Abstract
Emotional voice conversion aims to convert the spectrum and prosody to change the emotional patterns of speech, while preserving the speaker identity and linguistic content. Many studies require parallel speech data between different emotional patterns, which is not practical in real life. Moreover, they often model the conversion of fundamental frequency (F0) with a simple linear transform. As F0 is a key aspect of intonation that is hierarchical in nature, we believe that it is more adequate to model F0 in different temporal scales by using wavelet transform. We propose a CycleGAN network to find an optimal pseudo pair from non-parallel training data by learning forward and inverse mappings simultaneously using adversarial and cycle-consistency losses. We also study the use of continuous wavelet transform (CWT) to decompose F0 into ten temporal scales that describes speech prosody at different time resolution, for effective F0 conversion. Experimental results show that our proposed framework outperforms the baselines both in objective and subjective evaluations.

Index Terms: emotional voice conversion, non-parallel data, CycleGAN, continuous wavelet transform

1. Introduction
Emotion, as an essential component of human communication, can be conveyed by various prosodic features, such as pitch, intensity, and speaking rate [1]. It plays an important role as a manifestation at semantic and pragmatic level of spoken languages. An adequate rendering of emotion in speech is critically important in expressive text-to-speech [2, 3], personalized speech synthesis, and intelligent dialogue systems, such as social robots and conversational agents.

Emotional voice conversion is a voice conversion (VC) technique for converting the emotion from the source utterance to the target utterance, while preserving the linguistic information and the speaker identity, as illustrated in Figure 1. It shares many similarities with conventional voice conversion. Both of them aim to convert non-linguistic information through mapping features from source to target. They are also different because conventional voice conversion techniques consider prosody-related features as speaker-independent. As speaker identity is thought to be characterized by the physical attributes of the speaker, which are strongly affected by the spectrum and determined by the voice quality of the individual [4], conventional VC studies mainly focus on spectrum conversion.

Figure 1: An emotional voice conversion system is trained on speech data of different emotional patterns from the same speaker. At run-time, the system takes the speech of one emotion as input, and converts to that of another [7–9]. On the other hand, emotion is inherently supra-segmental and hierarchical in nature [5, 6], that is manifested both in spectrum and prosody. Therefore, emotion cannot be handled simply at frame level, as it is insufficient to just convert the spectral features frame-by-frame.

Early studies of VC marked a success by training the spectral mapping on parallel speech data between source and target speaker [10, 11]. Many statistical approaches have been proposed in the past decades, such as Gaussian Mixture Model (GMM) [12], and Partial Least Square Regression (PLSR) [13]. Other VC methods, such as Non-negative Matrix Factorization (NMF) [14] and exemplar-based sparse representation schemes [15–17] were designed to address the over-smoothing problem in VC.

With the advent of deep learning, the performance of VC systems has been markedly improved. Neural Network (NN) -based methods, such as Restricted Boltzmann Machine (RBM) [18], Feed Forward NN [19], Deep Neural Network (DNN) [20], and Recurrent Neural Network (RNN) [21] have helped VC systems to achieve a higher level in terms of modeling the relationship between source and target features. More recently, some approaches have been proposed to eliminate the need of parallel data for VC such as Deep Bidirectional Long-Short-Term Memory (DBLSTM) with i-vector [22], variational autoencoder [23], DBLSTM with model using Phonetic Posterior-grams (PPGs) [24], and GANs [25–28]. The successful practice of these deep learning methods became the source of inspiration for this study.

The early studies on emotional VC [29, 30] only focused on prosody conversion by using a classification and regression tree to decompose the pitch contour of the source speech into a hierarchical structure, then followed by GMM and regression-based clustering methods. One attempt to handle both spectrum and prosody conversion [31–33] was the GMM-based technique [7]. Another approach is a combination of Hidden Markov Model (HMM), GMM and F0 segment selection method for
transforming F0, duration and spectrum, which was proposed in [34].
More recently, exemplar-based emotional VC approach based on NMF [8] and other NN-based models such as DNN [35],
Deep Belief Network (DBN) [36] and DBLSTM [37] were also proposed to perform spectrum and prosody mapping.

