Deep Neural Network based Voice Conversion with A Large Synthesized Parallel Corpus

Zhengqi Wen[*1], Kehuang Li[†], Jianhua Tao[*] and Chin-Hui Lee[†]
[*] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
E-mail: {zqwen, jhtao}@nlpr.ia.ac.cn Tel: +86-15001385087
[†] School of ECE, Georgia Institute of Technology, Atlanta, GA. 30332-0250, USA
E-mail: kehle@gatech.edu, chl@ece.gatech.edu

***Abstract*— we propose a voice conversion framework to map the speech features of a source speaker to a target speaker based on deep neural networks (DNNs). Due to a limited availability of the parallel data needed for a pair of source and target speakers, speech synthesis and dynamic time warping are utilized to construct a large parallel corpus for DNN training. With a small corpus to train DNNs, a lower log spectral distortion can still be seen over the conventional Gaussian mixture model (GMM) approach, trained with the same data. With the synthesized parallel corpus, a speech naturalness preference score of about 54.5% vs. 32.8% and a speech similarity preference score of about 52.5% vs. 23.6% are observed for the DNN-converted speech from the large parallel corpus when compared with the DNN-converted speech from the small parallel corpus.**

I. INTRODUCTION

Voice conversion (VC) is a technology that modifies a source speaker's utterance to sound like a target speaker. There are plenty of techniques proposed in literature to realize voice conversion, such as vector quantization [1], Gaussian mixture models (GMM) [2], pitch-synchronous overlap addition [3], artificial neural network [4], and multiple function [5]. Among these methods, GMM is one of the most popular methods. However it often suffers from the low quality problems, such as over-smoothing [6] and over-fitting [7].

We believe there are two key issues need to be addressed for a high quality VC. First, the mapping function for transforming the speech features from the source to the target speakers. In contrast to probabilistic GMM approaches, regression [8] and classification [9-10] based deep neural networks (DNNs) have recently attracted a lot of attention due to its great modeling capabilities. Besides automatic speech recognition it has also been adopted in voice conversion. Desai *et al*. [11] utilized a mapping function based on artificial neural network (ANN). Chen *et al*. [12-14] used restricted Boltzmann machine (RBM) and Bernoulli bidirectional associative memory to construct a global nonlinear mapping. Nakashika *et al*. [15] proposed two deep belief networks (DBNs) and an ANN for conversion. Xie *et al*. [16] trained an ANN with a sequence error minimization criterion for the speech features. Mohammadi *et al*. [17] proposed ANN based conversion from a deep auto-encoder. All these systems were usually built with a limited number of parallel utterances which is often too small to train a good DNN for the high-quality voice conversion.

The other critical issue is that the required parallel corpus is often not easy to construct. The widely used CMU ARCTIC corpus [18] has only 1132 utterances for every speaker, too small to train a high-quality voice conversion operation. There are also some conversion methods using nonparallel corpora, e.g., phonetic information based alignment [19] and vocal tract length normalization [20]. As summarized in [21], the more similar the corresponding source and target speakers are, the less speaker-dependent information can be taken advantage of.

In this paper, we propose a DNN-based voice conversion function which is trained on a highly-desired, large synthesized parallel corpus obtained with the proposed speech synthesis [22-24] and dynamic time warping (DTW) techniques [25]. The DNN is stacked with pre-trained RBMs and fine-tuned with a minimum mean square error (MMSE) criterion. The proposed DNNs generate a lower log spectral distortion (LSD) when compared with the conventional GMM approach with a small training data. With the large synthesized parallel corpus, the objective and subjective experimental results also demonstrate an even better performance over the same DNN-based conversion systems, but trained with only a small corpus.

