INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The USTC System for Voice Conversion Challenge 2016: Neural Network Based Approaches for Spectrum, Aperiodicity and F_0 Conversion

Ling-Hui Chen[1,2], *Li-Juan Liu*[2], *Zhen-Hua Ling*[1], *Yuan Jiang*[2], *Li-Rong Dai*[1]

[1]National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, P.R.China
[2]iFLYTEK Research, Hefei, P.R. China

{chenlh,zhling,lrdai}@ustc.edu.cn, {ljliu,yuanjiang}@iflytek.com

Abstract

This paper introduces the methods we adopt to build our system for the evaluation event of Voice Conversion Challenge (VCC) 2016. We propose to use neural network-based approaches to convert both spectral and excitation features. First, the generatively trained deep neural network (GTDNN) is adopted for spectral envelope conversion after the spectral envelopes have been pre-processed by frequency warping. Second, we propose to use a recurrent neural network (RNN) with long short-term memory (LSTM) cells for F0 trajectory conversion. In addition, we adopt a DNN for band aperiodicity conversion. Both internal tests and formal VCC evaluation results demonstrate the effectiveness of the proposed methods.
Index Terms: voice conversion, frequency warping , DNN, RNN, LSTM

1. Introduction

Voice conversion (VC) is a technique that modifies the speech characteristic of a source speaker in order to make it sounds like being uttered by the target speaker. This technique can also be applied to many other relevant areas, such as speech enhancement [1], foreign language learning [2], and so on.

Most of the mainstream voice conversion methods consists of two parts: spectral conversion and prosody conversion. As spectra directly convey most of the timbre characteristics of a speaker, many approaches that focused on spectral conversion have been proposed. In the early methods, such as codebook mapping based methods [3], the generated speech was always discontinuous due to the hard clustering and discrete mapping of spectral features. In order to cope with this problem, Gaussian mixture model (GMM) based methods were proposed for soft clustering and continuous mapping of spectral features [4, 5]. In these methods, the features of source speaker's spectral sequence were converted frame by frame. Thus, they did not consider the continuous sequential nature of spectral features. Considering this, Toda et al. proposed to model the spectral features with dynamic components [6]. In this method, maximum likelihood parameter generation (MLPG) method was adopted to generate the converted spectral feature trajectories. Global variances (GVs) of feature trajectories were also considered and modeled in order to alleviate the over-smoothing problem of converted acoustic features.

This work was partially funded by the National Nature Science Foundation of China (Grant No.61273032) and the Electronic Industry Development Fund of Ministry of Industry and Information Technology (Grant No. [2014]425).

In recent years, deep neural networks (DNNs) have been successfully applied to many areas of speech processing, such as automatic speech recognition (ASR) [7] and statistical parametric speech synthesis (SPSS) [8]. DNNs have the advantage of modeling complex non-linear functions. Therefore, they have also been applied to voice conversion to describe the mapping relationship between source and target speakers. Chen et al. proposed to use restricted Boltzmann machines (RBMs) to model the joint distribution of source and target feature spaces [9]. A combination of deep belief networks (DBNs) and neural networks was introduced by Nakashika et al. in order to build the spectral mapping in a high-order eigenspace [10]. Nakashika et al. also proposed to use recurrent temporal RBM to model the temporal correlations across sequential frames [11]. Sun et al. proposed to use the bi-directional LSTM based RNN to model temporal dependencies among frames in a time sequence [12].

On the other hand, prosody feaures, such as fundamental frequency (F_0), duration, intensity, etc., also contain important speaker characteristics. Since it is difficult to model these features without linguistic information, most of the current voice conversion methods only use a simple F_0 conversion, which is a Gaussian normalization process that linearly converts F_0 values of the source speaker in log-scale to match that of the target speaker. However, this simple conversion is insufficient to convert speaker's prosody characteristics.

