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

Speech Augmentation via Speaker-Specific Noise in Unseen Environment

Ya'nan Guo[1,2], *Ziping Zhao*[3,1,*], *Yide Ma*[2], *Björn Schuller*[1,4]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]School of Information Science and Engineering, Lanzhou University, China
[3]College of Computer and Information Engineering, Tianjin Normal University, China
[4]GLAM – Group on Language, Audio & Music, Imperial College London, UK

guoyn10@gmail.com

Abstract

Speech augmentation is a common and effective strategy to avoid overfitting and improve on the robustness of an emotion recognition model. In this paper, we investigate for the first time the intrinsic attributes in a speech signal using the multi-resolution analysis theory and the Hilbert-Huang Spectrum, with the goal of developing a robust speech augmentation approach from raw speech data. Specifically, speech decomposition in a double tree complex wavelet transform domain is realized, to obtain sub-speech signals; then, the Hilbert Spectrum using Hilbert-Huang Transform is calculated for each sub-band to capture the noise content in unseen environments with the voice restriction to 100−4000 Hz; finally, the speech-specific noise that varies with the speaker individual, scenarios, environment, and voice recording equipment, can be reconstructed from the top two high-frequency sub-bands to enhance the raw signal. Our proposed speech augmentation is demonstrated using five robust machine learning architectures based on the RAVDESS database, achieving up to 9.3 % higher accuracy compared to the performance on raw data for an emotion recognition task.

Index Terms: Emotion Recognition, Speech Augmentation, Speech decomposition, Bidirectional LSTM−Attention

1. Introduction

Data augmentation [1] is a common and widely accepted strategy to enrich the diversity of training data by artificially constructed additional training samples using various signal/data processing techniques. Increasing the quantity and enriching content of training data has been consistently demonstrated to be beneficial to prevent the overfitting of models and improve the overall robustness of automatic speech recognition models [2]. However, as audio data is more sensitive than images, an affine-transformation motivated data augmentation strategy is not suitable for audio data. Especially for emotional data such as speech, emotional expressions are particularly delicate [3], so speech data is easy to be polluted. The immoderate speech enhancement might lead to the concealment of the vital emotional information, so the fidelity and integrity is a concern for speech augmentation techniques. To date, few audio augmentation methods have been proposed, the mainstream approaches include feature-level technique using vocal tract length perturbation (VTLP) [4] and stochastic features, speed perturbation [5], voice transformation [6], noise addition [7], artificial copies [8] and some combined models [9, 10]. However, the performance of all of these algorithms always varies dramatically with the task and database, therefore, there is still large room for improvement in this field.

To counter unnecessary interference from speech augmentation via noise addition, we start to design the audio augmentation model artificially from original data. After a deep exploration of sub-speech signals' energy-time-frequency Hilbert spectrum via Hilbert-Huang Transform (HHT) [11], we find that each sample has very different interference that is affected by the speaker individual, scenarios, environment and further factors. We name this kind of interference as speaker-specific noise in unseen environment. To be specific, we first accomplish the speech decomposition using the Dual-Tree Complex Wavelet Transform (DT-CWT) [12], locate the inferences in each sub-speech signal, and aggregate them to reconstruct the noise-like interference with the constraint of typical human voice frequency range approximately in 100-4000 Hz [13]. The final augmented speech can be expressed as the superposition of the raw one and the noise-like interference. Ultimately, the proposed speech augmentation algorithm is respectively tested on five speech emotion recognition models using context-based Mel Frequency Cepstral Coefficients (MFCCs), SliCQ-nonstationary Gabor transformation (SliCQ-NSGT) and ComParE features representations, yielding 9.3 % accuracy improvement at most comparing with the result on original data.

This paper is organized as follows. Section 2 introduces the proposed method, Section 3 describes the experimental results and discuss, conclusions are presented in Section 4.

2. Methodology

In this section, we first introduce the speech decomposition method, analyse what kind of interference exists in speech, then propose the speech argumentation method; finally, the acoustic representations and adopted models are introduced.