Inspired by the success of sequence-to-sequence models in text-to-speech synthesis, a sequence-to-sequence encoder-decoder based model [38] was also investigated to transform the intonation of a human voice, and can convert the emotion of neutral utterances effectively. Rule-based emotional VC approaches such as [39] are capable of controlling the degree of emotion using dimensional space such as arousal and valence.

We note that the training of most of the emotional VC systems relies on parallel training data, which is not practical in real life applications. Motivated by that, more recently, a style transfer auto-encoder [9] was proposed, which can learn from non-parallel training data. A source-target pair in non-parallel dataset represents source and target emotions. But, unlike those in parallel dataset, they can carry different linguistic content, that may make data collection much easier.

Prosody conveys linguistic, para-linguistic and various types of non-linguistic information, such as speaker identity, emotion, intention, attitude and mood. It is observed that prosody is influenced by short-term as well as long-term dependencies [40, 41]. We note that F0 is an essential prosodic factor with respect to the intonation in speech, describing the variation of the vocal pitch over different time domains, from the syllables to the entire utterance. Therefore, it should be represented with hierarchical modeling [42–44], for example, in multiple time scales. The early studies on emotional voice conversion use a Logarithm Gaussian (LG)-based linear transformation method [7–9, 29, 30] to convert F0. Such single pitch value of F0 representation does not characterize speech prosody well [5, 6, 41]. Continuous Wavelet Transform (CWT) decomposes a signal into frequency components and represent it with different temporal scales, that becomes an excellent instrument. CWT has already been applied for speech voice conversion frameworks such as DKPLS [40] and exemplar-based conversion [44, 45]. It has been also shown to be effective for emotional voice conversion such as NMF-based approach [42, 43] and DBLSTM-based approach [37]; and for emotional speech synthesis have been investigated in [46–48].

In this paper, we propose an emotional VC framework with CycleGAN that is trained on non-parallel data to map a speaker’s speech from one emotion to another. We use melspectrum to represent the acoustic features and CWT coefficients for prosodic features. Our framework does not rely on either parallel training data or any other extra modules such as speech recognition or time alignment procedures.

The main contributions of this paper include: 1) we propose a parallel-data-free emotional voice conversion framework; 2) we show the effect of prosody for emotional voice conversion; 3) we effectively convert spectral and prosodic features with CycleGAN; 4) we investigate different training strategies for spectrum and prosody conversion such as separate training and joint training; and 5) we outperform the baseline approaches, and achieve quality converted voice.

This paper is organized as follows: In Section 2, we describe the details of CycleGAN and CWT decomposition of F0. In Section 3, we explain our proposed spectrum and prosody conversion for emotional VC framework. Section 4 reports the experimental results. Conclusion is given in Section 5.

2. Related Work

2.1. CycleGAN

Recently, generative adversarial learning has become very popular in machine learning applications, such as computer vision [49–52] and speech information processing [53, 54]. In this paper, we focus on a GAN-based network called CycleGAN, which is capable of learning a mapping between source $x \in X$ and target $y \in Y$ from non-parallel training data. It is based on the concept of adversarial learning [55], which is to train a generative model to find a solution in a min-max game between two neural networks, called as generator $(G)$ and discriminator $(D)$. CycleGAN was first proposed for computer vision [56,57], and then extended to various fields including speech synthesis and voice conversion [26,27,58].

