Emotional Voice Conversion Using Neural Networks with Different Temporal Scales of F0 based on Wavelet Transform

Zhaojie Luo[1], *Jinhui Chen*[1], *Toru Nakashika*[2], *Tetsuya Takiguchi*[1], *Yasuo Ariki*[1]

[1]Graduate School of System Informatics, Kobe University, Japan
{luozhaojie, ianchen}@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp
[2]Graduate School of Information Systems, University of Electro-Communications, Japan
nakashika@uec.ac.jp

Abstract

An artificial neural network is one of the most important models for training features of voice conversion (VC) tasks. Typically, neural networks (NNs) are very effective in processing nonlinear features, such as mel cepstral coefficients (MCC) which represent the spectrum features. However, a simple representation for fundamental frequency (F0) is not enough for neural networks to deal with an emotional voice, because the time sequence of F0 for an emotional voice changes drastically. Therefore, in this paper, we propose an effective method that uses the continuous wavelet transform (CWT) to decompose F0 into different temporal scales that can be well trained by NNs for prosody modeling in emotional voice conversion. Meanwhile, the proposed method uses deep belief networks (DBNs) to pretrain the NNs that convert spectral features. By utilizing these approaches, the proposed method can change the spectrum and the prosody for an emotional voice at the same time, and was able to outperform other state-of-the-art methods for emotional voice conversion.

Index Terms: emotional voice conversion, continuous wavelet transform, F0 features, neural networks, deep belief networks,

1. Introduction

Recently, the study of Voice Conversion (VC) has attracted wide attention in the field of speech processing. This technology can be widely applied in various application domains. For instances, emotion conversion [1], speaking assistance [2], and other applications [3] [4]. Therefore, the need for this type of technology in various fields has continued to propel related researches each year. Many statistical approaches have been proposed for spectral conversion during the last decades [5] [6]. Among these approaches, a Gaussian Mixture Model (GMM) is widely used, and a number of improvements have been proposed [7] [8] for GMM-based voice conversion. Other VC methods, such as approaches based on non-negative matrix factorization (NMF) [9] [2] have also been proposed. The NMF and GMM methods are based on linear functions. For performing voice conversion better, the VC technique needs to train more complex nonlinear features such as Mel Cepstral Coefficients (MCC) [10] which are widely used in automatic speech and speaker recognition, some approaches construct non-linear mapping relationships using neural networks (NNs) to train the mapping dictionaries between source and target features [11], or using deep belief networks (DBNs) to achieve non-linear deep transformation [12]. The results have shown that these deep architecture models can perform better than shallow conversion in some complex voice features conversion.

However, most of the related works in respect to VC focus on the conversion of spectral features, rather than fundamental frequency (F0) conversion. The spectral features and F0 features obtained from STRAIGHT [13] can affect the voice's acoustic features and emotional features, respectively. F0 features are one of the most important parameters for representing emotional speech, because it can clearly describe the variation of voice prosody from one pitch period to another. But F0 features extracted from STRAIGHT are low-dimensional features that cannot be processed well by deep models such as NMF models or DBN models. Therefore, F0 features are usually converted by logarithm Gaussian normalized transformation (LG) [14] in these models. However, it has been proved that prosody conversion is affected by both short term dependencies as well as long term dependencies, such as the sequence of segments, syllables, words within an utterance, lexical and syntactic systems of a language [15]. The LG-based method is insufficent to convert the prosody effectively due to the constraints of their linear models and low dimensional F0 features [16]. Since the CWT can effectively model F0 in different temporal scales and significantly improve the speech synthesis performance [17]. Ming *et.al.* [16] used CWT in F0 modeling within the NMF model for emotional voice conversion and obtained a better result than the LG method in F0 conversion.

In this paper, inspired by deep learning models' ability to perform well in complex nonlinear feature conversion [12] and CWT's ability to improve F0 features conversion [16], we propose a novel method that uses NNs to train the CWT-F0 for converting the prosody of the emotional voice. Different from [16], we decompose the F0 into 30 temporal scales which contain more specifics of different temporal scales and train them by NNs which can perform better compared to the logarithm Gaussian model and NMF-based model. Since the DBNs are effective to spectral envelope conversion, for spectral features conversion, we train the MCC features by using DBNs proposed by Nakashika *et.al.* [12]. The reason we choose different models to separately convert the spectral features and F0 features is that although the wavelet transform decomposed F0 features to more complex features, they can be trained enough by NNs, while the more complex spectral features need a deeper architecture.

