Emotional Voice Conversion Using Deep Neural Networks with MCC and F0 Features

Zhaojie Luo, Tetsuya Takiguchi, Yasuo Ariki
Graduate School of System Informatics, Kobe University, Japan 657–8501
Email: luozhaojie@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

***Abstract*—An artificial neural network is one of the most important models for training features in a voice conversion task. Typically, Neural Networks (NNs) are not effective in processing low-dimensional F0 features, thus this causes that the performance of those methods based on neural networks for training Mel Cepstral Coefficients (MCC) are not outstanding. However, F0 can robustly represent various prosody signals (*e.g.*, emotional prosody). In this study, we propose an effective method based on the NNs to train the normalized-segment-F0 features (NSF0) for emotional prosody conversion. Meanwhile, the proposed method adopts deep belief networks (DBNs) to train spectrum features for voice conversion. By using these approaches, the proposed method can change the spectrum and the prosody for the emotional voice at the same time. Moreover, the experimental results show that the proposed method outperforms other state-of-the-art methods for voice emotional conversion.**

I. 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 to various application domains. For instances, voice conversion [1], emotion conversion [2], speaking assistance [3], and other applications [4] [5] 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 decades [6] [7]. Among these approaches, a Gaussian Mixture Model (GMM) is widely used. However, there are several shortcomings with the GMM spectral conversion method. First, GMM-based spectral conversion is a piece-wise linear transformation method, but the mapping relationship between humans voice conversion is generally non-linear, so non-linear voice conversion is more compatible with voice conversion. Second, the features which are trained using GMMs are usually low-dimensional features which may lost some important spectral details for speech spectra. The high-dimensional features, such as Mel Cepstral Coefficients (MCC) [8] which are widely used in automatic speech and speaker recognition, are more compatible with deep architecture learning.

A number of improvements have been proposed in order to cope with these problems such as integrating dynamic features and global variance (GV) into the conventional parameter generation criterion [9], using Partial Least Squares (PLS) to prevent the over-fitting problem encountered in standard multivariate regression [10]. 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 [11], using a conditional restricted Boltzmann machine (CRBM) to model the conditional distributions [12], or using deep belief networks (DBNs) to achieve non-linear deep transformation [13].

These models improve the conversion of spectrum features. Nevertheless, almost 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 deep architecture NNs well. But F0 is one of the most important parameters for representing emotional speech, because it can clearly describe the variation of voice prosody from one pitch period to another. For emotional voice conversion, some prosody features, such as pitch variables (F0 contour and jitter), and speaking rate have already been analyzed [14]. There were approaches forced on the simulation of discrete basic emotions. But, these methods are not compatible with the complex human emotional voices which are non-linear convert. There are also some works using a GMM-based VC technique to change the emotional voice [15] [16]. As abovementioned, recently acoustic voice conversion usually uses the non-linear suitable models (NNs, CRBMs, DBNs, RTRBMs) to convert the spectrum features, it is difficult to use the GMM to deal with F0 made by these frameworks. To solve these problems, we propose 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 belief networks (DBNs) to train MCC features for constructing the mapping relationship of spectral envelopes between source and target speakers. Then, we 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 [13], in the proposed model, we train the MCC features by using two DBNs for the source speaker and the target speaker, respectively, then using the NNs to connect the two DBNs for converting the individuality abstractions of the speakers. As it has been shown that the bottleneck features are effective to improve the accuracy and naturalness of synthesized speech [17], we construct the three-

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

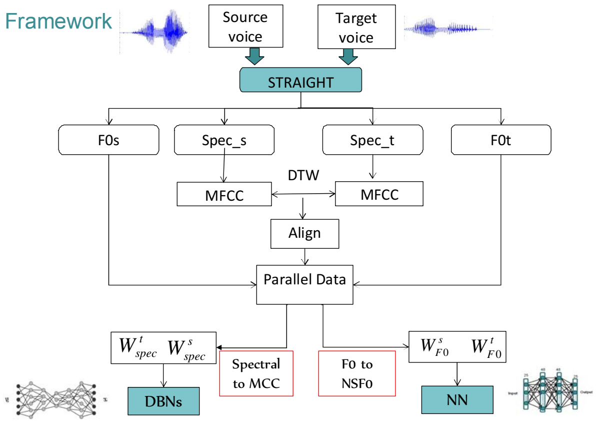


Fig. 1. Emotional voice conversion framework. $Spec\_s$ and $Spec\_t$ mean the spectral envelopes of source and target voice obtained from the STRAIGHT. $F0s$ and $F0t$ are the basic frequency of source and target speech. W^s_{spec}, W^t_{spec}, W^s_{F0} and W^t_{F0} are dictionaries of source spectrum, target spectrum, source F0 and target F0, respectively.

