Evaluation of the Pain Level from Speech: Introducing a Novel Pain Database and Benchmarks

Zhao Ren[1], *Nicholas Cummins*[1], *Jing Han*[1], *Sebastian Schnieder*[2,3], *Jarek Krajewski*[3,4], *Björn Schuller*[1,5]

[1] ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2] Industrial Psychology, Hochschule für Medien, Kommunikation und Wirtschaft Berlin, Germany
[3] Institute of Safety Technology, Human Factors and Diagnostics, University of Wuppertal, Germany
[4] Engineering Psychology, Rheinische Fachhochschule Cologne, Germany
[5] GLAM – Group on Language, Audio & Music, Imperial College London, UK
Email: `zhao.ren@informatik.uni-augsburg.de`

Abstract

In many clinical settings, the evaluation of pain is achieved through a manual diagnostics procedure relying heavily on verbal descriptions from the patient. Such procedures can be time-consuming, costly, liable to subjective biases and therefore often inaccurate. The automatic evaluation of pain based on paralinguistic speech cues has the potential to enable objective methodologies for improving the objectivity and accuracy of pain diagnosis. In this regard, we herein introduce a novel audiovisual pain database, the Duesseldorf Acute Pain Corpus, in which 844 recordings were collected from 80 subjects whose speech was collected while they undertook a cold pressor pain induction paradigm. The database is split into speaker independent training/development/test sets for a three-class level of pain classification task and we provide a comprehensive set of benchmark experimental results. The feature representations tested include functionals and bag-of-audio-words from three feature sets: the Computational Paralinguistics Challenge (ComParE) features, mel-frequency ceptral coefficients, and deep spectrum representations. We use support vector machines and long short-term memory recurrent neural networks (LSTM-RNN) as the classifiers. The best result, 42.7 % unweighted average recall on the test set, is obtained by LSTM-RNN working on the deep spectrum representations.

1 Introduction

Pain, a neural perception within the human brain, is a highly important reaction from an individual in terms of their psychological and physical health [1, 2]. In clinical practice, pain analysis plays a valued auxiliary role in the diagnosis of many health conditions. For example, the diagnosis of certain pathologies, as well as assessing their severities, can benefit from pain information, such as cancer [3] and Alzheimer [4]. Typical pain information solicited during clinical examinations includes pain location, type, time, length, and level [5]. Notwithstanding the importance of pain analysis, clinical examinations rely heavily on pain self-reports like questionnaires or drawings from patients such as the *Numerical Rating Scale* and the *Visual Analogue Scale* [6].

A particular disadvantage of conventional methods to evaluate of pain level is subjectivity, especially in relation to age and gender. This can potentially introducing biases, from either the patients or the clinician, during the pain evaluation procedure. In this context, the automatic and objective evaluation of pain levels aims at creating a more unified, comprehensive and efficient standard for the measurement of pain, thus enabling more accurate and efficient diagnostic procedures. Systems for the automatic detection of pain have started to be proposed and developed in the relevant literature. In particular, these approaches are based on facial expressions [7–9], body gestures, or motion descriptors [10, 11]. In addition, the voice is also an important parameter to evaluate pain, as it contains considerable information which is not only related to the physiological health like the cardiovascular system [12], but also the mental health such as depression [13].

While a large number of research works have been undertaken on various voice pathology detection paradigms, e. g., glottal cancer, laryngeal disorders, vocal cord paralysis, etc. [14], to the best of the authors' knowledge, there is very little research on the evaluation of pain level from the voice. Recently, a database of the speech for pain detection was collected and analysed [15]. However, the practicality of applying state-of-the-art deep learning topologies is limited by the small size of this corpus; 400 short samples from 27 participants. The newly collected *Duesseldorf Acute Pain (DAP) Corpus*, on the other hand, contains twice as many samples, as well as detailed annotations relating to the speaker's level of pain. The database contains 844 samples recorded from 80 subjects, with a total length of approximately 3 hours. The participants performed different read and free-form speech tasks, during which pain was induced using a cold pressor test.