In the rest of this paper, we describe features processing about MCC and CWT in Sec. 2. The DBNs and NNs used in our proposed method are introduced in Sec. 3. In Sec. 4, we describe the framework of our proposed emotional voice conversion system. Sec. 5 gives the detailed stages of process in experimental evaluations, and conclusions are drawn in Sec. 6.

2. Feature extraction and processing

To extract features from a speech signal, the STRAIGHT is frequently used. Generally, the smoothing spectrum and instantaneous-frequency-based F0 are derived as excitation features for every 5ms from the STRAIGHT [13]. To have the same number of frames, a dynamic time wraping method is used to align the extracted features (spectrum and F0) of source voice and target voice. Then, the aligned spectral features are translated into MCC. The F0 features produced by STRAIGHT are one dimensional and discrete. It is difficult to model the variations of F0 in all temporal scales using linear models. Inspired by the work in [16], before training the F0 features by NNs, we adopted CWT to decompose the F0 contour into several temporal scales that can be used to model different prosodic levels ranging from micro-prosody to the sentence level. The steps for processing details are as follows:

1) In order to explore the perceptual relevant information, F0 contour is transformed from linear scale to logarithmic semitone scale, which is referred to as logF0. As shown in Fig. 1(A), the logF0 is discrete. As the wavelet method is sensitive to the gaps in the F0 contours, we need to fill in the unvoiced parts in the logF0 with linear interpolation to reduce discontinuities in voice boundaries. Finally, normalize the interpolated logF0 contour to zero mean and unit variance. An example of an interpolated pitch contour is depicted in Fig. 1(B)

2) The continuous wavelet transform of F0 is defined by

$$W(\tau,t)=\tau^{-1/2}\int_{-\infty}^{\infty} f_0(x)\,\psi\left(\frac{x-t}{\tau}\right)dx \qquad (1)$$

$$\psi(t)=\frac{2}{\sqrt{3}}\pi^{-1/4}\left(1-t^2\right)e^{-t^2/2}, \qquad (2)$$

where $f_0(x)$ is the input signal and ψ is the Mexican hat mother wavelet. We decompose the continuous logF0 with 30 discrete scales, each one third octave apart. Our F0 is thus represented by 30 separate components given by

$$W_i(f_0)(t)=W_i(f_0)(2^{(i/3)+1}\tau_0,t)\left((i/3)+2.5\right)^{-5/2}, \qquad (3)$$

where i = 1,...,30 and τ_0=5 ms. As shown in Fig. 2, the top figure is the interpolated log-normalized F0 of the source voice. And the second pan to sixth pan show several examples of separate components which can represent the utterance, phrase, word, syllable and phone levels, respectively.

3. Training model

3.1. NNs

Neural networks (NNs) are trained on a frame error (FE) minimization criterion and the corresponding weights are adjusted to minimize the error squares over the whole source-target, stereo training data set. As shown in Eq. 4, the error of mapping is given by

$$\epsilon=\sum_t ||y_t-G(x_t)||^2, \qquad (4)$$

$G(x_t)$ denotes the NNs mapping of x_t and is defined as:

$$G(x_t)=(G^1\circ G^2\circ\cdots\circ G^L)=\bigodot_{l=1}^{L}G^{(l)}(x_t) \qquad (5)$$

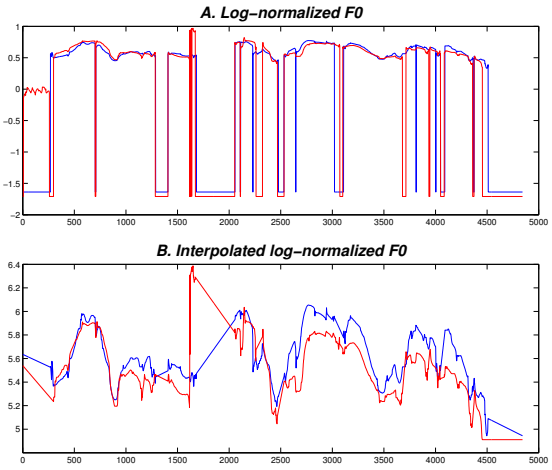

Figure 1: Log-normalized F0 (A) and interpolated log-normalized F0 (B). The red curve: target F0; The blue curve: source F0.

