# Deep Bidirectional LSTM Modeling of Timbre and Prosody for Emotional Voice Conversion

*Huaiping Ming*[1], *Dongyan Huang*[1], *Lei Xie*[2], *Jie Wu*[2], *Minghui Dong*[1]*and Haizhou Li*[1]

[1]Institute for Infocomm Research, A$^\star$STAR, Singapore
[2]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
{minghp, huang, mhdong, hli}@i2r.a-star.edu.sg, {lxie, jiewu}@nwpu.edu.cn

## Abstract

Emotional voice conversion aims at converting speech from one emotion state to another. This paper proposes to model timbre and prosody features using a deep bidirectional long short-term memory (DBLSTM) for emotional voice conversion. A continuous wavelet transform (CWT) representation of fundamental frequency (F0) and energy contour are used for prosody modeling. Specifically, we use CWT to decompose F0 into a five-scale representation, and decompose energy contour into a ten-scale representation, where each feature scale corresponds to a temporal scale. Both spectrum and prosody (F0 and energy contour) features are simultaneously converted by a sequence to sequence conversion method with DBLSTM model, which captures both frame-wise and long-range relationship between source and target voice. The converted speech signals are evaluated both objectively and subjectively, which confirms the effectiveness of the proposed method.

**Index Terms**: voice conversion, prosody, long short-term memory, recurrent neural networks

## 1. Introduction

Emotion, in everyday speech, is any relatively brief conscious experience characterized by intense mental activity. Emotion is often intertwined with mood, temperament, personality, disposition, and motivation, which play an important role in social interaction and decision making [1, 2, 3]. Though emotions are complex, vocal expression almost always accompanies an emotional state to communicate reaction and intention of actions, and speech signal contains rich information about emotion states [4]. Emotional voice conversion is a task of converting speech from one emotion state into another one, while keeping the basic linguistic and speaker information.

The most common speech characteristics include timbre and prosody. The voice timbre is characterized by spectral features, and prosody is concerned with those elements of speech that are not individual phonetic segments (vowels and consonants) but are properties of syllables and larger units of speech [5]. Recently, there has been tremendous research in emotional voice conversion and synthesis. Researchers are particularly interested in the prosodic factors of speech, since prosody reflects various features of the speaker including the emotional state. Tao *et al.* [6] tried to model F0 contour using a linear modification model, a Gaussian mixture model (GMM) and a classification regression tree model for prosody conversion from neutral speech to emotional speech. Zeynep *et al.* [7] built a system that converted spectrum, F0 and duration for transforming the emotion in speech. The F0 contour was modeled and generated by context-sensitive syllable HMMs, duration was transformed using phone-based relative decision trees, and spectrum was converted using a GMM-based method or a codebook selection approach. Aihara *et al.* proposed GMM-based emotional voice conversion applying both spectrum and prosody features [8]. Later, they proposed an exemplar-based emotional voice conversion approach based on non-negative matrix factorization (NMF) [9], where parallel exemplars were introduced to encode the source speech signal and synthesize the target speech signal. Ming *et al.* proposed to use CWT in F0 modeling for emotional voice conversion [10]. Specifically, they decompose the fundamental frequency contour into five temporal scales, which represent temporal changes ranging from micro-prosody to the utterance level. The feature representation of F0 and spectrum are simultaneously converted in a unified exemplar-based voice conversion framework.

Prosody features, including F0 and energy contour, are essential factors which significantly contribute to the underlying emotional state of a speech. It is widely agreed that prosody is inherently supra-segmental and hierarchical in nature [5, 11, 13, 14, 15]. Prosody conveys information that goes beyond the sequence of segments, syllables, and words found within an utterance, as well as beyond the lexical and syntactic systems of a language [5, 16]. As prosody is affected by both short term dependencies and long term dependencies, it is hard to model the variations of F0 in all temporal scales using linear models. There were many attempts to explore multiple temporal domains in prosody modeling, which seeks to model F0 at different units like phone, syllable and phrase levels [12, 17, 18, 19]. Recently, the CWT has been proposed for the analysis and modeling of F0 in task of text to speech synthesis [16, 20, 21] and voice conversion [10, 22]. The CWT can effectively model F0 in different temporal scales and significantly improve the system performance.