2.1. Speech decomposition

To dig into the intrinsic attributes in speech, we decompose the original speech into several sub-speech signals using DT-CWT, which is demonstrated to be efficient because of its shift-variance, low-directional selectivity in high dimensions as well as perfect-reconstruction characteristics [12]. In fact, DT-CWT is a specific case of CWT proposed by Kingsbury in 1998, which can be realised by different low-pass and high-pass filters based on two parallel Discrete Wavelet Transformations (DWT) [14]. The audio decomposition process of DT-CWT is exhibited in Figure 1. Seen from decomposed results, it can lead to twice the number of DWT wavelet coefficients and these wavelet coefficients are almost shift invariant, so a small change on the input signal cannot change the distribution of the energy of DT-CWT coefficients at different scales [15], hence, it holds

[*] Corresponding author: zhaoziping@tjnu.edu.cn

http://dx.doi.org/10.21437/Interspeech.2019-2712

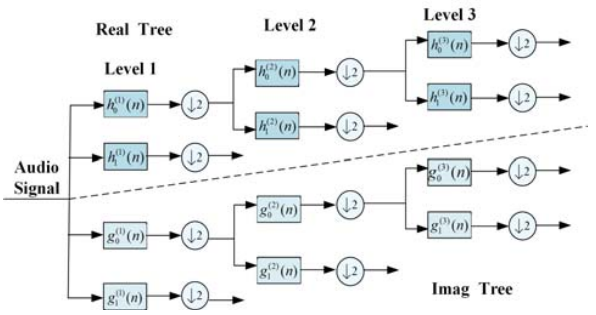


Figure 1: *DT-CWT of an audio signal.* $h_0(n)$ *and* $h_1(n)$ *indicate the low-pass and high-pass filter pair for the upper filter bank, respectively;* $g_0(n)$ *and* $g_1(n)$ *denote the low-pass and high-pass filter pair for the lower filter bank, respectively.*

promises for audio signal analysis. The wavelet function of DT-CWT can be expressed as:

$$\psi(t) = \psi_h(t) + i\psi_g z(t) \ , \tag{1}$$

where $\psi_h(t)$ and $\psi_g z(t)$ denote two real wavelets, i denotes the complex unit, and $\psi_h(t)$ and $\psi_g z(t)$ are a pair of Hilbert transforms. This complex function guarantees the better performance of DT-CWT for non-stationary signals, especially for speech.

In this work, 2-level DT-CWT decomposition is first carried out to obtain three sub-speech signals with the frequencies ranging from low to high, letting the speech information uniformly distribute at different sub-bands, and then, human voice and other disturbances can be distinguished from them according to human voice range. The subsequent speech enhancement is proposed based on the in-depth analysis of content in each sub-speech signal.

2.2. Speech Augmentation

The exploration of intrinsic properties in speech is based on the human voice frequency and energy-time-frequency Hilbert spectrum via HHT. It is generally accepted that 100-4000 Hz is usually referred to as the "voice frequency" (VF). So here, we think of VF as a reference to position the real voice and inferences in each sub-speech signal. The Hilbert spectrum based on HHT has strong signal analysis capability realised by empirical mode decomposition [16] and Hilbert spectral analysis, especially for non-stationary and non-linear data. Moreover, HHT is able to resolve frequency accurately and time it precisely without Heisenberg uncertainty [11], so here, HHT-based Hilbert spectrum is calculated to track the instantaneous variation in frequency distribution for all the sub-bands and the original data.