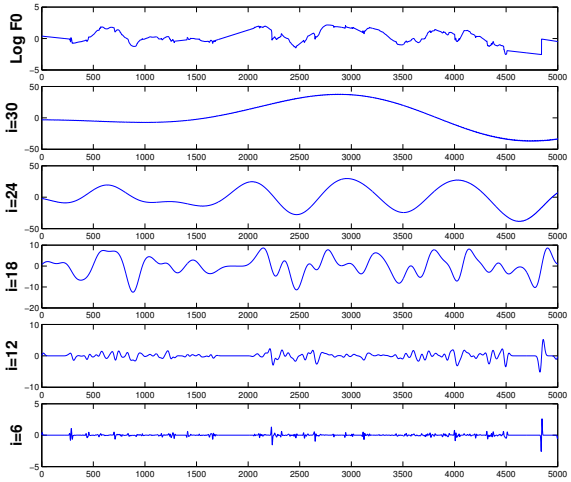

Figure 2: Interpolated log-normalized F0 and five wavelet transforms (i=30, i=24, i=18, i=12, i=6)

$$G^l(x_t)=\sigma(W^l x_t) \qquad (6)$$

Here, $\bigodot_{l=1}^{L}$ denotes composition of L functions. For instance, $\bigodot_{l=1}^{2}W^{(l)}(z)=\sigma(W^{(2)}\sigma(W^{(1)}(x_t))$. $W^{(l)}$ represents the weight matrices of layer l in NNs. σ denotes a standard tanh function which is defined as:

$$\sigma(x)=\tanh(x)=\frac{e^{2x}-1}{e^{2x}+1}, \qquad (7)$$

As shown in the training model of Fig. 3, we use a 4-layer NN model for prosody training. w1, w2 and w3 represent the weight matrices of first, second and third layers of NN, respectively.