A CycleGAN is incorporated with three losses: adversarial loss, cycle-consistency loss, and identity-mapping loss, learning forward and inverse mapping between source and target. Adversarial loss measures how distinguishable between the data distribution of converted data and source or target data. For the forward mapping, it is defined as:

$$L_{ADV}(G_{X\rightarrow Y}, D_Y, X, Y) = \mathbb{E}_{y \sim p(y)}[D_Y(y)] + \mathbb{E}_{x \sim p(x)}[\log(1 - D_Y(G_{X\rightarrow Y}(x)))] \quad (1)$$

The closer the distribution of converted data with that of target data, the smaller $L_{ADV}$ becomes. The adversarial loss only tells us whether $G_{X\rightarrow Y}$ follows the distribution of target data but does not help to preserve the contextual information. In order to guarantee that the contextual information of $x$ and $G_{X\rightarrow Y}$ will be consistent, the cycle-consistency loss is given as:

$$L_{Cyc}(G_{X\rightarrow Y}, G_{Y\rightarrow X}) = \mathbb{E}_{x \sim p(x)}[\|G_{Y\rightarrow X}(G_{X\rightarrow Y}(x)) - x\|_1] + \mathbb{E}_{y \sim p(y)}[\|G_{X\rightarrow Y}(G_{Y\rightarrow X}(y)) - y\|_1] \quad (2)$$

This loss encourages $G_{X\rightarrow Y}$ and $G_{Y\rightarrow X}$ to find an optimal pseudo pair of $(x, y)$ through circular conversion. To preserve the linguistic information without any external processes, an identity mapping loss is introduced as below:

$$L_{ID}(G_{X\rightarrow Y}, G_{Y\rightarrow X}) = \mathbb{E}_{x \sim p(x)}[\|G_{Y\rightarrow X}(x) - x\|_1] + \mathbb{E}_{y \sim p(y)}[\|G_{X\rightarrow Y}(y) - y\|_1] \quad (3)$$

We note that CycleGAN is well-known for achieving remarkable results without parallel training data in many fields from computer vision to speech information processing. In this paper, we propose to use CycleGAN for spectrum and prosody conversion for emotional voice conversion with non-parallel training data.

2.2. Continuous Wavelet Transform (CWT)

It is well-known that emotion can be conveyed by various prosodic features, such as pitch, intensity and speaking rate. F0 is an essential part with respect to the intonation. We note that the modeling of F0 is a challenging task as F0 is discontinuous due to the unvoiced parts, and hierarchical in nature. As a multi-scale modeling method, CWT makes it possible to decompose F0 to different variations over multiple time scales.

Wavelet transform provides an easily interpretable visual representation of signals. Using CWT, a signal can be decomposed into different temporal scales. We note that CWT has
been successfully used in speech synthesis [59, 60] and voice conversion [41, 45].

Given a bounded, continuous signal $k_0$, its CWT representation $W(k_0)(\tau, t)$ can be written as:

$$W(k_0)(\tau, t) = \tau^{-1/2} \int_{-\infty}^{+\infty} k_0(x) \psi\left(\frac{x - \tau t}{\tau}\right) dx$$

(4)

where $\psi$ is the Mexican hat mother wavelet. The original signal $k_0$ can be recovered from the wavelet representation $W(k_0)$ by inverse transform, given as:

$$k_0(t) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} W(k_0)(\tau, x) \tau^{-5/2} \psi\left(\frac{t - x}{\tau}\right) dx d\tau$$

(5)

However, if all information on $W(k_0)$ is not available, the reconstruction is incomplete. In this study, we fix the analysis at ten discrete scales, one octave apart. The decomposition is given as:

$$W_i(k_0)(t) = W_i(k_0)(2^{i+1} \tau_0, t)(i + 2.5)^{-5/2}$$

(6)

The reconstructed $k_0$ is approximated as:

$$k_0(t) = \sum_{i=1}^{10} W_i(k_0)(t)(i + 2.5)^{-5/2}$$

(7)

where $i = 1, \ldots, 10$ and $\tau = 5\text{ms}$. These timing scales were originally proposed in [61] and in prosody model [62, 63]. We believe that the prosody of emotion is expressed differently at different time scales. With the multi-scale representations, lower scales capture the short-term variations and higher scales capture the long-term variations. In this way, we are able to model and transfer the F0 variants from the micro-prosody level to the whole utterance level for emotion pairs. In Figure 2, we use an example to compare two utterances with the same content but different emotion across time scales.

