Emotion recognition in public speaking scenarios utilising an LSTM-RNN approach with attention

Alice Baird, Shahin Amiriparian, Manuel Milling, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Baird, Alice, Shahin Amiriparian, Manuel Milling, and Björn W. Schuller. 2021. "Emotion recognition in public speaking scenarios utilising an LSTM-RNN approach with attention." In *Spoken Language Technology Workshop (IEEE SLT 2021), Shenzhen, China, 19-22 January 2021*, edited by Zhijian Ou, Lei Xie, and Hsin-Min Wang, 397–402. New York, NY: IEEE. https://doi.org/10.1109/slt48900.2021.9383542.

EMOTION RECOGNITION IN PUBLIC SPEAKING SCENARIOS UTILISING AN LSTM-RNN APPROACH WITH ATTENTION

Alice Baird [1], *Shahin Amiriparian* [1], *Manuel Milling* [1], *Björn W. Schuller* [12]

[1] Chair of Embedded Intelligence for Healthcare and Wellbeing, University of Augsburg Germany
[2] GLAM – Group on Language, Audio, & Music, Imperial College London, UK

ABSTRACT

Speaking in public can be a cause of fear for many people. Research suggests that there are physical markers such as an increased heart rate and vocal tremolo that indicate an individual's state of wellbeing during a public speech. In this study, we explore the advantages of speech-based features for continuous recognition of the emotional dimensions of arousal and valence during a public speaking scenario. Furthermore, we explore biological signal fusion, and perform cross-language (German and English) analysis by training language-independent models and testing them on speech from various native and non-native speaker groupings. For the emotion recognition task itself, we utilise a *Long Short-Term Memory - Recurrent Neural Network* (LSTM-RNN) architecture with a self-attention layer. When utilising audio-only features and testing with non-native German's speaking German we achieve at best a *concordance correlation coefficient* (CCC) of 0.640 and 0.491 for arousal and valence, respectively – demonstrating a strong effect for this task from non-native speakers, as well as promise for the suitability of deep learning for continuous emotion recognition in the context of public speaking.

***Index Terms*—** public speaking, affective computing, long short-term memory, recurrent neural networks

1. INTRODUCTION

In modern society, public dissemination is a useful tool for knowledge-sharing. However, having a fear of public speaking means that some individuals avoid this opportunity. Public speaking can provoke disorders, including *Generalised Anxiety Disorder* (GAD), and acute stress [1], both having a substantial effect on short-term wellbeing [2]. Physical markers of such disorders are prominently observed in speech [3]. Furthermore, cultural differences in regards to an individual's response to the fear of public speaking have been researched, with markers including varied heart and speech rates [1].

To this end, observing emotional states during public speaking allows for a strong indication of the overall state of wellbeing [4], particularly as research has shown that an individual's typical emotion production can change during public speaking [5]. With this in mind, biological signals are not readily observable and require rather invasive methods to be continuously captured. Audio, however, can be observed non-invasively, and has shown to be a reliable indicator for an individual's state or trait [6, 7]. However, research has shown that the '*illusion of transparency*' can mean that alterations in speech are more prominent to the speaker than the audience [8], suggesting that biological signals may be more valuable for observation during a public speech.

For the current study, we have two core research goals, i) to evaluate if speech-based audio features are useful for recognition of emotion during a public speaking scenario ii) to understand the impact of fusing audio and biological signals for speech emotion recognition. To explore these goals we implement a deep learning-based approach, utilising a long short-term memory - recurrent neural network (LSTM-RNN) architecture with self-attention, to predict continuous emotion (arousal and valence). We apply an attention mechanism as this has shown to improve results for most sequence-based tasks, including emotion recognition [9, 10], as well as healthcare tasks such as continuous detection of sleepiness [11]. We train a series of models on various acoustic features as well as fusing the biological signals of 'blood volume pulse' (BVP) and 'skin conductance' (SC). The dataset utilised in this current work includes individuals speaking in both German and English during a public speaking scenario. Moussu et al. have shown that speaking in front of others in ones non-native language may cause more fear [12]. Motivated by this research and given the bilingual nature of the current dataset, we further organise the dataset into nativeness groupings to explore this effect.