Figure 2 illustrates the waveform of the original and decomposed sub-speech signals, and their Hilbert spectrum for 2-level DT-CWT speech decomposition results. Here, 'Sub-speech1', 'Sub-speech2' and 'Sub-speech3' separately denote the decomposed sub-bands speech with frequency from high to low. Red line represents the upper boundary of VF, namely $0.4 * 10^4$HZ, and the green line marks the lower one, namely $0.01 * 10^4$HZ. Visualising from Hilbert spectrum, we can find that almost all the voice-related information can be located in the low-frequency sub-band, namely, 'Sub-speech3'; 'Sub-speech2' and 'Sub-speech3' with the high frequency mainly store the interference information that is related to the unseen environment, so it is a good implication to separate the noise-like interference from the raw one. However, simply removing these high-frequency information might lead to information loss or incomplete emotional expression. After all, not all the noises are unrelated to emotion recognition. In contrast, noisy data can enhance the robustness of the model. Motivated by the speech augmentation of noise addition, we start employing the high-frequency sub-bands in the DT-CWT domain to reestablish the noise that is affected by the unseen environment. In reality, considering that each speech is recorded from different speakers with variou atmosphere, the reconstructed noise is a speaker-specific noise, which varies with the individual, scenarios, circumstance, and external equipment, etc. To be specific, so called speaker-specific noise can be achieved by fusing the top two high-frequency sub-speeches based on the perfect reconstruction capability of the DT-CWT, as is shown in figure 3. The final augmented speech can be defined as the superimposition of the original data and reconstructed noise.

To track the effectiveness of the proposed speech enhancement, we show the waveform and Hilbert spectrum of the speaker-specific noise, raw and augmented speech in figure 4. HHTs' better local analysing ability in both the time and frequency domain provides an excellent visualisation of emotion content and noise-inference. Analysed from their waveforms, we can realise that there is no greater difference in waveform shape for the original and enhanced speech except for some differences in the middle of the speech signal. So the proposed method can be regarded as a local enhancement algorithm, which is wise enough to use the self-noise to enhance itself, avoiding vital features to be polluted. At the same time, voice-related content is retained very well. Seen from Hilbert spectrum representation, noise can just be found in high-frequency region, and the speaker voice content is globally intact compared to the spectrum of the original data.

2.3. Representations

In our work, three robust acoustic representations are introduced to give a better description of speech emotions, namely context-based MFCCs [17], SliCQ-NSGT [12], and ComParE features [18].

Context-based MFCCs are the most stable acoustic features, which take human perception sensitivity with respect to frequencies and loudness into consideration. Existing studies have already validated its excellent representation description [17]. Another acoustic representation is calculated in the SliCQ-NSGT[1] domain. SliCQ-NSGT is the enhanced version of the constant-Q non-stationary Gabor transform, so it usually outperforms the 'classical' spectrogram for audio signals [19, 20, 21]. Here, we choose SliCQ-NSGT coefficients as acoustic representations for the emotion recognition task. Figure 5 exhibits the context-based MFCCs and SliCQ-NSGT representations. Their visual saliency guarantees the features distinctiveness of emotion expression. Finally, we use our toolkit openSMILE [22] to extract the large scale temporal and spectral *INTERSPEECH ComParE* feature set, which has been proven to be efficient on the task in [23]. Here, we set the configuration file as "*ComParE2016.conf*" for our purposes, and a total 6373 features are extracted from each input speech.

2.4. Model architectures

The aim of this work is to study the speech augmentation method, so we pay lower attention to the construction of neural networks. Here, a stable bidirectional-LSTM-attention(Bi-LSTM/A) model is designed to continue the following audio augmentation validation experiments. The structure of this