In this paper, we introduce the methods we used to build our system for Voice Conversion Challenge 2016 (VCC 2016). The task of VCC 2016 consists of speaker conversion of 25 source-target speaker pairs. We use neural network based methods for both spectral conversion and excitation conversion. The generatively trained deep neural network (GTDNN) is utilized for spectral conversion. A bilinear transformation based frequency warping method is adopted as a pre-processing of source spectral envelopes to improve conversion performance. The excitation conversion includes F_0 conversion and aperiodicity conversion. A combination of the LSTM-RNN based trajectory conversion and the Gaussian normalization method is proposed for converting F_0 sequences. And a DNN is adopted to convert the aperiodic components. Experimental results demonstrate the improvement obtained by using these methods.

2. Methods

2.1. Spectral Conversion

We adopt the generatively trained DNN (GTDNN) for spectral envelope conversion in our system. Although previous

http://dx.doi.org/10.21437/Interspeech.2016-456

work [13] showed that GTDNN can significantly improve the performance of spectral conversion comparing with conventional GMM based method, it is still difficult to obtain a good spectral conversion function for source-target pairs that have large differences on spectral characteristics. For instance, the degraded conversion performance is usually obtained in the cases of cross-gender conversions. In order to improve the conversion performances for such cases, we propose a combination of frequency warping and GTDNN-based conversion. The frequency warping is conducted as a pre-processing to the spectra of source speaker in order to reduce the spectral differences between input spectra and target spectra for GTDNN. Therefore, it would be relatively easier for the GTDNN model to learn the mapping function between the warped source spectra and the target spectra. The warping factor of the bilinear transformation [14] based frequency warping is optimized by a grid searching in the mel-cepstral space, i.e., we traverse all frequency warping factors and use the best value that can minimize the mel-cepstral distortion between the warped mel-cepstra and the target ones.

2.2. Excitation Conversion

2.2.1. F_0 Conversion

The Gaussian normalization based method for F_0 conversion has the advantage of preserving the natural F_0 trajectory contour of source speech, leading to highly natural prosody in converted speech. However, this method is only a simple frame-wise linear model for $\log F_0$ conversion. It can only convert the range of F_0 and cannot reshape F_0 contours, which contains a lot of speaker characteristics. Besides, the F_0 features extracted from source speakers always contain errors, such as half frequency and double frequency. The Gaussian normalization method can not handle this issue when generating the F_0 features of target speakers.

In order to cope with the problems, we propose to use RNNs with LSTM units (LSTM-RNN) for F_0 trajectory transformation. The long-term dependencies in the F_0 trajectory can be captured by RNNs because of the recurrent characteristic of hidden units. Comparing with the Gaussian normalization method, this model can reshape F_0 trajectories and convert the local characteristics in F_0 contours. In addition, F_0 values predicted by LSTM-RNNs would be more robust to the errors of F0 extraction, because the model can achieve temporal smoothing when generating output sequences due to the cross-frame dependency modeling. In this paper, an LSTM-RNN with linear output layer for regression is used to predict static F_0 values together their corresponding delta and acceleration components. Then the MLPG algorithm was applied to generate the converted F_0 sequences.

In our experiments we observed that the F_0 trajectories generated using the LSTM-RNN model tend to be over-smoothed. This greatly affects the naturalness of converted speech. This issue may be caused by the insufficient number of training samples in voice conversion for training the LSTM-RNN model. In order to address this issue, we propose to combine the Gaussian normalization and the LSTM-RNN conversion method. It is implemented by adding a constraint describing the difference between the generated F_0 trajectory and the Gaussian normalized F_0 trajectory into the cost function for MLPG. Supposing the mapping function constructed by the LSTM-RNN is denoted as $\boldsymbol{Y} = g(\boldsymbol{X})$, where $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, ..., \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, ..., \boldsymbol{Y}_T^\top]^\top$ are the input and output feature sequences, respectively. Meanwhile, let $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \hat{\boldsymbol{y}}_2^\top, ..., \hat{\boldsymbol{y}}_T^\top]^\top$ represents the static F_0 sequence generated by Gaussian normalization method. Then, the F_0 trajectory generated by the proposed method is

$$\boldsymbol{y}^* = \arg\min_{\boldsymbol{y}} (-\log P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) + \alpha \cdot |\boldsymbol{y} - \hat{\boldsymbol{y}}|^2), \quad (1)$$

where $\boldsymbol{Y} = \boldsymbol{M}\boldsymbol{y}$, and $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ is a Gaussian distribution

