The INTERSPEECH 2020 Computational Paralinguistics Challenge: Elderly Emotion, Breathing & Masks

Björn W. Schuller[1,2,3], *Anton Batliner*[2,4], *Christian Bergler*[4], *Eva-Maria Messner*[5], *Antonia Hamilton*[6], *Shahin Amiriparian*[2,3], *Alice Baird*[2], *Georgios Rizos*[1], *Maximilian Schmitt*[2], *Lukas Stappen*[2], *Harald Baumeister*[5], *Alexis Deighton MacIntyre*[6], *Simone Hantke*[3]

[1]GLAM – Group on Language, Audio & Music, Imperial College London, UK
[2]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[3]audEERING GmbH, Gilching, Germany
[4]Pattern Recognition Lab, FAU Erlangen-Nuremberg, Germany
[5]Department of Clinical Psychology and Psychotherapy, University of Ulm, Germany
[6]Institute of Cognitive Neuroscience, University College London, UK

schuller@IEEE.org

Abstract

The INTERSPEECH 2020 Computational Paralinguistics Challenge addresses three different problems for the first time in a research competition under well-defined conditions: In the *Elderly* Emotion Sub-Challenge, arousal and valence in the speech of elderly individuals have to be modelled as a 3-class problem; in the *Breathing* Sub-Challenge, breathing has to be assessed as a regression problem; and in the *Mask* Sub-Challenge, speech without and with a surgical mask has to be told apart. We describe the Sub-Challenges, baseline feature extraction, and classifiers based on the 'usual' ComParE and BoAW features as well as deep unsupervised representation learning using the auDeep toolkit, and deep feature extraction from pre-trained CNNs using the Deep Spectrum toolkit; in addition, we partially add deep end-to-end sequential modelling, and, for the first time in the challenge, linguistic analysis.

Index Terms: Computational Paralinguistics, Challenge, Elderly Emotion, Breathing, Speech under Mask

1. Introduction

In this INTERSPEECH 2020 COMputational PARAlinguistics challengE (ComParE) – the twelfth since 2009 [1], we address three new problems within the field of Computational Paralinguistics [2] in a challenge setting: In the **Elderly Emotion** Sub-Challenge (**ESC**), arousal (**A**) and valence (**V**) in the speech of elderly individuals is each modelled as a 3-class problem. In the advent of an ageing population and the associated challenges in health care, digital solutions are discussed and need to be provided to assist the elderly in managing their health and wellbeing [3]. To date, no public speech data set of elderly emotion has been released for such scientific purposes. In the **Breathing** Sub-Challenge (**BSC**), the task is to sequentially predict a temporal signal of recorded breath, which is measured using a piezoelectric respiratory belt worn by the speaker. Breathing patterns provide medical doctors and speech therapists vital information about an individual's respiratory and speech planning [4], insight into human affective states [5, 6], as well as cognitive and neurological health [7, 8]. Computational methods that automatically detect breathing events purely by analysing recorded speech can greatly facilitate such processes [9, 10]. Finally, in the **Mask** Sub-Challenge (**MSC**), the task is to tell apart whether a speaker wears a surgical mask or not. Modelling speech when the speaker wears a face mask is important for forensics and communication between surgeons. [11] report only a small effect of surgical masks on speech understanding by human listeners; this is corroborated for automatic speech understanding in [12]. In [13, 14], different types of masks seemed not to have a great impact on speaker identification. Based on these pilot studies, we might expect a not too high performance when trying to tell apart speech with or without surgical mask.

