INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Emotional Voice Conversion with Adaptive Scales F0 based on Wavelet Transform using Limited Amount of Emotional Data

Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, Yasuo Ariki

Graduate School of System Informatics, Kobe University, Japan
{luozhaojie, ianchen}@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

Deep learning techniques have been successfully applied to speech processing. Typically, neural networks (NNs) are very effective in processing nonlinear features, such as mel cepstral coefficients (MCC), which represent the spectrum features in voice conversion (VC) tasks. Despite these successes, the approach is restricted to problems with moderate dimension and sufficient data. Thus, in emotional VC tasks, it is hard to deal with a simple representation of fundamental frequency (F0), which is the most important feature in emotional voice representation, Another problem is that there are insufficient emotional data for training. To deal with these two problems, in this paper, we propose the adaptive scales continuous wavelet transform (AS-CWT) method to systematically capture the F0 features of different temporal scales, which can represent different prosodic levels ranging from micro-prosody to sentence levels. Meanwhile, we also use the pre-trained conversion functions obtained from other emotional datasets to synthesize new emotional data as additional training samples for target emotional voice conversion. Experimental results indicate that our proposed method achieves the best performance in both objective and subjective evaluations.

Index Terms: voice conversion, F0 features, emotion, continuous wavelet transform, deep learning

1. Introduction

Voice conversion (VC) has been widely used in many speech processing tasks, such as speaking assistance [1], speech enhancement [2] and other applications [3], [4]. Therefore, the need for this type of technology in various fields has continued to propel related studies each year. Recently, deep learning has dramatically improved the performance of VC systems through learning hierarchies of features optimized for the task at hand. However, deep learning models are restricted to problems with moderate dimensions and sufficient data, so most deep learning-based VC works focus on the conversion of spectral features, which mainly affect the voice acoustics of a voice, rather than on the conversion of F0 features, which mainly affect the prosody of a voice, because F0 features extracted from STRAIGHT [5] are low-dimensional features that cannot be processed well by deep learning models. One example of deeper VC methods is proposed by Desai *et al.* [6] based on Neural Networks (NNs). Nakashika *et al.* [7] also proposed a VC method using speaker-dependent restricted Boltzmann machines (RBMs) or deep belief networks (DBNs [8]) to achieve non-linear deep transformation. F0 features are usually converted by logarithm Gaussian (LG) normalized transformation [9] in these models.

As mentioned above, in VC tasks, the spectral and F0 features can affect the voices acoustic and prosodic features, respectively. The prosody plays an important role in conveying various types of non-linguistic information, such as the identity, intention, attitude, and mood, which represent the emotions of the speaker. However, previous studies have shown that prosody conversion is affected by both short- and long-term dependencies, such as the sequence of segments, syllables, and words within an utterance, as well as lexical and syntactic systems of a language [10]. The LG-based method is insufficient to convert prosody effectively owing to constraints of their linear models and low-dimensional F0 features. Recently, it has been shown that CWT can effectively model F0 in different temporal scales and significantly improve the speech synthesis performance [11]. For this reason, Suni *et.al.* [12] applied CWT for intonation modeling in hidden Markov model (HMM) speech synthesis. Ming *et.al.* used CWT in F0 modeling within the NMF model [13] or DBLSTM model [14] for emotional VC, and our earlier work [15] also decomposed the F0 into 30 temporal scales containing more specifics of different temporal scales by CWT, and trained them with NN models while using DBNs to train the spectral features.

In this paper, inspired by the deep learning models' ability to perform well in complex nonlinear feature conversion [7] and CWT's ability to improve F0 features conversion [13], we propose a novel method that uses adaptive scales CWT (AS-CWT) to decompose F0 to several scales and train them by NNs. Different from the research [14] or [15], which decomposed the F0 by 10 discrete scales, each one octave apart, or more scales up to 30, each one third octave apart, this approach systematically captures the F0 features of different temporal scales by adaptive scales, which can then represent different prosodic levels ranging from micro-prosody to the sentence levels, but better optimized. Moreover, to overcome the difficulty of a limited amount of training data, we also propose the use of an adaptive method, which enables us to synthesize new data along the conversion function pre-trained by other emotional data-sets. For instance, when performing the emotion conversion from an angry voice to a neutral voice, we can process an additional angry voice in advance by converting other data, such as happy and sad voices, to an angry voice. Given that the DBNs can effectively perform spectral envelope conversion, we use MCC features to train the spectral conversion function with DBNs proposed by Nakashika *et.al.* [7]. We chose different models to separately convert the spectral features and F0 feature. This is because, although the wavelet transform decomposed F0 features to more complex features, they can be trained adequately by NNs, whereas the more complex spectral features require a deeper architecture.

