Emotional Speech Conversion Using Deep Neural Networks *

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

1 Introduction

Recently, the study of Voice Conversion (VC) is being widely attracted attention in the field of speech processing. This technology can be widely applied in various application domains. For instances, speech enhancement, emotion conversion, speaking assistance , and other applications [1] are related to VC. Therefore, the need for this type of technology in various fields has continued to propel related research forward each year. Many statistical approaches have been proposed for spectral conversion during the last. Among these approaches, a Gaussian Mixture Model (GMM) is widely used. However, the features trained by GMM are usually low-dimensional features which may lost some important spectral details during making the speech spectra. The high-dimensional features, such as Mel Frequency Cepstral Coefficents MFCCs which are widely used in automatic speech and speaker recognition, are not compatible with GMM. There are also some approaches to construct non-linear mapping relationships, such as using artificial neural networks (ANNs) to train the mapping dictionaries between source and target features, using deep belief networks (DBNs) to achieve non-linear deep transformation [2]. These models improve the conversion of spectrum features. Nevertheless, most of the related works in respect to VC focus on the conversion of spectrum features, yet the seldom of those focus on F0 conversion, because F0 cannot be processed by a high-dimensional model neural networks (NNs) well. But F0 is one of the most important parameters for representing emotional speech.For emotional voice conversion, some prosody features, such as pitch variables and speaking rate have already been analyzed. There are also some works using a GMM-based VC technique to change the emotional voice [3]. As above-mentioned, recently acoustic voice conversion usually uses the non-linear suitable models (NNs, CRBMs, DBNs, RTDBNs) to convert the spectral features, it is difficult to use the GMM to deal with F0 made by these frameworks. To solve these problems, we proposed a new approach.

In this paper, we focus on the F0 features conversion and transformation of the spectrum features. We propose a novel method that uses the deep neural networks (DBNs) to train MFCC features for constructing the mapping relationship of spectral envelopes between source and target speakers. Then, adopt the neural networks (NNs) to train the normalized-segment-F0 features (NSF0) for converting the prosody of the emotional voice. Since the deep brief networks are effective to spectral envelopes converting, in the proposed model, we train the MFCC features by using the model combined with two different DBNs and concatenating NNs proposed by Nakashika [2]. For the prosody conversion, we use the F0 features. Because the F0 features extracted from the STRAIGHT are one-dimension features, which are not suitable for the NNs. Hence, in this study, we propose the normalized-segment-F0 (NSF0) features to transform the one-dimension F0 features into multiple-dimensions features. By so doing, the NNs can robustly process prosody signals that is presented on F0 features so that the proposed method can obtain high-quality emotional conversion results, which form the main contribution of this paper.

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

The proposed model consists of two parts. One part is the transformation of spectral features using the DBNs, the other is the F0 conversion using the NNs. The emotional voice conversion framework transforms both the excitation and the filter features from the source voice to the target voice is shown in Fig. 1. In this section, we briefly review the process based on STRAIGHT for extracting features from the source voice signal and the target voice signal, while we introduce the spectral conversion part and F0 conversion part.

* ディープニューラルネットワークを用いた感情声質変換, 羅兆傑, 滝口哲也, 有木康雄 (神戸大)

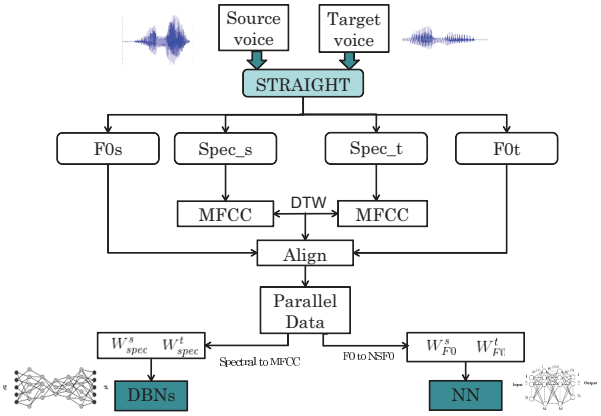

Fig. 1 The emotional voice conversion framework. In the framework *Spec_s* and *Spec_t* mean the spectral envelopes of source and target voice obtained from the STRAIGHT . *F0s* and *F0t* are the basic frequency of source and target speech. W^s_{spec} ,W^t_{spec} , W^s_{F0} and W^t_{F0} are dictionaries of source spectrum, target spectrum , source F0 and target F0, respectively.

