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

☆ Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki (Kobe University)

1 Introduction

Recently, deep learning has dramatically improved the performance of voice conversion (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 acoustics of a voice, rather than on the conversion of fundamental frequency (F0) features, which mainly affect the prosody of a voice, because F0 features extracted from STRAIGHT are low-dimensional features that cannot be processed well by deep learning models. As mentioned above, in VC tasks, the spectral and F0 features can affect the voice ’ s 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-term 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. And it has been shown that CWT can effectively model F0 in different temporal scales and significantly improve the speech synthesis performance [1]. For this reason, our earlier work [2] decomposed the F0 into 30 temporal scales features containing more specifics of different temporal scales by CWT, and trained them with NN models.

In this paper, we propose a novel method that 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 than the earlier method [2]. Moreover, to overcome the difficulty of a limited amount of training data, we also propose an adaptive training model, 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.

2 Adaptive Scales CWT

In our earlier work [2], 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 calculate means and standard deviations of the duration of sentence, phrase, and word levels, we first perform segmentation in the extra neutral voice 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 [3], the average duration of non-emphasized syllables was found to be 50ms and 180ms, and that of phone levels was 20ms to 40ms. Therefore, we set the mean

[*] This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2) and JSPS KAKENHI (Grant No. 17H01995).

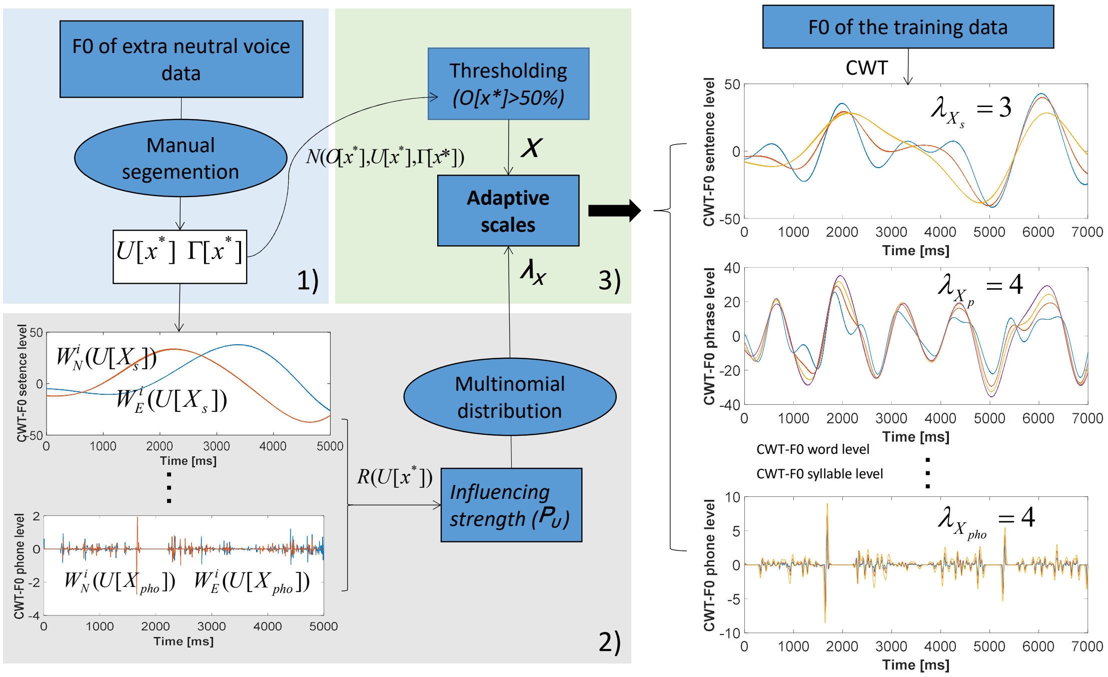

Fig. 1 Illustration of calculating the adaptive scales CWT and using them to decompose the F0 features. The left part of the figure shows the three main steps of calculating the adaptive scales, and the right part shows the samples of CWT-F0 features decomposed by adaptive scales CWT.

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.

2) Next, we calculate each temporal level's influencing strength 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:

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

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

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

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

where $\tau_0 = 1ms$, 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:

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

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 strengths*, 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 calculated 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

$$D(\mathbb{I}_{x^*}) = min(x^*) + \frac{max(x^*) - min(x^*)}{\lambda_{x^*}} * \mathbb{I}_{x^*}$$

$$\mathbb{I}_{x^*} = (0, ..., \lambda_{x^*}) \tag{7}$$

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

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

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

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

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 CWT-F0 using the NNs. In the first stage, we apply the training model used in our earlier work [2] that first transformed aligned spectral features of source and target voices to 24-dimensional MCC features. Then, we used these MCC features of the source and target voice as the input-layer data and output-layer data for the DBNs. Finally, we connected them using NNs for deep training. In the second stage, we used the high-dimension CWT-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

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

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. The method can be formulated as follows:

$$\arg\min_{G_{N\to A}} \ \|G_{N\to A}(X_N) - Y_A\|^2$$

$$\arg\min_{G_{S\to A}} \ \|G_{S\to A}(X_S) - Y_A\|^2$$

$$\arg\min_{G_{H\to A}} \ \|G_{H\to A}(X_H) - Y_A\|^2$$

$$X_R = [G_{N\to A}(X_N), G_{S\to A}(X_S), G_{H\to A}(X_H)]^T \tag{12}$$

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 functionwith the goal of converting the angry voice to a neutral voice as shown below:

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

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. 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.

- **LG (M1):** This system proposed by Nakashika *et al.* converts spectral features using DBNs, and converts the F0 features through the LG method.
- **NMF (M2):** Using non-negative matrix factorization (NMF) to convert five-scale CWT-F0 features.
- **CWT (M3):** This is our previous work [2] that uses DBNs to convert spectral features while using the NNs to convert the 30-scale CWT-F0 features.
- **AS-CWT (M4 proposed method):** 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), A lower F0-RMSE value indicates smaller predicting error. 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 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
M1	76.1	73.5	85.2	76.3	72.0	99.3
M2	69.4	66.9	74.3	70.4	62.3	75.2
M3	61.6	**62.2**	75.9	39.5	40.1	64.5
M4	**51.2**	64.1	**64.4**	**37.8**	**35.9**	**62.1**

5 Conclusions

In this paper, we propose the 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.

References

[1] M. Vainio *et al.*, “Continuous wavelet transform for analysis of speech prosody,” in TRASP 2013-Tools and Resources for the Analysys of Speech Prosody, 2013.

[2] Z. Luo *et al.*, “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.

[3] T. Toda et al., “Interlanguage phonology: Acquisition of timing control and perceptual categorization of durational contrast in japanese,” 2013.