For all tasks, a target value/class has to be predicted for each case. Contributors can employ their own features and machine learning algorithms; standard feature sets and procedures are provided. Participants have to use the predefined partitions for each Sub-Challenge. They may report results obtained from the Train(ing)/Dev(elopment) set – preferably with the supplied evaluation setups, but have only five trials to upload their results on the Test set per Sub-Challenge, whose labels are unknown to them. Each participation must be accompanied by a paper presenting the results, which undergoes peer-review and has to be accepted for the conference in order to participate in the Challenge. The organisers preserve the right to re-evaluate the findings, but will not participate in the Challenge. As evaluation measure, we employ: (1) in the **ESC** and the **MSC**, **Unweighted Average Recall (UAR)** as used since the first Challenge from 2009 [1], especially because it is more adequate for (unbalanced) multi-class classifications than Weighted Average Recall (i. e., accuracy) [2, 15]; (2) in the **BSC**, **Pearson's Correlation Coefficient** r [16], because the target variable is interval-scaled. Ethical approval for the studies has been obtained from the pertinent committees. In section 2, we describe the challenge corpora. Section 3 details baseline experiments, metrics, and baseline results; concluding remarks are given in section 4.

2. The Three Sub-Challenges

2.1. The Elderly Emotion Sub-Challenge (ESC)

For the **ESC**, the **Ulm State of Mind in Speech-elderly (USOMS-e) corpus** is used. Due to some technical constraints, we employ a subset of the whole database consisting of 87 subjects (55 f, 32 m, age 60–95 years, mean 71.01 years, std. dev. 9.14 years) and two negative and one positive spontaneous narratives per speaker. **A** and **V** [17] were assessed by the speakers (self-assessment) after each narrative, and post festum by 4 experts (expert-assessment) on a scale from 0-10; reference values are the mean of self-assessment and (the mean of) expert-

assessments. The questions were for for **A**: "On a scale from zero (very sleepy) to ten (very excited), how activated do you feel at this moment?", and for **V**: "On a scale of zero (very bad) to ten (very good), how negative or positive do you feel at this moment?". This yields global scores, given for a longer period of time, where emotions surely fluctuate [18]; more fine-grained attentional shifts towards own emotions would, however, change the subject's perception [19]. The general principle is mood congruency; positive core affect shifts attention to positive material, negative core affect to negative material, and vice versa [20]. The stereo audio from the video recordings were converted to mono, 16 kHz, 16 bit. Segments of 5 sec each were created from the cleaned recordings in an automatic way, resulting in 7 478 chunks. We provide both manually and automatically[1] created orthographic transcriptions on narrative-level; the former is used for the linguistic baseline. To create the three-class classification task, the raw values for the scores have been mapped onto (i) (**L**)ow: 0-6, (ii) (**M**)edium: 7-8, and (iii) (**H**)igh: 9-10.

2.2. The Breathing Sub-Challenge (BSC)

For the **BSC**, we employ a subset of the **UCL Speech Breath Monitoring (UCL-SBM)** database. All recordings took place in a quiet office space. Here, we use only spontaneous speech recordings that pose a greater challenge in terms of respiratory planning [4], and recordings from one (MLT1132, ADInstruments, Castle Hill, Australia) of the two piezoelectric respiratory belts worn by the subjects. The belt is positioned approximately four centimetres below the collarbone to record chest breathing, and produces a linear voltage reading in response to changes in thoracic circumference associated with respiration. Speech was recorded via an AKG model C555L head-mounted condenser microphone at a distance of approximately three centimetres from the mouth. All signals were sampled at 40 kHz; speech was downsampled to 16 kHz and breath belts to 25 Hz in post-processing. The breath signal was further normalised by dividing each value by the maximum recorded value across the dataset. All 49 speakers (29 f, 20 m) reported English as a primary language, but ranged in regional accents (e. g., American, Irish, etc.), as well as sociolect; ages range from 18 to approximately 55 years old (mean age 24 years; std. dev. = ~10 years). From each speaker, we recorded some five minutes of spontaneous speech; they were invited to answer a series of questions about their experience of visiting or living in the city of London; however, it was up to their discretion to choose another topic to speak about. The recordings were edited at a common duration of four minutes for conformity, as well as to avoid background noise or the experimenter's instructions. Each breath belt signal is a sequence of 6 000 continuous values.