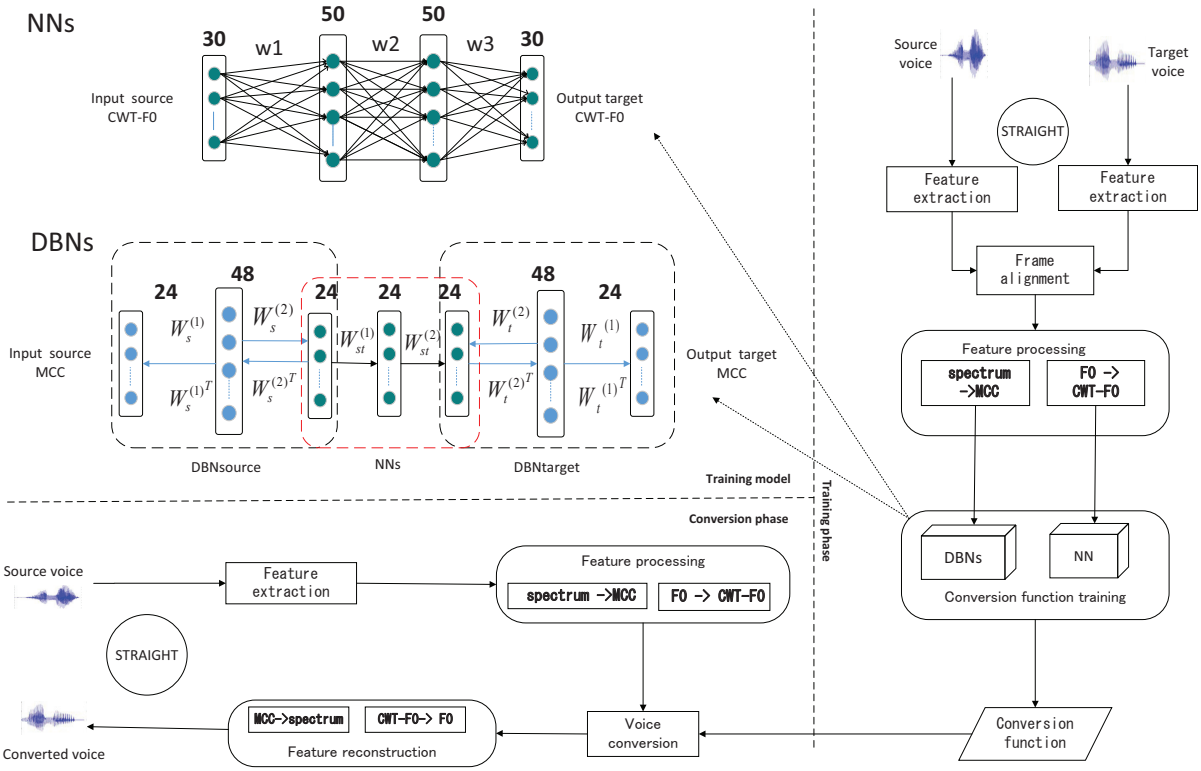

Figure 3: Framework of the proposed method

3.2. DBNs

Deep belief networks (DBNs) have an architecture that stacks multiple Restricted Boltzmann Machines (RBMs) which are composed of a visible layer and a hidden layer with full, two-way inter-layer connections but no intra-layer connections. As an energy-based model, the energy of a configuration (v, h) is defined as :

$$E(v,h) = -a^T v - b^T h - v^T W h, \tag{8}$$

where, $W \in R_{I\times J}$, $a \in R_{I\times 1}$, and $b \in R_{J\times 1}$ denote the weight parameter matrix between visible units and hidden units, a bias vector of visible units, and a bias vector of hidden units, respectively. The joint distribution over v and h is defined as:

$$P(v,h) = \frac{1}{Z} e^{-E(v,h)}. \tag{9}$$

The RBM has the shape of a bipartite graph, with no intra-layer connections. Consequently, the individual activation probabilities are obtained via

$$P(h_j = 1|v) = \sigma\left(b_j + \sum_{i=1}^{m} w_{i,j} v_i\right); \tag{10}$$

$$P(v_i = 1|h) = \sigma\left(a_i + \sum_{j=1}^{n} w_{i,j} h_j\right). \tag{11}$$

In DBNs, σ denotes a standard sigmoid function, $(\sigma(x) = 1/(1+e^{-x}))$. For parameter estimation, RBMs are trained to maximize the product of probabilities assigned to some training set data. To calculate the weight parameter matrix, we use the RBM log-likelihood gradient method as defined:

$$L(\theta) = \frac{1}{N}\sum_{n=1}^{N} log P_\theta\left(v^{(n)}\right) - \frac{\lambda}{N}\|W\|. \tag{12}$$

Here, $P_\theta\left(v^{(n)}\right)$ is the probability of visible vectors in the inner model with the model parameters $\theta = (W, a, b)$. To differentiate the $L(\theta)$ via Eq. 13, we can obtain W when making the $L(\theta)$ be the largest.

$$\frac{\partial L(\theta)}{\partial W_{ij}} = E_{P_{data}}[v_i h_j] - E_{P_\theta}[v_i h_j] - \frac{2\lambda}{N} W_{ij}. \tag{13}$$

where, $E_{P_{data}}\langle\cdots\rangle$ and $E_{P_\theta}\langle\cdots\rangle$ represent averages of input data and the inner model, respectively. As shown in the training model of Fig. 3, our proposed method has two different DBNs for source speech and target speech (DBNsource and DBNtarget). This is intented to capture the speaker-individuality information and connect them by the NNs. The numbers of each node from input x to output y are [24 48 24] for DBNsource and DBNtarget, respectively. And the connected NN is a 3-layers model. The whole training process of the DBNs was conducted with the following steps.

1) Train two DBNs for source and target speakers. In the training of DBNs, the hidden units computed as a conditional probability $(P(h|v))$ in Eq. 10 are fed to the following RBMs, and trained layer-by-layer until the highest layer is reached.

2) After pre-training the two DBNs separately, we connect them by the NNs. The weight parameters of NNs are estimated so as

to minimize the error between the output and the target vectors. 3) Finally, the entire network (DBNsource, DBNtarget and NNs) is fine-tuned by back-propagation using the MCC features.

4. Framework of proposed method

Our proposed framework, as shown in Fig. 3, transforms both the excitation and the filter features from the source voice to the target voice. As described in Sec. 2, we extracted spectral features and F0 features from both source voice and target voice by the STRAIGHT and use DTW to align them. We then process the aligned F0 features into CWT-F0 features for NNs and transform the aligned spectral features into the MCC features, respectively. The conversion function training of our proposed method has two parts. One part is the conversion of CWT-F0 using the NNs, the other is the MCC conversion using the DBNs.

