Arbitrary-scales continuous wavelet transform for emotional voice conversion *

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

1 Introduction

Recently, the study of voice conversion (VC) has attracted wide attention in the field of speech processing. Many statistical approaches have been proposed for spectral conversion in the past few decades. Among these approaches, a Gaussian Mixture Model (GMM) has been commonly used. Other VC methods, such as those based on non-negative matrix factorization (NMF), have also been proposed. The NMF and GMM methods are based on linear functions. For better VC performance, the VC technique needs to train more complex nonlinear features, some approaches construct non-linear mapping relationships using neural networks (NNs) to train the mapping dictionaries between the source and target features [1], whereas others use deep belief networks (DBNs) to achieve non-linear deep transformation [2]. Results have shown that these deep architecture models have better performance than shallow conversion in some complex voice feature conversion.

In our recent work [3], inspired by the ability of deep learning models to perform well in complex nonlinear feature conversion [2] and the ability of CWT to improve F0 feature conversion, we proposed a novel method that used NNs to train the CWT-F0 for converting the prosody of the emotional voice. In the current paper, we extend our earlier work [3] to systematically capture the F0 features of different temporal scales, which can then represent different prosodic levels ranging from micro-prosody to the sentence levels. We achieve this by using the Arbitrary-scales-CWT(AS-CWT) method to decompose the F0 contour into several temporal scales, which can more approximately represent each level of individual prosodics. Given that the DBNs can effectively perform spectral envelope conversion, we train the MCC features for spectral feature conversion by using DBNs proposed by Nakashika *et.al.* [2]. 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 enough by NNs, whereas the more complex spectral features require a deeper architecture.

In the remainder of this paper, we describe the proposed method in Sec. 2. Sec. 3 gives the detailed stages of process in experimental evaluations and conclusions are drawn in Sec. 4

2 Proposed method

2.1 Feature extraction and processing

The F0 features produced by STRAIGHT are one dimensional and discrete. Modeling the variations of F0 in all temporal scales using linear models is difficult. In this paper, we apply the AS-CWT method to decompose F0 features before training them. The steps for processing details are described below.

1) To explore the perceptually relevant information, F0 contour is transformed from the linear to logarithmic semitone scale, which is referred to as logF0.

2) Next, we calculate the scales of different prosodic levels ranging from sentence level to micro-prosody using AS-CWT method. In order to find the scales of sentence, phrase, and word levels, we first perform segmentation in some sentences of the training data. As shown in Figure 1, the ranges of duration vary in sentence, phrase, and word. We use the Gaussian function to separately calculate the probability densities of the duration in the sentence, phrase, and word using

$$\begin{aligned} S(x) &= N(x, \mu_s, \sigma_s^2) \\ P(x) &= N(x, \mu_p, \sigma_p^2) \\ W(x) &= N(x, \mu_w, \sigma_w^2), \end{aligned} \tag{1}$$

where $S(x)$, $P(x)$ and $W(x)$ represent the probability density of duration in the sentence, phrase and word, separately. The means and standard deviations of the durations in the sentence, phrase, and word are calculated from the pre-segmented training data, as shown in Figure 1. We choose the parts over

* Arbitrary-scales continuous wavelet transform for emotional voice conversion, by Zhaojie Luo, Tetsuya Takiguchi, and Yasuo Ariki (Kobe Univ.)

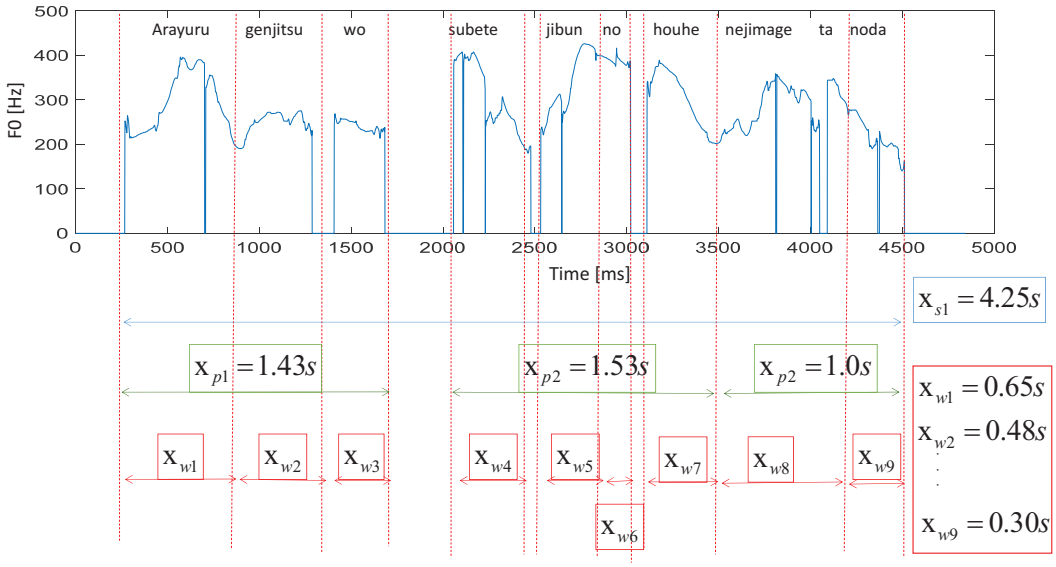

Fig. 1 Example of performing segmentation in the training data. Here, X_s, X_p and X_w represent the durations of sentence, phrase and word, respectively.