In this paper, we propose to convert the spectrum, energy contour and fundamental frequency (F0) simultaneously under a deep bidirectional LSTM recurrent neural network (RNN) framework of emotional voice conversion. The DBLSTM approaches have shown promising results in sequence modeling tasks, such as languages learning [23], speech recognition [24], TTS synthesis [25] and voice conversion [26]. Prosody features including CWT and logarithmic representation of F0 and energy contour are explored, and some interesting phenomena are fond. Prosody and spectral features are converted simultaneously by a sequence to sequence DBLSTM conversion method. The bidirectional recurrent connections attempt to learn the contextual information in both forward and backward directions. The memory blocks and peephole connections make it possible to access long-range contextual, which elegantly capture both frame-wised and long-range relationship between source and

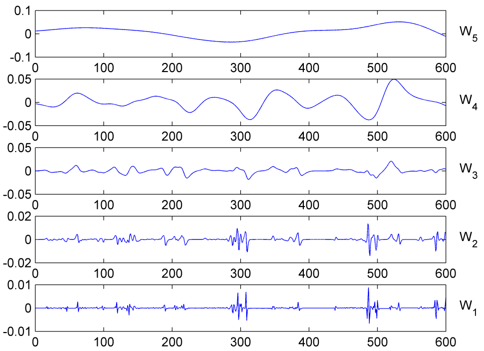

Figure 1: An example of the five-scale representation of F0.

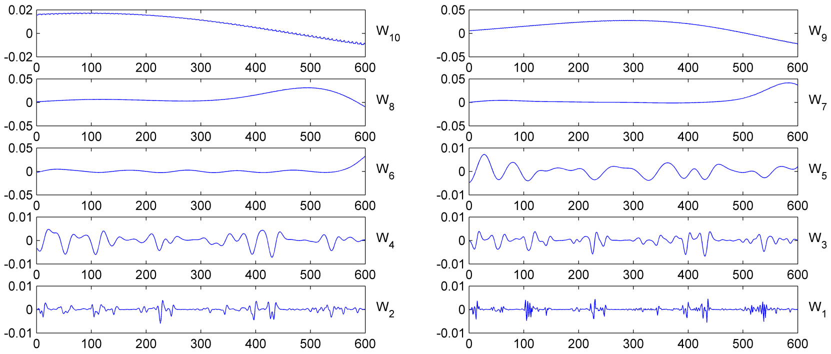

Figure 2: An example of the ten-scale representation of energy contour.

target prosody and spectral features. Due to the limited amount of training data, we also propose to use an adaptation method, which is to apply large amount of parallel data to train an average model to initialize the neural network.

The rest of this paper is organized as follow: In section 2, we introduce the prosody and spectral features, which will be used in conversion system. In section 3, we describe our method for emotional voice conversion. In section 4, the objective and subjective experiment results are presented. Conclusions are drawn in section 5.

## 2. Spectral and Prosody Features

The speech parameters including spectrum, aperiodicity component and fundamental frequency are extracted by applying STRAIGHT [27] analysis method. Denote the spectrum as $\mathbf{SP} \in \mathbb{R}^{F\times M}$, the energy of each frame is defined as

$$\mathbf{e}_m = \sqrt{\sum_{i=1}^{F} \mathbf{SP}_{i,m}^2} \quad , m = 1, ..., M, \tag{1}$$

where $F$ and $M$ are feature dimension and number of frames respectively. By calculating the energy for each frame of a speech signal, we can obtain the energy contour vector $\mathbf{e} \in \mathbb{R}^{1\times M}$.

It is well known that prosody is influenced both at a supra-segmental level, by long-term dependencies, and at a segmental-level, by short-term dependencies. We adopt continuous wavelet transform to decompose the F0 and energy contour into several temporal scales that model prosody at different temporal levels. The wavelet method is sensitive to the gaps in the F0 contour, therefore we apply a linear interpolation method to obtain a continuous F0 trajectory. The linear scale F0 and energy contour are transformed to the logarithmic scale, and then normalized to zero mean and unit variance as required by wavelet analysis.