layers DNBs (24-48-24) for both the source voice and target speakers. Hereby, the unit of middle-layer (48) is larger than the input-layer (24) and output-layer (24). We adopt the two three-layers DNBs and the connect NNs to build the six-layer deep architecture learning model.

For the prosody conversion, F0 features are used. Although many researchers have adopted the F0 features for emotional VC [18][19], the F0 features used in these approaches were mostly extracted by the STRAIGHT [20]. Since the F0 features extracted from the STRAIGHT were one-dimension features, which were 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. II. Sec. III gives the detailed stages of process in experimental evaluations and conclusions are drawn in Sec. IV.

II. PROPOSED METHOD

The proposed model consists of two parts. One part is the transformation of spectral features using the DBNs, and 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 as 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.

A. 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 5ms [20] from the STRAIGHT. As shown in Fig. 1, the spectral features are translated into Mel Frequency Cepstral Coefficents (MFCC) [21], which are known as working well in many areas of speech technologies [9][22]. To have the same number of frames between the source and target, a Dynamic Time Wrapping (DTW) method is used to align the extracted features (MFCC and F0) of source and target voices. Finally, the aligned features that have been processed by Dynamic Programming are used as the parallel data. Before training them, we need to transform the MFCC features to MCC features for the DBNs model and transform the F0 features to the normalized-segment-F0 features (NSF0), respectively. We will describe the transform methods and the training models of spectral and F0 in Sec. II.B and Sec. II.C.

B. Spectral features conversion

In this section, we will introduce the spectral conversion conducted by DBNs. DBNs have an architecture that stacks multiple Restricted Boltzmann Machines (RBMs) which compose a visible layer and a hidden layer. For each RBM, there are not connections among visible units or hidden units, yet it is connected by the bidirectional connections between the visible unit and hidden unit. As an energy-based model, the energy of a configuration (v, h) is defined as:

$$E\left(v,h\right) = -a^T v - b^T h - v^T W h, \tag{1}$$

where $W \in R_{I\times J}$, $a \in R_{I\times 1}$, and $b \in R_{J\times 1}$ denote the weight parameter matrix between visible units and hidden units, a bias vector of visible units, and a bias vector of hidden units, respectively. The joint distribution over v and h is defined as:

$$P\left(v,h\right) = \frac{1}{Z}e^{-E(v,h)}. \tag{2}$$

The RBM has the shape of a bipartite graph, with no intra-layer connections. Consequently, the individual activation probabilities are obtained via

$$P\left(h_j = 1|v\right) = \sigma\left(b_j + \sum_{i=1}^{m} w_{i,j} v_i\right), \tag{3}$$

$$P\left(v_i = 1|h\right) = \sigma\left(a_i + \sum_{j=1}^{n} w_{i,j} h_j\right). \tag{4}$$

In our model, σ denotes a standard sigmoid function, *i.e.*, $(\sigma\left(x\right) = 1/(1+e^{-x}))$. For parameter estimation, RBMs are trained to maximize the product of probabilities assigned to some training set data V (V is a matrix, each row of that is treated as a visible vector v). To calculate the weight parameter matrix, we use the RBM log-likelihood gradient method as follows:

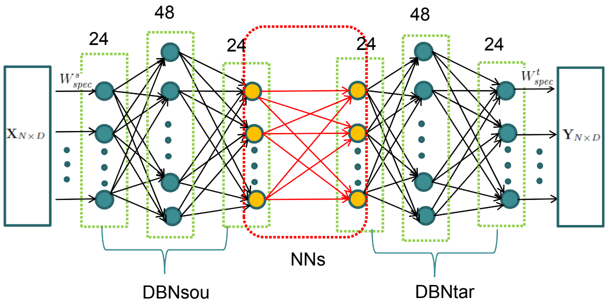


Fig. 2. DBNs model

$$L(\theta) = \frac{1}{N}\sum_{n=1}^{N} logP_{\theta}\left(v^{(n)}\right) - \frac{\lambda}{N}\left\|W\right\|. \tag{5}$$