The rest of our paper is organised as follows. First, in Section 2, we outline some related computational approaches in this area. We then describe the database used in our experiments in Section 3. Following this, we outline our experimental settings for the task of emotion prediction from speech and biological signals in Section 4. Our results are then discussed in Section 5, and we further perform a brief acoustic analysis to more deeply observe the machine learning result in Section 6. Finally, we provide conclusions and future work plans in Section 7.

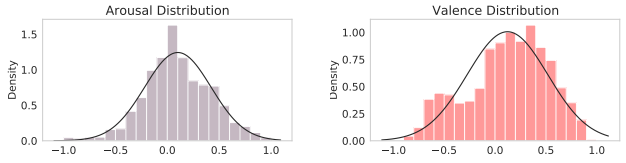

Fig. 1. Distribution of the gold-standard ratings for arousal (left) and valence (right) across all speakers in the BIOS-DB , including the normalised distribution curve.

2. BACKGROUND

In the field of affective computing, there are numerous computational methods for continuous recognition of emotional states, and the success of speech emotion recognition (SER) has led to a growing research community [13, 14]. Promising results for SER based on end-to-end learning [15], as well as unsupervised representation learning with sequence to sequence autoencoders [16] and image-to-audio transfer learning systems with pre-trained convolutional neural networks (CNNs) (e. g. , DEEP SPECTRUM) show promise [17]. However, many studies still find success with more hand-crafted approaches [18, 19]

Recently, Zhao et al. [20] applied an artificial neural network consisting of convolution-based local feature learning blocks and recurrent layers with LSTM cells to improve previous results on the Berlin EmoDB database [21]. In regards to biological signal prediction from speech, there have been minimal works [22], however, in [23], the authors use an end-to-end approach based on 1D convolutional and recurrent layers for the prediction of emotions from biological signals, achieving competitive results on the Audio Visual Emotion Challenge (AVEC) 2016 data [24].

With the current contribution focused on emotion prediction during a public speaking scenario, we found that there has been minimal research in this area. In [25], the authors utilise conventional machine learning classifiers such as support vector machines (SVMs) to infer emotion from non-verbal vocalisations, during a public speaking exercise. Similarly, in [26], the authors explore various window lengths for extracting standard audio features, i. e. , jitter, pitch, and MFCCs, to predict states of stress during public speaking. With these works in mind, to the best of the authors' knowledge, this contribution is the first of its kind, to utilise deep learning methods in the context of public speaking for emotion recognition, as well as exploring the advantages of biological signal fusion.

3. THE BIOSPEECH DB

The BioSpeech DB (BIOS-DB) was first introduced in [22], and the interested reader is referred to this publication for further details. In this contribution, we present an updated version of the BIOS-DB . This version includes higher resolution biological signals, which may be of more use to a wider variety of research fields[1].

The currently used version of the BIOS-DB contains 42 speakers (17 female), with a mean age of 26.76 years, and a standard deviation of 6.62 years. From these speakers, 30 are native German, and 12 are from a variety of foreign countries. The number of speakers in the dataset is typical for the field of computational paralinguistics [28], due to the time-cost related to quality data collection and annotation [29]. Each participant was asked to speak a text ("The North Wind and the Sun") out loud in front of a minimum of 4 observers, speaking in German and English. During their speech, 3 of the observers were using joysticks to continually rate the emotion of the speakers in regards to the arousal and valence dimensions, where the Y axis is arousal (i. e. , strength of the emotion), and the X axis is the valence of the emotion (i. e. , negative or positive). During their speech, two channels of audio were captured, one from a lapel microphone and one from a room microphone placed on the table in front of the speaker. Furthermore, two sensors were placed on the finger of the participants to capture blood volume pulse (BVP) as a % of blood volume pressure, and skin conductance (SC) measured in microSiemens (μS), at a sampling frequency of 2 048 Hz and 256 Hz, respectively.