[1] http://www.univie.ac.at/nonstatgab/slicq

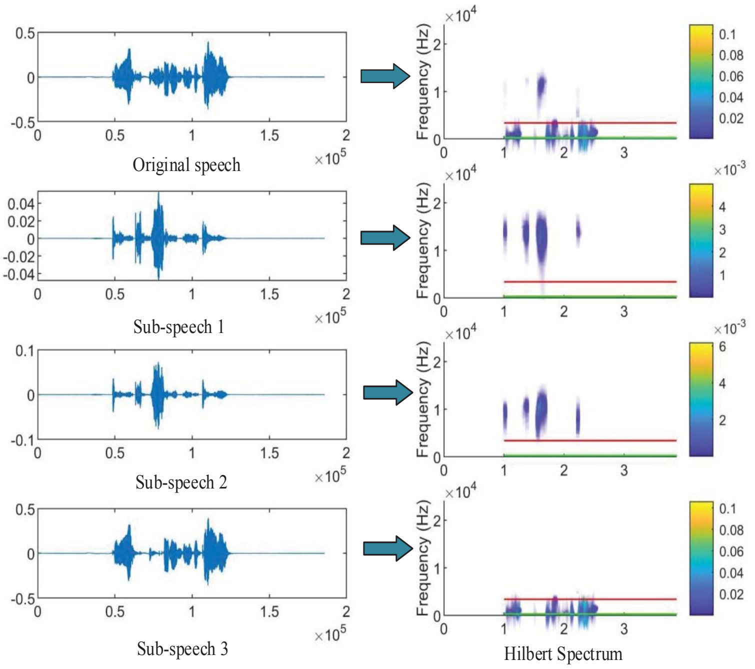


Figure 2: *Waveforms of speech decomposition and their Hilbert spectrum based on HHT: Sub-speech 1, Sub-speech 2 and Sub-speech 3 separately denote the sub-bands with frequency from high to low.*

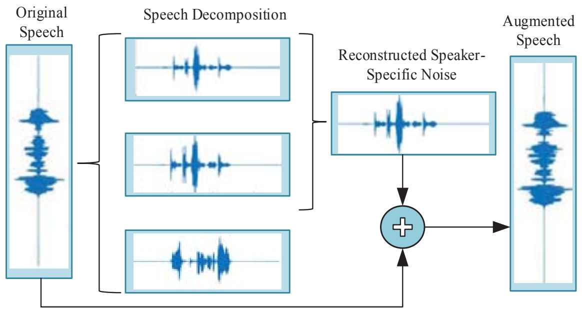


Figure 3: *Proposed speech augmentation method*

model is illustrated in Figure 6. A dropout is used to accelerate the fitting of the model. More details about this model can be found in [24, 25]. The bidirectional LSTM without an attention model is also designed as a comparison model to demonstrate the effectiveness of attention. Context-based MFCCs and SliCQ-NSGT representations of the raw and enhanced data are separately fed into the bidirectional-LSTM-attention and bidirectional-LSTM model with diverse units to explore the accuracy improvement for the emotion recognition task. Another robust combination of the *ComParE* feature set and SVMs is also employed to test the proposed method.

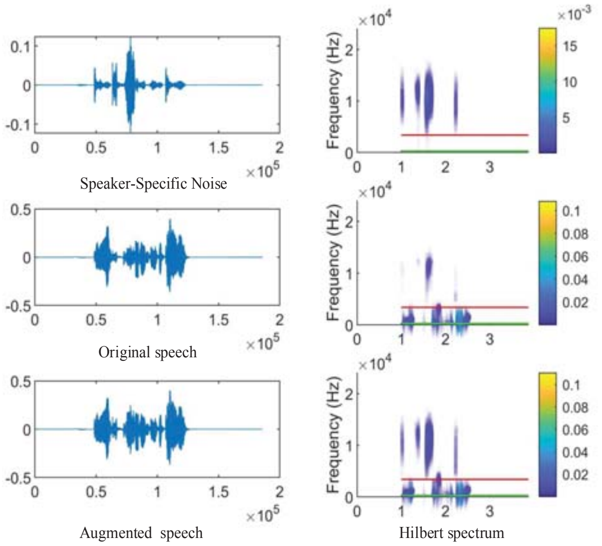


Figure 4: *Waveform and HHT Hilbert spectrum of speaker-specific noise, original and enhanced speech.*