2.3. The Mask Sub-Challenge (MSC)

In the **MSC**, the **Mask Augsburg Speech Corpus (MASC)** is used. It comprises recordings of 32 German native speakers, wearing the surgical mask from Lohmann and Rauscher, type Sentinex Lite (16 f, 16 m, age from 20 to 41 years, mean age 25.6 years, std. dev. 4.5 years); the recordings took place in a sound-proof audio studio, using the large diaphragm condenser microphone C4500 BC from AKG; audio was sampled at a rate of 48 kHz with 24 bit, downsampled and converted to 16 kHz and mono/16 bit; the total duration is 10 h 9 min 14 sec. The participants performed different tasks without a mask and while wearing the mask: They answered some questions, read words known for their usage in medical operation rooms, drew a picture and talked about it, and described pictures, e. g., sport activities, families, kids, food, or locations. The task is to recognise whether the speaker was recorded while wearing a mask or not. The recordings were segmented into chunks of 1 sec duration without overlap.

[1] https://cloud.google.com/speech-to-text

Table 1: *Databases: Number of instances per class in the Train/Dev/Test splits: USOMS-e: # of narratives, per L/M/H for A/V; UCL-SBM: # of speakers; MASC: # of chunks. Test split distributions are blinded during the ongoing challenge and will be given in the final version.*

#	Train	Dev	Test	Σ
Ulm State-of-Mind in Speech-elderly (USoMS-e) corpus				
L	33/13	40/18	—	—
M	30/44	28/50	—	—
H	24/30	19/19	—	—
Σ	87/87	87/87	87/87	261/261
UCL Speech Breath Monitoring (UCL-SBM) corpus				
Σ	17	16	16	49
Mask Augsburg Speech Corpus (MASC)				
no-mask	5 353	6 666	—	—
mask	5 542	7 981	—	—
Σ	10 895	14 647	11 012	36 554

3. Experiments and Results

For all corpora, the segmented audio was converted to single-channel 16 kHz, 16 bits PCM format. Table 1 shows the number of cases for Train, Dev, and Test for the three databases; partitions were gender-balanced. For the **ESC**, in the acoustic analysis, chunks of 5 sec were processed for **A** and for **V**; later, the majority votings for **A** and for **V** are averaged for each narrative; this constitutes the baseline. In the linguistic analysis, the whole narrative was processed for **A** and for **V**, and accordingly, the mean of these two measures serves as baseline. For the regression task in the **BSC**, the number of speakers is given, and for the classification task in the **MSC**, we display the number of items (chunks of 1 sec).

3.1. Approaches

ComParE Acoustic Feature Set: The official baseline feature set is the same as has been used in the seven previous editions of the ComParE challenges, starting from 2013 [21]. It contains 6 373 static features resulting from the computation of functionals (statistics) over low-level descriptor (LLD) contours [22, 21]. A full description of the feature set can be found in [23]. For the **BSC**, preliminary experiments included frame-level extraction (40 msec hop size) of the 65 ComParE feature set low-level descriptors (LLDs), as well as their first derivation (delta), resulting in a 130 dimensional LLD feature set. However, due to the Support Vector Machine (SVM) paradigm which is used for the other ComParE features baselines, results in this case were at best .221 and .389 r for development and test, respectively. Through further evaluation, a 1 sec hop size for ComParE functionals with cubic spline interpolation to estimate the datapoints at a rate of 40 ms during training found meaningful improvements and is therefore chosen for the ComParE features baseline. We assume that the reason for performance improvement by the latter approach is that the temporal speech patterns that are informative towards the prediction of the breath signal are in general longer than the 25 Hz upper

belt signal frequency. In addition to these features provided in the baseline package, participants can also extract the according LLDs from the openSMILE configuration. Combined with a sequence classification utilising, e. g., Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs), similar as in following sections, this may show substantial improvements.