To differentiate the $L(\theta)$ via (6), we can obtain W when making the $L(\theta)$ be the largest.

$$\frac{\partial L(\theta)}{\partial W_{ij}} = E_{P_{data}}\left[v_i h_j\right] - E_{P_{\theta}}\left[v_i h_j\right] - \frac{2\lambda}{N}W_{ij}. \tag{6}$$

In this study, we use the 24-dimentional MCC 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 MCC features. Meanwhile, we respectively use the MCC 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 (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, respectively. $X_{N\times D}$ and $Y_{N\times D}$ are defined in (7) (D=24).

$$\begin{aligned} X_{N\times D} &= \left[x_1, ..., x_m, ..., x_N\right], x_m = \left[x_1, ..., x_D\right]^{\mathrm{T}} \\ Y_{N\times D} &= \left[y_1, ..., y_m, ..., y_N\right], y_m = \left[y_1, ..., y_D\right]^{\mathrm{T}}. \end{aligned} \tag{7}$$

In summary for the above discussions, the whole training process of the DBNs can be conducted as follows three steps.
1) Train two DBNs for source and target speakers. In the training of DBNs, the hidden units computed as a conditional probability $(P(h|v))$ in (3) are fed to the following RBMs, and trained layer-by-layer until the highest layer is reached.
2) After training two DBNs, we connect the DBNsou and DBNtar and train them by using NNs. Weight parameters of NNs are estimated so as to minimize the error between the output and the target vectors.
3) Finally, each parameter of the whole networks (DBNsou, DBNtar and NNs) is fine-tuned by back-propagation using the MCC features.

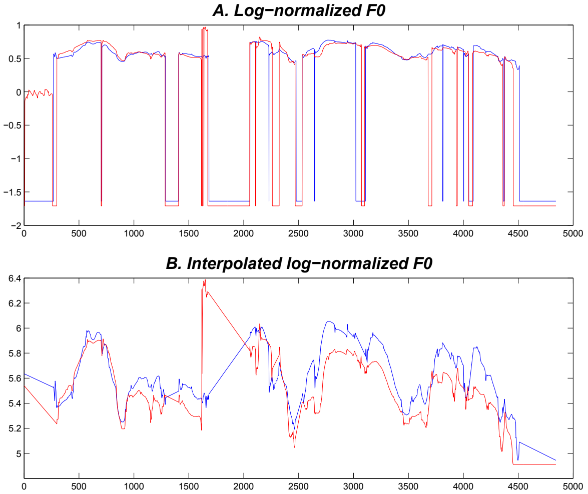


Fig. 3. Log-normalized F0 (A) and interpolated log-normalized F0 (B). The red curve: target F0; The blue curve: source F0.

C. F0 features conversion

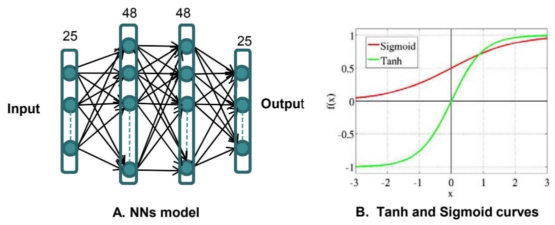
For prosody conversion, F0 features are usually adopted. In conventional methods, a logarithm Gaussian normalized transformation [23] is used to transform the F0 from the source speaker to the target speaker as follows:

$$\log\left(f0_{conv}\right) = \mu_{tgt} + \frac{\sigma_{tgt}}{\sigma_{src}}\left(\log\left(f0_{src}\right) - \mu_{src}\right) \tag{8}$$