60% to process the scales of the sentence, phrase, and word levels. Each temporal duration is defined by

$$\begin{aligned} s_i &= min(x_s) + \frac{max(x_s) - min(x_s)}{\lambda} * i \\ p_i &= min(x_p) + \frac{max(x_p) - min(x_p)}{\lambda} * i \\ w_i &= min(x_w) + \frac{max(x_w) - min(x_w)}{\lambda} * i, \end{aligned} \tag{2}$$

where s_i, p_i and w_i represent the durations of sentence, phrase, and word, respectively; x_s, x_p and x_w are the values when probability densities $S(x)$, $P(x)$, and $W(x)$, respectively, are over 60%; $i = 0, ..., \lambda$; and λ is the number of scales in sentence, phrase, and word. The average duration of non-emphasized syllables was found between $50ms$ and $180ms$ [4], and that of phone level is $20ms$ to $40ms$. Thus, the durations of syllable and phone can be represented as $syl_i = 50 + ((180 - 50)/\lambda) * i$ and $pho_i = 20 + ((40 - 20)/\lambda) * i$, respectively.The scales can then be represented by

$$\begin{aligned} \theta_i &= \log_2(D_i/\tau_0) \\ D_i &\in \{s_i, p_i, w_i, syl_i, pho_i\}, \end{aligned} \tag{3}$$

Where, $\tau_0 = 5ms$, and $\{D_i\}_{i=0,...,\lambda}$ represents all the durations of the sentence, phrase, word, syllable, and phone levels. θ_i represents each scale calculated by Eq. 3.

3) After calculating the scales that can model prosody at different temporal levels, we adopt CWT to decompose the F0 contour with several temporal scales. The continuous wavelet transform of F0 is defined by

$$W(\tau, t) = \tau^{-1/2} \int_{-\infty}^{\infty} f_0(x)\, \psi\left(\frac{x - t}{\tau}\right) dx \tag{4}$$

$$\psi(t) = \frac{2}{\sqrt{3}} \pi^{-1/4} \left(1 - t^2\right) e^{-t^2/2}, \tag{5}$$

where $f_0(x)$ is the input signal and ψ is the Mexican hat wavelet. The original signal f0 can be recovered from the wavelet representation $W(f0)$ by inverse transform:

$$f_0(t) = \int_{-\infty}^{\infty} \int_{0}^{\infty} W(f_0)(\tau, x) \tau^{-5/2} \psi\left(\frac{x - t}{\tau}\right) dx d\tau \tag{6}$$

As described in [5], the reconstruction is incomplete if all information on $W(f0)$ is not available. In that study, the authors performed the decomposition and reconstruction by choosing ten scales, all of which are one octave apart. In our recent work in [3], we decomposed the continuous logF0 with 30 discrete scales, each separated by one third of an octave. Increasing the number of scales can result in a better reconstruction after the decomposition. However, we want to select the features to better represent the utterance, phrase, word, syllable, and phone levels,

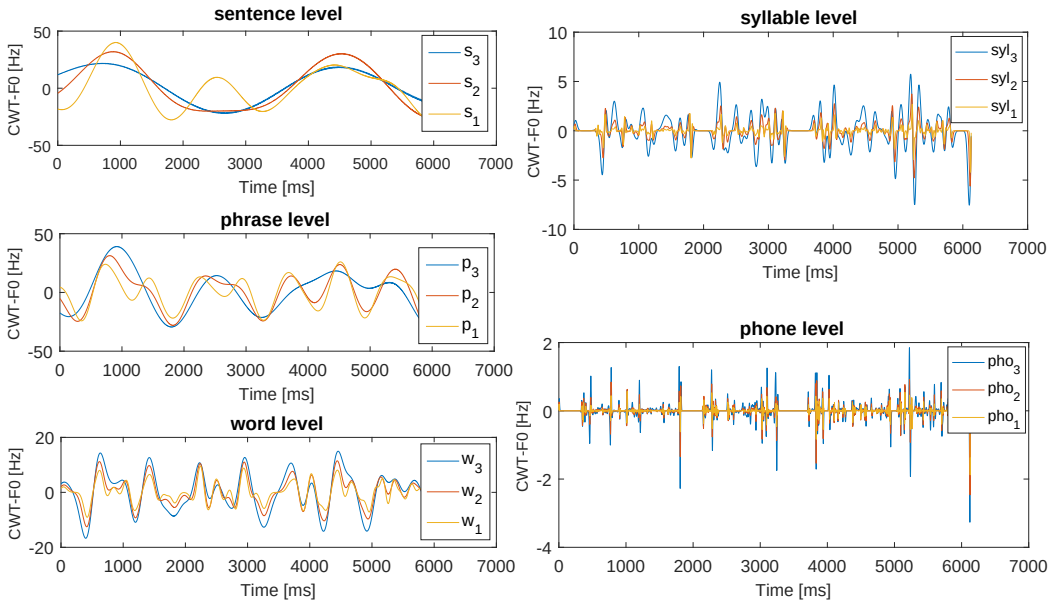

Fig. 2 Example of sentence, phrase, word, syllable, and phone level scales when the number of each level (λ) is set to 3. The red, bule, and yellow curves represent the scales in each level when temporal duration (D_i) is calculated with i=1, i=2, and i=3, respectively.