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) = N(\cdot; g(\boldsymbol{X}), \boldsymbol{\Sigma}_{\boldsymbol{g}}), \quad (2)$$

whose mean vector is given by the output of the LSTM-RNN and its covariance matrix is diagonal and given by the global variance of target F_0 features. α is a factor that needs to be tuned. The proposed method can be viewed as a trade-off between the Gaussian normalization and the LSTM-RNN based conversion.

The characteristics of F_0 features are strongly correlated with linguistic information. Since transcriptions and phoneme alignment are not available for most voice conversion tasks as well as in this challenge, the bottleneck features (BNs), which are initially designed for ASR [15] and are assumed to be strongly related to phone categories, are used as a part of the input features for the construction of the LSTM-RNN based F_0 conversion model in this paper. In our implementation, the DNN for extracting BNs was trained with a 2000-hour speech corpus [16]. The BN layer in the DNN was the last hidden layer and the dimensionality of BN features was 50.

2.2.2. Aperiodicity Conversion

On the other hand, the aperiodicity (AP) components of the excitation also conveys important information of speaker individuality. There are normally two ways to handle AP in voice conversion. The first one directly uses the AP of source speaker to synthesize converted speech [17, 12, 18]. The second one performs model based band aperiodicity (BAP) conversion, e.g., GMM based mapping. In our system a DNN is adopted as a regression model to capture the non-linear mapping relationships between BAPs of source speaker and target speaker. The DNN is trained in a conventional way using the back propagation (BP) algorithm.

3. Internal Experiments

3.1. Experimental Setup

The VCC 2016 event provided an English corpus consisting of 5 source speakers (3 female and 2 male, named as SF1, SF2, SF3, SM1, SM2) and 5 target speakers(2 female and 3 male, named as TF1, TF2, TM1, TM2 and TM3). Each speaker uttered the same sentence set consisting of 162 sentences. The waveforms were recorded with 16kHz/16bit format. We randomly chose 150 sentences for training and the remaining 12 sentences were left for test in our internal experiments. The STRAIGHT [19] vocoder was used to extract acoustic features, including F_0, BAP, and spectral envelope, and to synthesize speech waveform using converted acoustic features. The acoustic features were extracted at a frame shift of 5ms. The length of FFT was set to 1024, leading to 513-dimension spectral envelopes. The alignment between acoustic feature sequences of the source speaker and the target speaker was obtained using dynamic time warping (DTW) algorithm.

For spectral conversion, the GTDNN-based spectral conversion module was constructed following the configuration in [13]. The original F_0 sequences were interpolated to continuous

Table 1: Preference scores (%) on naturalness (Nat.) and similarity (Sim.) of systems with and without frequency warping (FW) based pre-processing.

		w/o FW	w. FW	N/P	p-value
Nat.	F2F	32.58	25.76	41.66	0.307
	F2M	13.64	**31.82**	54.54	**0.002**
	M2F	17.50	**33.33**	49.17	**0.014**
	M2M	17.42	**37.12**	45.46	**0.002**
Sim.	F2F	25.76	22.73	51.51	0.619
	F2M	16.67	25.00	58.33	0.138
	M2F	14.17	**31.67**	54.16	**0.004**
	M2M	22.73	34.85	42.42	0.066

F_0 sequences using an exponential decay function [20] for training LSTM-RNN. The input features consisted of 50-order BN features and 3-order logF_0 (i.e. the static, delta and acceleration components) of the source speaker. The network outputs the converted logF_0 sequences together with their delta and acceleration components. Both the input and output features were normalized to zero mean and unit variance before training. An LSTM-RNN with one uni-directional hidden recurrent layer of 32 LSTM cells was used in our experiments It was trained using back propagation through time (BPTT) algorithm with 5 frame delay to consider future dependencies. The learning rate was 0.01. α in (1) was set to 0.3. The DNN for BAP conversion had two hidden layers, each of which had 128 nodes. The input feature consisted of 16-order mel-cepstra, logF_0 and 5-order BAP of the source speaker. The original logF_0 values without interpolation were used directly as input in order to incorporate the voicing/unvoicing information for BAP conversion. In addition, a context window of 3 frames (i.e. previous, current and next) was adopted for composing input features for DNN training.