Bag-of-Audio-Words (BoAWs): These have been applied successfully for, e. g., acoustic event detection [24] and speech-based emotion recognition [25]. Audio chunks are represented as histograms of acoustic LLDs, after quantisation based on a codebook. One codebook is learnt for the 65 LLDs from the ComParE feature set, and another one for the 65 deltas of these LLDs. In Table 2, results are given for different codebook sizes. Codebook generation is done by *random sampling* from the LLDs/deltas in the training data. Each LLD/delta is assigned to the 10 audio words from the codebooks with the lowest Euclidean distance. Both BoAW representations, one from the LLDs and one from their deltas, are concatenated. Finally, a logarithmic term frequency weighting is applied to compress the numeric range of the histograms. LLDs are extracted with the openSMILE toolkit, BoAW are computed using openXBOW [26]. As with the ComParE acoustic features, for the **BSC**, we extract BoAWs with a frame size of 1 second, and apply interpolation at a rate of 40 ms during training.

Deep Spectrum: The feature extraction Deep Spectrum toolkit[2] is applied to obtain first deep representations from the input audio data utilising pre-trained convolutional neural networks (CNNs) [27]. Deep Spectrum features have been shown to be effective including for speech processing [28] and sentiment analysis [29]. First, audio signals are transformed into mel-spectrogram plots using a Hanning window of width 32 ms and an overlap of 16 ms. From these, 128 Mel frequency bands are computed. The generated spectrograms are then forwarded through ResNet50 [30], a pre-trained CNN, and the activations of the 'avg_pool' layer of the network are extracted, resulting in a 2 048 dimensional Deep Spectrum feature set.

auDeep: Another feature set is obtained through unsupervised representation learning with recurrent sequence to sequence autoencoders, using the auDeep toolkit[3] [31, 32]. These, in particular, explicitly model the inherently sequential nature of audio with RNNs within the encoder and decoder networks [31, 32]. In the auDeep approach, Mel-scale spectrograms are first extracted from the raw waveforms in a data set. In order to eliminate some background noise, power levels are clipped below four different given thresholds in these spectrograms, which results in four separate sets of spectrograms per data set. Subsequently, a distinct recurrent sequence to sequence autoencoder is trained on each of these sets of spectrograms in an unsupervised way, i. e., without any label information. The learnt representations of a spectrogram are then extracted as feature vectors for the corresponding instance. Finally, these feature vectors are concatenated to obtain the final feature vector. For the results shown in Table 2, the autoencoders' hyperparameters were not optimised.

Linguistic Modelling: It is well known that **V** cannot be optimally modelled by acoustic features only; both semantic denotations of lexemes (e. g., negations) and connotations of words and phrases are important additional information [33]. To this aim, we developed a lightweight LinguistIc Feature Extractor (LiFE) pipeline to extract and train linguistic features for USOMS-e[4]. Transformer language embeddings, such as BERT [34], recently showed tremendous success over a wide range of Natural Language Processing tasks. At first, our pipeline utilises a frozen German BERT model to extract a 768-dimensional context embedding vector for each word of a story. The sequence of encoded words is then fed into a feature compression block to encode a single feature vector for the entire story. The pipeline provides two ways to do this, either Global Maximum pooling (GMax) or bidirectional LSTM RNNs with an attention module (BLAtt), followed by two 512 dimensional Rectified Linear Unit (ReLU) and sigmoid feedforward layers. In addition, the pipeline provides the option of a German Part-Of-Speech (POS) tagging which can be used to train a part-of-speech embedding supported by an auxiliary loss. The output of this last layer is used as feature input for the reproducible SVM evaluation.

End-to-End Deep Sequence Modelling: Our End2End baseline[5] utilises a CNN to extract high-level, shift-invariant features from the raw time wave representation, and a subsequent RNN