3.1. Data Processing

Audio data was captured at a 44.1 kHz sampling rate with 16 bit resolution in MONO WAV format, and has been converted to 16 kHz, 16-bit WAV for use with popular feature extraction tool kits. As our fusion strategy requires an identical sample frequency of the given modalities, we choose 16 Hz for audio feature extraction as well as for resampling of the biological signals. We observe the expected loss of information caused by the resampling of the biological signals to be minimal, and to support this we performed peak analysis across the BVP signals of all speakers, finding that the mean distance between peaks is 0.668 s and 0.656 s, for 2 048 Hz and 16 Hz, respectively

A gold-standard for the emotion labels was calculated between the three raters utilising the Evaluator Weighted Estimator (EWE) method. EWE is described with more detail in [30] and has been applied repeatedly on emotion-based corpora [31]. The mean inter-rater agreement across all speakers in the BIOS-DB , from the three annotators was 0.47 and 0.36 (based on a range of [0,1]) for arousal and valence, respectively. For the machine learning experiments the gold-standard emotion labels were re-scaled to [-1,1] based on the maximum possible value. In Figure 1, the distribution of the gold-standard ratings for both emotional dimensions, across all speakers used in our experiments is given.

[1]Follow the DOI for a link to the current data repository [27]

Table 1. Subset of the original BIOS-DB - speaker independent partitions, for German (GER) and NonGerman (NonGER) speakers (#).

#	Train	Test	$\sum$
GER	25	5	30
NonGER	7	5	12
$\sum$	32	10	42

4. EXPERIMENTAL SETTINGS

A brief overview of our recurrent attention-based approach for the task of emotion recognition (arousal and valence) from speech with biological signal fusion is given in Figure 2. We evaluate the data in several language-based groups, training language-dependent models, and testing on various native and non-native subsets. An overview of the data distribution across the speaker-independent partitions is given in Table 1.

4.1. Audio Features

We extract conventional hand-crafted speech features, as well as a deep learning approach in which spectrogram-based data representations are extracted from the speech signals.

As a conventional and well established speech approach, the 88 dimensional EGEMAPS feature set [32], is used given the advantages found in similar paralinguistic tasks [3]. From each audio instance, EGEMAPS acoustic features are extracted with the OPENSMILE toolkit [33]. The default parameter settings from OPENSMILE are used, and features are extracted with a window size of 62.5 ms.

For the unsupervised approach, we extract a 4 096-dimensional feature set of deep data-representations using the DEEP SPECTRUM toolkit [34][2]. DEEP SPECTRUM has shown success for similar audio- and speech-based tasks [35], and extracts features from the audio data using pre-trained image convolutional neural networks (CNNs). For this study, we extract *Viridis* colour map spectrograms (cf. Figure 3 for colour map), and we extract features with the same window length of 62.5 ms, using the default VGG16 pre-trained network. During training, the DEEP SPECTRUM features are standardised by subtracting the mean and scaling to unit variance.

4.1.1. Data Augmentation

To infer the effect of data quantity in this context, we apply data augmentation to the spectrogram features. For this we extract the spectrogram representations of each audio file at a window length of 1 second and a hop size of 62.5 ms. We then apply the SPECAUGMENT approach [36] which masks portions of frequency and time from each extracted spectrogram. With this approach we duplicate the training data by a factor of 2. We then utilise the methodology provided by the DeepSpectrum toolkit to extract feature vectors from the augmented spectrogram images directly.

[2] https://github.com/DeepSpectrum/DeepSpectrum

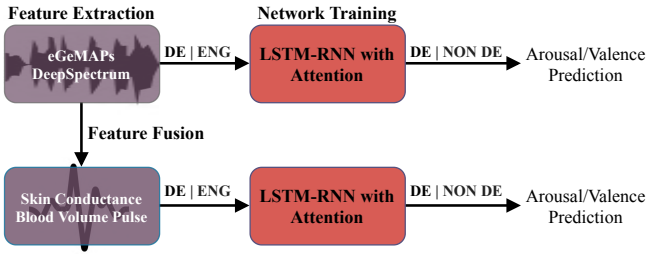

Fig. 2. Overview of the experimental settings applied in this study for the task of emotion recognition (arousal and valence) from speech and biological signal fusion.

4.2. LSTM-RNN with Attention

In order to address the sequential nature of both the audio and biological-based input features, we implement an LSTM-RNN based architecture. The network consists of one recurrent layer with 128 LSTM cells. We then add a self-attention sequence layer, with a sequence-wide attention window, and sigmoid as the attention activation. The output is fed into a feed-forward layer to provide the predictions.