The continuous wavelet transform of a input signal $f(x)$ is defined by

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

where $\psi$ is the Mexican hat mother wavelet. We fix the analysis at 10 discrete scales, each one octave apart. Then $f(x)$ is represented by 10 separate components given by

$$W_i(f,t) = W_i(f)(2^{i+1}\tau_0, t)(i+2.5)^{-5/2}, \tag{3}$$

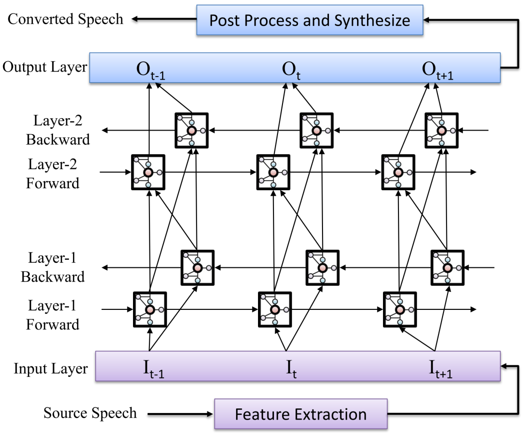

Figure 3: The DBLSTM-RNN based voice conversion framework.

where $i = 1, ..., 10$ and $\tau_0 = 5$ ms. The original signal is approximately reconstructed by the following *ad hoc* reconstruction formula:

$$f(t) = \sum_{i=1}^{10} W_i(f,t)(i+2.5)^{-5/2}. \tag{4}$$

For F0, attempting to relate the wavelet transform scales to levels of linguistic structure [16, 21], adjacent scales are combined, which result in a five-scale representation defined by

$$\mathbf{w}_i = W_{2i-1}(f,t) + W_{2i}(f,t), \tag{5}$$

where $i = 1, ..., 5$.

We denote the five-scale representation of F0 and ten-scale representation of energy as $\mathbf{F0}_{\text{cwt}}$ and $\mathbf{E}_{\text{cwt}}$ respectively. An example of the representation of F0 and energy contour are shown in Fig. 1 and Fig. 2 respectively. The lower scales (high frequencies) capture short-term variations and that higher scales (low frequencies) capture long-term variations. To evaluate the proposed method, we also propose to convert logarithmic scale F0 contour $\mathbf{F0}$ and logarithmic scale energy contour $\mathbf{E}$ under the proposed framework.

## 3. LSTM Based Voice Conversion System

### 3.1. Basic Framework

Conventional RNNs can access only a limited range of context because of the vanishing gradient problem. Long short-term memory (LSTM) take advantage of specially designed memory cells which store information to overcome this limitation. The proposed framework for prosody and spectral feature conversion is shown in Fig. 3. The network architecture is a combination of bidirectional RNNs and LSTM memory block, which can learn long-range contextual in both forward and backward directions. By stacking multiple hidden layers, a deep network architecture is built to capture high level representation of voice features. In this system, features including Mel-cepstral coefficients (MCEPs), F0 and energy contour features are concatenated to build a feature vector, and then the features in this vector are simultaneously converted by the DBLSTM-RNN model.

For bidirectional RNNs, the iteration functions for forward sequence $\overleftarrow{h}$ and backward sequence $\overrightarrow{h}$ are as follows:

$$\overrightarrow{h}_t = \mathcal{H}(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}), \tag{6}$$

$$\overleftarrow{h}_t = \mathcal{H}(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}), \tag{7}$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_y, \tag{8}$$

where $x, y, h, t$ are the input, output, hidden state and time sequence respectively. For LSTM memory block, the recurrent hidden layer function $\mathcal{H}$ is implemented according to the following equations:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i), \tag{9}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{ci}c_{t-1} + b_f), \tag{10}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c), \tag{11}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o), \tag{12}$$

$$h_t = o_t \tanh(c_t), \tag{13}$$

where $i, f, o, c$ refer to the input gate, forget gate, output gate and the element of cell $C$ respectively, and $\sigma$ is the sigmoid function.

### 3.2. Average Model Training and Adaptation

It is well known that deep neural network based methods require large amount of training data to obtain good results. However, in some application scenarios such as voice conversion, it is hard to obtain large parallel data for training. Therefore, we proposed to train an average model, which is to utilize parallel data from other speakers to pre-train the network. The obtained parameters should be at the neighborhood of the optimal solution, thus can be used as the initial values for training a network which has limited training data.