II. DNN-BASED VOICE CONVERSION

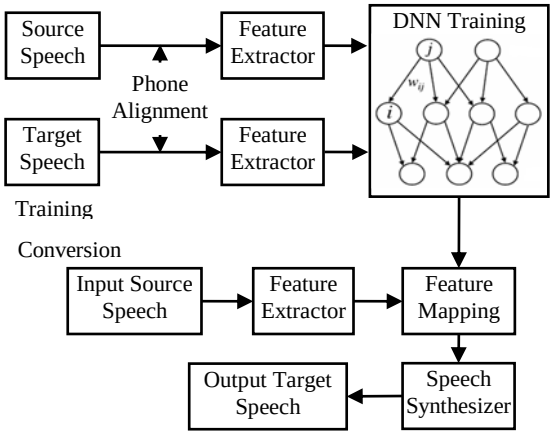

Figure 1: *A flow chart of DNN-based voice conversion.*

The workflow of the DNN based voice conversion system is described in Fig. 1. DNN is used as a regression tool to construct the nonlinear mapping for the speech features from the source to the target speakers.

DNN training is split into two steps [26]: pre-training and fine-tuning which are shown in the left and right of Fig. 2, respectively. In the pre-training stage, a number of RBMs are trained with a contrastive convergence (CD) criterion. The input of the first layer is normalized with zero mean and unity variance so the pre-trained

[1] Work done while visiting Georgia Tech in 2014-2015

DNN is stacked as the Gaussian-Bernoulli RBM and the rest of Bernoulli-Bernoulli RBMs.

The pre-trained RBMs are next fine-tuned with the MMSE criterion. The input source feature is propagated as in Eq. (1) and the MSE is defined as the Euclidean distance between the generated and the original target features in Eq. (2).

$$\hat{y} = \tilde{g}\left(g(\dots g(W_1, b_1, x))\right) \quad (1)$$

where W and b are the weight matrix and bias vector, g is the sigmoid function, $\tilde{g}$ is a linear function, x is the input speech feature and $\hat{y}$ is the generated target feature.

$$D_{MSE}(\hat{y}, y) = \frac{1}{T}\sum_{t=1}^{T}(\hat{y} - y)^2 \quad (2)$$

where T is the frame number, y is the target speech feature and $\hat{y}$ is the generated target speech feature.

A stochastic gradient descent algorithm is performed in mini-batches to update the weights in Eq. (3). The mini-batch size is set as 256 and learning rate is set as 0.0005 in the following DNN training.

$$(W_l, b_l) \leftarrow (W_l, b_l) + \lambda \frac{\partial D_{MSE}(\hat{y}, y)}{\partial (W_l, b_l)} \qquad 0 \le l \le L \quad (3)$$

where L is the layer number and λ is the learning rate.

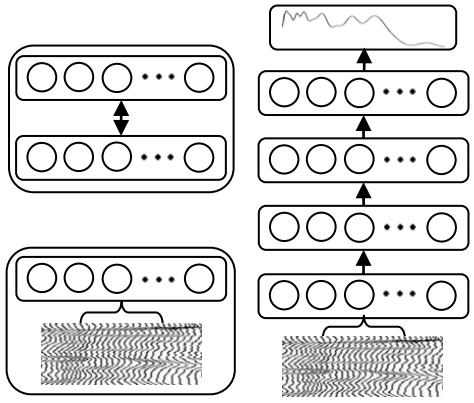

Figure 2: *Pre-training and fine-tuning with* stacked *RBMs.*

III. PROPOSED PARALLEL CORPUS

In a conventional voice conversion system, some features are first extracted from speech waveform and then a mapping function is established to map the features from the source to the target speakers. The features usually reflect the same content and thus a parallel corpus is needed. Furthermore, the features should be aligned in the time domain due to varying lengths of the same sound uttered by different speakers. It is not easy to fulfill these requirements especially for a large parallel corpus. In the following we propose to construct a parallel corpus with a limited unparalleled corpus using speech synthesis methods with dynamic time warping based phone alignments.