In the remaining part of this paper, we describe our AS-CWT method in Section 2. The training models used in our proposed method are introduced in Section 3. Section 4 gives the detailed stages process of experimental evaluations, and Section 5 presents our conclusions.

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-984

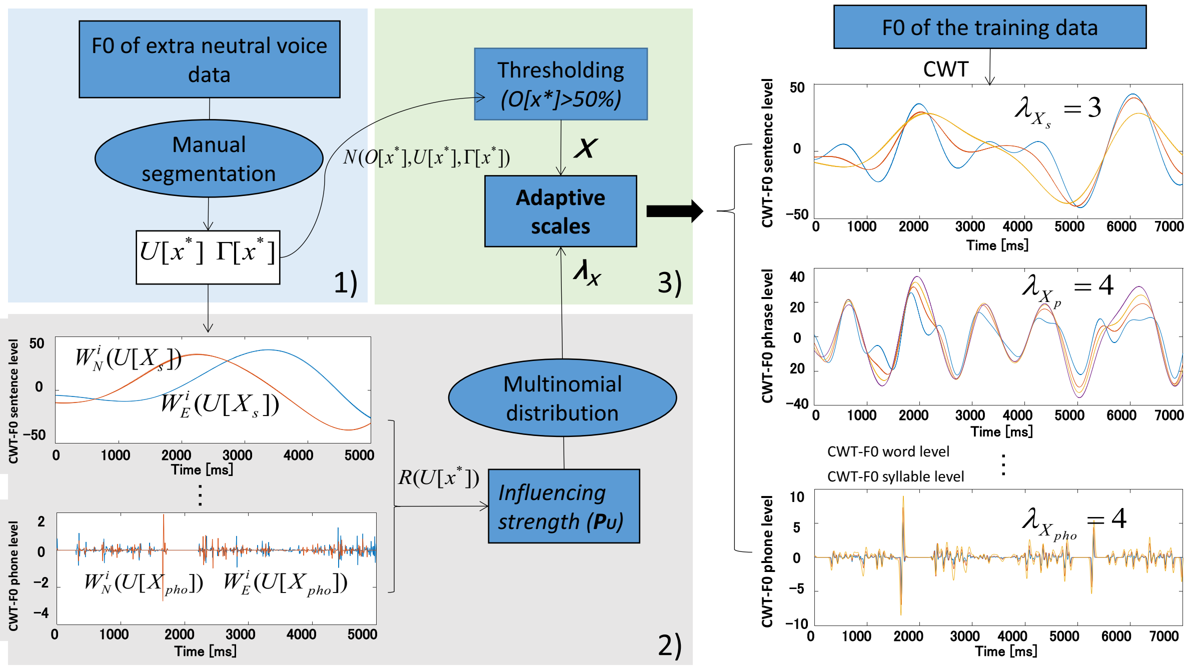

Figure 1: *Illustration of calculating the adaptive scales CWT and using them to decompose the F0 features. The left part of the figure shows the three main steps of calculating the adaptive scales, and the right part shows the samples of CWT-F0 features decomposed by adaptive scales CWT.*

2. Adaptive Scales CWT

In our earlier work [15], we adopted CWT to decompose the F0 contour into 30 temporal scales before training the F0 features using NNs. The decomposed 30-dimensional features are linearly spaced scales, each separated by one-third of an octave. However, only the features that can represent the utterance, phrase, word, syllable, and phone levels are useful for training. Thus, in the current paper, we apply an adaptive scales method to decompose F0 features by wavelet transform before training them. As shown in the left part of Figure 1, there are three main steps in calculating the adaptive scales. 1) Calculate the optimized duration for each temporal level using the extra data. 2) We investigate the variability in each temporal level as a rich source of information for studying the degree of impact of every level in emotion conversion as a function of $influencing\ strength$, and 3) calculate adaptive scales with the $influencing\ strength$ and optimized duration of each temporal level obtained in 1) and 2). The steps for processing details are described below.