For prosody training, we use the 30-dimentional CWT-F0 features for emotional voice features training. To achieve this, we transfered the parallel data which consist of the aligned F0 features of source and target voices to CWT-F0 features. Then use the 4-layers NN models to train the CWT-F0 features. The numbers of nodes from the input layer to output layer are [30 50 50 30]. For spectral features training, we transform aligned spectral features of source and target voices to 24-dimentional MCC features. We then used these MCC features of the source and target voice as the input-layer data and output-layer data for DBNs. Then we connect them by the NNs for deep training. The conversion phase of Fig. 3 shows how our trained conversion function can be applied. The source voice is processed into spectral features and F0 featurs by the STRAIGHT, which are then transformed to MCC and CWT-F0 features, respectively. These features can then be fed into the conversion function to convert the features. Finally, we convert them back to spectrum and F0, and use these features to reconstruct the waveform with STRAIGHT.

5. Experiments

5.1. Experimental Setup

To evaluate the proposed method, we compared the results with several state-of-the-art methods as follows:

- **DBNs+LG:** This system proposed by Nakashika *et al.* converts spectral features by DBNs and converts the F0 features by the logarithm Gaussian method [12], which can be expressed with the following equation:

$$\log\left(f0_{conv}\right) = \mu_{tgt} + \frac{\sigma_{tgt}}{\sigma_{src}}\left(\log\left(f0_{src}\right) - \mu_{src}\right) \tag{14}$$

  where μ_{src} and σ_{src} are the mean and variance of the F0 in logarithm for the source speaker, μ_{tgt} and σ_{tgt} are those for the target speaker. $(f0_{src})$ is the source speaker pitch and $(f0_{conv})$ is the converted fundamental frequency for the target speaker.

- **DBNs+NMF:** Using the DBNs to convert spectral features while using the non-negative matrix factorization (NMF) to convert five-scales CWT-F0 features.

- **DBNs+NNs (proposed method):** This is the proposed system that uses the DBNs to convert spectral features while using the NN to convert the 30-scale CWT-F0 features.

We used a database of emotional Japanese speech constructed in [18]. And the waveforms used were sampled at 16 kHz. Input and output have the same speaker but different emotions. We made the datasets as happy voices to neutral voices, angry voices to neutral voices and sad voices to neutral voices. For each dataset, 50 sentences were chosen as training data and 10 sentences were choosen for evaluation voice.

Table 1: MCD and F0-RMSE results for different emotions. A2N, S2N and H2N represent the datasets angry to neutral voice, sad to neutral voice and happy to neutral voice, respectively.

	MCD			F0-RMSE		
	A2N	S2N	H2N	A2N	S2N	H2N
Source	6.03	5.18	6.30	76.8	73.7	100.4
DBNs+LG	5.47	4.77	5.92	76.1	73.5	85.2
DBN+NMF	5.46	4.78	5.93	69.4	66.9	74.3
DBN+NN	5.47	4.77	5.93	61.6	64.2	75.9

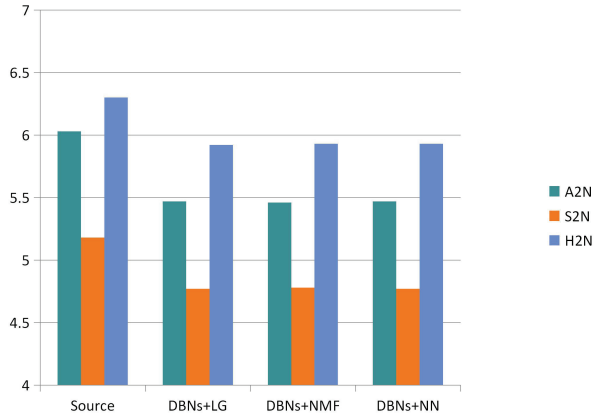

Figure 4: Mel-cepstral distortion evaluation of spectral features conversion

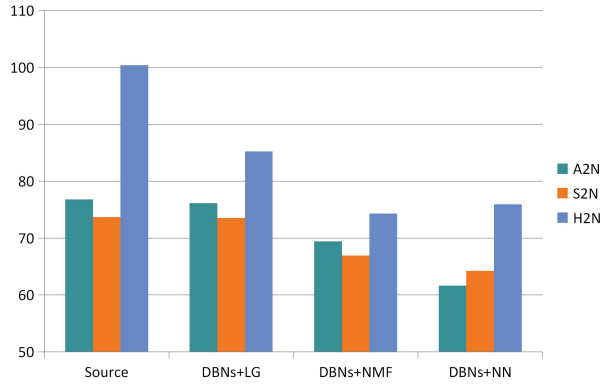

Figure 5: Root mean squared error evaluation of F0 features conversion

5.2. Objective Experiment

Mel cepstral distortion (MCD) was used for the objective evaluation of spectral conversion, which is defined as:

$$MCD = (10/\ln 10)\sqrt{2\sum_{i=1}^{24}(mc_i^t - mc_i^c)^2} \qquad (15)$$

where mc_i^t and mc_i^c represent the target and the converted mel-cepstral, respectively.
To evaluate the F0 conversion, we used the Root Mean Squar Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((F0_i^t) - (F0_i^c))^2} \qquad (16)$$

where $F0_i^t$ and $F0_i^c$ denote the target and the converted F0 features, respectively. A lower MCD and F0-RMSE value indicate smaller distortion or predicting error. Unlike the RMSE evaluation function used in [16], which evaluated the F0 conversion by calculating logarithmic scaled F0, we used original target F0 and converted F0 for calculating the RMSE values. Since our RMSE function evaluates complete sentences that contain both voiced and unvoiced F0 features instead of the voiced logarithmic scaled F0, the RMSE values will be high. For emotional voices, the unvoiced features also include some emotional information. Therefore, we choose the F0 of complete sentences for evaluation instead of the voiced logarithmic scaled F0.

The average MCD and F0-RMSE results over all evaluation pairs are reported in Table 1. The MCD results are presented in the left part of Table 1. Comparing DBNs with source, DBNs decrease the the value of MCD. As shown in Fig. 4, among DBN+LG, DBN+NMF and DBN+NN, MCD decreases or increases slightly, it proves that the conversion of F0 does not affect the spectral features conversion too much. The F0-RMSE results are presented in the right part of Table 1. As shown in Table 1 and Fig. 5, the conventional linear conversion logarithm Gaussian can affect the conversion of happy voice to neutral, but affect slightly on the conversion of angry voice and sad voice to neutral voice. The NMF method and proposed method can both affect the conversion of all emotional voice datasets, and the proposed method can get a better conversion result as a whole.

Fig. 6 shows the example of source emotion F0, Fig. 7 and Fig. 8 show the target F0 and converted F0, respectively. Here, we can see that after converted by the proposed method, F0 is much similar to the tareget neutral vocie.

5.3. Subjective Experiment

We conducted a subjective emotion evaluation by a mean opinion score test. The opinion score was set to a five-point scale (the emotion of sample voice sounded more similar to the target speech and different from source speech, the larger point will be given). In each test, 50 utterances (10 for source speech, 10 for target speech and 30 for converted speech by each method) are selected and 10 listeners are involved. Each subject listened to source and target speech. Then the subject listened to the speech converted by the three methods and give the point to them. As shown in Table 2 and Fig. 5, the angry voice to neutral voice and sad voice to neutral voice can obtain a better result than the happy voice to neutral voice by the method DBN-NMF and DBN-NN. But, the conventional Gaussian method is proved to be poorly in conversion of angery voice to neutral voice, and the DBN-NN(proposed method) obtained a better score than the other two methods in each emotional voice conversion.

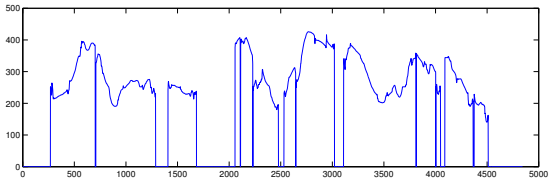

Figure 6: Example of F0 spoken with source anger emotion

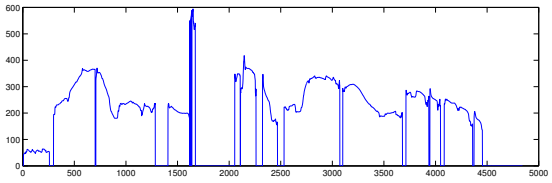

Figure 7: Example of F0 spoken with target neutral emotion

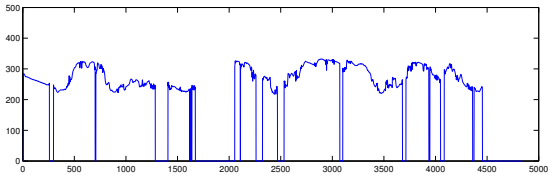

Figure 8: Example of converted F0

Table 2: MOS results for different emotions. A2N, S2N and H2N represent the datasets angry to neutral voice, sad to neutral voice and happy to neutral voice, respectively.