A. Parallel Corpus Construction

Hidden Markov model (HMM) [27] based speech synthesis [22] and unit-selection based speech synthesis [23] are two prevalent approaches commonly used. Speech generated from HMM-based speech synthesis is flexible, but with a smooth quality while speech generated from unit-selection based speech synthesis is not as smooth but exhibiting a higher quality. These two methods will be both adopted and a series of comparison experiments will be carried out in Section 4.

In HMM-based speech synthesis, two systems should first be built with the limited unparalleled data for the source and the target speakers. The correlation between text information and the speech features, such as duration, spectral parameter and fundamental frequency (F0), is constructed by decision trees with a maximum likelihood (ML) criterion at the HMM state level. In the synthesis stage, the input text is transmitted to label sequences which are then put into the decision trees to find the corresponding state-level duration, spectrum and F0 parameters. The speech features are then generated with the ML parametric generation (MLPG) algorithm [28] from the corresponding GMMs in Eq. (4),

$$c = \left(W^T \hat{U} W\right)^{-1} W^T \hat{U} \hat{\mu} \quad (4)$$

where W is the window matrix including dynamic features, $\hat{U}$ is the covariance matrix and $\hat{\mu}$ is the mean vector.

In unit-selection based speech synthesis, we adopted our previous proposed method called hybrid speech synthesis [24] to construct the parallel corpus. The proposed technique can generate speech with a mean opinion score (MOS) of 3.8 and get a higher preference score over the traditional hybrid speech synthesis systems.

Here the generated speech is directly selected from the original corpus according to the maximum likelihood criterion. Assuming a sentence contains N syllables, $(\lambda_1, \lambda_2, \dots, \lambda_N)$, then λ_n^i and λ_n^f are defined as the trained initial and final models for every Mandarin syllable. The corresponding speech, $u_n = (u_{n,1}, u_{n,2}, \dots, u_{n,T})$, for the nth syllable is also split into $u_{n,t}^i$ and $u_{n,t}^f$ for the initial and final. So the likelihood of the candidate for the nth syllable is defined as follows:

$$LL(u_n, \lambda_n) = LL(u_n^i, \lambda_n^i) + LL(u_n^f, \lambda_n^f) \quad (5)$$

$$LL(u_n^i, \lambda_n^i) = \log P(u_n^i | \lambda_n, Q_n^i) + \log P(T_n^i | \lambda_n^{i,dur}) \quad (6)$$

$$LL(u_n^f, \lambda_n^f) = \log P(u_n^f | \lambda_n, Q_n^f) + \log P(T_n^f | \lambda_n^{f,dur}) \quad (7)$$

where Q_n^i and Q_n^f stand for the state allocations. $\lambda_n^{i,dur}$ and $\lambda_n^{f,dur}$ are the duration models. T_n^i and T_n^f are the frame numbers.

The optimal syllable sequence u^* is then solved as follows:

$$u^* = \arg\max_u \sum_{n=1}^{N} LL(u_n, \lambda_n) \quad (8)$$

Searching is expanded into a two-dimension space. One is the syllable sequence and the other is the candidates for every syllable. It can be realized by dynamic programming. Before that, the cost in Eq. (5) is converted into the traditional form of a sum of "target cost" and "concatenation cost" as follows:

$$u^* = \arg\min_u \{\sum_{n=1}^{N} TC(u_n) + \sum_{n=1}^{N} CC(u_{n-1}, u_n)\} \quad (9)$$

where $TC(u_n)$ indicates the weighted sum of likelihood in u_n from the second frame to last but one frame, and $CC(u_{n-1}, u_n)$ calculates the sum of likelihood of the last frame in u_{n-1} and the first frame in u_n.

B. Features Alignment

DTW is used in most typical utterance alignment algorithm. Speech features are aligned in a minimum distance constraint through a Viterbi search algorithm. The result of DTW depends on the distance measure and is sensitive to noise. However, the aligned result is very vital to train the mapping function. So we adopt a reliable alignment method described below.

