EMOTIONAL VOICE CONVERSION WITH WAVELET TRANSFORM USING DUAL SUPERVISED ADVERSARIAL NETWORKS *

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

1 Introduction

Emotional voice conversion (VC) is a kind of VC technique for converting prosody in speech, which can represent different emotions, while keeping the linguistic information unchanged. Recently, deep learning has dramatically improved the performance of VC systems. However, deep learning models are restricted to problems with moderate dimensions and sufficient data, so most deep learning based VC works do not focuse on the emotional VC, which is mainly affected by low-dimensional F0 features with insufficient training VC data.

In this paper, to overcome the problem of moderate dimensions, we used our earlier work [1] to systematically capture the F0 features of different temporal scales with the adaptive scales continuous wavelet transform (AS-CWT) method. To deal with the difficulty of a limited amount of training data, we propose dual supervised adversarial networks (dual-SANs), which combine the GANs [2] with dual supervised learning [3] to train the MCC and CWT-F0 features. The effectiveness of GANs 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, and it has also begun to be applied in speech synthesis. Dual supervised learning is a technique for supervised learning tasks in dual forms. The dual learning method uses one agent to represent the model for the primal task and another agent to represent the model for the dual task. Then they are asked to teach each other through a reinforcement learning process. In emotional VC task, we need to convert the voice from one emotion to another. They are dual tasks with labeled emotional voice data. Thus, due to dual supervised learning's ability to improve the training performances and GANs' ability to mitigate the over-smoothing problem caused in the low-level data space when converting the CWT-F0 and MCC features, we develope a novel dual-SANs model, which enables emotional VC to be trained from each of two sets of labeled emotional voices from two domains. In dual-SANs architecture, the primal GAN learns to convert voices from emotion A to those of emotion B, while the dual GAN is trained for inverse conversion. The closed loop made by the primal and dual tasks allows the voice to be converted from emotion A to emotion B, and also obtain the inverse conversion from emotion B to emotion A. Moreover, the two dual tasks have intrinsic connections to each other because their source and target voice are in the same labeled emotional data sets. Using dual supervised learning can mutually benefit and regularize the training process.

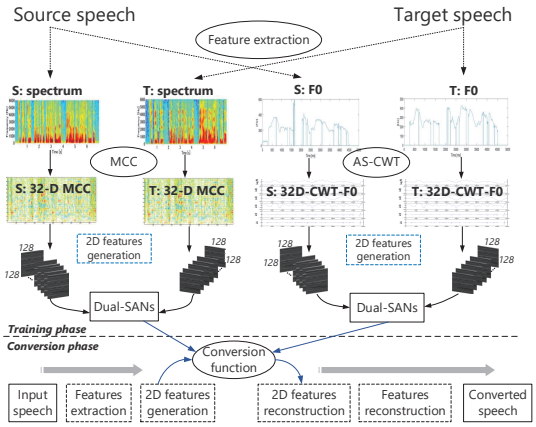

Fig. 1 Framework of our proposed emotional voice conversion.

2 Emotional VC using Dual-SANs

2.1 Framework Overview

Our proposed framework is shown in Fig. 1. In the training phase, we extracted the spectral features and F0 features from both the source voice and target voice by the STRAIGHT. Next, we transformed spectral features of the source and target voices to 32-dimensional MCC features. Then, we used the AS-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 will lead to 32 dimensional CWT-F0 features. Next,

*Dual Supervised Adversarial Networks を用いた感情声質変換, 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).

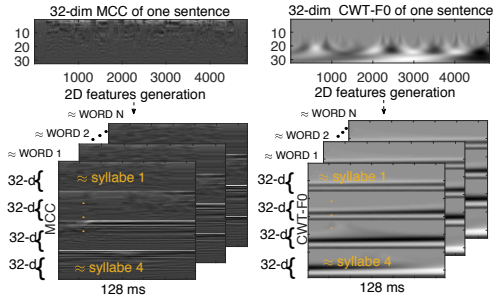

Fig. 2 Examples of 128*128 size 2D-MCC (left) and 2D-CWT-F0 (right) textures reshaped from the 32-dimensional MCC and CWT-F0 features .

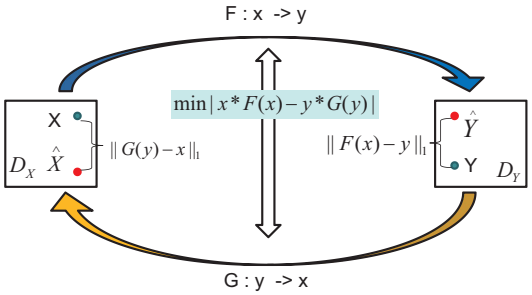

Fig. 3 Illustration of calculating the loss of dual-SANs.

we used dynamic time warping (DTW) to align the MCC and CWT-F0 features of the source and target voices, respectively. The conversion function training for MCC features and CWT-F0 features uses the proposed dual-SANs, which will be described in Section 2.2. GANs have been successful in image generation. So, before training in the dual-SANs models, we reshaped the aligned MCC matrix and CWT-F0 matrix to 2D features of 128×128 size which are shown in Fig. 2. As described in [1], the average duration of syllables was found to be 50 ms to 180 ms, and the words from 300 ms to 650 ms. Thus, one 2D feature (128×128) can approximately represent one word (32×512) made up by four syllables ($32 \times 128 \times 4$). By doing this, a higher overall accuracy of the dual-SANs can be achieved.