	A2N	S2N	H2N
DBNs+LG	2.03	2.63	2.76
DBN+NMF	3.37	3.02	2.94
DBN+NN	3.57	3.59	3.40

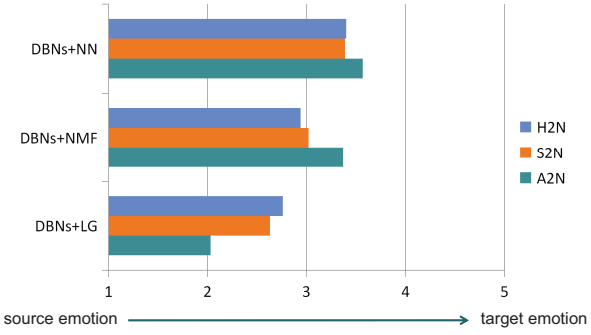

Figure 9: MOS evaluation of emotional voice conversion

6. Conclusions and future work

In this paper, we proposed a method using DBNs to train the MCC features to construct mapping relationship of the spectral envelopes, while using NNs to train the CWT-F0 features which are conducted by the F0 features for prosody conversion between source and target speakers. Comparison between the proposed method and the conventional methods (logarithm Gaussian, NMF) have shown that our proposed model can effectively change the acoustic and the prosody for the emotional voice at the same time. In this paper, we only coverted the emotional voices to neutral voices and the model needs to conduct the parallel speech data which will limit the conversion only one to one. In the future work, we will do experiments about neutral to emotional voices conversion. Also, there are researches using the raw waveforms for deep neural networks training [19] [20]. We will apply the new DBNs model which can straightly use the raw waveform features. It will let the emotional voice conversion model be widely used for practical applications in the future.

7. References

[1] S. Mori, T. Moriyama, and S. Ozawa, "Emotional speech synthesis using subspace constraints in prosody," *in ICME*, pp. 1093–1096, 2006.

[2] R. Aihara, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders using dictionary selective non-negative matrix factorization," *in SLPAT*, pp. 29–37, 2014.

[3] J. Krivokapić, "Rhythm and convergence between speakers of american and indian english," *Laboratory Phonology*, vol. 4, no. 1, pp. 39–65, 2013.

[4] T. Raitio, L. Juvela, A. Suni, M. Vainio, and P. Alku, "Phase perception of the glottal excitation of vocoded speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] Z.-W. Shuang, R. Bakis, S. Shechtman, D. Chazan, and Y. Qin, "Frequency warping based on mapping formant parameters," in *Ninth International Conference on Spoken Language Processing*, 2006.

[6] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion." *in Interspeech*, pp. 1965–1968, 2007.

[7] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[8] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[9] R. Takashima, T. Takiguchi, and Y. Ariki, "Exemplar-based voice conversion in noisy environment," *in Spoken Language Technology Workshop (SLT)*, pp. 313–317, 2012.

[10] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *in ICASSP*, pp. 137–140, 1992.

[11] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," *in ICASSP*, pp. 3893–3896, 2009.

[12] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets." *in INTERSPEECH*, pp. 369–372, 2013.

[13] H. Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical science and technology*, vol. 27, no. 6, pp. 349–353, 2006.

[14] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," *in Fuzzy Systems and Knowledge Discovery*, vol. 4, pp. 410–414, 2007.

[15] M. S. Ribeiro and R. A. Clark, "A multi-level representation of f0 using the continuous wavelet transform and the discrete cosine transform," *in ICASSP*, pp. 4909–4913, 2015.

[16] H. Ming, D. Huang, M. Dong, H. Li, L. Xie, and S. Zhang, "Fundamental frequency modeling using wavelets for emotional voice conversion," *in Affective Computing and Intelligent Interaction (ACII)*, pp. 804–809, 2015.

[17] M. Vainio, A. Suni, D. Aalto *et al.*, "Continuous wavelet transform for analysis of speech prosody," *in TRASP 2013-Tools and Resources for the Analysys of Speech Prosody*, 2013.

[18] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis." *IEEE Trans Speech Audio Proc*, pp. 2401–2404, 2003.

[19] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] M. Bhargava and R. Rose, "Architectures for deep neural network based acoustic models defined over windowed speech waveforms," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.