SPECTRAL CONVERSION USING DEEP NEURAL NETWORKS TRAINED WITH MULTI-SOURCE SPEAKERS

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

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

ljliu037@mail.ustc.edu.cn, {chenlh, zhling, lrdai}@ustc.edu.cn

ABSTRACT

This paper presents a method for voice conversion using deep neural networks (DNNs) trained with multiple source speakers. The proposed DNNs can be used in two ways for different scenarios: 1) in the absence of training data for source speaker, the DNNs can be treated as source-speaker-independent models and perform conversions directly from arbitrary source speakers to certain target speaker; 2) the DNNs can also be used as initial models for further fine-tuning of source-speaker-dependent DNNs when parallel training data for both source and target speakers are available. Experimental results show that, as source-speaker-independent models, the proposed DNNs can achieve comparable performance to conventional source-speaker-dependent models. On the other hand, the proposed method outperforms the conventional initialization method with restricted Boltzmann machines (RBMs).

Index Terms— voice conversion, deep neural networks, source-speaker-independent mapping

1. INTRODUCTION

Voice conversion (VC) is a technique that changes speaker characteristic of the speech of source speaker in order to make it sounded like that of the target speaker. Many approaches have been proposed for spectral conversion in voice conversion during the past decades [1, 2, 3, 4, 5, 6, 7]. Statistical methods are popular nowadays because of their stable performance in different conversion pairs. Among these statistical methods, Gaussian mixture model (GMM) based methods became the mainstream methods, especially after the method of utilizing joint density GMM (JDGMM) with dynamic features and parameter generation considering global variance (GV) was proposed [8].

However, there are several problems in the JDGMM-based approaches. One of them is that the spectral conversion described by JDGMM is a piece-wise linear transformation function, which is insufficient to model the nonlinear mapping relationship between two speakers. Recently, neural network (NN) based approaches have attracted much research attention. In some of these approaches, the NNs are used as generative models, e.g., restricted Boltzmann machine (RBM) [9], conditional RBM (CRBM) [10] and generatively trained deep neural network (DNN) [11], to derive the conditional distributions for generating converted spectral features. The other approaches directly use NNs to learn nonlinear feature mapping functions between the spectral features of source and target speakers [12] [13]. DNN has the ability to model highly nonlinear mapping relationship between spectra of two speakers because it usually contains several hidden layers with nonlinear activation functions, such as sigmoid, tanh. The second problem is that most of the models in conventional approaches, including NN-based approaches, are built for conversions of certain source and target speakers, which means that we need to construct a conversion model for each pair of conversion. This makes the use of conventional methods inflexible.

In this paper, we propose a new training strategy for DNN in order to learn a source-speaker-independent mapping function. The source-speaker-independent model is learned from the training data of multiple source speakers with a global input layer. Benefitting from the strong modeling ability, DNN can learn the distribution of acoustic space containing many speakers and perform automatic speaker interpolation for the input of unknown speakers at conversion stage. Two structures of DNNs are adopted in this paper according to the number of target speakers employed in the output layer: single target DNN (ST-DNN) and multiple targets DNN (MT-DNN). MT-DNN is proposed in the light of multi-task learning [14]. In MT-DNN, the learning of mapping function to one target can help to improve those to other targets. The proposed DNNs are trained using pre-stored parallel utterances of multiple speakers. In the application cases where no training data of source speaker is available, the estimated source-speaker-independent model can be used directly to convert the spectral features of source speaker, which are unseen in the training set, to those of the target speaker. Further, if parallel training data of source and target speakers are available, the pre-trained DNNs can be used as initial models for further fine-tuning of source-speaker-dependent model instead of the conventional pre-training method using RBMs [15]. Experimental results show the effectiveness of the proposed methods.

This paper is organized as follows. We give a brief introduction of the conventional JDGMM based method in section 2. Detailed technique descriptions of our proposed methods are presented in section 3. In section 4, we show the experimental conditions and results. A conclusion of this paper is made in the end.

2. SPECTRAL CONVERSION USING JDGMM

Let $\boldsymbol{X}_t$, $\boldsymbol{Y}_t$ represent spectral features of source and target speakers, respectively. The joint distribution of $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ is described by a GMM in JDGMM system:

$$P(\boldsymbol{Z}_t; \eta^{(z)}) = \sum_{m=1}^{M} \beta_m N(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \ \sum_{m=1}^{M} \beta_m = 1, \qquad (1)$$

This work was partially funded by the National Nature Science Foundation of China (Grant No. 61273264) and the electronic information industry development fund of China (Grant No. 2013-472).

where $\eta^{(z)} = \{\boldsymbol{\beta}_m, \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}, m = 1, 2, ..., M\}$ represents the parameter set of GMM, M is the number of mixture components, β_m,$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}$, $\boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}$ are the weight, mean vector and covariance matrix of the m-th mixture. The parameters are estimated with estimation-maximum(EM) algorithm.