2.1 Feature extraction

To extract features from a speech signal, the STRAIGHT model speech is frequently adopted. Generally, the pitch-adaptive-time-frequency smoothing spectrum and instantaneous-frequency-based F0 are derived as excitation features for every 5 ms [4] from the STRAIGHT. As shown in Fig. 1, the spectral features are translated into MFCCs. To have the same number of frames, a Dynamic Time Wrapping (DTW) method is used to align the extracted features (MFCC and F0) of source and target speeches. Finally, the aligned features that have been processed by Dynamic Programming are saved as the parallel data.

2.2 Spectral features conversion

In this section, we will convert the MFCCs by DNNs model [2] .

In this study, we use the 24-dimentional MFCC features for spectral training. As shown in Fig. 1, we transfer the parallel data which concludes the aligned spectral features of source and target voices to MFCC features. Meanwhile, we respectively use the MFCC features of the source and target voice as the input-layer data and output-layer data for DBNs. Fig. 2 shows the architecture of the DBNs convert spectral features, which indicates two different DBNs for source speech and target speech

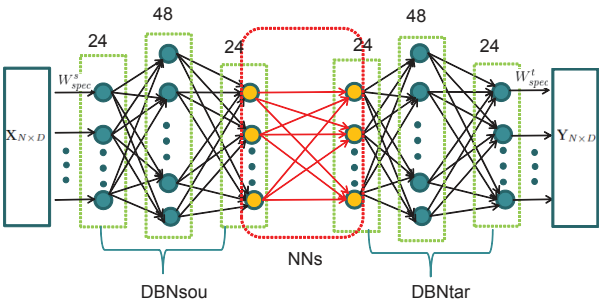

Fig. 2 DBNs model

(DBNsou and DBNtar) so as to capture the speaker-individuality information and connect them by the NNs. The numbers of each node from input x to output y in Fig. 2 were [24 48 24] for DBNsou and DBNtar. $X_{N\times D}$ and $Y_{N\times D}$ represent N examples of D-dimensional source feature and target feature training vectors. $X_{N\times D}$ and $Y_{N\times D}$ are defined in Eq. 1, where D=24.

$$\begin{aligned} X_{N\times D} &= [x_1, ..., x_N], x = [x_1, ..., x_D]^{\mathrm{T}} \\ Y_{N\times D} &= [y_1, ..., y_N], y = [y_1, ..., y_D]^{\mathrm{T}}. \end{aligned} \quad (1)$$

2.3 F0 features conversion

For prosody conversion, F0 features are usually adopted, and while it needs to be transformed. Relative methods used a logarithm Gaussian normalized transformation to transform the F0 from the source speaker to the target speaker as indicated in the Eq. 2 below.

$$\log(f0_{conv}) = \mu_{tgt} + \frac{\sigma_{tgt}}{\sigma_{src}}(\log(f0_{src}) - \mu_{src}) \quad (2)$$

As mentioned in the introduction section, non-linear conversion models are more compatible with the complex human emotional speeches. We use the NN models to train the F0 features in our proposed methods. The F0 features conducted from STRAIGHT are one-dimensional features and discrete. Before training the F0 features by NNs, we need to transform F0 features into high-dimension data. To process the high-dimensional features, we adopt continuous wavelet transform (CWT)to decompose the F0 contour into several temporal scales that can be used to model different prosodic levels ranging from micro-prosody to the sentence level [5]. Then,transform the decomposed features into the segment-level features. Detail processing are the following steps.