4.2.1. Network training

We train each model for 5 epochs with a batch size of 64 using the Adam optimiser with a learning rate of 0.001. Due to the temporal nature of both the signals, we reshaped the input data in sequences of 20 feature vectors (equal to 1.25 seconds). Training of the network is made speaker independently, and the model is updated iteratively per speaker. To avoid potential speaker bias caused by this training method, for each trained model, we shuffle the order of speakers.

Two types of language-based models (German and English) are trained with the two feature sets described (EGEMAPS and DEEP SPECTRUM), as well as fusing the BVP and SC signals, to evaluate the effect of this fusion strategy. For model testing, we group the speakers into Native-Germans speaking German (GER-GER), Native-Germans speaking English (GER-ENG), Non-Germans speaking German (NonGER-GER), and Non-Germans speaking English (NonGER-ENG). We also report results from all test speakers together (All). To evaluate the prediction accuracy, we utilise concordance correlation coefficient (CCC) as our evaluation metric. CCC is well established in the field of SER [15, 37], and considers offset and scaling variance better than the conventional Correlation Coefficient metric [38].

Table 2. Continuous (A)rousal and (V)alence recognition results from the BIOS-DB . Results obtained from the mean of all test speakers in that grouping, across the 5 best-performing models trained on both English (ENG) and German (GER) languages (Lang.). † indicates the use of data augmentation. Reported is *concordance correlation coefficient* (CCC) from audio features only, as well as from the fusion of blood volume pulse (BVP) and skin conductance (SC). Emphasised results for arousal indicate >0.3 CCC, and for valence >0.1 CCC. Results with * are discussed in Section 5.

		GER-GER		GER-ENG		NonGER-GER		NonGER-ENG		All	
Feature Set	Lang.	A	V	A	V	A	V	A	V	A	V
eGeMAPS	ENG	.072	.045	.165	**.102***	**.403**	**.279***	**.582***	**.175***	.269	**.130***
	GER	.075	.069	.147	.072	**.382**	**.148**	**.370**	.088	.203	.084
DeepSpectrum	ENG	.046	.074	.117	.045	.233	.089	**.349**	.063	.147	.037
	GER	-.003	.064	-.004	.021	**.471**	.078	**.339**	.010	.114	.028
DeepSpectrum†	ENG	.172	.056	.018	**.393**	.283	**.308**	.156	**.244**	.158	**.226**
	GER	-.050	**.192**	.096	**.419**	**.640**	**.491**	**.387**	**.296**	**.344**	**.334**
eGeMAPS + BVP	ENG	.127	.059	.260*	**.112**	**.423**	**.347***	**.518***	**.100**	.269	**.145***
	GER	.084	.053	.160	.018	**.307**	.036	**.358**	.035	.194	.029
DeepSpectrum + BVP	ENG	.077	.055	.169	.034	.292	.088	**.361**	.015	.155	.034
	GER	.034	.076	.057	.040	**.491**	**.103**	**.342**	.017	.167	.042
eGeMAPS + SC	ENG	.082	.077	.166	.086	**.438**	**.165**	**.468***	.099	.228	.094
	GER	.047	.051	.096	**.129**	**.333**	**.298**	.171	**.126**	.114	**.137**
DeepSpectrum + SC	ENG	.027	.066	.065	.001	**.376**	.006	**.353**	.027	.145	.010
	GER	.056	.051	.099	.078	.271	**.141**	.165	.176	.178	.078

5. DISCUSSION OF RESULTS

An overview of the results for all experimental paradigms is given in Table 2. Where we discuss significant differences, this is based on the predictions from all speakers and a mean of all models and proceeds an evaluation of normality using a Shapiro-Wilktest [39]; we then perform a two-tailed T-test, and reject the null hypothesis at a level of $p < 0.05$.

The results in Table 2 show that language appears to play a notable role for emotion recognition in this context. The best Native-German correlation for arousal is 0.260 CCC for the model trained on English, and tested on Germans speaking English (GER-ENG). As well as this, The German only models (GER-GER) have consistently negligible correlations as compared to NonGER-GER.