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 for the target speaker. $f0_{src}$ is the source speaker pitch and $f0_{conv}$ is the converted pitch frequency for the target speaker. As mentioned in the introduction section, non-linear conversion models are more compatible with the complex human emotional voices. Therefore, we use the NNs models to train the F0 features in our proposed methods. The reason why we choose different models for F0 conversion and spectral conversion is that the spectral features and F0 features are not closely correlated and the F0 features are not as complex as spectral features. As shown in Fig. 3, the F0 feature obtained from STRAIGHT is one dimensional feature and discrete. Before training the F0 features by NNs, we need to transform the F0 features into the Normalized Segment F0 features (NSF0). We can transform F0 features into high-dimension data through the following two steps.
1) Normalizing the F0 features by Z-score normalization model, we can obtain the rescaled features that are normalized by the mean and variance $(0, 1)$. The standard score of the samples is calculated as follows:

$$z = \frac{x - \mu}{\sigma}, \tag{9}$$

where μ is the mean and σ is the standard deviation.
2) Transform the normalized F0 features to the segment-level features which are high-dimension ones. We form the segment-



A. NNs model

B. Tanh and Sigmoid curves

Fig. 4. NNs model and curves of activation function

level feature vector by stacking features in the neighboring frames as follows:

$$\begin{aligned} X_{N\times(2w+1)} &= [x_1, ..., x_m, ..., x_N]^{\mathrm{T}}, \\ x(m) &= [z(m-w), ..., z(m), ..., z(m+w)]^{\mathrm{T}}, \end{aligned} \tag{10}$$

where w is the window size on each side. (10) represents N examples of $2w+1$-dimensional source features. In the proposed model, we set $w = 12$. 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 convert the 25-dimentional NSF0 features by NNs. As shown in Fig.4A, we used the 4-layers NNs model to train the NSF0 features. The numbers of nodes from the input layer x to the output layer are [25 48 48 25]. Fig.3 shows that the curve of the F0 features are changed sharply during the whole time. Unlike the smooth curve of the spectral features, we adopt the tanh activation function:

$$f(x) = \tanh(x) = \frac{e^{2x}-1}{e^{2x}+1}, \tag{11}$$

which is different from the sigmoid function used in the DBNs with spectral features training models. As shown in Fig.4B, the tanh function has stronger gradient and the values are in the range $[-1, 1]$. These mean that the tanh function is more compatible to the sharply changed curve of F0 features.

III. Experiments

A. Database

We used a database of emotional Japanese speech constructed in [24]. 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.

B. Spectral features conversion

For the training and validation sets, we resampled the acoustic signals to 16kHz, extracted STRAIGHT parameters and used a Dynamic Time Wrapping (DTW) method to align the extract features. The aligned F0 features and MFCC (conducted by spectral features) were used as the parallel data. In our proposed method, we used the MCC features for training the DBNs models. Since the NNs model [11] proposed by Desai is the well-known voice conversion method based on Artificial Neural Network and the recurrent temporal restricted Boltzmann machines (RTRBMs) model [25] is the new and effective approach about voice conversion. We used NNs model and RTRBMs model to train the MCC features from the emotional voices to neutral voices for comparison. DBNs, NNs and RTRBMs are trained by using the MCC features of all datasets because considering the different emotion from FUM to the neural emotion of FON may influence the spectral conversion.

C. F0 features conversion

We used 4-layers NNs to convert the aligned NSF0 features. For comparison, we also used the Gaussian normalized transformation method to convert the aligned F0 features extracted from parallel data. The datasets are the different emotional voices from FUM to the neural voice of FON (angry to neural, happy to neural and sad to neural). For making the training data, each set concludes 50 sentences. For the validation, 10 sentences were arbitrarily selected from the database.

D. Results and discussion

Mel Cepstral Distortion (MCD) was used for the objective evaluation of spectral conversion:

$$MCD = (10/\ln 10)\sqrt{2\sum_{i=1}^{24}\left(mc_i^t - mc_i^e\right)^2} \tag{12}$$

where mc_i^t and mc_i^e represent the target and the estimated mel-cepstral, respectively. Fig.5 shows the result of the MCD test. As shown in this figure, our proposed DBNs model can convert the spectral features better than the NNs, and no significant difference with the RTRBMs. But the training time of the DBNs method is much faster than 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 little 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):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log\left(F0_i^t\right) - \log\left(F0_i^c\right)\right)^2} \tag{13}$$

where $F0_i^t$ and $F0_i^c$ denote the target and the converted F0 features, respectively. Fig. 6 shows that our proposed method obtains a better result than the traditional Gaussian normalized transformation method in the all datasets. (angry to neutral, happy to neutral, sad to neutral.)