Our initial analysis on the DAP Corpus is a three-class (mild/moderate/severe) level-of-pain classification task, in which we set benchmark accuracies using three popular computational paralinguistic paradigms. In particular, we use speech representations in the form of the INTERSPEECH COMPUTATIONAL PARALINGUISTICS CHALLENGE (COMPARE) features [16], Mel-Frequency Cepstral Coefficients (MFCCs), and image-based DEEP SPECTRUM features learnt via transfer learning from Image Neural Networks (ImagNet) [17]. In addition, Bag-of-Audio-Words (BoAW) features [18] are utilised on the aforementioned low-level features as well. Given the links between pain and emotion [2], these are feature representations are chosen as they have frequently been shown to capture emotions in speech [19–22]. For classification, we consider static Support Vector Machines (SVM), and temporal contextual Long Short-term Memory Recurrent Neural Networks (LSTM–RNN), both of which have also been successfully employed in speech emotion recognition [20–22].

The remainder of this paper is structured as follows: the DAP Corpus is introduced in Section 2; the experiments, key settings and results are then presented in Section 3; and finally, our conclusions and future work directions are given in Section 4.

ISBN 978-3-8007-4767-2

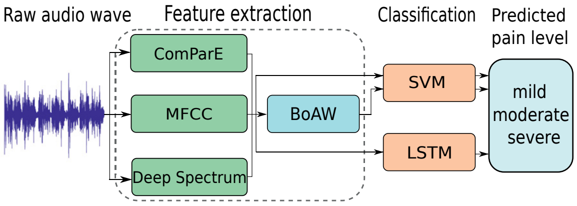


Figure 1: The pipeline of the approaches used in the benchmarks. The COMPARE, MFCC, and DEEP SPECTRUM features and the BoAW derived from these are fed into SVMs or LSTM–RNNs for classification. The audio wave is from the file 'dev_0.wav'.

2 Pain Data Collection

2.1 Participants

As already mentioned, the DAP corpus consists of speech recorded from 80 participants (41 m, 39 f), with a total of 844 audio recordings. Their ages vary from 19 to 64 years old – on average, 35.3 years, with a standard deviation of 14.9 years. For participation in the study, the following inclusion criteria was used: age between 18 and 70 years. the exclusion criteria used included: psychotropic drugs, beta-blockers or analgesic medication (24 h before measurements), and decreased vascular perfusion. All of the subjects gave their written as well as oral consent to participate. They were informed that they could discontinue the study whenever they wanted and without giving any reason for their decision. A cold pressor test was used as the pain stimulus source. The left hand was immersed up to the wrist in ice-chilled water (0.5–1.5°C). The water tub (2.8 l) was shaken manually by the experimenter every 30 s to prevent the water from warming up around the skin.

2.2 Collection Paradigm

All recordings were made in quiet rooms with a microphone/headset/hardware setup, the tasks were presented on a computer in front of the participants. Audio files were recorded with a 44.1 kHz sample rate, and down-sampled to 16 kHz with a quantisation of 16 bit. The speech material consisted of different reading passages and speaking tasks while performing an experimental pain induction procedure (The Cold Pressor Test, duration 15 minutes). The participants were asked to read aloud sentences regarding voice commands in German as used for driver assistance systems (How can I reach the fastest the Czech embassy in the Perle-Baumgaertner Street?) ["Wie komme ich am zügigsten zur tschechischen Botschaft auf der Perle-Baumgärtner-Straße?"] and a German short story "The North Wind and the Sun" (widely used within phonetics, split into 2 parts of nearly equal length). Furthermore, spontaneous dialogue speech was elicited by asking subjects to book a doctor's appointment. Within each experimental pain induction session of about 15 minutes, the voice commands were repeated 5 times, the short story and doctor appointment twice.

2.3 Annotations

To annotate the data, participants rated their level of pain on the clinically reliable and valid 11-point *Numeric Pain Rating Scale* (NPRS [23]). The NPRS was used to capture the subjects' level of pain. The scale is anchored with the phrase "0 = no pain" and "10 = worst imaginable pain".