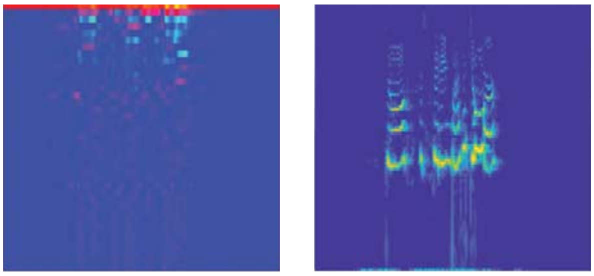
3. Experiments and Results

3.1. Database

The RAVDESS [26] speech dataset consists of 24 professional actors (12 female, 12 male), vocalising two lexically-matched statements in a neutral North American accent sentences.The speech recordings consist of 8 emotions, including neutral, calm, happy, sad, angry, fearful, disgust and surprise, Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only, Audio-Video, and Video-only. In this work, we just focus on the emotion recognition task based on Audio-only, and 1440 speech files are included.

3.2. Experimental set-up

In this work, we regard the speaker with different emotion/intensity as an another new speaker, divide the RAVDESS speech corpus as train, develop and test set according to the ratio of 4:1:1, and ensure classes balance for each set. Robust context-based MFCCs, SliCQ-NSGT and ComParE features are



(a)Context-based MFCCs (b)SliCQ-NSGT coefficients

Figure 5: *Illustration of context-based MFCCs and SliCQ-NSGT map*

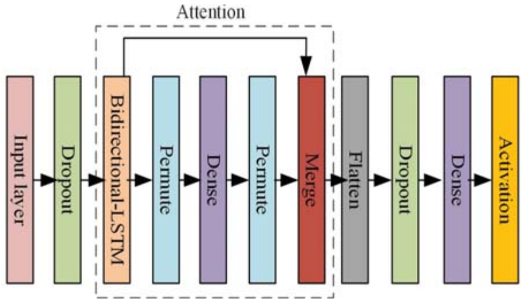


Figure 6: *Illustration of bidirectional-LSTM-attention model*

calculated on raw and enhanced data respectively. All these descriptions are fed into the Bi-LSTM-attention, Bi-LSTM and SVMs models for emotion recognition. All acoustic features are respectively standardised to zero mean and unit variance before the training phase. For all experiments, we employ batches of 60 samples with the Adadelta optimiser for batch gradient descent. Instead of standardising the number of epochs during training, we use an early stopping strategy to end training once accuracy does not improve within 10 epochs. Thus, every model is potentially trained for a different number of epochs but yields the best performance. 50 % dropout and batch-normalisation modules are used to avoid the over-fitting, Hyperbolic tangent and Sigmoid activation are used for classification results in our deep learning models. In addition, moderate ComParE features are fed into linear kernel SVMs with the complexity of $[1e-4, 1e-3, \ldots, 1e-0]$ to evaluate the model performance.

3.3. Results and discussion

Table 1 displays the accuracy of different models for emotion recognition based on raw and augmented data. Once augmented, the emotion recognition performance is evidently improved, the maximum accuracy improvements are respectively 9.5 %, 11.1 %, 2.9 %, 5.0 % and 4.6 % for *MFCC/Bi-LSTM*, *MFCC/Bi-LSTM/A*, *SliCQNSGT/Bi-LSTM*, *SliCQNSGT/Bi-LSTM/A* and *ComParE/SVM* models, hence the proposed algorithm is effective enough to enhance data, at the same time avoiding data pollution.

Besides, experimental results indicate that MFCCs-motivated models can reach greater improvement than SliCQNSGT-motivated models after data augmentation, but it seems that SliCQNSGT coefficients have stronger ability to capture the distinguished emotion features because of powerful signal analysis capability of SliCQ-NSGT, so they reach relatively higher accuracy. The introduction of an attention module brings a further accuracy gain, effectively perfecting our model for emotion recognition. Luckily, even the robust *ComParE/SVM* model also achieves 4.6 % increase in accuracy.