1) In order to find the optimized duration of sentence, phrase, and word levels, we first perform segmentation in the extra neutral voice data. As shown in Figure 2, means and standard deviations of the duration of the sentence, phrase, and word can be calculated from the pre-segmented data. We denote by $U[x^*]$ and $\Gamma[x^*]$ the mean and standard deviation of duration of each temporal level x^*, $x^* \in X$, and X is the set $\{X_s, X_p, X_w, X_{syl}, X_{pho}\}$, which represents the duration of five temporal levels. According to [16], the average duration of non-emphasized syllables was found to be $50ms$ to $180ms$, and that of phone levels was $20ms$ to $40ms$. Therefore, we set the mean of the syllable level $U[X_{syl}]$ to $115ms$, the middle values between $50ms$ and $180ms$, and phone level $U[X_{pho}]$ to $30ms$. The standard deviation $\Gamma[X_{syl}]$ is set to $65ms$ and $\Gamma[X_{pho}]$ is $10ms$.

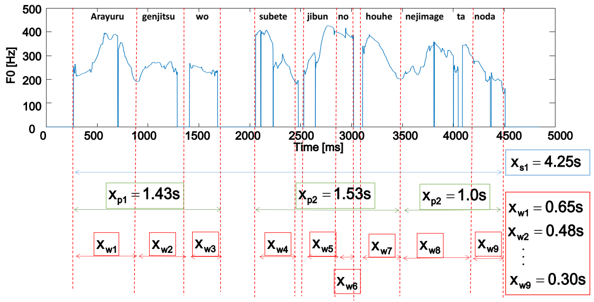

Figure 2: *Example of performing segmentation in the training data. Here, X_s, X_p and X_w represent the durations of sentence, phrase and word, respectively.*

2) Next, we calculate each temporal level's which can represent the proportion of the influence among all the temporal levels in the emotional VC. We first calculate the relative distance between the emotional voice and neutral voice in each temporal level as shown below:

$$R(U[x^*]) = \frac{\sqrt{\sum_{i=1}^{n}(W_E^i(U[x^*]) - W_N^i(U[x^*]))^2}}{n * U[x^*]} \quad (1)$$

where the mean $U[x^*]$ of each level is obtained in the first step, and n is the number of training data in each emotional voice data set. $W_E^i(U[x^*])$ and $W_N^i(U[x^*])$ represent the continuous wavelet transform function of F0 using the emotional and neutral input signal, in different temporal level x^*. The transform functions are defined by

$$W_E^i\left(U[x^*]\right) = \tau^{-1/2}\int_{-\infty}^{\infty} F_{E0}\left(x_i\right)\psi\left(\frac{x-U[x^*]}{\tau}\right)dx$$
$$W_N^i\left(U[x^*]\right) = \tau^{-1/2}\int_{-\infty}^{\infty} F_{N0}\left(x_i\right)\psi\left(\frac{x-U[x^*]}{\tau}\right)dx \tag{2}$$

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

where $\tau = 1ms$, ψ is the Mexican hat wavelet, $F_{E0}\left(x_i\right)$ and $F_{N0}\left(x_i\right)$ represent the emotional and neutral input signal, respectively. Then, the $influencing\ strength$ of each temporal level can be ranked by

$$P_{U[x^*]} = \frac{R(U[x^*])}{\sum_{x^*\in X} R(U[x^*])} \tag{4}$$

Then, we can draw the optimized number of scales for CWT in each temporal level with the $influencing\ strength$ from a multinomial distribution:

$$\begin{aligned} \lambda_{\mathbf{X}} &\sim Multinomial(N, \mathbf{P_U}) \\ \lambda_{x^*} \in \lambda_{\mathbf{X}} &= (\lambda_{X_s}, \lambda_{X_p}, \lambda_{X_w}, \lambda_{X_{syl}}, \lambda_{X_{pho}}) \\ P_{U[x^*]} \in \mathbf{P_U} &= (P_{U[X_s]}, P_{U[X_p]}, P_{U[X_w]}, P_{U[X_{syl}]}, P_{U[X_{pho}]}) \end{aligned} \tag{5}$$

where N is the total number of scales, which can be set in different values, vectors $\mathbf{P_U}$ are made up of all the $influencing\ strength$s, and $\lambda_{\mathbf{X}}$ represents the number of scales in all the temporal levels. Therefore, the λ_{x^*} can represent the number of scales in each temporal level.

3) The third step is using the $influencing\ strength$ and optimized duration to calculate the adaptive scales of each temporal level. First, we use the Gaussian function to separately calculate the probability densities of the duration in each temporal level using

$$O[x^*] = N(O[x^*], U[x^*], \Gamma[x^*]) \tag{6}$$

where $O[x^*]$ represents the probability density of duration in each temporal level. Then, we set a threshold to draw the valid values x^*, when probability density $O[x^*]$ is over 50%. The optimized duration can then be represented by

$$\begin{aligned} D(\mathbb{I}_{x^*}) = min(x^*) + \frac{max(x^*) - min(x^*)}{\lambda_{x^*}} * \mathbb{I}_{x^*} \\ \mathbb{I}_{x^*} = (0, ..., \lambda_{x^*}) \end{aligned} \tag{7}$$

where λ_{x^*} represents the optimum number of scales for CWT in each temporal level calculated in Eq. 5, and x^* is the valid value of duration in each temporal level. Finally, the adaptive scales can then be represented by

$$\theta_{\mathbb{I}_{x^*}} = \log_2(D(\mathbb{I}_{x^*})/\tau_0) \tag{8}$$

where $\tau_0 = 1ms$. After calculating the scales that can model prosody at different temporal levels, we adopt CWT to decompose the F0 contour with these adaptive scales and our F0 is represented by separate components given by

$$W_{\theta_{\mathbb{I}_{x^*}}}(f_0)(t) = W_{\theta_{\mathbb{I}_{x^*}}}(f_0)(2^{\theta_{\mathbb{I}_{x^*}}+1}\tau_0, t)\left(\theta_{\mathbb{I}_{x^*}} + 2.5\right)^{-5/2} \tag{9}$$

The original signal is approximately recovered by

$$f_0 = \sum_{\mathbb{I}_{x^*}=0}^{\lambda_{x^*}} \sum_{x^*\in X} W_{\theta_{\mathbb{I}_{x^*}}} f_0(t)(\theta_{\mathbb{I}_{x^*}} + 2.5)^{-5/2} + \epsilon(t) \tag{10}$$

where $\epsilon(t)$ is the reconstruction error.

3. Training Model

The conversion function training of our proposed method has two stages. The first stage is the MCC conversion using the DBNs, the other is the conversion of F0 features using the NNs. In the first stage, we apply the training model used in our earlier work [15] that first transformed aligned spectral features of source and target voices to 24-dimensional MCC features. Then, train these MCC features by the 7-layers DBNs. In the second stage, we used the high-dimension F0 features for prosody features training. To achieve this, we transfer the parallel data consisting of the aligned F0 features of the source and target voices to CWT-F0 features by using the AS-CWT method. Then we used the 4-layer NN models to train the CWT-F0 features. Neural networks 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. The learning problem is to find an optimized mapping function $G_{E\to N}$ that satisfies

$$\underset{G_{E\to N}}{\arg\min} \ \|G_{E\to N}(X_E) - Y_N\|^2 \tag{11}$$

where, X_E represents the input CWT F0 features, and Y_N is the target CWT F0 features. However, to train such a regression model, a large corpus with different emotions is required. For this paper's scope with only a limited amount of emotional voice data, NNs may suffer from an insufficient amount of training data, leading to poor performance. To address the problem, we propose a NNs model using the other emotional data sets to synthesize new emotional data as additional training samples for target emotional voice conversion. An example of converting angry voice to neutral voice can be formulated as follows:

$$\begin{aligned} &\underset{G_{N\to A}}{\arg\min} \ \|G_{N\to A}(X_N) - Y_A\|^2 \\ &\underset{G_{S\to A}}{\arg\min} \ \|G_{S\to A}(X_S) - Y_A\|^2 \\ &\underset{G_{H\to A}}{\arg\min} \ \|G_{H\to A}(X_H) - Y_A\|^2 \\ X_R &= [G_{N\to A}(X_N), G_{S\to A}(X_S), G_{H\to A}(X_H)]^T \end{aligned} \tag{12}$$

where Y_A represents the anger voice data set, and X_N, X_S and X_H represent the input neutral, sad, and happy voice data sets, respectively. Thus, $G_{N\to A}$, $G_{S\to A}$ and $G_{H\to A}$ represent the networks that are trained for converting the other voice datasets to an angry voice data set. X_R represents the synthesized new angry voice data. Then, we concatenated X_A with the synthesized angry voice data X_R in Eq. 12 to calculate the conversion function with the goal of converting the angry voice to a neutral voice as shown below:

$$\underset{G_{A\to N}}{\arg\min} \ \|G_{A\to N}\begin{pmatrix} X_R \\ X_A \end{pmatrix} - Y_N\|^2 \tag{13}$$

Other emotional voice conversion can also be conducted by the proposed method using pre-trained conversion functions to synthesize new data as additional training samples for target voice conversion. Since there are sufficient neutral voice data, there is no need to synthesize the neutral voice in the proposed method.

4. Experiments

We used a database of emotional Japanese speech constructed in a previous study [17]. The waveforms used were sampled at 16 kHz. Input and output data had the same speaker but expressing different emotions. We classified the six data sets into the following: happy to neutral voices, angry to neutral voices, and sad to neutral voices, as well as their inverse conversion from neutral voices to each emotion voices. For each data set, 50 sentences were chosen as training data and 10 sentences were chosen for the VC evaluation.

To evaluate the proposed method, we compared the results with several state-of-the-art methods listed below.

- **DBNs+LG:** This system proposed by Nakashika *et al.* converts spectral features using DBNs, and converts the F0 features through the LG method [7].
- **NMF:** Using non-negative matrix factorization (NMF) to convert five-scale CWT-F0 features.
- **DBNs+CWT:** Our previous work [15] that uses DBNs to convert spectral features while using the NNs to convert the 30-scale CWT-F0 features.
- **AS-CWT:** This is the proposed system that uses DBNs to convert spectral features while using NNs to convert the CWT-F0 features decomposed by AS-CWT method.

4.1. Objective Experiment

To evaluate the F0 conversion, we used the root-mean-square error (RMSE),

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((F0_i^t) - (F0_i^c))^2} \qquad (14)$$

where $F0_i^t$ and $F0_i^c$ denote the target and the converted F0 features, respectively. A lower F0-RMSE value indicates smaller predicting error. Unlike the RMSE evaluation function used in [13], which evaluated the F0 conversion by calculating logarithmic scaled F0, we used the original target F0 and converted the F0 to calculate the RMSE values. Given that our RMSE function evaluates complete sentences that contain both voiced and unvoiced F0 features instead of the voiced logarithmic scaled F0, the RMSE values are expected to 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 F0-RMSE results from emotional to neutral pairs and their inverse conversion are reported in Table 1. As shown in Table 1, the conventional linear conversion LG can affect the conversion of happy to neutral, but only slightly affect the conversion of angry voices and sad voices to neutral voices. The NMF method, previously proposed CWT method, and the new proposed AS-CWT method can affect the conversion of all emotional voice datasets. In addition, the proposed method can obtain significant improvement in F0 conversion as a whole.