In the parallel corpus constructed in the speech synthesis methods, the source and the target speakers utter the same text which means the two corresponding phone sequences are expected to be the same. The uttered phone lengths can get from the speech synthesis methods directly. For example, the phone's duration from the HMM-based speech synthesis is decided by the decision tree and the phone's duration from hybrid speech synthesis can then be obtained from phone segmentation of the original corpus. The alignment between the source and the target speakers can be done by interpolating the

target speaker's speech feature to the same length as the source speaker for every phone considered.

IV. Experiments and Discussion

In this section, the effectiveness of the proposed techniques for voice conversion is evaluated. Detail about the experiments is described in Section 4.1. The proposed DNN-based voice conversion system is set up in Section 4.2. Then in Section 4.3, DNN-based VC is compared with GMM-based VC with the same training data. The effectiveness of the proposed VC with the synthesized corpus is evaluated in Section 4.4.

A. Experiment Setup

The parallel corpora used in the following experiments were from a female talker and a male talker both speaking Mandarin with the same content for about three hours. And the rest of the unparalleled sentences were about four hours. The speech parameters used in these experiments were line spectral pair (LSP) [29] with dynamic features as the spectral parameters which were extracted from the STRAIGHT spectrum [30] and logarithmic fundamental frequency (LF0) as the excitation parameter. In synthesis stage, maximum likelihood parametric generation (MLPG) [28] algorithm is adopted to generate speech parameters with the dynamic features. Most of the speaker's characteristics have been included in the spectrum, so in this study we only considered the LSP conversion from the source speaker to the target speaker. The LF0 conversion was realized in the following equation,

$$\mathrm{LF0_T} = (\mathrm{LF0_S} - \overline{\mathrm{LF0_S}}) * \frac{\mathrm{Var_T}}{\mathrm{Var_S}} + \overline{\mathrm{LF0_T}} \qquad (10)$$

where $\overline{\mathrm{LF0_S}}$ and $\overline{\mathrm{LF0_T}}$ are the LF0's mean of the source speaker and target speaker, respectively.$\mathrm{Var_S}$ and $\mathrm{Var_T}$ are the LF0's variance of the source speaker and target speaker, respectively. $\mathrm{LF0_S}$ and $\mathrm{LF0_T}$ are the input source and the converted target speaker's LF0.

The quality of the converted speech was verified through log spectral distortion (LSD in dB) [31] and ABX preference tests [32] in naturalness and similarity. In the ABX preference tests, listeners will be asked to listen to two versions of converted speech and choose one which sounds much better than the other in the naturalness tests or much closer to the original speech in the similarity tests. The better one will get a preference score of "1" or no preference (N/P) score of "1". The final scores were calculated by the mean value of the scores given by the listeners in our lab. Fifteen native speakers, who had previously undergone listening tests in our lab, took part in our experiments and listened to about twenty sentences for every converted method.

B. Preliminary DNN Experiments

The first type of DNN was trained with the existing parallel corpus of 3 hours without using any synthesis. This network contained four hidden layers with 1024 units. The input and the output formed a pair of the source and target speaker's LSPs. The DNNs (DNN-0 … DNN-10) trained with different window lengths of contextual information [22] were compared in Fig. 3 and Table 1, where DNN-0 contains only one frame and DNN-10 contains one central frame with 10 left and 10 right frames. It shows that the LSD between the generated and the target spectra decreases as the context window length increases. When the window length is beyond 3, the LSD decreases very slowly and stops improving after the window length of 6. So considering the computing complexity and quality, the window length for contextual information was fixed at 6 for all the following series of experiments. An ABX preference test of naturalness was conducted to confirm the effectiveness of the LSD measure for describing the synthesized speech's quality. The experimental result in Table 1 declares a clear improvement of increasing the window length from a shorter DNN-0 with a 0.145 score to a longer DNN-6 with a 0.812 score. The third column of 0.043 is the no preference score.

