Speech Prosody Conversion using Sequence Generative Adversarial Nets with Continuous Wavelet Transform F0 Features *

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

1 Introduction

In the voice conversion tasks, the spectral and F0 features can affect the acoustic and prosodic features, respectively. Particularly in emotional VC tasks where the prosody plays an important role in conveying various types of non-linguistic information that represent the mood of the speaker, such as identity, intention and, attitude. Previous studies (*ex.* [1]) have shown that prosody conversion is affected by both short- and long-term dependencies in different temporal levels such as the phones, syllables, and words, within an utterance. The LG-based method is insufficient to convert prosody effectively because of constraints of their linear models and low-dimensional F0 features.

In recent years, it has been shown that the CWT method can effectively model F0 in different temporal scales and significantly improve speech synthesis performance. Our earlier work [2] systematically captures the F0 features of different temporal scales using AS-CWT, which transforms F0 features into high-dimensional CWT-F0 features containing more specifics. Thus, building on top of the success of using CWT-F0 features for prosody conversion, in this study, we want to go one step further to generate emotional voice more similar to target emotion using a generative model.

In this study, inspired by the success of Generative Adversarial Networks (GAN) model in VC tasks, we propose an emotional VC framework that using the GAN model. The effectiveness of GAN is due to the fact that an adversarial loss forces the generated data to be indistinguishable from real data. This is particularly powerful for image generation tasks. So, before training in the GAN model, we segmented the voice features such as the spectral features and CWT-F0 features to the suitable size matrices for training.

However, a generative adversarial model only discriminates between "real" and "fake" features, it has limitations when the goal is for generating sequences affected by both short- and long-term dependencies in different temporal levels. Because, GANs can only give the score/loss for an entire sequence when it has been generated; for a partially generated sequence, such as the word level and syllable level, it is non-trivial to balance how well as it is for word level and the syllable level as the entire sequence. As described above, we segmented the features to suitable size matrices, for normal GAN models, they were converted to the suitable size matrices independently from each other and do not address the continuity of the resulting parameters between matrices. In this paper, we designed sequence GAN architectures to address the continuity of each matrix in a completed sentence, which can train both a supra-segmental level by long-term dependencies and a segmental-level by short-term dependencies.

2 Related work

2.1 Continuous wavelet transform

The continuous wavelet transform of F0 is defined by

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

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

where $f_0(x)$ is the input signal and ψ is the Mexican hat mother wavelet. We decompose the continuous F0 with 32 discrete scales, each one third of an octave apart. Our F0 is thus represented by 32 separate components given by

$$W_i(f_0)(t) = W_i(f_0)(2^{(i/3)+1}\tau_0, t) \quad (3)$$

where i = 1, ..., 32 and τ_0=1 ms. Fig. 1 shows several CWT-F0 feature examples of decomposed components, which can represent the utterance, phrase, word, syllable, and phone levels, respectively.

---

*Sequence Generative Adversarial Networks を用いた感情音声変換, 羅兆傑, 滝口哲也, 有木康雄 (神戸大)

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2) and JSPS KAKENHI (Grant No.JP17H01995).

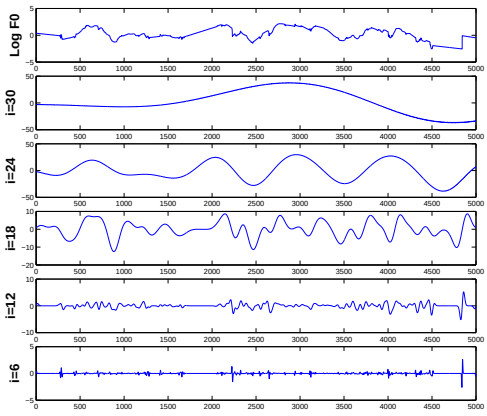

Fig. 1 Interpolated log-normalized F0 and five wavelet transforms (i=30, i=24, i=18, i=12, i=6)

2.2 Generative Adversarial Networks

GAN has obtained impressive results for image generation. The key to the success of the GAN is learning a generator distribution $P_G(x)$ that matches the true data distribution. It consists of two networks: a generator, G, that transforms noise variables $z \sim P_{Noise}(z)$ to data space $x = G(z)$ and a discriminator D that assigns probability $p = D(x)$ when x is a sample from the $P_{Data}(x)$ and assigns probability $1-p$ when x is a sample from the $P_G(x)$. In a GAN, D and G play the following two-player minimax game with the value function $V(G, D)$:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{x\sim p_z(z)}[\log(1 - D(G(z)))] \quad (4)$$

This enables the discriminator, D, to find the binary classifier that provides the best possible discrimination between true and generated data and simultaneously enables the generator, G, to fit $P_{Data}(x)$. Both G and D can be trained using back-propagation.