2.4 Sample Processing

The length of all recorded samples taken together, is approximately 3 hours in total. The time length of the individual clips varies from 3.5 s to 66.9 s with 12.8 s on average. Details of the recordings are described in Table 1. The database is split into three speaker independent datasets (train/development/test), considering the balance of gender and age among each. The three classes are consistent intervals according to the pain level: 1) mild: 0–2; 2) moderate: 3–5; 3) severe: 6–10.

3 Experiments and Results

In this section, the feature sets, experimental set-up and results are given. The feature sets include COMPARE, MFCC, and DEEP SPECTRUM representations. In addition, BoAW features are extracted from the three feature sets accordingly. Both SVM and LSTM–RNN classifiers are used in this work. An overview of the experimental framework is given in Figure 1.

3.1 Feature Representations

3.1.1 The COMPARE Acoustic Feature Set

The COMPARE feature set, as the official baseline features employed in the INTERSPEECH COMPARE challenges [16], is considered in this work. Through the computation of various functionals over spectral and prosodic low-level descriptor (LLD) contours, a 6 373 dimensional feature vector is generated for each audio instance when using this feature set. We extract these features using the open source OPENSMILE toolbox and the scripts provided as part of the COMPARE 2016 challenge [16]. Full details of the feature set and OPENSMILE can be found in [24]. It is worth noting, given the aforementioned links between pain and emotion [2], the (considerably smaller) extended Geneva Minimalistic Acoustic Parameter Set (EGEMAPS) [25] was also initially considered in this work. However, preliminary testing indicated that COMPARE were better suited to the task of pain recognition.

3.1.2 MFCCs Acoustic Feature Set

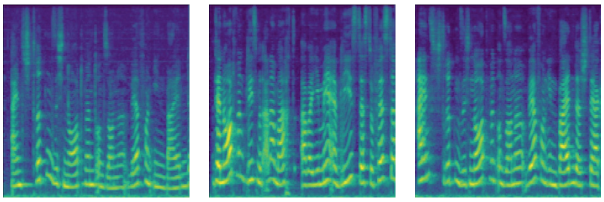
MFCCs are employed in this work due to their success in speech emotion recognition [20] and in tasks related to pain detection, such as infant cry [26] analysis. We use the adapted version of the OPENSMILE toolkit from the COMPARE 2016 challenge to extract the features; first, MFCCs 1–14 are extracted as LLDs, and then 100 functionals are generated from each LLD, according to the definitions provided in the 2016 Script [16]. Thus, in total, $1\,400 = 14 \times 100$ MFCCs features are yielded per audio instance.

3.1.3 DEEP SPECTRUM Features

DEEP SPECTRUM features, generated from time-frequency images – spectrograms or scalograms – of audio instances by pre-trained image neural networks, have been successfully applied for a range of acoustic and speech classification tasks [21, 27–29]. In this work, the DEEP SPECTRUM representations are obtained by processing mel-spectrogram images of the audio instances through the VGG16 neural network [30]. VGG16 is a Convolutional Neural Network trained for the 1 000-class image classification task proposed in the 2015 ImageNet Large Scale Visual Recogni-

Table 1: An overview of partitioning of the DAP Corpus. The speech recordings have been divided into three classes according to the Numeric Pain Rating Scale (NPRS): mild (0–2); moderate (3–5); and, severe (6–10).

Dataset	*Total*	*mild*	*mod.*	*severe*	*Durations(s)*			
					Total	*Min*	*Max*	*Avg*
Training	526	320	127	79	6704.4	3.5	66.9	12.7
Development	163	95	42	26	2062.7	3.7	39.7	12.7
Test	155	108	25	22	2018.0	3.8	36.0	13.0
Total	844	523	194	127	10785.0	3.5	66.9	12.8



(a) mild (b) moderate (c) severe

Figure 2: The mel-spectrogram images of speech samples with different labels (mild/moderate/severe). The images are extracted from the first 3.5 s segment of the audio files recorded from the same person: (a) 'train_250.wav', (b) 'train_53.wav', (c) 'train_337.wav'.

tion Competition (ILSVRC) [17]. We obtained the parameters of VGG16 from Pytorch[1].