Table 1: *Accuracy for different models. Complexity: Units/ Complexity; '+': combination;'/A': with attention module*

Accs [%]	Accuracy =		
Model	Complexity	Original	Augmented
MFCC + Bi-LSTM	100	38.6	39.5
	200	32.1	42.6
	400	33.6	37.1
	800	30.8	36.4
MFCC + Bi-LSTM/A	100	46.0	50.8
	200	46.9	56.1
	400	41.9	53.0
	800	43.4	50.3
SliCQNSGT + Bi-LSTM	100	59.5	61.1
	200	60.2	61.7
	400	44.6	47.5
	800	35.5	38.0
SliCQNSGT + Bi-LSTM/A	100	55.1	58.3
	200	56.6	59.4
	400	57.6	58.0
	800	59.0	64.0
ComParE2016 + SVM	e-0	68.8	71.7
	e-1	69.2	73.8
	e-2	69.6	72.9
	e-3	63.8	64.2
	e-4	50.8	52.5

Figure 7 exhibits the maximum accuracy values as well as their improvement with/without data augmentation for all architectures. A remarkable accuracy increase occurs on the *MFCC/Bi-LSTM/A* model, achieving 9.3 % gain. The best performance is 73.8 %, obtained from the *ComParE/SVM* model with 4.2 % increase. In total, all these experiments demonstrate the effectiveness of our method, and it should be a promising way to ameliorate speech for the emotion recognition task.

Although we achieve greater performance improvement on several models for emotion recognition, it is just based on a single database, so more emotion databases should be tested in future work. Also, the local features in sub-bands should be payed more attention to in future acoustic analysis tasks.

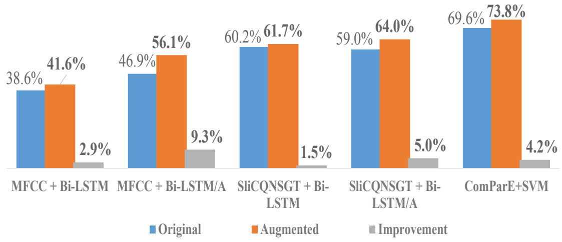


Figure 7: *Accuracy improvement for different models.*

4. Conclusions

In summary, we have performed an experimental study of speech augmentation for emotion recognition. Context-based MFCCs, SliCQ-NSGT, as well as ComParE representations were calculated to feed the different machine learning emotion-recognition architectures on original and augmented speech data. The experimental results have shown the effectiveness of the proposed method, and it should be recommended to enhance and enrich speech databases for emotion recognition tasks in the future.

5. Acknowledgements

The work presented in this paper was substantially supported by the China Scholarship Council (CSC) and the Key Program of the Natural Science Foundation of Tianjin (Grant No. 18JCZDJC36300). We thank the help from our colleagues Manuel Milling as well as Dr. Judith Dineley.

6. References

[1] Z. Zhang, N. Cummins, and B. Schuller, "Advanced data exploitation in speech analysis: An overview," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 107–129, Jul. 2017.

[2] T. Fukuda, R. Fernandez, A. Rosenberg, S. Thomas, B. Ramabhadran, A. Sorin, and G. Kurata, "Data augmentation improves recognition of foreign accented speech," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 2409–2413.

[3] Z. Zhao, Y. Zhao, Z. Bao, H. Wang, Z. Zhang, and C. Li, "Deep spectrum feature representations for speech emotion recognition," in *Proc. ASMMC-MMAC*. Seoul, Korea: ACM, 2018, pp. 27–33.

[4] N. Jaitly and G. G.E. Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. the 30th ICML Workshop on Deep Learning for Audio, Speech and Language*, vol. 117, Georgia, 2013, no pagination.

[5] W. Hartmann, T. Ng, R. Hsiao, S. Tsakalidis, and R. Schwartz, "Two-stage data augmentation for low-resourced speech recognition," in *Proc. INTERSPEECH*, San Francisco, California, USA, 2016, pp. 2378–2382.

[6] A. Sorin, S. Shechtman, and A. Rendel, "Semi parametric concatenative tts with instant voice modification capabilities," in *Proc. INTERSPEECH*, Stockholm, Sweden, 2017, pp. 1373–1377.