Table 2: *Results for the three Sub-Challenges. The* ***official baselines*** *for Test are highlighted (bold and greyscale); there are* ***no*** *official baselines for Dev. C: Complexity parameter of the SVM/SVR, optimised for all from 10^{-5} to 1. N: Codebook size for Bag-of-Audio-Words (BoAW) splitting the input into two codebooks (*ComParE-*LLDs*/ComParE-*LLD-deltas) of the same given size, with 10 assignments per frame.. $ResNet50$: pre-trained CNN used for extraction of* Deep Spectrum *features. X: Threshold power levels for $S2SAE$ under which was clipped.* LiFE*: Linguistic feature extraction pipeline and SVM. End2End with hidden units N_h.* ***UAR****: Unweighted Average Recall. r: Pearson's correlation coefficient.* **E***: Elderly,* ***A/V*** *(Arousal/Valence as baseline);* **B***: Breathing;* **M***: Mask.*

	E		**B**		**M**	
	UAR [%]		r		UAR [%]	
	Dev (A/V)	Test (A/V)	Dev	Test	Dev	Test
C	**openSMILE: ComParE functionals+SVM**					
10^{-5}	39.1/33.3	47.9/33.3	.244	.442	56.8	59.8
10^{-4}	38.7/37.5	41.8/34.5	.234	.435	60.3	67.7
10^{-3}	34.1/40.1	38.7/36.9	.175	.333	62.3	67.8
10^{-2}	26.4/45.7	36.9/35.4	.081	.212	62.6	66.9
N	**openXBOW: ComParE BoAW+SVM**					
125	37.2/39.4	40.6/36.4	.185	.357	59.8	58.7
250	34.1/41.8	49.1/31.5	.201	.349	61.5	62.7
500	37.8/35.5	46.6/33.6	.209	.367	63.1	65.0
1000	30.5/46.0	42.2/31.0	.226	.366	63.6	66.1
2000	43.2/42.1	42.2/35.7	.215	.355	64.2	67.7
Network	**DeepSpectrum+SVM**					
ResNet50	37.8/36.2	**49.8**/38.9	–	–	63.4	70.8
X [dB]	**auDeep: S2SAE+SVM**					
-30	36.2/34.8	43.2/32.4	–	–	60.1	57.4
-45	37.7/28.1	43.7/33.8	–	–	61.3	60.3
-60	37.2/31.2	40.1/36.0	–	–	61.9	61.6
-75	42.4/25.6	42.2/34.3	–	–	61.6	62.2
Fused	36.8/23.3	46.6/32.0	–	–	64.4	66.6
Block	**LiFE: Transformer+SVM**					
GMax	39.6/54.2	37.9/41.3	–	–	–	–
BLAtt	40.6/49.2	44.0/**49.0**	–	–	–	–
BLAtt+POS	33.3/51.9	34.3/44.5	–	–	–	–
Fused	34.1/56.1	34.3/44.5	–	–	–	–
N_h RNN	**End2End: CNN+LSTM RNN**					
128	–	–	.498	.727	–	–
256	–	–	.507	**.731**	–	–
	Fusion of Best					
	–	49.1/38.4	–	.621	–	**71.8**