Table 1: *F0-RMSE results for different emotions. A2N, S2N and H2N represent the datasets angry , sad and happy voice to neutral voice, respectively. N2A, N2S and N2H represent their inverse conversion*

	E2N			N2E		
	A2N	S2N	H2N	N2A	N2S	N2H
Source	76.8	73.7	100.4	76.8	73.7	100.4
DBNs+LG	76.1	73.5	85.2	76.3	72.0	99.3
NMF	69.4	66.9	74.3	70.4	62.3	75.2
DBNs+CWT	61.6	**62.2**	75.9	39.5	40.1	64.5
AS-CWT	**51.2**	64.1	**64.4**	**37.8**	**35.9**	**62.1**

4.2. Subjective Experiment

We conducted subjective evaluations using a 5-scale MOS test. The opinion score was set to a five-point scale (the more similar to the emotion of the sample voice the target speech sounded, the larger the point given). Here, we tested the emotional to neutral pairs (H2N, S2N, A2N) and their inverse conversion (N2H, N2S, N2A). In each test, 50 utterances (10 for source speech, 10 for target speech, and 30 for converted speech by the three methods) were selected, and 10 listeners were involved. Each subject listened to the source and target speeches. Then, the subject listened to the speech converted using the three methods and asked to give each conversion a point. Figure 3 shows the result of MOS test, the error bar shows the 95% confidence interval. As the figure shown, the conventional LG method shows poor performance in the conversion of anger voice to neutral voice. The AS-CWT (proposed method) obtained a better score than the LG method and NMF in every emotional VC. The difference between AS-CWT and CWT is not statistically significant when dealing with the conversion from emotional voice to netural voice, but obtained a better score when coverting the neutral voice to emotional voice.

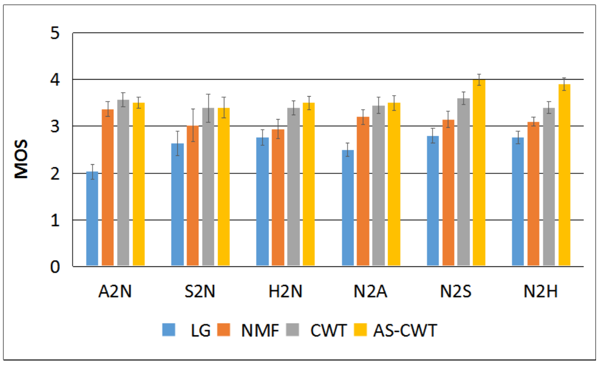

Figure 3: *MOS evaluation of emotional voice conversion*

5. Conclusions

In this paper, we propose the adaptive scales continuous wavelet transform (AS-CWT) method to systematically capture the F0 features of different temporal scales. Meanwhile, we also use the pre-trained conversion functions to synthesize new emotional data as additional training samples for target emotional voice conversion. A comparison between the proposed method and the conventional methods (logarithm Gaussian, NMF) shows that our proposed model can effectively change the prosody of the emotional voice.

6. Acknowledgements

This work was supported in part by PRESTO, JST under grant JPMJPR15D2.

7. References

[1] 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.

[2] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," vol. 1, pp. 285–288, 1998.

[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] 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.

[6] 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.

[7] 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.

[8] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[9] 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.

[10] 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.

[11] 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.

[12] A. S. Suni, D. Aalto, T. Raitio, P. Alku, M. Vainio *et al.*, "Wavelets for intonation modeling in hmm speech synthesis," in *8th ISCA Workshop on Speech Synthesis, Proceedings, Barcelona, August 31-September 2, 2013*, 2013.

[13] 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.

[14] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional lstm modeling of timbre and prosody for emotional voice conversion," *in Interspeech*, 2016.

[15] Z. Luo, J. Chen, T. Nakashika, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with different temporal scales of f0 based on wavelet transform," in *9th ISCA Speech Synthesis Workshop*, pp. 140–145.

[16] T. Toda *et al.*, "Interlanguage phonology: Acquisition of timing control and perceptual categorization of durational contrast in japanese," 2013.

[17] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "GMM-based voice conversion applied to emotional speech synthesis." *In: Proc. European Conf. on Speech Communication and Technology (Eurospeech 03)*, pp. 2401–2404, 2003.