[7] D. Pearce and J. Picone, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," *Inst. for Signal & Inform. Process., Mississippi State Univ., Tech. Rep*, 2002, no pagination.

[8] H. Liao, "Speaker adaptation of context dependent deep neural networks," in *Proc. ICASSP*, Vancouver, Canada, 2013, pp. 7947–7951.

[9] G. Keren, J. Han, and B. Schuller, "Scaling speech enhancement in unseen environments with noise embeddings," in *Proc. Workshop on Speech Processing in Everyday Environments(CHiME 2018)*, Hyderabad, India, 2018, no pagination.

[10] G. Keren, J. Deng, J. Pohjalainen, and B. Schuller, "Convolutional neural networks with data augmentation for classifying speakers' native language." in *Proc. INTERSPEECH*, San Francisco, 2016, pp. 2393–2397.

[11] N. Huang, *Hilbert-Huang transform and its applications*. World Scientific, 2014, vol. 16, no pagination.

[12] Y. Guo, J. Han, Z. Zhang, B. Schuller, and Y. Ma, "Exploring a new method for food likability rating based on DT-CWT theory," in *Proc. the 20th International Conference on Multimodal Interaction (ICMI)*, Boulder, Colorado, USA, 2018, pp. 569–573.

[13] T. Teng, L. Sze, and O. Yeng, "Abnormal sound analytical surveillance system using microcontroller," in *Proc. the 14th IEEE Colloquium on Signal Processing and its Applications (CSPA)*, Malacca, Malaysia, 2016, pp. 162–166.

[14] Selesnick, I. W, Baraniuk, R. G, Kingsbury, and N. G, "The dual-tree complex wavelet transform," *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 123–151, Nov. 2005.

[15] X.-C. Yuan, C.-M. Pun, and C. P. Chen, "Robust mel-frequency cepstral coefficients feature detection and dual-tree complex wavelet transform for digital audio watermarking," *Information Sciences*, vol. 298, pp. 159–179, Mar. 2015.

[16] B. Dong, Z. Zhang, and B. Schuller, "Empirical mode decomposition: A data-enrichment perspective on speech emotion recognition," in *Proc. the 6th International Workshop on Emotion and Sentiment Analysis (ESA), satellite of LREC*, Portoroz, Slovenia, 2015, pp. 71–75.

[17] G. Liu, "Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech," *CoRR*, vol. abs/1806.09010, 2018, no pagination.

[18] B. Schuller, S. Steidl, and e. A. Batliner, "The INTERSPEECH 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *Proc. INTERSPEECH*, San Francisco, 2016, pp. 2001–2005.

[19] G. Velasco, N. Holighaus, M. Dörfler, and T. Grill, "Constructing an invertible constant-q transform with non-stationary gabor frames," in *Proc. the 14th International Conference on Digital Audio Effects (DAFx-11)*, Paris, France, 2011, pp. 19–23.

[20] T. Liu, S. Yan, and W. Zhang, "Time–frequency analysis of non-stationary vibration signals for deployable structures by using the constant-q nonstationary gabor transform," *Mechanical Systems and Signal Processing*, vol. 75, pp. 228–244, Jun. 2016.

[21] Y. Xiao, Y. Hong, X. Chen, and W. Chen, "The application of dual-tree complex wavelet transform (DTCWT) energy entropy in misalignment fault diagnosis of doubly-fed wind turbine (DFWT)," *Entropy*, vol. 19, no. 11, p. 587, Nov. 2017.

[22] F. Eyben, Florian, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proc. the 21st ACM International Conference on Multimedia (MM)*, Barcelona, Spain, 2013, pp. 835–838.

[23] B. Schuller, S. Steidl, and e. a. A. Batliner, "The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in *Proc. INTERSPEECH*, vol. 5, Hyderabad, India, 2018, no pagination.

[24] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proc. the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 2, Berlin, Germany, 2016, pp. 207–212.

[25] Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based Bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 272–276.

[26] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one, Public Library of Science*, vol. 13, May 2016, no pagination.