3.2. Subjectitve Evaluation on Frequency Warping Based Pre-processing

A preference listening test was conducted to evaluate the performance of the frequency warping based pre-processing for spectral envelope conversion. All the internal listening tests in this paper were conducted on the Amazon Mechanical Turk (AMT) [1], a crowd-source platform. Four conversion pairs, each of which was randomly chosen from female-to-male (F2M), female-to-female (F2F), male-to-male (M2M), and male-to-female (M2F) conversion types respectively, were used for the test in this experiment. 10 subjects participated in each of these tests. The results on similarity and naturalness are presented in Table. 1. It can be seen that frequency warping can help to improve both the similarity and naturalness of the speech generated using GTDNN-based spectral conversion. The similarity of M2F was significantly improved and all naturalness was significantly improved except for F2F. We found that the estimated frequency warping function for F2F conversions were close to linear mapping in our experiments.

3.3. Subjective Evaluation on BAP Mapping

In this section, two systems were compared: the one with BAP conversion and the one that didn't use BAP for synthesis speech. 10 paid native English speakers from the AMT listening platform participated in these tests. Results in Table 2 showed that the naturalness was improved for all conversion types by adopting BAP mapping, even if the improvement was only significant for M2M. The proposed BAP mapping can improve the similarity of converted speech except M2F, and the improvement was significant for F2M and M2M.

[1] https://www.mturk.com

Table 2: Preference scores (%) for the evaluation of BAP mapping.

		w/o BAP.	w/ BAP conv.	N/P	p-value
Nat.	F2F	34.85	37.12	28.03	0.759
	F2M	25.76	31.06	43.18	0.421
	M2F	31.67	40.00	28.33	0.282
	M2M	13.33	**44.17**	42.50	**0.000**
Sim.	F2F	33.33	36.36	30.31	0.678
	F2M	17.42	**29.55**	53.03	**0.041**
	M2F	26.67	25.00	48.33	0.800
	M2M	14.17	**38.33**	47.50	**0.000**

Table 3: RMSEs (Hz) of two F_0 conversion methods between each source speaker (row) and each target speaker (column).

		SF1	SF2	SF3	SM1	SM2
Gauss.	TF1	46.38	40.03	39.79	35.06	45.63
	TF2	44.46	32.91	37.39	39.20	35.80
	TM1	22.18	21.17	19.61	22.34	17.92
	TM2	15.16	16.56	14.34	15.04	15.11
	TM3	23.45	21.58	21.89	22.19	24.15
LSTM	TF1	34.68	35.24	35.73	31.45	36.16
	TF2	33.93	31.18	36.29	35.11	32.27
	TM1	16.77	18.89	19.35	18.48	17.46
	TM2	12.95	14.63	14.33	12.96	13.42
	TM3	18.47	19.83	21.25	19.77	20.49

3.4. Evaluation on LSTM-RNN Based F_0 Conversion

3.4.1. Objective evaluation

Firstly, we compared the root mean square errors (RMSEs) of the LSTM-RNN based F_0 trajectory mapping and the conventional Gaussian normalization based conversion method. The results of all the 25 conversion pairs are presented in Table 3. It can be seen that the RMSEs of LSTM-RNN is consistently decreased compared with the conventional method except for the SM2-TM1 conversion pair. The average RMSE reduction after using the LSTM-RNN based method is 3.45 Hz, relatively 12.54% reduction from 27.49 Hz to 24.04 Hz.