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

[3] https://github.com/auDeep/auDeep

[4] https://github.com/lstappen/USOMS-e_LiFE

[5] https://github.com/glam-imperial/ComParE2020-Breathing-End2End

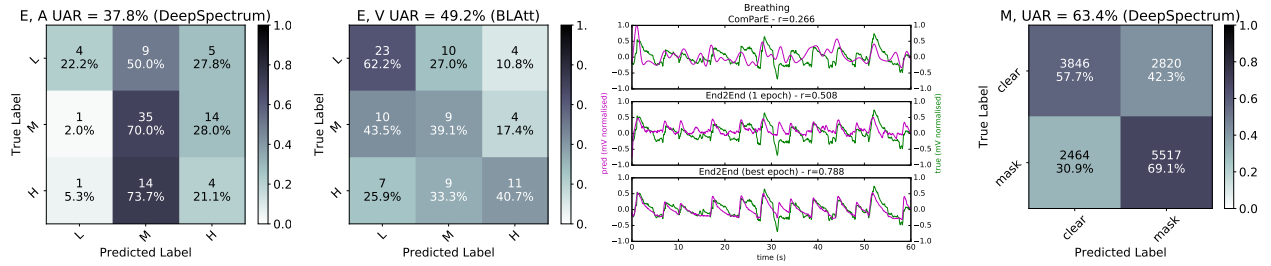

Figure 1: *Confusion matrices for **E**, **A**, and **V**, left; and **M**, right; overall number of instances per task given in Table 1. Middle: exemplary reference breath contour for **B** in green, predicted contour in mangenta; above with* ComParE*; middle with End2End, worse prediction; bottom with End2End, better prediction. For **E** and **M**, the individual approach/hyperparameters performing on Dev for the best Test result (without fusion) were chosen: for **E**, **A** Deep Spectrum+SVM, for **E**, **V** LiFE: Transformer+SVM, BLAtt; for **M** Deep Spectrum. In the cells, absolute number of cases is given, and percent of 'classified as' of the class displayed in the respective row; percentage also indicated by colour-scale: the darker, the higher.*

with LSTM cells which performs the final prediction. The former consists of three stacked one-dimensional CNN layers and the latter of two stacked LSTM-RNN layers. The number of filters and widths of the convolutional layers are 64-128-256 and 8-6-6, respectively, and each one is followed by a max pooling layer that undersamples at a stride of 10-8-8. The hidden units of the RNN layers are equal to N_h. This model provides us with a sequence of 6 000 hidden states, each of which is passed through a linear layer to provide the breath belt signal prediction. The training loss is the r calculated between the flattened true and predicted signals. Adam is used as optimiser with a learning rate of 0.002, and 100 epochs of training. Instead of early stopping, we again use the model parameters of the best Dev performance to evaluate on test. This is usually at 50 epochs – i. e., no further Dev improvement was observed after that. Such an architecture has been successful in the task of continuous-valued, sequential emotion recognition [35, 36].

3.2. Challenge Baselines and Interpretation

For the sake of transparency and reproducibility of the baseline computation, in line with previous years, we use an open-source implementation of SVMs or Support Vector Regression (SVR) with linear kernels. The provided scripts employ the scikit-learn toolkit with its classes LinearSVC and LinearSVR, respectively, for the classification based on functionals, BoAW, auDeep, and Deep Spectrum features. All feature representations were scaled to zero mean and unit standard deviation (MinMaxScaler of scikit-learn), using the parameters from the respective training set (when Train and Dev were fused for the final classifier, the parameters were calculated on this fusion). The complexity parameter C was always optimised during the development phase. For the acoustic approaches in the **ESC**, we upsampled the minority classes by a natural factor to balance the three classes in Train and Dev. Each Sub-Challenge package includes scripts that allow participants to reproduce the baselines and perform the testing in a reproducible and automatic way (including pre-processing, model training, model evaluation on Dev, and scoring by the competition and further measures). This year, we provide the six approaches outlined above. The same way as in the last three years, we chose the highest results on Test for defining the baselines, irrespective of the corresponding results on Dev, in order to prevent participants from surpassing the official baseline by simply repeating or slightly modifying other constellations that can be found in Table 2. A fusion of the best five (**E**) or four (**M**) configurations (each different approach with its best parameters) with *Majority Voting* is given in the last row; in the case of **B**, a mean calculation of the best three configurations is given due to the regression nature of the task. As can be seen in Table 2, for **E**, the baseline is **UAR** $= 49.4\,\%$ as mean over **A**, UAR $= 49.8\,\%$ (Deep Spectrum), and **V**, UAR $= 49.0\,\%$ (BLAtt); for **B**, it is Pearson's $r = .731$ (End2End); for **M**, it is **UAR** $= 71.8\,\%$ (fusion of best). Figure 1, left, displays the confusion matrices for Dev corresponding to the best result on Test, for **E** (**A** and **V**). It is reassuring that 'bad' confusions, i. e., Low with High and vice versa, are not that frequent. As expected, **V** can be modelled better with linguistics than **A** and vice versa. In the middle, for **B**, we display an identical exemplary reference contour for **B** in green, and predicted contour in magenta. It is obvious that the coarse quantisation to 1 Hz and subsequent feature interpolation do not yield optimal results, cf. Table 2; moreover, End2End is by design a sequential approach that considers the temporal proximity of samples, due to the usage of RNNs. To the right, we find the confusion matrix for **M**. For such a 2-class problem, it is difficult to tell whether the fact that mask is better predicted than without mask (clear) can be interpreted or it is simply owed to the approach.