To extract the Deep Spectrum features, firstly, the audio clips are segmented into chunks whose length and overlap depend on the classifier being used (cf. Section 3.2.1 and Section 3.2.2 for details), then, the mel-spectrogram image for each chunk is generated using 128 mel filterbanks and plotted with the *viridis* colour map. The images are then resized to 224 × 224 pixels to be compatible with VGG16. The rescaled images are then processed by VGG16, which is constructed from 13 ([2, 2, 3, 3, 3]) convolutional layers, five maxpooling layers, three fully connected layers, and a soft-max layer. The Deep Spectrum representations are finally created from the activations of the first fully connected layer.

3.1.4 Bag-of-Audio-Words

BoAW extracts sparse histograms of occurrences as audio representations; each LLD vector in a given audio clip is assigned to an audio word from a codebook learnt from some training data and the histograms are created by counting the number of assignments for each audio word. BoAW has been applied for a series of acoustic tasks, notably for speech emotion recognition where it has achieved state-of-the-art accuracies [22]. We create BoAW features for each of the aforementioned data partitions using the openXBOW toolbox [18], applying an adapted version of the baseline script from the 2017 INTERSPEECH ComParE Challenge [31].

For the ComParE features, we quantise 65 extracted LLDs (see [24] for details) and similarly, for the MFCC, we quantise the 14 extracted LLDs. Noting that there is no LDD representations of the Deep Spectrum features and we extract multiple representations per utterance by cutting each clip into non-overlapping chunks of 3.5 seconds, we then quantised the Deep Spectrum features from each chunk. For all features, the codebook generation was done by openXBOW default *random sampling* setting, and we conducted an iterative search to identify the optimal *codebook size* ($Cs \in \{125, 250, 500, 1\,k, 2\,k\}$), with the number of assignments consistently set to 10.

3.2 Classification Set-up

In this sub-section, we introduce the experimental setup for the two classifiers.

3.2.1 Support Vector Machines

Due to the well established ability of SVMs to generalise, especially in relation to small and unbalanced dataset, they are commonly used to set the benchmark accuracies e. g., [16, 31]. In this study, SVM is also used as our benchmark classifier. Both the functionals and BoAW feature set of ComParE, MFCC, as well as the original and bagged Deep Spectrum representations are fed into the SVM models (cf Figure 1). For the actual implementation we use the WEKA toolkit[2], using linear kernels and tuning the complexity parameter $C \in [10^{-6}; 10^{-1}]$ on the development partition.

Note that, for SVM classification we extract multiple Deep Spectrum representations by cutting each clip into non-overlapping chunks of 3.5 s (as per the BoAW features). We then use Margin Sampling Value (MSV) for late-fusion to generate a single prediction per clip. MSV is designed to identify the most confident predicted label, which has the highest difference between the first and second highest probabilities of predictions for one corresponding sample [32]. We have previously successfully used MSV for fusing multiple models in a similar experimental set-up for an acoustic scene classification task [27].

3.2.2 Long Short-Term Memory Recurrent Neural Networks

LSTM neural networks, as a particular type of RNN that facilitate learning over many time steps, have already been widely used for a manifold applications, such as speech emotion recognition [33], voice conversion [34], and speech synthesis [35]. An LSTM is constructed by a forget gate, an input gate, and an output gate that allow each neuron to learn when data can enter, leave or be deleted memory during system training.

In our experiments, we vary the number of layers (M) for the LSTM–RNNs: we set the number of neurons in our one layer LSTM–RNNs as 60; the numbers of neurons

[1] http://pytorch.org/

[2] https://www.cs.waikato.ac.nz/ml/weka/

Table 2: Performances comparison of speech-based pain detection approaches evaluated on the development and test sets of the DAP corpus. The experimental results are evaluated by the Unweighted Average Recall (UAR).