3.4.2. Subjective evaluation

Preference listening tests were conducted to evaluate the performance of the proposed LSTM-RNN based method and the Gaussian normalization method. 10 conversion pairs, covering all four conversion types , were selected for evaluation. Both naturalness and similarity were evaluated in the tests. 10 paid native English speakers participated in these tests via the AMT platform. Results presented in Table 4 show that the proposed method achieved better speech naturalness on M2F and M2M conversion pairs, while the listeners preferred to voices converted by the baseline method on F2F and F2M. It

Table 4: Preference scores (%) for comparing two F_0 conversion methods.

		Gauss.	LSTM	N/P	p-value
Nat.	F2F	36.36	35.99	27.65	0.942
	F2M	30.81	23.74	45.45	0.056
	M2F	23.11	**35.99**	40.90	**0.006**
	M2M	26.04	**34.38**	39.58	**0.029**
Sim.	F2F	29.92	**44.70**	25.38	**0.005**
	F2M	31.82	24.75	43.43	0.061
	M2F	22.73	**23.48**	53.79	**0.014**
	M2M	27.16	29.69	43.15	0.638

can also be seen that the similarity of the voices converted by the proposed method were significantly improved on F2F and M2F. Although there were improvements on M2M, the results were not significant. In addition, the baseline method got better speaker similarity on F2M.

4. Evaluation of VCC 2016

According to our internal experimental results, we applied different strategies for different conversion pairs. Firstly, BAP conversion was only applied to the pairs with male target; for the others, BAP features were not used. Secondly, LSTM-RNN based F_0 conversion was adopted for the pairs with female source speaker; Gaussian normalization was adopted for other pairs.

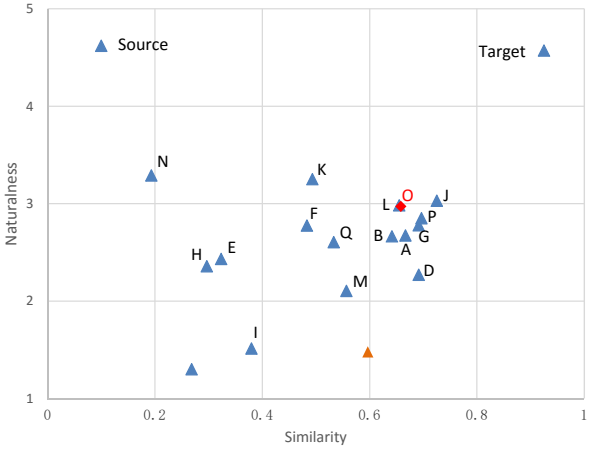

Figure 1: Scatter plot of overall similarity and naturalness scores for all systems in VCC 2016.

The overall naturalness and similarity scores announced by event organizers are plotted in Figure 1. For the convenience of analysis, only the binary similarity scores are used. It can be seen that our system (system O) locates in the area of the top systems, which proves the effectiveness of our method. According to the results significance analysis, system K and system N are significantly better than our system on naturalness. However their similarity scores are relatively lower. There is no significant difference among our system and system J and L. In the similarity evaluation, although system A, D, G, J and P get higher similarity score than our system, significance analysis shows no significant difference between our system and each of these systems.

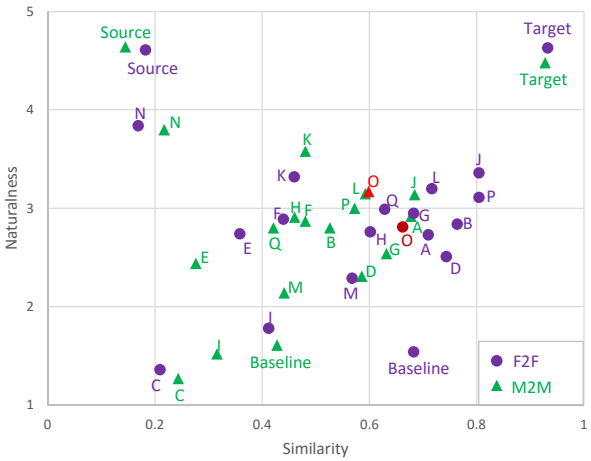

Figure 2: Scatter plot of scores of intra-gender conversion.

