8079.

[19] F. Eyben, M. Wollmer and B. Schuller, 'OpenEAR - introducing the Munich open-source emotion and affect recognition toolkit', in *International Conference on Affective Computing and Intelligent Interaction*, 2009, pp. 1–6. DOI: 10.1109/ACII.2009.5349350.

[20] B. Schuller, S. Steidl and A. Batliner, 'The Interspeech 2009 Emotion Challenge', in *Interspeech*, 2009, pp. 312–315. [Online]. Available: https://www.isca-speech.org/archive/interspeech_2009/i09_0312.html.

[21] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho and K. Evanini, 'The Interspeech 2016 Computational Paralinguistics Challenge: Deception, sincerity and native language', in *Interspeech*, 2016, pp. 2001–2005. DOI: 10.21437/Interspeech.2016-129.

[22] F. Eyben, K. R. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan and K. 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, 2015. DOI: 10.1109/TAFFC.2015.2457417.

[23] 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.