4. Concluding Remarks

This year's challenge is new by three new tasks (elderly emotion, breathing, and speech with/without a mask), all of them highly relevant for applications. Besides the by now 'classic' approaches **ComParE** and **Bag-of-Audio-Words (BoAWs)**, we further featured sequence-to-sequence autoencoder-based audio features by the **auDeep** toolkit, **Deep Spectrum**, a LinguistIc Feature Extractor (**LiFE Transformer**) as well as **End2End Deep Sequence Modelling**. For all computation steps, scripts are provided that can, but need not be used by the participants. We expect participants to obtain better performance measures by employing novel (combinations of) procedures and features including such tailored to the particular tasks.

5. Acknowledgements

We acknowledge funding from the EU's HORIZON 2020 Grant No. 115902 (RADAR CNS), the Leverhulme Trust by Grant No. RPG-2016-251, and the EPSRC Grant No. 2021037. We thank the sponsor of the Challenge, audEERING GmbH.

6. References

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learnt from the First Challenge," *Speech Communication*, vol. 53, pp. 1062–1087, 2011.

[2] B. Schuller and A. Batliner, *Computational Paralinguistics – Emotion, Affect, and Personality in Speech and Language Processing*. Chichester, UK: Wiley, 2014.

[3] W. A. Rogers and T. L. Mitzner, "Envisioning the future for older adults: Autonomy, health, well-being, and social connectedness with technology support," *Futures*, vol. 87, pp. 133 – 139, 2017.

[4] A. Capellan and S. Fuchs, "The interplay of linguistic structure and breathing in German spontaneous speech," in *Proc. Interspeech*, Lyon, France, 2013, pp. 2014–2018.

[5] F. Goldman-Eisler, "Speech-breathing activity – a measure of tension and affect during interviews," *British Journal of Psychology*, vol. 46, no. 1, p. 53, 1955.

[6] E. Heim, P. H. Knapp, L. Vachon, G. G. Globus, and S. J. Nemetz, "Emotion, breathing and speech," *Journal of Psychosomatic Research*, vol. 12, no. 4, pp. 261–274, 1968.

[7] A. I. Gillespie, "The relationship between voice and breathing in the assessment and treatment of voice disorders," Perspectives of the ASHA Special Interest Groups, 2016, https://doi.org/10.1044/persp1.SIG3.94.

[8] J. Hlavnička, R. Čmejla, T. Tykalová, K. Šonka, E. Růžička, and J. Rusz, "Automated analysis of connected speech reveals early biomarkers of parkinson's disease in patients with rapid eye movement sleep behaviour disorder," *Scientific reports*, vol. 7, p. 12, 2017.

[9] D. Ruinskiy and Y. Lavner, "An Effective Algorithm for Automatic Detection and Exact Demarcation of Breath Sounds in Speech and Song Signals," *IEEE Transactions on Audio Speech and Language Processing*, vol. 15, pp. 838–850, 2007.

[10] V. S. Nallanthighal, A. Härmä, and H. Strik, "Deep Sensing of Breathing Signal During Conversational Speech," in *Proc. Interspeech*, Graz, Austria, 2019, pp. 4110–4114.

[11] L. L. Mendel, J. A. Gardino, and S. R. Atcherson, "Speech Understanding Using Surgical Masks: A Problem in Health Care?" *Journal of the American Academy of Audiology*, vol. 19, pp. 686–695, 2008.

[12] M. Ravanelli, A. Sosi, M. Matassoni, M. Omologo, M. Benetti, and G. Pedrotti, "Distant Talking Speech Recognition in Surgery Room: the DOMHOS Project," in *Proc. AISV*, Venice, Italy, 2013, p. 13 pages.