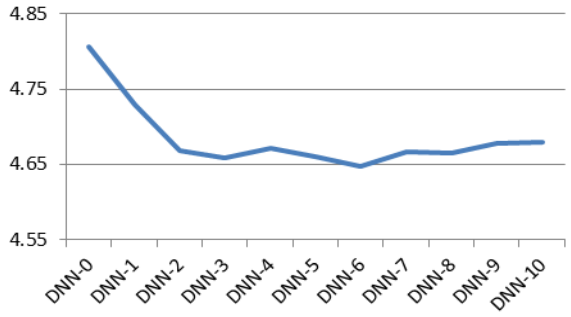

Figure 3: *Log spectral distortion (LSD) with different window lengths for contextual information.*

Table 1: Naturalness preference score with a 0.05 confidence interval comparison converted between DNN-0 and DNN-6

System	DNN-0	DNN-6	N/P
Naturalness Preference	0.145	**0.812**	0.043

C. Comparison with GMM-based Method

A GMM-based voice conversion system [6] was trained on the existing parallel corpus and 2047 GMMs were kept for the conversion. To enhance the synthesized speech's quality, the global variance (GV) of frequency domain delta LSP [33] was trained for overcoming the over-smoothing problem.

The comparison results are listed in Tables 2 and 3. The scores on naturalness declare a clear preference of DNN-6 with 47.8% over GMM with 9.7%. As for the similarity preference scores, the difference between DNN-6 (at 32.3%) and GMM (at 6.8%) is also significant. LSD in Table 3 also gives DNN-6 a 0.117 dB advantage over GMM. It can be concluded that the proposed DNN-based approach is much better than the GMM-based methods in voice conversion.

Table 2: Preference scores between the converted speech from the DNN-6 and GMM with a 0.05 confidence interval

	DNN-6	GMM	N/P
Naturalness	**0.478**	0.097	0.425
Similarity	**0.323**	0.068	0.609

Table 3: LSDs between the natural speech and converted speech by the DNN-6 and GMM

Methods	LSD in dB
GMM	4.764
DNN-6	4.647

We believe there are two key reasons: (1) DNN's great nonlinear approximation power, and (2) easy incorporation of contextual information in DNN. When the window length sets as 0 as shown in Fig. 3 and Table 3, the LSD between DNN-0 and GMM is actually very small. But when the window length increases, the discrepancies between these two methods become obvious.

D. *Experiments based on Synthesized Large Corpora*

The second type of VC DNN, in contrast to the one discussed in Section 4.2, was trained with 50 hours of enlarged parallel data using the proposed synthesis framework discussed in Section 3. Two parallel corpora, one generated from HMM-based (Corpus1) and the other from hybrid synthesis (Corpus2), were produced.

First, the two synthesis methods were compared in LSD in Table 4. Speech generated from the HMM-based synthesis method takes smaller LSD than that in hybrid speech synthesis. We believe it is mainly attributed to the flexibility of HMM-based synthesis. For example, the duration for every Mandarin syllable is much easier to control in HMM-generated speech. On the other hand the desired lengths are modified a great deal in hybrid speech synthesis resulting in higher LSD as shown in Table 4.

Next, two DNN-based nonlinear voice conversion functions from one female speaker to another male speaker were trained with the two abovementioned enlarged parallel corpora using our proposed VC technique discussed in Section 2. These DNNs took six hidden layers with 2048 nodes in each hidden layer. The VC-generated speech signals with the two synthesized corpora were compared for voice conversion in Table 5, which is one of our main interests in this study. As seen, Corpus1-DNN's 4.79 dB in LSD declares a clear advantage over Corpus2-DNN's 5.91 dB in the converted speech's spectra. It is consistent with the result listed in Table 4. The other comparison in Table 5 is that the LSDs for Corpus1-DNN and Corpus2-DNN were larger than the LSD for DNN-6. This is also in line with our expectation with that speech used for training DNN-6 was the original not the synthesized signals used in obtaining Corpus1-DNN and Corpus2-DNN.