	ComParE		MFCC		Deep Spec.	
UAR [%]	*Dev*	*Test*	*Dev*	*Test*	*Dev*	*Test*
C	**Functionals + SVM**					
10^{-6}	47.3	**42.0**	39.9	41.0	37.0	33.7
10^{-5}	45.2	38.3	39.9	41.0	37.6	31.8
10^{-4}	40.3	39.1	39.2	39.0	32.9	32.6
10^{-3}	42.2	36.8	39.0	38.5	31.3	32.5
10^{-2}	40.8	34.5	35.3	35.1	31.6	30.9
10^{-1}	40.8	34.5	33.5	36.7	31.1	34.7
Cs	**BoAW + SVM**					
125	46.8	35.2	39.6	30.7	44.0	35.7
250	47.9	38.9	40.6	40.0	41.7	28.6
500	46.0	33.8	45.7	38.4	40.8	30.5
1000	43.9	39.4	43.3	38.8	40.0	35.5
2000	50.3	33.0	40.2	**41.3**	—	—
M	**LSTM–RNNs**					
1	36.1	31.7	38.5	31.2	38.6	38.4
2	39.3	36.9	39.1	34.8	40.0	**42.7**
3	34.5	31.0	39.1	33.6	36.9	37.3

in our two layers LSTM–RNNs as 120–60; and the numbers of neurons in our three layers are 480–120–60. The LSTM–RNNs are then followed by a highway network layer and a softmax layer for classification; highway networks have been shown to perform better than fully connected layers for very deep neural networks [36].

To generate a suitable sequential input for the LSTM–RNNs, all audio clips are segmented into smaller chunks with length of 0.6 s and overlap of 0.3 s before feature extraction. Then, the features with time steps are fed into the different LSTM–RNNs topologies. The predicted label of the last 'frame' (small audio chunk with the length of 0.6 s) is the output of the neural networks. The final predicted label is chosen from the results of multiple audio clips by the MSV strategy as well as mentioned in Section 3.2.1.

The LSTM–RNNs are realised by the deep learning library TFLearn[3]. The learning rate is 0.0001, and the batch size is set to 128. The results in Table 2 are the strongest among the iteration number *epoch* $\in [30;90]$.

3.3 Results and Analysis

All experimental results are given in Table 2. Performance is evaluated using the Unweighted Average Recall (UAR) rather than the Weighted Average Recall as the pain dataset is highly unbalanced. Note that the reported 'BoAW + SVM' results for each parameter value of Cs is the strongest from a search on a set of C values ($C \in [10^{-6};10^{-1}]$) on the development set. Further, no result is given for the set-up Deep Spectrum and BOAW with codebook size $Cs = 2\,000$. This is because the number of training samples was less than 2 000.

The results indicate that both the ComParE and MFCC feature sets perform stronger when combined with an SVM.

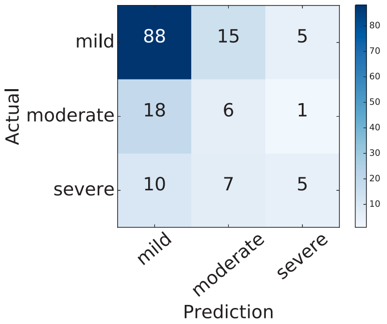


Figure 3: Confusion matrix of the best performance of 42.7% on the test set. The result is obtained by the LSTM–RNNs from the Deep Spectrum features.

The Deep Spectrum feature set, on the other hand, performs better than the ComParE and MFCC features when using LSTM–RNNs. However, z-tests conducted on the strongest performing systems (as marked in bold in Table 2) indicated this was not at a significant level. The best result of 42.7 % UAR on the test set is obtained on the Deep Spectrum under LSTM–RNNs with two hidden layers. The confusion matrix of this result is given in Figure 3. This plot indicates that the *mild* class is the easiest to recognise. However, the *severe* class is difficult to be classify correctly; we speculate this is due to the unbalanced nature of the dataset.