The conversion phase in Fig. 1 shows how our trained conversion function can be applied. The source voice is processed into 32-dimensional MCC and CWT-F0 features. These features can then be fed into the conversion function to be converted to target features. Finally, we transformed them back to spectrum and F0, and used these features to reconstruct the waveform using STRAIGHT.

2.2 Proposed Dual-SANs Model

When dealing with training of emotional VC, each two sets of labeled and paired 128×128 features were sampled from domains of source emotional voice X and target emotional voice Y, respectively. As shown in Fig. 3, our model contains two conversion functions $F : x \rightarrow y$ and $G : y \rightarrow x$. The primal task of Dual-SANs is to learn a coversion function $F : x \rightarrow y$ that converts the emotional voice $x \in X$ to the target emotional voice $y \in Y$, while the dual task is to train an inverse conversion function $G : y \rightarrow x$. To realize this, we apply two GANs.

The primal GAN learns the conversion function F and a discriminator D_Y that discriminates between converted outputs of F and real members of domain Y. Analogously, the dual GAN learns the conversion function G and a discriminator D_X. We apply adversarial losses to both conversion functions. For the conversion function $F : X \rightarrow Y$ and its discriminator D_Y, we express the objective as:

$$L_F(F, D_Y, X, Y) = E_{y \sim P_{data(y)}}[\log D_Y(y)] + E_{x \sim P_{data(x)}}[\log(1 - D_Y(F(x))] \quad (1)$$

The goal of emotional VC is to learn a converted emotional voice distribution $P_{F(x)}$ that matches the target emotional voice distribution $P_{data(y)}$. Eq. (1) enables D_Y to find the binary classifier that provides the best possible discrimination between true and converted voice and simultaneously enables the function F to fit $P_{data(y)}$. $L_F(F, D_Y, X, Y)$ is maximized and minimized with respect to D_Y and F, respectively.

$$F^* = arg \max_{D_Y} \min_{F} L_F(F, D_Y, X, Y) \quad (2)$$

We use the similar adversarial loss for the inverse conversion function $G : Y \rightarrow X$ and its discriminator D_X as well:

$$G^* = arg \max_{D_X} \min_{G} L_G(G, D_X, Y, X) \quad (3)$$

$$L_G(G, D_X, Y, X) = E_{x \sim P_{data(x)}}[\log D_X(x)] + E_{y \sim P_{data(y)}}[\log(1 - D_X(G(y))] \quad (4)$$

In supervised dual tasks, the primal task's source emotional voice X is the dual task's target, and the target emotional voice Y is the task's source. For any $x \in X$, $y \in Y$, if the two models are learned separately by minimizing their own loss functions,

there is no guarantee that the intrinsic connections between the two dual tasks will be utilized. Unlike dual supervised learning [3] applied in machine translation, which uses the duality of joint probability to regularized the training process, in emotional VC, owning to the true value conversion and the duality of the paired emotional voice dataset, we used the true values of x, y and their converted value $F(x)$, $G(y)$ to keep the training process regularized. The point in dual supervised task is to learn the two models, F and G, by minimizing their loss functions subject to the dual constraint:

$$L_{dual}(F,G) = \\ E_{x\sim P_{data(x)},y\sim P_{data(x)}}[\|x * F(x) - y * G(y))\|_1] \quad (5)$$

By doing so, the intrinsic connection between F and G is explicitly strengthened, which is supposed to push the learning process towards the right direction and keep a balance when learning the two dual tasks. Combining the dual loss in dual supervised learning and adversarial loss in GANs, our full objective function is:

$$L(F,G,D_X,D_Y) = L_F(F,D_Y,X,Y) \\ + L_G(G,D_X,Y,X) + \lambda L_{dual}(F,G) \quad (6)$$

where λ controls the relative importance of the two objectives. We aim to solve:

$$F^*,G^* = arg \max_{D_Y,DX} \min_{F,G} L(F,G,D_Y,DX) \quad (7)$$

Throughout the training process, conversion functions F and G are optimized to learn the converted emotional voice which cannot be distinguished from target emotional voice by corresponding discriminators D_Y and D_X, as well as to minimize the dual loss $\|x * F(x) - y * G(y))\|$. In Section 3, we also descirbe an experiment comparing the full dual-SANs model against the original GANs trained with the adversarial loss $L_F(F,D_Y,X,Y)$ alone.

3 Experiments

In our 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 the speaker was expressing different emotions. We classified the three data sets into the following: angry to neutral voices (A2N), sad to neutral voices (S2N), and happy to neutral voices (H2N). For each data set, 50 sentences were chosen as training data and 10 sentences were chosen for the VC evaluation.