In this work, the CMU-ARCTIC [28] database and some data from corpus described in [29] are used for training the average model. The number of training data is 13398 sentences and the number of validation data is 2364 sentences. Then the parameters of the trained average model are used to initialize the neural network for emotion state conversion. A machine learning library named CURRENNT [30] is used to train the DBLSTM model. The number of units in each layer is $[64, 96, 128, 96, 64]$ for prosody feature with CWT representation, and it is $[51, 96, 128, 96, 51]$ for logarithmic scale prosody feature respectively.

### 3.3. Spectral and Prosody Feature Conversion

In the conversion stage, input spectral and prosody features are concatenated to build a feature vector. The features are normalized to zero mean and unit variance, and then converted by the BDLSTM-RNN methods described in 3.1. We do not convert the aperiodicity component and directly copy it to target to synthesize the converted speech.

Table 1: Features converted by different systems.

| Conversion Method | Converted Features | Adaptation |
|---|---|---|
| NMF | $\mathbf{SP}, \mathbf{F0}_{\text{cwt}}$ | NA |
| LSTM-1 | $\mathbf{SP}$ | Yes |
| LSTM-2 | $\mathbf{SP}, \mathbf{F0}$ | Yes |
| LSTM-3 | $\mathbf{SP}, \mathbf{E}$ | Yes |
| LSTM-4 | $\mathbf{SP}, \mathbf{F0}_{\text{cwt}}$ | Yes |
| LSTM-5 | $\mathbf{SP}, \mathbf{E}_{\text{cwt}}$ | Yes |
| LSTM-6 | $\mathbf{SP}, \mathbf{F0}_{\text{cwt}}, \mathbf{E}_{\text{cwt}}$ | Yes |
| LSTM-7 | $\mathbf{SP}, \mathbf{F0}_{\text{cwt}}, \mathbf{E}_{\text{cwt}}$ | No |

By reforming the components of converted features, we obtain the converted spectral and prosody features $\mathbf{SP}^c$, $\mathbf{E}^c_{\text{cwt}}$, and $\mathbf{F0}^c_{\text{cwt}}$. The logarithmic scale converted F0 and energy contour are reconstructed according to Eq. (4). Then the mean and variance of the converted logarithmic scale F0 and energy contour are normalized to those of the target speaker. Finally, the exponential value of the logarithmic scale F0 and energy contour are calculated to obtain the converted F0 and energy contour. In order to make the energy contour of converted spectrum more close to that of the target, we take advantage of the information about converted energy $\mathbf{e}^c$. Firstly, we take the converted spectrum $\mathbf{SP}^c$ as input to calculate it's energy contour $\mathbf{e}^t$ according to Eq. (1). Then the energy ratio for each speech frame is calculated as

$$\mathbf{r} = \frac{\mathbf{e}^t}{\mathbf{e}^c}, \tag{14}$$

where the divisions are element-wise, and $\mathbf{r} \in \mathbb{R}^{1\times M}$. By replicating the energy ratio vector, we get an energy ration matrix $\mathbf{R} \in \mathbb{R}^{F\times M}$. Finally, the energy contour improved spectrum is given by

$$\mathbf{SP}^c = \frac{\mathbf{SP}^c}{\mathbf{R}}, \tag{15}$$

where the divisions are also element-wise.

## 4. Experiments

### 4.1. Experimental Setup

The emotional speech corpus [29] is used in our experiment. To collect high quality data, a professional actress was selected for emotional data collection. The waveform is 16-bit quantization at 16 kHz sampling rate. We use the speech of four emotions, which are neural, happy, fear and sad from the corpus. The task is to convert the neutral speech to another emotional speech. For each conversion pair, 77 parallel utterances are randomly selected as the training data, 13 utterances as the validation set and another 10 utterances as the evaluation set. There is no overlapping between the three sets.

The exemplar-based emotional voice conversion method (denoted as NMF) described in [10] is applied as a baseline system. Considering different feature combinations, the conducted experiments on different systems are summarized in Table 1. In systems where F0 features are not listed, F0 is linear converted by normalize the mean and variance of source to target.