Furthermore, if we look only at the Non-German results, we see a promising increase in CCC, particularly for the English trained models. Indeed, in this paradigm, we see our best valence result comes from NonGER-GER, with 0.279 CCC, which is then significantly ($p < 0.05$) increased through the fusion of BVP biological signals up to 0.347 CCC. This result suggests that the positive to negative aspect of emotion (which is typically a challenge for audio modelling) is captured more easily when individuals are speaking in their non-native language. A result which is slightly agreed upon with the GER-ENG valence score of 0.102 CCC, and even from the NonGER-ENG valence results of 0.175 CCC – as within the dataset there are only 2 native English speakers.

For the prediction of arousal, the best correlation is 0.582 CCC, and comes from the audio-only English model, when testing on NonGER-ENG, a result which is significantly ($p < 0.05$) stronger than results obtained from fusion of BVP and SC signals. For the fusion results in general, we see that for arousal there is little to no benefit; however, valence does show improvement particularly with BVP, and with the EGEMAPS result for the SC German model (0.298 CCC).

Across most of the testing paradigms, the handcrafted EGEMAPS features perform better than DEEP SPECTRUM for this task. However, through the use of data augmentation, we see more stable results across all, with our best result for arousal of .640 CCC obtained in the NonGER-GER grouping. This suggests that the data augmentation approach is suitable to improve the robust nature of the model, and further establishes the findings in regards to the language groupings, as we do see similar patterns of behaviour between the feature sets.

In this regard, for the GER-GER model trained on English there is an improvement with data augmentation, which would also point to the language dependency of our models and the task itself. For further research, it would be of interest to perform feature selection with the EGEMAPS features, to explore which features from this set perform highly in this context. As well as this various audio-based augmentation approaches, such as additive noise, may also improve the robustness of these results.

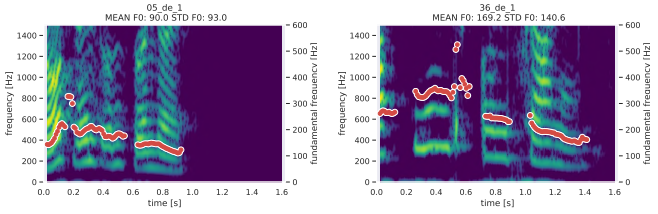

Fig. 3. Spectrogram representation of speaker #5 GER-GER (left), and speaker #36 NonGER-GER (right). Of note, we see that there is a shorter duration of speech in GER- samples, and more pitch (F0) variance in NonGER- samples.

6. ACOUSTIC ANALYSIS

To explore the machine learning results further, we perform a brief acoustic analysis on the audio samples. We extract the standard deviation (STD) of Pitch F0 (Hz), and intensity (dB), from a short (ca. 2 seconds) segment containing the phrase "The North Wind and the Sun" in English or "Der Nordwind und die Sonne" in German. We additionally extract spectrograms from the samples, plotting F0 to visually observe any acoustic phenomena present in the samples (cf. Figure 3 for a selection of spectrogram representations.).

6.1. Fundamental Frequency

When analysing the fundamental frequency (F0), we find that native German speakers have an overall lower mean F0. Germans speaking German and English show a mean F0 of 105.87 Hz and 113.72 Hz, respectively, and a mean F0 standard deviation (STD) of 43.11 Hz, and 45.90 Hz. However, Non-Germans speaking German and English have a mean of 145.23 Hz (F0 STD: 52.15 Hz), and 150.52 Hz (F0 STD: 45.95 Hz), respectively. This leads us to assume that F0 plays a role in this task, particularly as the STD in the speech signal is higher for Non-German speakers. We also see as expected that female speakers show a higher F0 than males, 168.05 Hz (mean STD 49.51 Hz) compared to 87.05 Hz (mean STD 42.39 Hz), potentially this F0 variance may have aided emotion recognition for Non-German speakers as there is a slight imbalance in the dataset (7:8 Non-German:German, females), although this would require further investigation.