Finally, voice converted speech was evaluated in another set of ABX tests for the Corpus-1 based VC system which showed lower LSD in Table 5 than that for Corpus-2 based VC. The preference scores on naturalness and similarity listed in Table 6 mean that converted speech from the synthesized large corpora (Corpus-1), although with larger LSD in Table 5, is much preferred to VC based speech generated from the small corpus (DNN-6). For example, for the similarity scores listed in the bottom row in Table 6, we can see the effectiveness of the proposed technique in voice conversion with a large synthesized parallel corpus could be confirmed at 52.5% of the time when compared with the VC system at a score of 0.236 obtained with a smaller parallel corpus but with natural speech. We believe more research is needed.

Table 4: LSDs between the natural speech and synthesized speech by HMM-based speech synthesis method and hybrid speech synthesis method

Methods	LSD in dB
HMM-based speech synthesis	5.01
Hybrid speech synthesis	5.64

Table 5: LSDs between the natural speech and converted speech by the DNN-6, Corpus1-DNN and Corpus2-DNN

Methods	LSD in dB
DNN-6	4.65
Corpus1-DNN	4.79
Corpus2-DNN	5.91

Table 6: Preference comparisons between the converted speech with Corpus1-DNN and DNN-6 with a 0.05 confidence interval

	Corpus1-DNN	DNN-6	N/P
Naturalness	**0.545**	0.328	0.127
Similarity	**0.525**	0.236	0.239

V. CONCLUSION AND FUTURE WORK

We propose a DNN-based voice conversion framework. Hybrid speech synthesis method and HMM based speech synthesis method are used to synthesize the large parallel corpora needed for DNN training to improve the quality of converted speech. Utterances from these corpora are phone-aligned with dynamic time warping to facilitate the nonlinear mapping required in DNN training. The experimental results show that speech converted by the proposed DNN-based approach takes a lower log spectrum distortion than the GMM-based method and the preference scores also declare a clear DNN advantage in listening tests. Meanwhile improved preference scores are obtained when more synthesized parallel corpora are incorporated.

In future work, we would investigate the issue of how big an original speech corpus size is needed in order to generate a usable parallel corpus for DNN-based voice conversion. The current size of three hours is too big in practice. We would also attempt to use the proposed DNN-based VC technique to speaker adaptation and to mimic a speaker's individuality in speech synthesis. ASR-based classification techniques could also be utilized in training DNNs.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61403386, No.61273288, No.61233009, No.61203258, No.61305003, No. 61332017, No.61375027), and the Major Program for the National Social Science Fund of China (13&ZD189).

REFERENCES

[1] [1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," J. Acoust. Soc. Japan. (E), vol. 11, no.2, pp. 71–76, 1990.

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

[3] [3]H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," Speech Communication, vol. 11, no. 2–3, pp. 175–187, 1992.

[4] [4] M. Narendranath, H. A. Murthy, S. Rajendran, and B. Yegnanarayana, "Transformation of formants for voice conversion using artificial neural networks," Speech Communication, vol. 16, no. 2, pp. 207–216, 1995.

[5] [5] N. Iwahashi and Y. Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation withweighting by radial basis function networks," Speech Communication, vol. 16, no. 2, pp. 139–151, 1995.