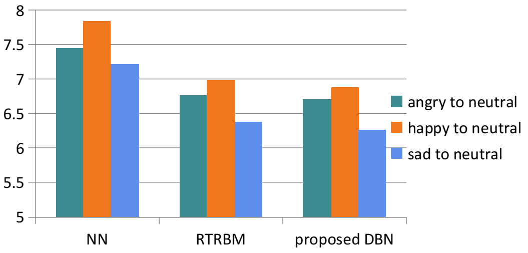


Fig. 5. Mel-cepstral distortion evaluation of spectral features conversion

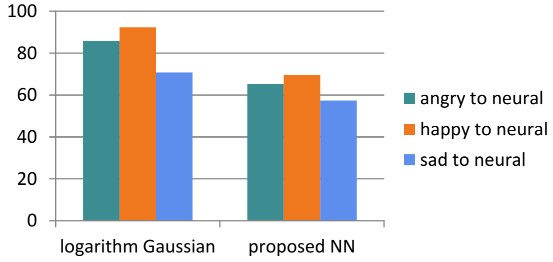


Fig. 6. Root mean squared error evaluation of F0 features conversion

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a method using DBNs to train the MCC 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 conventional methods (NNs and GMM) has shown that our proposed model can effectively change the acoustic voice and the prosody for the emotional voice at the same time.

There are still some problems in our proposed VC method. This method needs to conduct the parallel speech data that will limit the conversion only one to one. Recently, there are researches using the raw waveforms for deep neural networks training [26][27]. In the future work, we will apply the DBNs model which can straightly use the raw waveform features.

REFERENCES

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1. IEEE, 1998, pp. 285–288.

[2] S. Mori, T. Moriyama, and S. Ozawa, "Emotional speech synthesis using subspace constraints in prosody," in *Multimedia and Expo, 2006 IEEE International Conference on*. IEEE, 2006, pp. 1093–1096.

[3] R. Aihara, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders using dictionary selective non-negative matrix factorization," *ACL 2014*, p. 29, 2014.

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

[5] T. Raitio, L. Juvela, A. Suni, M. Vainio, and P. Alku, "Phase perception of the glottal excitation of vocoded speech," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[6] Z.-W. Shuang, R. Bakis, S. Shechtman, D. Chazan, and Y. Qin, "Frequency warping based on mapping formant parameters," in *Ninth International Conference on Spoken Language Processing*, 2006.

[7] D. Erro and A. Moreno, "Weighted frequency warping for voice conversion." in *Interspeech*, 2007, pp. 1965–1968.

[8] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, vol. 1. IEEE, 1992, pp. 137–140.

[9] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, 2007.

[10] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 912–921, 2010.

[11] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 3893–3896.

[12] Z. Wu, E. S. Chng, and H. Li, "Conditional restricted boltzmann machine for voice conversion," in *Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on*. IEEE, 2013, pp. 104–108.

[13] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets." in *INTERSPEECH*, 2013, pp. 369–372.

[14] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, "Approaching automatic recognition of emotion from voice: a rough benchmark," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[15] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1145–1154, 2006.

[16] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Gmm-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 134–138, 2012.

[17] Z. Wu and S. King, "Minimum trajectory error training for deep neural networks, combined with stacked bottleneck features," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[18] Š. Beňuš, U. D. Reichel, and J. Šimko, "F0 discontinuity as a marker of prosodic boundary strength in lombard speech," 2015.

[19] M. Ma, K. Evanini, A. Loukina, X. Wang, and K. Zechner, "Using f0 contours to assess nativeness in a sentence repeat task," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[20] 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.

[21] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various mfcc implementations on the speaker verification task," in *Proceedings of the SPECOM*, vol. 1, 2005, pp. 191–194.

[22] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[23] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin," in *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*, vol. 4. IEEE, 2007, pp. 410–414.

[24] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "Gmm-based voice conversion applied to emotional speech synthesis." *IEEE Trans Speech Audio Proc*, vol. 7, pp. 2401–2404, 2003.

[25] T. Nakashika, T. Takiguchi, and Y. Ariki, "High-order sequence modeling using speaker-dependent recurrent temporal restricted boltzmann machines for voice conversion," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[26] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[27] M. Bhargava and R. Rose, "Architectures for deep neural network based acoustic models defined over windowed speech waveforms," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.