6.2. Intensity

We additionally extract the sound intensity (dB) from each of the audio files. In this case, we see that speakers speaking German have a lower mean intensity than those speaking English, 63.09 dB, 66.92 dB, respectively. We see further that Non-Germans speaking English have a higher mean intensity as compared to all other groupings tested of 68.24 dB. This observation leads us to assume that intensity of the audio signal is also meaningful, as we again see a more prominent variance in Non-German speakers, although as with the finding for F0 this would require a more in-depth analysis.

7. CONCLUSIONS AND FUTURE WORK

In this contribution we presented results from experiments focusing on emotion recognition in a public speaking scenario. We utilised an LSTM-RNN with an attention mechanism and evaluated various audio-based feature sets as well as fusion with biological signals. Of high interest, findings suggest that speech variances from the Non-German speakers may have aided modelling of emotion for this grouping. A finding which is further established through a continued trend across results in each grouping (even after data augmentation), e. g. , the large disparity in results when testing on German speakers speaking German, although this would need further more specific evaluation.

We find that audio features are suitable alone for the task of predicting emotion in this context. Although fusing biological signals with audio has shown only minimal CCC improvement in most cases, valence does appear to be modelled better with this fusion strategy. However, this behaviour was not consistent across feature sets and so would also require a deeper analysis. To this point, it would be of great interest to explore temporal and frequency domain feature extraction of the biological signals in a future study, as well as conducting a more close analysis of the acoustic findings relating to speaker nativeness during public speaking.

8. ACKNOWLEDGEMENTS

This work is funded by the Bavarian State Ministry of Education, Science and the Arts in the framework of the Centre Digitisation.Bavaria (ZD.B).

9. REFERENCES

[1] G. D. Bodie, "A racing heart, rattling knees, and ruminative thoughts: Defining, explaining, and treating public speaking anxiety," *Communication education*, vol. 59, no. 1, pp. 70–105, 2010.

[2] T. J. Davis, M. Morris, and M. M. Drake, "The moderation effect of mindfulness on the relationship between adult attachment and well-being," *Personality and Individual Differences*, vol. 96, pp. 115–121, 2016.

[3] A. Baird, , N. Cummins, S. Schnieder, J. Krajewsk, and B. Schuller, "An Evaluation of the Effect of Anxiety on Speech - Computational Prediction of Anxiety from Sustained Vowels," in *Proc. INTERSPEECH 2020*. Shanghai, China: ISCA, 2020, p. [to appear]].

[4] N. S. Schutte and J. M. Malouff, "Emotional intelligence mediates the relationship between mindfulness and subjective well-being," *Personality and individual differences*, vol. 50, no. 7, pp. 1116–1119, 2011.

[5] A. N. Niles and M. G. Craske, "Incidental emotion regulation deficits in public speaking anxiety," *Cognitive Therapy and Research*, vol. 43, no. 2, pp. 419–426, 2019.

[6] B. W. Schuller and A. M. Batliner, *EMOTION, AFFECT AND PERSONALITY IN SPEECH AND LANGUAGE PROCESSING*. Wiley Online Library, 1988.

[7] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 Emotion Challenge," in *Proceedings INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association*, ISCA. Brighton, UK: ISCA, September 2009, pp. 312–315.

[8] A. M. Goberman, S. Hughes, and T. Haydock, "Acoustic characteristics of public speaking: Anxiety and practice effects," *Speech communication*, vol. 53, no. 6, pp. 867–876, 2011.

[9] Z. Zhao, Z. Bao, Y. Zhao, Z. Zhang, N. Cummins, Z. Ren, and B. Schuller, "Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition," *IEEE Access*, vol. 7, pp. 97 515–97 525, 2019.

[10] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2227–2231.

[11] S. Amiriparian, P. Winokurov, V. Karas, S. Ottl, M. Gerczuk, and B. W. Schuller, "A novel fusion of attention and sequence to sequence autoencoders to predict sleepiness from speech," *arXiv preprint arXiv:2005.08722*, 2020.

[12] L. Moussu and E. Llurda, "Non-native english-speaking english language teachers: History and research," *Language Teaching, 2008, vol. 41, núm. 3, p. 315-348*, 2008.

[13] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[14] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, 2020.