4 Conclusions

The misdiagnosis of pain levels through associated subjective biases can increase the unnecessary costs and risks associated with superfluous medical treatments. The monitoring of behavioural signals such as speech can potentially aid pain analysis by providing more objective evidence. In this regard, the Duesseldorf Acute Pain Corpus, a novel pain speech database collected as 844 recordings from 80 subjects was introduced. We split the corpus into training/development/test partitions and performed a set of three-class level-of-pain classification tasks. The strongest result of 42.7 % UAR was achieved using Deep Spectrum features combined with a LSTM–RNN classifier.

In future work, the generation of adaptive 'self-shaping' Deep Spectrum features will be investigated using evolutionary learning, and the data augmentation will be developed to improve the performance of classification. We also plan to annotate the data for emotions and explore multitask classification paradigms.

5 Acknowledgements

This work was partially supported by the Horizon H2020 Marie Skłodowska-Curie Actions Initial Training Network European Training Network (MSCA-ITN-ETN) project under grant agreement No. 766287 (TAPAS), and the European Union's Seventh Framework under grant agreement No. 338164 (ERC StG iHEARu).

[3] http://tflearn.org/

References

[1] A. Apkarian, M. Bushnell, R. Treede, and J. Zubieta, "Human brain mechanisms of pain perception and regulation in health and disease," *Journal of Pain*, vol. 9, pp. 463–484, Jan. 2005.

[2] A. Craig, "A new view of pain as a homeostatic emotion," *Trends in Neurosciences*, vol. 26, pp. 303–307, June 2003.

[3] P. Mantyh, "Cancer pain and its impact on diagnosis, survival and quality of life," *Nature Reviews Neuroscience*, vol. 7, pp. 797–809, Oct. 2006.

[4] E. Scherder, A. Bouma, M. Borkent, and O. Rahman, "Alzheimer patients report less pain intensity and pain affect than non-demented elderly," *Psychiatry*, vol. 62, pp. 265–272, Fall 1999.

[5] S. Arner and B. Meyerson, "Lack of analgesic effect of opioids on neuropathic and idiopathic forms of pain," *Pain*, vol. 33, pp. 11–23, Apr. 1988.

[6] J. Lee, M. Lee, J. Kim, H. Kim, S. Park, J. Tae, and S. Choi, "Pain relief scale is more highly correlated with numerical rating scale than with visual analogue scale in chronic pain patients.," *Pain Physician*, vol. 18, pp. E195–200, Mar. 2015.

[7] F. Tsai, Y. Hsu, W. Chen, Y. Weng, C. Ng, and C. Lee, "Toward development and evaluation of pain level-rating scale for emergency triage based on vocal characteristics and facial expressions," in *Proc. INTERSPEECH*, (San Francisco, CA), pp. 92–96, 2016.

[8] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. Matthews, "Painful data: The UNBC-Mcmaster shoulder pain expression archive database," in *Proc. AFGR*, (Santa Barbara, CA), pp. 57–64, 2011.

[9] J. Kappesser and A. C. de C. Williams, "Pain and negative emotions in the face: Judgements by health care professionals," *Pain*, vol. 99, pp. 197–206, Sep. 2002.

[10] M. S. Aung *et al.*, "The automatic detection of chronic pain-related expression: Requirements, challenges and the multimodal EmoPain dataset," *Transactions on Affective Computing*, vol. 7, pp. 435–451, July 2015.

[11] T. A. Olugbade, M. Aung, N. Bianchi-Berthouze, N. Marquardt, and A. Williams, "Bi-modal detection of painful reaching for chronic pain rehabilitation systems," in *Proc. ICMI*, (Istanbul, Turkey), pp. 455–458, 2014.

[12] B. Schuller, F. Friedmann, and F. Eyben, "Automatic recognition of physiological parameters in the human voice: Heart rate and skin conductance," in *Proc. ICASSP*, (Vancouver, Canada), pp. 7219–7223, 2013.

[13] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, July 2015.

[14] J. Rafael Orozco Arroyave, J. Francisco Vargas Bonilla, and E. Delgado Trejos, "Acoustic analysis and non linear dynamics applied to voice pathology detection: A review," *Recent Patents on Signal Processing*, vol. 2, pp. 96–107, Sep. 2012.