3 Proposed method

The framework of our proposed emotional VC system is shown in Fig. 2. As shown in the figure, we extracted the spectral features and F0 features from both source voice and target voice by STRAIGHT [3]. Next, we transformed spectral features of the source and target voices to 32-dimensional MCC features. We subsequently used the CWT method to transform one-dimensional F0 features into high-dimensional CWT-F0 features. Here, we set the total number of scales to 32, which

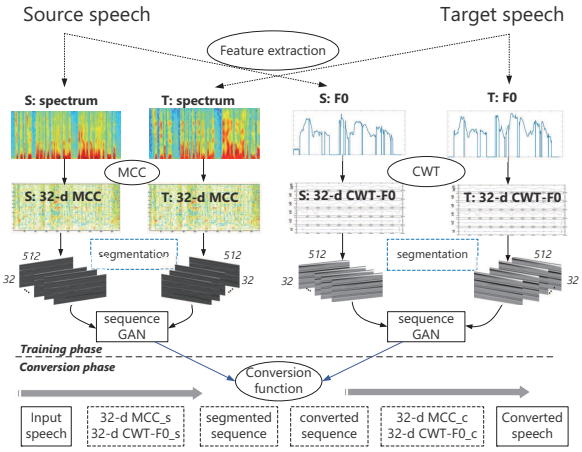

Fig. 2 Framework of our proposed emotional VC.

will lead to 32-dimensional CWT-F0 features. The CWT method can systematically capture 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. After obtaining the MCC and CWT-F0 features, we used dynamic time warping (DTW) to align them of the source and target voices.

To improve training of the conversion function of MCC features and CWT-F0 features with limited amounts of emotional voice data, we adopt the sequence-GAN training model.

In the sequence-GAN, we design the generator as a fully convolutional network (FCN), which has been proven to be able to solve the problem of continuity between frames in the sequence to sequence conversion [4]. We treat the acoustic features (MCC matrices and CWT matrices) as image-like data. Thus, before training in the sequence-GAN, we reshaped the aligned MCC matrices and the CWT-F0 matrices to a suitable size for CNN training. As shown in Fig. 3, the sentence features are segmented to 32×512-size sequence features, which represents a temporal dependency of 0.5 sec. The numbers of sequences m are dependent on the arbitrary length of the training sentences. In the primal task of the sequence-GAN, the input matrices ($x_{1:m}$) and output matrices ($y_{1:m}$) of one global sentence are parallel converted by generators ($G_{A(1:m)}$) and discriminated by discriminators ($D_{B(1:m)}$). Furthermore, m is set to be the minibatch size, which can be used as minibatch discrimination [5] for sequence-GAN. Then, the discriminator can be modeled by training not only the word level features, but also the

sentence level features which consist of the multiple sequences. By this way, the discriminator could potentially help the generator to obtain the conversion parameters of global sentence level.

The conversion phase in Fig. 2 shows how our trained conversion function can be applied. The source voice is processed into 32-dimensional MCC (32-d MCC_s) and CWT-F0 (32-d CWT-F0_s) features. These features can be segmented to 32×512-size features and then fed into the conversion function to be converted to target features. Finally, we transform them back to spectrum and F0 and used these features to reconstruct the waveform, using STRAIGHT.

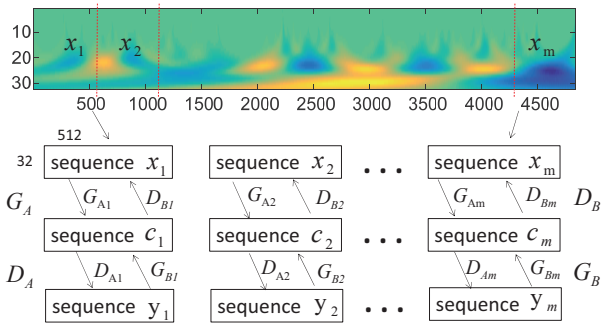

Fig. 3 Illustration of training the sentence features in the sequence-GAN. x_m, c_m and y_m represent the source, converted, and target segmental features (sequence of acoustic features with a duration of 512 frames), respectively. The number of sequences m for each sentence depends on the size of training sentences. G_A and D_A, G_B and D_B, represent the generator and discriminator in the sequence-GAN.

4 Experiments

We used a database of emotional Japanese speech constructed in a previous study [6]. The waveforms used were sampled at 16 kHz. In the database, 50 sentences from the ATR Japanese phonetically balanced text set were used in the experiments. These 50 sentences are designed to include a minimum phone set of Japanese. All the text were read by a professional narrator with neutral, angry, happy and sad voices. Input and output data had the same speaker, but expressing different emotions. We set the six datasets into the following: happy to neutral voice, angry to neutral voice, and sad to neutral voice, as well as their inverse conversion from neutral voice to each emotional voice.