[15] 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 *Proc. ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[16] S. Amiriparian, M. Freitag, N. Cummins, and B. Schuller, "Sequence to Sequence Autoencoders for Unsupervised Representation Learning from Audio," in *Proc. DCASE 2017*, Munich, Germany, 2017, pp. 17–21.

[17] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. W. Schuller, "Snore sound classification using image-based deep spectrum features." in *INTERSPEECH*, vol. 434, 2017, pp. 3512–3516.

[18] H. K. Palo, M. Chandra, and M. N. Mohanty, "Recognition of human speech emotion using variants of mel-frequency cepstral coefficients," in *Advances in Systems, Control and Automation*. Springer, 2018, pp. 491–498.

[19] M. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using mfcc," in *2017 international conference on wireless communications, signal processing and networking (WiSPNET)*. IEEE, 2017, pp. 2257–2260.

[20] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312 – 323, 2019.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[22] A. Baird, S. Amiriparian, M. Berschneider, M. Schmitt, and B. Schuller, "Predicting biological signals from speech: Introducing a novel multimodal dataset and results," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2019, pp. 1–5.

[23] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 985–990.

[24] 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 *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 3–10.

[25] T. Pfister and P. Robinson, "Speech emotion classification and public speaking skill assessment," in *International Workshop on Human Behavior Understanding*. Springer, 2010, pp. 151–162.

[26] M. Soury and L. Devillers, "Stress detection from audio on multiple window analysis size in a public speaking task," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 529–533.

[27] A. Baird, S. Amiriparian, and B. Schuller, "Bios-db: a multimodal database of individuals in a public speaking scenario, including emotional annotation," 2020. [Online]. Available: 10.5281/zenodo.4281253

[28] B. W. Schuller, A. Batliner, C. Bergler, E.-M. Messner, A. Hamilton, S. Amiriparian, A. Baird, G. Rizos, M. Schmitt, L. Stappen *et al.*, "The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks," *Proceedings INTERSPEECH. Shanghai, China: ISCA*, 2020.

[29] S. Hantke, A. Abstreiter, N. Cummins, and B. Schuller, "Trustability-based dynamic active learning for crowdsourced labelling of emotional audio data," *IEEE Access*, vol. 6, pp. 42 142–42 155, 2018.

[30] B. W. Schuller, *Intelligent audio analysis*. Springer, 2013.

[31] L. Stappen, A. Baird, G. Rizos, P. Tzirakis, X. Du, F. Hafner, L. Schumann, A. Mallol-Ragolta, B. W. Schuller, I. Lefter, E. Cambria, and I. Kompatsiaris, "Muse 2020 challenge and workshop: Multimodal sentiment analysis, emotion-target engagement and trustworthiness detection in real-life media," in *1st International Multimodal Sentiment Analysis in Real-life Media Challenge and Workshop, co-located with the 28th ACM International Conference on Multimedia*. ACM, 2020.

[32] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, 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, 2016.

[33] F. Eyben, F. Weninger, F. Groß, and B. Schuller, "Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor," in *Proc. ACM MM '13*. Barcelona, Spain: ACM, 2013, pp. 835–838.

[34] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, and B. Schuller, "Snore Sound Classification Using Image-based Deep Spectrum Features," in *Proc. INTERSPEECH 2017*. Stockholm, Sweden: ISCA, 2017, pp. 3512–3516.

[35] A. Baird, S. Amiriparian, N. Cummins, S. Sturmbauer, J. Janson, E.-M. Messner, H. Baumeister, N. Rohleder, and B. Schuller, "Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test," in *Proc. INTERSPEECH 2019*. Graz, Austria: ISCA, 2019, pp. 534–538.

[36] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[37] W. Li, Y. Zhang, and Y. Fu, "Speech emotion recognition in e-learning system based on affective computing," in *Third International Conference on Natural Computation (ICNC 2007)*, vol. 5. IEEE, 2007, pp. 809–813.

[38] F. Weninger, F. Ringeval, E. Marchi, and B. W. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio." in *IJCAI*, vol. 2016, 2016, pp. 2196–2202.

[39] J. Peat and B. Barton, *Medical Statistics: A Guide to Data Analysis and Critical Appraisal*. Malden, MA, USA: Blackwell Publishing Ltd, 2008.