### 3. Spectrum and Prosody Conversion for Emotional Voice Conversion

In this section, we propose an emotional VC framework that performs both spectrum and prosody conversion using Cycle-Consistent Adversarial Networks. As an essential component of prosody, we propose to use CWT to decompose one-dimensional F0 into 10 time-scales. The proposed framework is trained on non-parallel speech data, eliminating the need of parallel training data and effectively converts the emotion of source speaker from one state to another.

The training phase of the proposed framework is given in Figure 3. We first extract spectral and F0 features from both source and target utterances using WORLD vocoder [64]. It is noted that F0 features extracted from WORLD vocoder are discontinuous, due to the voiced/unvoiced parts within an utterance. Since CWT is sensitive to the discontinuities in F0, we perform the following pre-processing steps for F0: 1) linear interpolation over unvoiced regions, 2) transformation of F0 from linear to logarithmic scale, and 3) normalization of the resulting F0 to zero mean and unit variance. We then perform the CWT decomposition of F0 as given in Eq. (6) and Algorithm 1.

We train CycleGAN for spectrum conversion with 24-dimensional Mel-cepstral coefficients (MCEPs), and for prosody conversion with 10-dimensional F0 features for each speech frame. We note that the source and target training data are from the same speaker, but consist of different linguistic content and different emotions. By learning forward and inverse mappings simultaneously using adversarial and cycle-consistency losses, we encourage CycleGAN to find an optimal mapping between source and target spectrum and prosody features.

The run-time conversion phase is shown in Figure 4. We first use the WORLD vocoder to extract spectral features, F0, and aperiodicity (AP) from a given source utterance. Similar to that of training phase, we encode spectral features as 24-
dimensional MCEPs, and obtain 10-scale F0 features through CWT decomposition of F0, that is also reported in Algorithm 1. 24-dimensional MCEPs and 10-scale F0 are fed into the corresponding trained CycleGAN models to perform spectrum and prosody conversion separately. We reconstruct the converted F0 with CWT synthesis approximation method, that is given in Eq. (7) and Algorithm 2. Finally, we use WORLD vocoder to synthesize the converted emotional speech.

**Algorithm 1: CWT Decomposition**

**Input:**
- The signal, $k_0[n], m = 0, 1, ..., N - 1$
- Mother wavelet, $\psi(t)$
- Scale, $i$
- $T$, where $[-T, T]$ is the support of $\psi(t)$

**Output:**
- Wavelet transform: $W_i(k_0)[n], n = 0, 1, ..., N - 1$

**Parameter:**
- $\tau$: sampling interval, set to be 0.005
- $d_j$: space between discrete scales, set to be 0.5
- $s_0$: starting scale, is set to be $2 + \tau$

**Begin**
1. Let $\tilde{k} = [k_0, k_1, ..., k_{N-1}, (0, 0, 0)^T]$;
2. Let $\tilde{\psi}(t) = \frac{1}{2} \psi \left( \frac{T}{2} \right)$;
3. Let $h_0[n] = \tilde{\psi}(m), m = 0, 1, ..., 2iT$;
4. Let $\tilde{h}_i = [h_i[0], h_i[1], ..., h_i[2iT], 0, ..., 0]^T$;
5. $W_i[n] = ifft(fft(\tilde{k}) \ast fft(\tilde{h}_i))$, $n = 0, 1, ..., 2iT + N - 1$;
6. $W_i[k_0][n] = W_i[n + iT], n = 0, 1, ..., N - 1$;
**Return:** $W_i(k_0)[n], n = 0, 1, ..., N - 1$

**End**

### 4. Experiments

We conduct both objective and subjective experiments to assess the performance of our proposed parallel-data-free emotional VC framework. In this paper, we use the emotional speech corpus [65], which is recorded by a professional American actress, speaking English utterances with the same content in seven different emotions. We randomly choose four emotions, that are 1) neutral, 2) angry, 3) sad, and 4) surprise.