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

[7] [7] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5,pp. 912-921, 2010.
[8] [8] Y. Xu, J. Du, L.-R. Dai and C.-H. Lee, "A Regression Approach to Speech Enhancement Based on Deep Neural Networks," IEEE Trans. Audio, Speech and Language Proc., Vol. 23, No. 1, pp. 7-19, January 2015.
[9] [9] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.
[10] [10] G. E. Hinton, L. Deng, D. Yu, and G. E. Dahl, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Process. Magazine, vol. 29, no. 6, pp. 82–97, 2012.
[11] [11] S. Desai, A. W. Black, B. Yegnanarayana, "Spectrum Mapping Using Artificial Neural Networks for Voice Conversion," IEEE Trans. Audio, Speech, Lang. Process, vol. 18, No. 5, July 2010.
[12] [12] L. H. Chen, Z. H. Ling, Y. Song, L. R. Dai, "Joint Spectral Distribution Modeling Using Restricted Boltzmann Machines for Voice Conversion," in Proc. Interspeech, pp. 3052-3056, 2013.
[13] [13]L. H. Chen, Z. H. Ling, and L. R. Dai,"Voice Conversion Using Generative Trained Deep Neural Networks with Multiple Frame Spectral Envelopes," Proc. Interspeech, pp. 2313-2317, 2014.
[14] [14] L. H. Chen, Z. H. Ling, L. J. Liu, L. R. Dai, "Voice Conversion Using Deep Neural Networks With Layer-Wise Generative Training," IEEE/ACM Transactions on Audio, Speech, and Language Proc, vol.22, no.12, pp.1859--1872, Dec. 2014.
[15] [15]T. Nakashika, R. Takashima, T. Takiguchi, and Y. Ariki, "Voice Conversion in high-order eigen space using deep belief nets," Proc. Interspeech, pp. 369-372, 2013.
[16] [16] F. L. Xie, Y. Qian, Y. Fan, F. K. Song and H. Li, "Sequence Error (SE) Minimization Training of Neural Network for Voice Conversion," in Proc. Interspeech, pp. 2283-2287, 2014.
[17] [17] S. H. Mohammadi, A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," IEEE Spoken Language Technology Workshop, pp.19--23, Dec. 2014
[18] [18] J. Kominek and A. Black, "The CMU ARCTIC databases for speech synthesis," Tech. Report CMU-LTI-03-177, Language Technologies Institute, Carnegie Mellon University, 2003.
[19] [19] J. Tao, M. Zhang, J. Nurminen, J. Tian, and X. Wang, "Supervisory data alignment for text-independent voice conversion," IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 5, pp. 932–943, 2010.
[20] [20] D. Suendermann, H. Ney, and H. Hoege, "VTLN-Based cross-language voice conversion," Proc. ASRU'03, Virgin Islands, 2003.
[21] [21] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, and J. Hirschberg, "Text-independent cross-language voice conversion," Proc. ICSLP, 2006.
[22] [22] H. Zen, K. Tokuda, and A. Black, "Statistical Parametric Speech Synthesis," Speech Communication, 51(11), 1039-1064.
[23] [23] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. ICASSP, pp. 373-376, 1996.
[24] [24] R. Zhang, J.-H. Tao, Y. Li and Z.-Q. Wen, "A novel hybrid mandarin speech synthesis system using different base units for model training and concatenation," Proc. ICASSP, pp. 295-299, 2014.
[25] [25] C. S. Myers, and L. R. Rabiner, "A comparative study of several dynamic time-warping algorithms for connected word recognition," The Bell System Technical Journal, 60(7):1389-1409, September 1981.
[26] [26] G. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, vol. 18, pp. 1527-1554, 2006.
[27] [27] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. IEEE, 77 (2), pp. 257-286, Feb. 1989.
[28] [28] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP, pp.1315-1318, June 2000.
[29] [29] F. K. Soong, and B.-H. Juang, "Line spectrum pair (UP) and speech data compression", Proc. ICASSP, San Diego, Vol.1, pp. 1.10.1-1.10.4, May 1984.
[30] [30] 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, 27(5), 187–207, 1999.
[31] [31] Y. Ephraim and D. Malah, "Speech Enhancement using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-33, No. 2, pp. 443-445, 1985.
[32] [32] L. Blin, O. Boeffard and V. Barreaud, "WEB-based listening test system for speech synthesis and speech conversion evaluation," Proc. LREC (Marrakech (Morocco)), 2008.
[33] [33] S.-F. Pan, Y. Nankaku, K. Tokuda and J.-H. Tao, "Global variance modeling on frequency domain delta LSP for HMM-based speech synthesis," Proc. ICASSP, pp.4716-4719, 2011.