### 4.2. Objective Evaluation

We use Mel-cepstral distortion (MCD) to measure the spectral distortion and mean square error (MSE) of logarithmic scale F0 to measure the F0 distortion. The MCD between the converted

Table 2: MCD and F0-MSE results for different emotions.

| | MCD [dB] | | | F0-MSE | | |
|---|---|---|---|---|---|---|
| | Happy | Fear | Sad | Happy | Fear | Sad |
| Source | 6.48 | 6.36 | 6.29 | 36.96 | 62.61 | 18.21 |
| NMF | 6.04 | 6.15 | 5.31 | 17.55 | 13.43 | 4.87 |
| LSTM-1 | 5.35 | 5.28 | 4.72 | 20.00 | 13.17 | 4.29 |
| LSTM-2 | **5.26** | **5.11** | **4.60** | 14.18 | **7.78** | 3.42 |
| LSTM-3 | 8.96 | 7.41 | 4.79 | 20.00 | 13.17 | 4.29 |
| LSTM-4 | 5.33 | 5.20 | 4.68 | **11.10** | 9.08 | **3.34** |
| LSTM-5 | 5.29 | 5.29 | 4.73 | 20.00 | 13.17 | 4.29 |
| LSTM-6 | 5.32 | 5.25 | 4.69 | **11.10** | 9.08 | **3.34** |
| LSTM-7 | 5.45 | 5.43 | 4.89 | 20.55 | 15.17 | 4.76 |

Table 3: Subjective classification results for NMF system.

| Target / Perception | Happy | Fear | Sad | Neutral |
|---|---|---|---|---|
| Happy | **46.5**% | 20.5% | 1.8% | 31.2% |
| Fear | 32.4% | **40.0**% | 11.2% | 16.4% |
| Sad | 0% | 2.9% | **59.4**% | 37.7% |

and corresponding target Mel-ceptral is calculated as

$$MCD[dB] = \frac{10}{\log 10}\sqrt{2\sum_{d=1}^{24}(c_{m,d} - \widehat{c}_{m,d})^2}, \qquad (16)$$

where $c_m$ and $\widehat{c}_m$ are the $m$-th coefficients of the target and converted MCCs, respectively. The MSE between the converted and corresponding target logarithmic scaled F0 is calculated as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\Big(\log(F0_i) - \log(\widehat{F0}_i)\Big)^2, \qquad (17)$$

where $F0_i$ and $\widehat{F0}_i$ represent the $i$-th elements of target and converted vector of F0 respectively, and $N$ is the length of F0. A lower MCD and F0-MSE value indicates smaller distortion or prediction error.

The average MCD and F0-MSE results over all evaluation pairs are reported in Table 2. We calculate MCD and F0-MSE between the source speaker and the target speaker as a reference. For F0-MSE, we observe that converting the five-scale representation of F0 applying DBLSTM obtains best results for Happy and Sad, and converting the logarithmic F0 applying DBLSTM obtains best results for Fear. Comparing with linear conversion (LSTM-1) or exemplar-based conversion (NMF), we can say that the non-linear conversion (LSTM-2, LSTM-4) of F0 feature obtains much better results. Comparing LSTM-7 with LSTM-6, it is obvious that the proposed adaptation method significantly improves the accuracy of predicted F0.

For MCD, we observe that system LSTM-2 obtains the best results for all emotion states. System LSTM-3 which converts logarithmic energy contour obtains the worst results, where the MCD for Happy and Fear even bigger than the MCD between source and target. Comparing system NMF with LSTM based system, all LSTM based system get smaller MCD except system LSTM-3. This results confirms that the non-linear DNN method outperforms the linear exemplar-based method. Comparing system LSTM-6 with LSTM-7, the MCD decrease for all emotion states after adaptation. Comparing system LSTM-5 with LSTM-3, the MCD decrease for all emotion states with CWT representation of energy contour, which confirms the effectiveness of using CWT for energy contour modeling.