We perform CWT to decompose F0 into 10 different scales and train CycleGAN using non-parallel training data to learn the relationships of spectral and prosody features between different emotions of the same speaker. CycleGAN-based spectrum conversion framework, denoted as baseline, is used as the reference framework. In this framework, F0 is transformed through LG-based linear transformation method.

We are also interested in the effect of joint and separate training for spectrum and prosody features. In joint training, we concatenate 24 MCEPs and 10 CWT coefficients to form a vector for each frame to train the joint spectrum-prosody CycleGAN. In separate training, we train a spectrum CycleGAN with the MCEP features, and a prosody CycleGAN with the CWT coefficients separately. Hereafter, we denote the separate training as CycleGAN-Separate, and the joint training as CycleGAN-Joint. The comparison of the frameworks can be also seen in Table 1.

#### 4.1. Experimental Setup

The speech data in [65] is sampled at 16kHz with 16-bit per sample. The audio files for each emotion are manually segmented into 100 short parallel sentences (approximately 3 minutes). Among them, 90 and 10 sentences are provided as training and evaluation sets, respectively. In order to make sure that our proposed model is trained under non-parallel condition, the first 45 utterances are used for the source and the other 45 sentences are used for the target. 24 Mel-cepstral coefficients (MCEPs), fundamental frequency (F0), and APs are then extracted every 5 ms using WORLD vocoder [64]. As a pre-processing step, we normalize the source and target MCEPs per dimension.

We report the performance of three frameworks that use CycleGAN, namely 1) baseline 2) CycleGAN-Joint, and 3) CycleGAN-Separate. For the baseline, we extract 24-dimensional MCEPs and one-dimensional F0 features for each frame. For both CycleGAN-Separate and CycleGAN-Joint, each speech frame is represented with 24-dimensional MCEPs and 10-dimensional F0 features. We adopt the same network structure for all frameworks. We design the generators using a one-dimensional (1D) CNN to capture the relationship among the overall features while preserving the temporal structure.
The 1D CNN is incorporated with down-sampling, residual, and up-sampling layers. As for the discriminator, a 2D CNN is employed. For all frameworks, we set \( \lambda_{C2C} = 10 \). \( L_{1D} \) is only used for the first \( 10^5 \) iterations with \( \lambda_{1D} = 5 \) to guide the learning process.

We train the networks using the Adam optimizer with a batch size of 1. We set the initial learning rates to 0.0002 for the generators and 0.0001 for the discriminators. We keep the learning rate the same for the first \( 2 \times 10^5 \) iterations, which then linearly decays over the next \( 2 \times 10^5 \) iterations. The momentum term \( \beta_1 \) is set to be 0.5. As CycleGAN does not require source-target pair to be the same length, time alignment is not necessary.

### 4.2. Objective Evaluation

We perform objective evaluation to assess the performance of both spectrum and prosody conversion. In all experiments, we use 45-45 non-parallel utterances during training.

#### 4.2.1. Spectrum Conversion

We employ Mel-cepstral distortion (MCD) between the converted and target Mel-cepstra to measure the spectrum conversion, that is given as follows:

\[
MCD[dB] = (10/\ln 10) \left( 10 \left[ \frac{1}{N} \sum_{n=1}^{N} (\text{mceps}_c^n - \text{mceps}_t^n) \right] \right)
\]

where \( \text{mceps}_c^n \) and \( \text{mceps}_t^n \) represent the converted and target MCEPs sequences, respectively. A lower MCD indicates better performance.