To evaluate the proposed method in converting different emotional datasets, we compared the results with several state-of-the-art methods listed below.

- **DBNs+LG [7]:** This system proposed by Nakashika *et al.* converts spectral features using DBNs and converts the F0 features through the logarithm Gaussian (LG) method.
- **DBNs+NNs [2]:** This is a previously published method that uses the DBNs to convert spectral features, while using the NNs to convert the CWT-F0 features.
- **GAN:** This method uses basic GAN to train MCC and CWT-F0 features for conversion functions.
- **sequence-GAN:** This method uses our proposed sequence-GAN model to train MCC and CWT-F0 features for conversion functions.

We carried out a subjective emotion classification test, for the emotional to neutral pairs (H2N, S2N, and A2N) and their inverse conversion (N2H, N2S, and N2A) comparing different methods (DBNs+LG, DBNs+NNs, GAN, sequence-SANs). For each test model, 60 utterances (10 for angry, 10 for sad, 10 for happy and 30 for neutral) are selected, and 10 listeners are involved. The listeners are asked to label a converted voice as Angry, Sad, Happy or Neutral. As shown in Table 4 (a), when evaluating the original recorded emotional speech utterances, the classifier performed quite well, hence the corpus can be used in the emotion classification test. The classification results for the converted voices of DBNs+LG, DBNs+NNs, GAN and sequence-GAN are shown in (b), (c), (d) and (e) of Table 4, respectively.

As shown in Table 4 (b), the conventional DBNs+LG method shows poor performance in all emotional VC tasks, especially for the conversion of the emotional voice to neutral voice. Therefore the F0 features converted by the conventional logarithm Gaussian method are not enough for emotional VC under the normal VC framework.

Comparing the results of Table 4 (c) with Table 4 (d), it is clear that DBNs+NNs method obtains better classification results than the GAN model, although the GAN model yielded a slightly better result than the DBNs+NNs method in the objective experiment. This result confirms that the

Table 1 Results of classification for recorded voices and converted voice by different methods [%].

	(a) original recorded voice				(b) DBNs+LG				(c) DBNs+NNs				(d) GAN				(e) sequence-GAN			
Tar./Percept	Angry	Sad	Happy	Neutral	Angry	Sad	Happy	Neutral	Angry	Sad	Happy	Neutral	Angry	Sad	Happy	Neutral	Angry	Sad	Happy	Neutral
Angry	90	2	0	8	27	5	0	68	59	13	3	25	33	21	2	44	71	4	1	24
Sad	2	97	0	1	3	33	2	62	4	53	0	43	39	45	15	1	12	68	3	17
Happy	0	0	98	2	3	5	18	74	2	2	68	38	3	2	47	48	2	3	80	15
Neutral	0	0	5	95	24	30	27	19	12	15	17	56	18	22	23	37	12	17	11	60

GAN model training may lead to instability and deregularized processing of some converted samples.

With reference to the results of sequence-GAN shown in Table 4 (e), the proposed method yielded about 70% classification accuracy in average, which has indicated a better result than the other models. Especially for the conversion of neutral voice to emotional voice, the average classification accuracy of the converted emotional voice is about 75%, which is 15% higher than the converted neutral voice.

5 Conclusions

This study proposed an emotional VC method using sequence-GAN with MCC and CWT-F0 features. In order to obtain better training results, we segment MCC features and CWT features to sequences with computable size, and then separately trained them with the proposed sequence-GAN. A comparison between the proposed method and the conventionally used methods shows that our proposed model can effectively change the prosody of the emotional voice due to GAN's ability to mitigate the over-smoothing problem caused in the low-level data space. We also compared our proposed method to the model using original GAN, and the results show that sequence-GAN can strengthen and regularize the training process in the emotional VC.

参考文献

[1] 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.

[2] Z. Luo *et al.*, "Emotional voice conversion with adaptive scales F0 based on wavelet transform using limited amount of emotional data," Proc. Interspeech 2017, pp. 3399–3403, 2017.

[3] 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.

[4] T. Kaneko *et al.*, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," Proc. Interspeech 2017, pp. 1283–1287, 2017.

[5] T. Salimans *et al.*, "Improved techniques for training GANs," in *Advances in Neural Information Processing Systems*, 2016, pp. 2234–2242.

[6] H. Kawanami *et al.*, "GMM-based voice conversion applied to emotional speech synthesis," In Proc. European Conf. on Speech Communication and Technology (Eurospeech '03), pp. 2401–2404, 2003.

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