To evaluate our proposed **dual-SANs** method, we compared the results with several state-of-the-art methods. **DBNs+LG** proposed by Nakashika *et.al.* converted spectral features using DBNs, and converted the F0 features through the LG method [4]. **DBNs+NNs** is our previous work [1] that used DBNs to convert spectral features while using pretrained NNs to convert the CWT-F0 features decomposed by the AS-CWT method. We also compared with the original **GANs**.

3.1 Objective Experiment

To evaluate the spectral conversion and F0 conversion, we used Mel Cepstral Distortion (MCD) and root-mean-square error (RMSE), respectively. As shown in Table 1, comparing DBNs with the source, the DBNs decreased the value of MCD. Using GANs without dual supervised learning obtains slightly better results than the DBNs model, and the proposed dual-SANs model significantly decreases the MCD value. The average F0-RMSE results are shown in the right part of Table 1. The conventional linear conversion LG can only affect the conversion of happy to neutral, but only slightly affects the other conversions. The other three methods can affect the conversion of all emotional voice datasets. In addition, the GANs and proposed dual-SANs and can obtain significant improvement in F0 conversion.

Fig. 4 shows examples of source, target, and converted voice spectrograms. As discussed above, the MCD and RMSE values obtained using GANs without dual supervised learning are slightly better than DBNs models. However, as shown in Fig. 4, unlike normal object image generation, the dissimilarities between the source and target spectrogram images are very small. Without dual supervised learning, the source image is hard to be regularized to the target image due to the problem of insufficient data. We can clearly see that adding the dual supervised learning can obtain converted spectrograms very similar to the target. While the spectrograms obtained without dual supervised learning sometimes fail to be regularized to the target spectrogram image. Fig. 4 (D) shows a failed example of emotional voice converted by GANs.

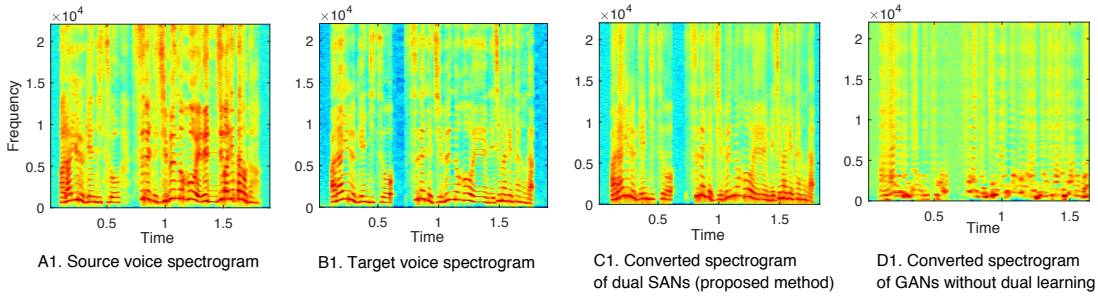

Fig. 4 Spectrograms of the source, target and converted emotional voices.

Table 1 MCD and F0-RMSE results for the conversion of emotional voice to neutral voice.

	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+NNs	5.47	4.77	5.92	51.1	52.1	64.4
GANs	5.58	4.51	5.99	39.9	50.3	61.1
dual-SANs	3.38	4.77	4.31	35.4	36.1	59.0

3.2 Subjective Experiment

We conducted a subjective emotion evaluation using a mean opinion score 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 higher the point value). Here, we tested the emotional to neutral pairs (H2N, S2N, A2N). 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 speech samples. Then, the subject listened to the speech converted using the three methods and then was asked to assign a point value to each conversion. Fig. 5 shows the result of the MOS test, the error bar shows the 95% confidence interval. As the figure shows, the conventional LG method shows poor performance in the conversion of anger voice to neutral voice. Although using GANs without dual supervised learning obtained a slightly better results than the DBNs+NNs method in the objective experiment, due to the instability and unregularized process of some converted features, it got worse scores in MOS test. The dual-SANs obtained best score in every emotional VC.

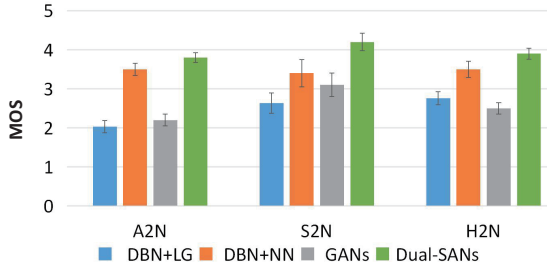

Fig. 5 MOS evaluation of emotional voice conversion

4 Conclusion

This paper proposed an emotional VC method using dual supervised adversarial networks (dual-SANs) with MCC and CWT-F0 features. In order to obtain a better training results, we process 2D MCC features and 2D AS-CWT features, and separately train them with the proposed dual-SANs. The results shows that the dual supervised learning has a great affect on the dual emotional VC tasks.

参考文献

[1] 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.

[2] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[3] Y. Xia *et al.*, "Dual supervised learning," arXiv preprint arXiv:1707.00415, 2017.

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