<table>
<thead>
<tr>
<th>Framework</th>
<th>Spectrum Conversion</th>
<th>Prosody Conversion (F0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>Spectrum CycleGAN</td>
<td>LG-based F0 linear transformation</td>
</tr>
<tr>
<td>CycleGAN-Joint</td>
<td>Joint Spectrum-Prosody CycleGAN</td>
<td></td>
</tr>
<tr>
<td>CycleGAN-Separate (proposed)</td>
<td>Spectrum CycleGAN</td>
<td>Prosody CycleGAN</td>
</tr>
</tbody>
</table>

Table 1: The comparison of the baseline, CycleGAN-Joint, and CycleGAN-Separate for spectrum and prosody conversion.

#### 4.2.2. Prosody Conversion

We use Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE) to report the performance of prosody conversion [41]. The RMSE between the converted F0 and the corresponding target F0 is defined as:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (F_{0c}^i - F_{0t}^i)^2}
\]

where \( F_{0c}^i \) and \( F_{0t}^i \) denote the converted and target interpolated F0 features, respectively. \( N \) is the length of F0 sequence. We note that a lower RMSE value represents better F0 conversion performance.

The PCC between the converted and target F0 sequences is given as:

\[
\rho(F_{0c}, F_{0t}) = \frac{\text{cov}(F_{0c}, F_{0t})}{\sigma_{F_{0c}} \sigma_{F_{0t}}}
\]

where \( \sigma_{F_{0c}} \) and \( \sigma_{F_{0t}} \) are the standard deviations of the converted F0 sequences (\( F_{0c}^i \)) and the target F0 sequences (\( F_{0t}^i \)), respectively. We note that a higher PCC value represents better F0 conversion performance.

Table 3 reports the RMSE and PCC values of F0 conversion for a number of settings in a comparative study. In this experiment, we conducted three emotional conversion settings: 1) neutral-to-angry, 2) neutral-to-sad, and 3) neutral-to-surprise. We also report the overall performance. As for RMSE results, first of all, we observe that the proposed prosody conversion, based on CycleGAN with CWT-based F0 decomposition outperforms the traditional baseline (denoted as \( \text{baseline} \)) where F0 is converted with LG-based linear transformation method. Secondly, the proposed separate training with CycleGAN for spectrum and CWT-based prosody conversion overall achieves better result (RMSE: 63.03) than separate training (RMSE: 65.05), which is also consistent with the objective evaluation. PCC results suggest that both joint and separate training of CWT-based F0 features achieve similar results.

We would like to highlight that the proposed CWT-based modeling for F0 always outperforms the baseline framework that uses LG-based linear transformation method.
### Subjective Evaluation

We further conduct two listening experiments to assess the proposed frameworks in terms of emotion similarity. We perform XAB test to assess the emotion similarity by asking listeners to choose the one which sounds more similar to the original target between A and B in terms of emotional expression. XAB test has been widely used in speech synthesis such as voice conversion [41], singing voice conversion [54] and emotional voice conversion [48]. In both experiments, 45-45 non-parallel utterances are used during training. We selected two emotion combinations for the listening experiments, that are 1) neutral-to-angry (N2A), and 2) neutral-to-surprise (N2S). 13 subjects participated in all the listening tests, each of them listens to 80 converted utterances in total.

We first conduct XAB test between the baseline and our proposed method to show the effect of our proposed framework that performs separate training of CycleGAN-based conversion for spectrum and CWT-based F0 modeling. Consistent with the previous experiments, our proposed framework is again denoted as CycleGAN-Separate. Listeners are asked to listen to the source utterances, the baseline, our proposed method and the reference utterances respectively. Then, they are asked to choose the one which sounds more similar to the reference in terms of emotional expression. We note that both frameworks perform spectral conversion in the same way, while our proposed framework performs a more sophisticated F0 conversion, which is modeling with CWT, and then converting with CycleGAN. The results are reported in Figure 5 for 2 different emotional conversion scenarios that are N2A and N2S. We observe that the proposed CycleGAN-Separate outperforms the baseline framework in both experiments, which shows the effectiveness of prosody modeling and conversion, for emotional voice conversion.