Table 4: Subjective classification results for LSTM-2 system.

| Target / Perception | Happy | Fear | Sad | Neutral |
|---|---|---|---|---|
| Happy | **43.5**% | 22.4% | 8.2% | 25.9% |
| Fear | 7.1% | **76.4**% | 12.9% | 3.6% |
| Sad | 1.8% | 1.8% | **75.9**% | 20.5% |

Table 5: Subjective classification results for LSTM-4 system.

| Target / Perception | Happy | Fear | Sad | Neutral |
|---|---|---|---|---|
| Happy | **66.7**% | 2.8% | 2.8% | 27.7% |
| Fear | 20.0% | **58.9**% | 9.4% | 11.7% |
| Sad | 0% | 1.1% | **83.3**% | 15.6% |

Table 6: Subjective classification results for LSTM-6 system.

| Target / Perception | Happy | Fear | Sad | Neutral |
|---|---|---|---|---|
| Happy | **77.1**% | 4.1% | 1.2% | 17.6% |
| Fear | 19.4% | **69.4**% | 5.3% | 5.9% |
| Sad | 1.8% | 0.6% | **70.6**% | 27.0% |

Table 7: Subjective classification results for LSTM-7 system.

| Target / Perception | Happy | Fear | Sad | Neutral |
|---|---|---|---|---|
| Happy | **55.6**% | 6.1 % | 1.1% | 37.2% |
| Fear | 14.4% | **56.7**% | 15.6% | 13.3% |
| Sad | 0.6% | 0.6% | **77.7** % | 21.1% |

### 4.3. Subjective Evaluation

We conduct a subjective emotion classification test. In each test, 30 utterances (10 for Happy, 10 for Fear and 10 for Sad) are selected and 18 experienced listeners are involved. The listeners are asked to label a converted voice as Happy, Fear, Sad or Neutral. As a sanity check, a subjective emotion classification test for original recorded emotional speech utterances is first conducted. Not surprisingly, all utterances are correctly classified by the 18 listeners. According to the objective evaluation results and small scale listening test, the results of some systems, such as LSTM-3, are obviously not good. Therefore, we only conduct subjective test for system NMF, LSTM-2, LSTM-4, LSTM-6 and LSTM-7, and the results are shown in Table 3, 4, 5, 6, 7 respectively.

Comparing NMF system with LSTM based systems, it is very clear that LSTM based methods obtains better results. It is interesting that different feature combinations obtain good results for different emotion states. Specifically, LSTM-2 obtains best results for Fear, LSTM-4 obtains best results for Sad and LSTM-6 obtains best results for Happy. Comparing with LSTM-7, LSTM-6 increases the classification rate for Happy and Fear significantly, but it slightly decreases the rate for Sad. Considering both objective and subjective results, we can see that the decrease of objective measures not always leads to better classification results for some emotion states, which is an interesting phenomenon.

## 5. Conclusions

We propose a method to convert spectral and prosody features simultaneously in a DBLSTM-RNN based voice conversion framework. An adaptation method is applied to improve the system performance when there is a limited amount of training data. The CWT representation is proved to be effective for modeling of F0 and energy. The adaptation method significantly improves system performance. Both objective and subjective experiments confirm the effectiveness of the proposed method. We will explore more prosody features in the future work.

# 6. References

[1] J. A. Russell, J. A. Bachorowski and J. M. Fernández-Dols, (2003). Facial and vocal expressions of emotion. Annual review of psychology, 54(1), 329-349.

[2] R. W. Picard and R. Picard, (1997). Affective computing (Vol. 252). Cambridge: MIT press.

[3] D. L. Schacter, (2011). Psychology Second Edition, 41 Madison Avenue, New York, NY 10010.

[4] K. R. Scherer, (2003). Vocal communication of emotion: A review of research paradigms. Speech communication, 40(1), 227-256.

[5] D. Hirst, and A. Di Cristo, (1998). Intonation systems: a survey of twenty languages. Cambridge University Press.

[6] J. Tao, Y. Kang and A. Li, (2006). Prosody conversion from neutral speech to emotional speech. IEEE Transactions on Audio, Speech, and Language Processing, 14(4), 1145-1154.

[7] Z. Inanoglu and S. Young, (2007, August). A system for transforming the emotion in speech: combining data-driven conversion techniques for prosody and voice quality. In INTERSPEECH (pp. 490-493).