so we apply the non-linear scales θ_i calculated in Eq. 3, which can better represent the duration (D_i) of all levels of linguistic structure. Therefore, our F0 is represented by separate components given by

$$W_{\theta_i}(f_0)(t) = W_{\theta_i}(f_0)(2^{\theta_i+1}\tau_0, t)\,(\theta + 2.5)^{-5/2}, \quad (7)$$

The original signal is approximately recovered by

$$f_0 = \sum_{i=0}^{\lambda} W_{\theta_i} f_0(t)(\theta_i + 2.5)^{-5/2} + \epsilon(t) \quad (8)$$

where $\epsilon(t)$ is the reconstruction error, and λ represents the number of scales in each temporal level.

2.2 Training model

The conversion function training of our proposed method has two stages. The first stage is the conversion of CWT-F0 using the NNs, the other is the MCC conversion using the DBNs. In the first stage, we used the high dimension CWT-F0 features for prosody features training. To achieve this, we transfered 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-layers NN models to train the CWT-F0 features. The numbers of nodes from the input layer to output layer are [5λ 10λ 10λ 5λ]. In the second stage, we 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 DBNs. Finally, we connected them by the NNs for deep training.

3 Experiments

3.1 Experimental Setup

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 set the three datasets into the following: happy to neutral voices, angry to neutral voices, and sad to neutral voices. For each dataset, 50 sentences were chosen as training data and 10 sentences were choosen for the VC evaluation. As described in Section 2.1, to obtain the optimum numbers of scales (λ) in each temporal level, We evaluated the accuracy of the reconstruction by decomposing and reconstructing several training sentences with different values of λ used in AS-CWT method and find eight scales is the most suitable numbers of sclaes. Hence, we select eight scales in each temporal levels in our proposed model .

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

- **DBNs+LG (M1):** This system converts spectral features by DBNs and converts the F0 features through the LG method [2], which can be expressed with the equation

$$\log\left(f0_{conv}\right) = \mu_{tgt} + \frac{\sigma_{tgt}}{\sigma_{src}}\left(\log\left(f0_{src}\right) - \mu_{src}\right) \tag{9}$$

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

- **DBNs+NMF (M2):** Using the DBNs to convert spectral features while using the non-negative matrix factorization (NMF) to convert five-scale CWT-F0 features.

- **DBNs+CWT (M3):** This is the proposed method in our previous work [3]that uses DBNs to convert spectral features while using the NNs to convert the 30-scales CWT-F0 features; each scale is separated by one third of an octave.

- **DBNs+AS-CWT (proposed method M4):** This is the proposed system that uses the DBNs to convert spectral features while using the NNs to convert the CWT-F0 features decomposed by AS-CWT method, each temporal level has eight scales (8*5 scales in total).

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
M1	5.47	4.77	5.92	76.1	73.5	85.2
M2	5.46	4.78	5.93	69.4	66.9	74.3
M3	5.47	4.77	5.93	61.6	64.2	75.9
M4	5.47	4.77	5.93	51.1	52.1	64.4

3.2 Result and discussion

Mel Cepstral Distortion (MCD) was used for the objective evaluation of spectral conversion, while using Root Mean Squar Error (RMSE) For F0 conversion evaluating. The average MCD and F0-RMSE results are reported in Table 1. Comparing DBNs with source, DBNs decrease the the value of MCD. However, among DBNs+LG, DBNs+NMF, DBNs+CWT and DBNs+AS-CWT, MCD decreases or increases slightly, proving that the conversion of F0 does not have a significant impact on the spectral feature conversion. The F0-RMSE results are presented in the right part of Table 1 has shonw that the proposed method can obtain significant improvement in F0 conversion as a whole.

4 Conclusions

In this paper, we proposed a method using DBNs to train the MCC features, while using NNs to train the CWT-F0 features, which are conducted by the F0 features with arbitrary scales for prosody conversion between the source and target speakers. A comparison between the proposed method and the conventional methods shows that our proposed model can effectively change the acoustic and the prosody for the emotional voice at the same time.

References

[1] S. Desai *et al.*, "Voice conversion using artificial neural networks," in ICASSP, pp. 3893–3896, 2009.

[2] T. Nakashika *et al.*, "Voice conversion in high-order eigen space using deep belief nets.," in INTERSPEECH, pp. 369–372, 2013.

[3] 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*, 2016.

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

[5] A. S. Suni *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.