Speech Augmentation via Speaker-Specific Noise in Unseen Environment

Ya’nan Guo1,2, Ziping Zhao3,1,*, Yide Ma2, Björn Schuller1,4
1ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
2School of Information Science and Engineering, Lanzhou University, China
3College of Computer and Information Engineering, Tianjin Normal University, China
4GLAM – 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 the 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 augmentation 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 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
promises for audio signal analysis. The wavelet function of DT-CWT can be expressed as:

$$\psi(t) = \psi_h(t) + \psi_g(t), \quad (1)$$

where $\psi_h(t)$ and $\psi_g(t)$ denote two real wavelets, $i$ denotes the complex unit, and $\psi_h(t)$ and $\psi_g(t)$ 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 $0.4 \times 10^3$ HZ, and the green line marks the lower one, namely $0.01 \times 10^3$ 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 various 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. HHT’s 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-NSGT1 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

---

1http://www.univie.ac.at/nonstatgab/sliqc
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 employed to test the proposed method.
Table 1: Accuracy for different models. Complexity: Units/Complexity; ‘+’: combination; ‘/A’: with attention module

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC + Bi-LSTM</td>
<td>81.0</td>
</tr>
<tr>
<td>MFCC + Bi-LSTM/A</td>
<td>84.8</td>
</tr>
<tr>
<td>SliCQNSGT + Bi-LSTM</td>
<td>89.0</td>
</tr>
<tr>
<td>SliCQNSGT + Bi-LSTM/A</td>
<td>92.0</td>
</tr>
<tr>
<td>ComParE/SVM</td>
<td>96.0</td>
</tr>
</tbody>
</table>

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 paid more attention to in future acoustic analysis tasks.

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