[13] R. Saeidi, T. Niemi, H. Karppelin, J. Pohjalainen, T. Kinnunen, and P. Alku, "Speaker Recognition For Speech Under Face Cover," in *Proc. Interspeech*, Dresden, F.R.G., 2015, pp. 1012–1016.

[14] R. Saeidi, I. Huhtakallio, and P. Alku, "Analysis of Face Mask Effect on Speaker Recognition," in *Proc. Interspeech*, San Francisco, CA, 2016, pp. 1800–1804.

[15] A. Rosenberg, "Classifying skewed data: Importance weighting to optimize average recall," in *Proc. Interspeech*, Portland, OR, 2012, pp. 2242–2245.

[16] K. Pearson, "Note on Regression and Inheritance in the Case of Two Parents," *Proc. Royal Society of London*, vol. 58, pp. 240–242, 1895.

[17] J. Russel, "Core affect and the psychological construction of emotions," *Psychological Review*, vol. 110, pp. 145–172, 2003.

[18] P. Koval and P. Kuppens, "Changing emotion dynamics: Individual differences in the effect of anticipatory social stress on emotional inertia," *Emotion*, vol. 22, pp. 256–267, 2012.

[19] P. Koval, E. A. Butler, T. Hollenstein, D. Lanteigne, and P. Kuppens, "Emotion regulation and the temporal dynamics of emotions: Effects of cognitive reappraisal and expressive suppression on emotional inertia," *Cognition and Emotion*, vol. 29, pp. 831–851, 2014.

[20] N. Shwarz and G. L. Clore, "Mood, misattribution, and judgments of well-being: Informative and directive functions of aggective states," *Journal of Personality and Social Psychology*, vol. 45, pp. 513–523, 1983.

[21] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The Interspeech 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in *Proc. Interspeech*, Lyon, France, 2013, pp. 148–152.

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

[23] F. Weninger, F. Eyben, B. Schuller, M. Mortillaro, and K. R. Scherer, "On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common," *Frontiers in Emotion Science*, vol. 4, pp. 1–12, 2013.

[24] H. Lim, M. J. Kim, and H. Kim, "Robust Sound Event Classification Using LBP-HOG Based Bag-of-Audio-Words Feature Representation," in *Proc. Interspeech*, Dresden, Germany, 2015, pp. 3325–3329.

[25] M. Schmitt, F. Ringeval, and B. Schuller, "At the Border of Acoustics and Linguistics: Bag-of-Audio-Words for the Recognition of Emotions in Speech," in *Proc. Interspeech*, San Francisco, CA, 2016, pp. 495–499.

[26] M. Schmitt and B. W. Schuller, "openXBOW – Introducing the Passau Open-Source Crossmodal Bag-of-Words Toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2017.

[27] 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, 2017, pp. 3512–3516.

[28] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, S. Pugachevskiy, and B. Schuller, "Bag-of-deep-features: Noise-robust deep feature representations for audio analysis," in *Proc. IJCNN*, Rio de Janeiro, Brazil, 2018, pp. 2419–2425.

[29] S. Amiriparian, N. Cummins, S. Ottl, M. Gerczuk, and B. Schuller, "Sentiment analysis using image-based deep spectrum features," in *Proc. ACIIW 2017*, San Antonio, TX, 2017, pp. 26–29.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[31] 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.

[32] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "auDeep: Unsupervised Learning of Representations from Audio with Deep Recurrent Neural Networks," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, 2018.

[33] S. G. Karadoğan and J. Larsen, "Combining Semantic and Acoustic Features for Valence and Arousal Recognition in Speech," *3rd International Workshop on Cognitive Information Processing (CIP), Baiona, Spain*, pp. 1–6, 2012.

[34] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[35] 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.

[36] P. Tzirakis, S. Zafeiriou, and B. W. Schuller, "End2you–the imperial toolkit for multimodal profiling by end-to-end learning," *arXiv preprint arXiv:1802.01115*, 2018.