We then conduct XAB test between joint and separate training to assess different training strategies for spectrum and prosody conversion. The results are reported in Figure 6 for two different emotional conversion scenarios N2A and N2S. We observed that the performance of separate training (denoted as CycleGAN-Separate) is much better than the joint training (denoted as CycleGAN-Joint). Our proposed method achieves 93.6% on N2A and 96.5% on N2S, which we believe are remarkable.

---

**Table 3:** A comparison of the RMSE and PCC results of the baseline, CycleGAN-Joint and CycleGAN-Separate for three different emotion combinations (neutral-to-angry, neutral-to-sad and neutral-to-surprise).

<table>
<thead>
<tr>
<th>Datasets</th>
<th>RMSE [Hz]</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>CycleGAN-Joint</td>
</tr>
<tr>
<td>Neutral→Angry</td>
<td>71.09</td>
<td>64.35</td>
</tr>
<tr>
<td>Neutral→Sad</td>
<td>62.99</td>
<td>57.46</td>
</tr>
<tr>
<td>Neutral→Surprise</td>
<td>77.89</td>
<td>73.10</td>
</tr>
<tr>
<td><strong>Overall mean</strong></td>
<td>70.62</td>
<td>65.05</td>
</tr>
</tbody>
</table>

---

**Figure 5:** The XAB preference results with 95% confidence interval between the baseline and CycleGAN-Separate in emotion similarity experiments.

**Figure 6:** The XAB preference results with 95% confidence interval between the CycleGAN-Joint and CycleGAN-Separate in emotion similarity experiments.

## 4.4. Joint vs. Separate Training of Spectrum and Prosody

We observe that the listeners prefer separate training much more than the joint training. We consider that prosody is manifested at different time scales, which also consists of content-dependent and content-independent elements. The joint training ties the CWT coefficients of F0 with the spectral features at the frame level, that assumes that prosody is content-dependent. With the limited number of training samples (45 pairs and around 3 minutes of speech), the CycleGAN model resulting from the joint training does not generalize well the emotional mapping for unseen content at run-time inference. With the separate training, the CycleGAN model is trained for spectrum and prosody separately. In this way, the prosody CycleGAN learns sufficiently well from the limited number of training samples between the emotion pairs in a content-independent manner. Therefore, separate training outperforms joint training in terms of emotion similarity.

---

## 5. Conclusion

In this paper, we propose a high-quality parallel-data-free emotional voice conversion framework. We perform both spectrum and prosody conversion based on CycleGAN. We provide a non-linear method which uses CWT to decompose F0 into different timing-scales. Moreover, we also study the joint and separate training of CycleGAN for spectrum and prosody conversion. We observe that separate training of spectrum and prosody can achieve better performance than joint training, in terms of emotion similarity. Experimental results show that our proposed emotional voice conversion framework can achieve better performance than the baseline with non-parallel training data.

---

## 6. Acknowledgments

This work is supported by Human-Robot Interaction Phase 1 (Grant No. 192 25 00054), National Research Foundation Singapore under the National Robotics Programme. It is also supported by National Research Foundation Singapore under the AI Singapore Programme (Award Number: A1SG-100E-2018-006), and Programmatic Grant No. A18A2B0046 (Human Robot Collaborative AI for AME) and A1687B0033 (Neuromorphic Computing) from the Singapore Government’s Research, Innovation and Enterprise 2020 plan in the Advanced Manufacturing and Engineering domain.
7. References


[49] Kenan Emir Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf Kassim, “Semantically consistent hierarchical text to fashion image synthesis with an enhanced-attentional generative adversarial network,” in Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.