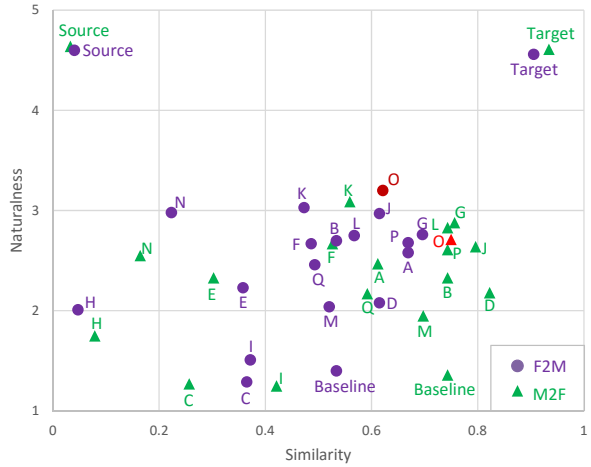

Figure 3: Scatter plot of scores of cross-gender conversion.

Figure 2 shows the scatter plot of intra-gender (F2F and M2M) conversions, and Figure 3 shows that of cross-gender (M2F and F2M) conversions. It shows that our system performs well on cross-gender conversion. In intra-gender conversion, our system also worked well on M2M, however it didn't work well on F2F. As introduced above, we didn't apply any proposed excitation conversion methods on F2F conversion according to the results of our internal listening test which may not use enough test sentences to get reliable results.

5. Conclusions

In this paper, we presented the details of our system for VCC 2016. We built a system that utilized neural networks for conversions of all acoustic features. GTDNNs were adopted for spectral envelope conversion. Frequency warping was used as a pre-processing of the source spectra in order to promote the conversion performance of GTDNNs. We also introduced to use LSTM-RNNs to realize F_0 trajectory conversion. Besides, DNNs were used for BAP conversion. Internal experiments and formal evaluation results showed the effectiveness of our system. However, our system didn't perform well on F2F conversion. To improve the robustness of our method will be the future work. We also plan to find a unified model to convert all features simultaneously instead of using separate models.

6. References

[1] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2505–2517, 2012.

[2] D. Felps, H. Bortfeld, and R. Gutierrez-Osuna, "Foreign accent conversion in computer assisted pronunciation training," *Speech communication*, vol. 51, no. 10, pp. 920–932, 2009.

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988, pp. 655–658.

[4] Y. Stylianou, O. Cappe, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Audio, Speech, and Lang. Process*, vol. 6, no. 2, pp. 131–142, mar. 1998.

[5] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proc. ICASSP*, 1998, pp. 285–288.

[6] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, and Lang. Process*, vol. 15, no. 8, pp. 2222–2235, nov. 2007.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[8] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7962–7966.

[9] L.-H. Chen, Z.-H. Ling, Y. Song, and L.-R. Dai, "Joint spectral distribution modeling using restricted Boltzmann machines for voice conversion," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 3052–3056.

[10] T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 369–372.

[11] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using rnn pre-trained by recurrent temporal restricted boltzmann machines," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 580–587, March 2015.

[12] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4869–4873.

[13] L.-H. Chen, Z.-H. Ling, L.-J. Liu, and L.-R. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 12, pp. 1859–1872, 2014.

[14] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 930 – 944, sep. 2005.

[15] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks." *Interspeech*, pp. 237–240, 2011.

[16] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text." in *LREC*, vol. 4, 2004, pp. 69–71.

[17] H. Siln, J. Nurminen, E. Helander, and M. Gabbouj, "Voice conversion for non-parallel datasets using dynamic kernel partial least squares regression," in *Interspeech*, 2013.

[18] X. Tian, Z. Wu, S. W. Lee, and N. Q. Hy, "Sparse representation for frequency warping based voice conversion," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015.

[19] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–208, 1999.

[20] C. J. Chen, R. A. Gopinath, M. D. Monkowski, M. A. Picheny, and K. Shen, "New methods in continuous mandarin speech recognition." in *Eurospeech*, 1997.