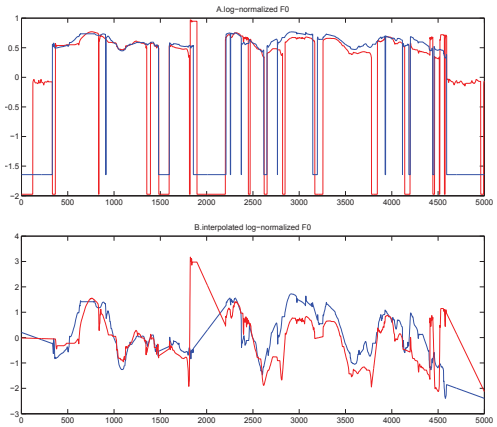

Fig. 3 log-normalized F0 (top) and interpolated log-normalized F0 (bottom). The red curve: target F0; The blue curve: source F0.

1)In order to explore the perceptual relevant information, the linear scale F0 contour is transformed to the logarithmic semitone scale. As shown in Fig.3, the log-normalized F0 is discrete. The wavelet method is sensitive to the gaps in the F0 contour, so we need to add the unvoiced parts to the logf0 with linear interpolation to reduce discontinuities in voicing boundaries. In addition, to alleviate edge artifacts, constant f0 was added prior to and after the utterance. The pre-utterance f0 value was set to the mean f0 value calculated over logf0 ; the post-utterance f0 was set to the minimum of logf0. Finally, the interpolated logf0 contour is normalized to zero mean and unit variance. An example of an interpolated pitch contour is depicted in the bottom pan of Fig.3.

2)The continuous wavelet transform of f0 is defined by

$$W(\tau,t)=\tau^{-1/2}\int_{-\infty}^{\infty}f_0(x)\,\psi\left(\frac{x-t}{\tau}\right)dx \qquad (3)$$

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

where $f_0(x)$ is the input signal and (ψ) is the Mexican hat mother wavelet. we decompose the continuous logf0 at 10 discrete scales, each one octave apart. Then f0 is represented by 10 separate components given by

$$W_i(f_0)(t)=W_i(f_0)(2^{i+1}\tau_0,t)\,(i+2.5)^{-5/2}, \qquad (5)$$

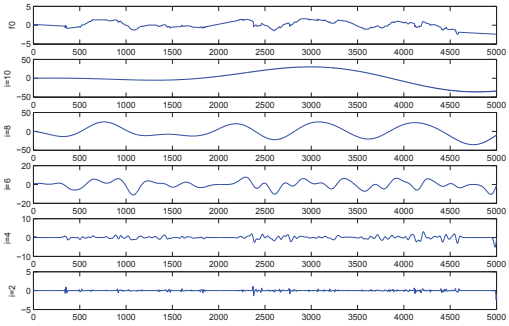

Fig. 4 Interpolated log-normalized F0 and five wavelet transforms(i=2, i=4, i=6, i=8, i=10)

where i = 1,...,10 and τ_0=5 ms. As shown in Fig.4,the first pan above is the interpolated log-normalized F0 of the source voice. And the second pan to sixth pan show the separate components of i=10, i=8, i=6, i=4, i=2 which can represent the utterance, phrase, word, syllable, phone levels respectively.

The original signal can be approximately reconstructed by the following reconstruction formula:

$$f_0(t)=\sum_{1}^{10}W_i(f_0)(t)\,(i+2.5)^{-5/2} \qquad (6)$$

3) Transform the processed 10-dimensions features to the segment-level features . We form the segment-level feature vector by stacking features in the neighboring frames as:

$$\begin{aligned}&X_{N\times w}=[x_1,...,x_N]^{\mathrm{T}},\\&x(m)=[z(m-w),...,z(m),...,z(m+w)]^{\mathrm{T}}\end{aligned} \qquad (7)$$

where w is the window size on each side. Eq. 7 represents N examples of w-dimensional source features. In the proposed model, we set the $w=3$ that the 10-dimensional normalized F0 features made up by the 10 separate components can be transformed to the 30-dimensional normalized-segment-F0 features(NSF0). To guarantee the coordination between the initial source and conversion signals, we adopt the same approach for the target features transformation.After transforming F0 features to the NSF0 features, we converted the 30-dimentional NSF0 features by NNs. we used the 4-layers NN models to train the NSF0 features. The numbers of nodes from input layer x to output layer are [30 50 50 30].