[8] R. Aihara, R. Takashima, T. Takiguchi and Y. Ariki, (2012). GMM-based emotional voice conversion using spectrum and prosody features. In American Journal of Signal Processing, 2(5), 134-138.

[9] R. Aihara, R. Ueda, T. Takiguchi and Y. Ariki, (2014, December). Exemplar-based emotional voice conversion using non-negative matrix factorization. In Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA) (pp. 1-7). IEEE.

[10] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong and H. Li, (2015). Fundamental frequency modeling using wavelets for emotional voice conversion. In 6th Affective Computing and Intelligent Interaction (ACII) Workshop on Affective Social Multimedia Computing.

[11] K. Yu, (2012, October). Review of F0 modelling and generation in HMM based speech synthesis. In IEEE 11th International Conference on Signal Processing (ICSP), (Vol. 1, pp. 599-604).

[12] C. H. Wu, C. C. Hsia, C. H. Lee and M. C. Lin, (2010). Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing, 18(6), 1394-1405.

[13] A. Wennerstrom, (2001). The music of everyday speech: Prosody and discourse analysis. Oxford University Press.

[14] Y. Xu, (2011). Speech prosody: A methodological review. Journal of Speech Sciences, 1(1), 85-115.

[15] J. J. O. H. Kawasaki-Fukumori, (1997). Alternatives to the sonority hierarchy for explaining segmental sequential constraints. In Language and its ecology: Essays in memory of Einar Haugen, 100, 343.

[16] M. S. Ribeiro and R. A.Clark (2015, April). A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (pp. 4909-4913).

[17] J. Latorre and M. Akamine, (2008, September). Multilevel parametric-base F0 model for speech synthesis. In INTERSPEECH (pp. 2274-2277).

[18] Y. Qian, Z. Wu, B. Gao and F. K. Soong, F. K. (2011). Improved prosody generation by maximizing joint probability of state and longer units. IEEE Transactions on Audio, Speech, and Language Processing, 19(6), 1702-1710.

[19] N. Obin, A. Lacheret and X. Rodet, (2011, August). Stylization and trajectory modelling of short and long term speech prosody variations. In INTERSPEECH.

[20] M. Vainio, A. Suni and D. Aalto, (2013). Continuous wavelet transform for analysis of speech prosody. In Tools and Resources for the Analysys of Speech Prosody, an INTERSPEECH 2013 satellite event.

[21] A. S. Suni, D. Aalto, T. Raitio, P. Alku and M. Vainio, (2013). Wavelets for intonation modeling in HMM speech synthesis. In 8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona.

[22] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong and H. Li, (2016). Exemplar-based Sparse Representation of Timbre and Prosody for Voice Conversion. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[23] F. A. Gers, and J. Schmidhuber, (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. IEEE Transactions on Neural Networks, 12(6), 1333-1340.

[24] H. Sak, A. W. Senior, and F. Beaufays, (2014, September). Long short-term memory recurrent neural network architectures for large vocabulary speech recognition. In INTERSPEECH (pp. 338-342).

[25] Y. Fan, Y. Qian, F. L., Xie and F. K. Soong, (2014, September). TTS synthesis with bidirectional LSTM based recurrent neural networks. In INTERSPEECH (pp. 1964-1968).

[26] L. Sun, S. Kang, K. Li and H. Meng, (2015, April). Voice conversion using deep Bidirectional Long Short-Term Memory based Recurrent Neural Networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (pp. 4869-4873).

[27] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino and H. Banno, (2008, March). TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (pp. 3933-3936).

[28] J. Kominek and A. W. Black, (2004). The CMU Arctic speech databases. In Fifth ISCA Workshop on Speech Synthesis.

[29] S. Liu, D. Y. Huang, W. Lin, M. Dong, H. Li and E. P. Ong, (2014, December). Emotional facial expression transfer based on temporal restricted Boltzmann machines. In Asia-Pacific Signal and Information Processing Association (APSIPA).

[30] F. Eyben, J. Bergmann, and F. Weninger. CURRENNT CUDA-enabled Machine Learning Library For Recurrent Neural Networks. *https://sourceforge.net/projects/currennt/*