[15] Y. Oshrat, A. Bloch, A. Lerner, A. Cohen, M. Avigal, and G. Zeilig, "Speech prosody as a biosignal for physical pain detection," in *Proc. SP*, (Boston, MA), pp. 420–424, 2016.

[16] B. Schuller *et al.*, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. INTERSPEECH*, (San Francisco, CA), pp. 2001–2005, 2016.

[17] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Journal of Computer Vision*, vol. 115, pp. 211–252, Dec. 2015.

[18] M. Schmitt and B. Schuller, "openXBOW – introducing the Passau open-source crossmodal bag-of-words toolkit," *Journal of Machine Learning Research*, vol. 18, pp. 1–5, Oct. 2017.

[19] W. Zheng, M. Xin, X. Wang, and B. Wang, "A novel speech emotion recognition method via incomplete sparse least square regression," *Signal Processing Letters*, vol. 21, pp. 569–572, May 2014.

[20] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *Journal of Smart Home*, vol. 6, pp. 101–108, Apr. 2012.

[21] N. Cummins, S. Amiriparian, G. Hagerer, A. Batliner, S. Steidl, and B. Schuller, "An image-based deep spectrum feature representation for the recognition of emotional speech," in *Proc. ACM Multimedia*, (Mountain View, CA), pp. 478–484, 2017.

[22] 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), pp. 495–499, 2016.

[23] W. Downie, P. Leatham, V. Rhind, V. Wright, J. Branco, and J. Anderson, "Studies with pain rating scales," *Annals of the rheumatic diseases*, vol. 37, pp. 378–381, August 1978.

[24] 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), pp. 835–838, 2013.

[25] F. Eyben, K. Scherer, B. Schuller, *et al.*, "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *Transactions on Affective Computing*, vol. 7, pp. 190–202, Apr. 2016.

[26] R. Vempada, S. Kumar B., and K. Rao, "Characterization of infant cries using spectral and prosodic features," in *Proc. NCC*, (Kharagpur, India), pp. 1–5, 2012.

[27] Z. Ren, V. Pandit, K. Qian, Z. Yang, Z. Zhang, and B. Schuller, "Deep sequential image features on acoustic scene classification," in *Proc. DCASE*, (Munich, Germany), pp. 113–117, 2017.

[28] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proc. INTERSPEECH*, (Stockholm, Sweden), pp. 3512–3516, 2017.

[29] Z. Ren, N. Cummins, V. Pandit, J. Han, K. Qian, and B. Schuller, "Learning image-based representations for heart sound classification," in *Proc. DH*, (Lyon, France), 2018. 5 pages.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, (San Diego, CA), 2015. no pagination.

[31] B. Schuller *et al.*, "The INTERSPEECH 2017 computational paralinguistics challenge: Addressee, cold & snoring," in *Proc. INTERSPEECH*, (Stockholm, Sweden), pp. 3442–3446, 2017.

[32] T. Scheffer, C. Decomain, and S. Wrobel, "Active hidden markov models for information extraction," in *Proc. IDA*, (Porto, Portugal), pp. 309–318, 2001.

[33] J. Deng, N. Cummins, J. Han, X. Xu, Z. Ren, V. Pandit, Z. Zhang, and B. Schuller, "The University of Passau open emotion recognition system for the multimodal emotion challenge," in *Proc. CCPR*, (Chengdu, China), pp. 652–666, 2016.

[34] R. Li, Z. Wu, Y. Ning, L. Sun, H. Meng, and L. Cai, "Spectro-temporal modelling with time-frequency LSTM and structured output layer for voice conversion," in *Proc. INTERSPEECH*, (Stockholm, Sweden), pp. 3409–3413, 2017.

[35] H. Zen, "Acoustic modeling in statistical parametric speech synthesis–from HMM to LSTM-RNN," in *Proc. MLSLP*, (Aizu-Wakamatsu, Japan), 2015. no pagination.

[36] R. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.