3 Experiments

3.1 Database

We used a database of emotional Japanese speech constructed in [6]. From this database, we selected the angry voices, happy voices and sad voices of speaker (FUM) for the source, and the neutral voices of speaker (FON) for target. For each emotional voice, 50 sentences were chosen as training data. We made the datasets as happy voices to neutral voices, angry voices to neutral voices and sad voices to neutral voices.

3.2 Result and discussion

Mel Cepstral Distortion (MCD) was used for the objective evaluation of spectral conversion. For comparison,we use the NN model, the RTRBM model and the DBNs model to convert spectral features respectively. As shown in the Fig 5, For emotional voice conversion DBNs model can convert the spectral features better than the NNs, and no significant difference with the RTRBMs. Although our training datasets are all from the FUM to FUN and the content of the sentences are the same. We can also see that the MCD evaluations from different emotional voices conversion to the neutral voice are a bit different. The result confirms that different emotions in the same speech can influence the spectral conversion and DNBs models proved to be the fast and effective method in the spectral conversion of emotional voice.

For evaluating the F0 conversion, we used the Root Mean Squar Error (RMSE).We used the Gaussian normalized transfromation method and proposed method to convert the F0 features for comparison. Fig.6 shows that our proposed method obtains a better result than the traditional Gaussian normalized transformation method in the all datasets. (angry to neural, happy to neural, sad to neural.)

4 Conclusions

In this paper, we proposed a method using DBNs to train the MFCC features to construct mapping relationship of the spectral envelopes between source and target speakers, using NNs to train the NSF0 features which are conducted by the F0 features for prosody conversion. Comparison between the proposed method and the past methods (NNs, GMM) has shown that our proposed model can effectively change the acoustic voice and the prosody for the emotional voice at the same time.

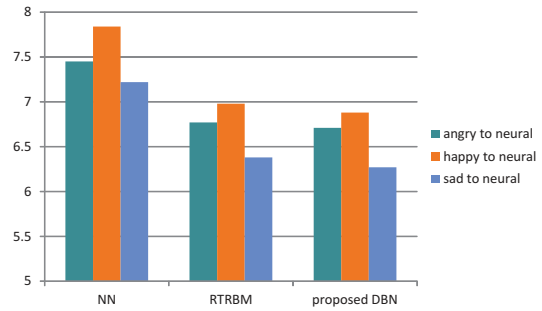

Fig. 5 Mel-cepstral disortion evaluation of spectral features conversion

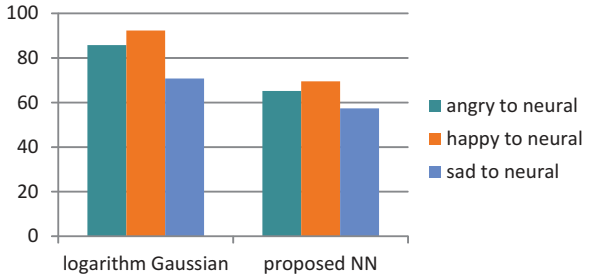

Fig. 6 Root mean squared error evaluation of F0 features conversion

References

[1] J. Krivokapić, "Rhythm and convergence between speakers of american and indian english," Laboratory Phonology, vol. 4, no. 1, pp. 39–65, 2013.

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

[3] R. Aihara *et al.*, "Gmm-based emotional voice conversion using spectrum and prosody features," American Journal of Signal Processing, vol. 2, no. 5, pp. 134–138, 2012.

[4] 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.

[5] H. Kruschke and M. Lenz, "Estimation of the parameters of the quantitative intonation model with continuous wavelet analysis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[6] H. Kawanami *et al.*, "Gmm-based voice conversion applied to emotional speech synthesis," 2003.