In [8], correlations between frames are considered by modeling the static and dynamic features together so as to improve the continuity of converted spectra. In this condition, the converted static speech sequence is generated under the maximum output probability parameter generation criterion:

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \eta^{(z)}), \tag{2}$$

$$s.t. \qquad \boldsymbol{Y} = \boldsymbol{U}\boldsymbol{y}, \tag{3}$$

where $\boldsymbol{X}$ is the source feature sequence augmented with dynamic features, $\boldsymbol{U}$ is a window matrix that is used to generate static and dynamic features from the static features.

3. PROPOSED METHOD

3.1. Spectral Conversion Using DNN

Feedforward DNNs can be directly used to estimate the spectral mapping function between source speaker and target speaker [12]. Since hidden nodes in NNs are usually characterized by nonlinear functions, such as sigmoid, tanh functions, DNN has the ability to model the nonlinear mapping relationship between spectra of the two speakers. Figure 1 illustrates the spectral conversion process using DNN. The bottom layer is the input layer of source spectral features while the top layer is the output layer of converted spectral features. In this paper, concatenated spectral features of the source speaker, which are three consecutive frames including the current, preceding and succeeding frames, are adopted as the input features. Only the frame of converted spectral feature corresponding to the current input frame is generated in the output layer .

Conventionally, DNN is used to model the mapping function for a certain speaker pair. The model is trained with parallel utterances of a single source and a single target speaker. However, this training strategy cannot realize flexible conversions because new DNNs have to be built for new speaker pairs. In order to cope with this problem, we propose a new training strategy for DNN-based voice conversion which is to learn a source-speaker-independent mapping function. Multiple source speakers are used in this training method. DNN is trained in attempt to map the input spectral features of these source speakers to those of the same target speaker. Due to the good generalization ability, DNN can capture the feature space of multiple source speakers well. Therefore, DNN is able to perform appropriate speaker interpolation for source speaker who is unseen in the training data set at conversion stage.

Two types of DNN are adopted for constructing the source-speaker-independent model in this paper, according to the structures with different numbers of target speakers in the output layer, i.e., single target DNN (ST-DNN) and multi-target DNN (MT-DNN). The training strategies for the conventional DNN and the proposed ST-DNN, MT-DNN are presented in the following sections.

3.2. DNN Trained with Single Source and Single Target

Conventional DNN is trained with parallel data set of single source and single target speaker. The training process includes two steps:

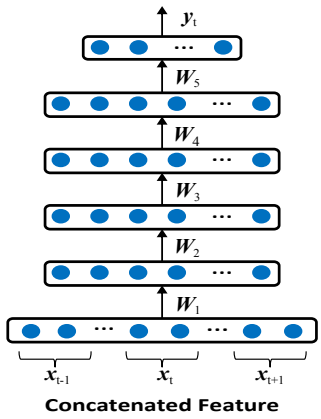

Fig. 1. The spectral conversion process using DNN. In this paper, the spectral features of 3 concatenated frames are used as the input while DNN outputs the converted feature vector corresponding to central frame.

1) Pre-training step: DNN trained with random initialization often gets trapped in a poor local optima. A pre-training method that train a deep belief network (DBN) as the initial model can be used to improve the performance of DNN [15]. The training process of a DBN is conducted by learning a stack of RBMs [16] using the training data of source speaker.

2) Fine-tuning step: feature sequences of source and target speakers are aligned to the same length in advance. Then the back propagation (BP) algorithm is used to estimate the parameters of the DNN using the minimum mean square error (MMSE) criterion. The model parameters are usually updated using the mini-batch gradient descent (MBGD) algorithm.

3.3. DNN Trained with Multi-source and Single Target

Parallel utterances of multiple source speakers and a certain target speaker are used for the training of ST-DNN. The training process also takes two steps: pre-training and fine-tuning. The pre-training process is conducted as the same as that of conventional DNN except that all the training data of multiple source speakers are used.

During the fine-tuning process, feature sequence of each source speaker is aligned to that of the target speaker in advance. The ST-DNN is optimized by minimizing the summation of mean square errors of utterances from all source speakers in the training set. The objective function is defined as follows:

$$L(\theta) = \sum_{s=1}^{S} \sum_{t=1}^{T} ||\boldsymbol{y}_t - f(\boldsymbol{x}_{t,s}, \theta)||^2, \tag{4}$$

where $\boldsymbol{y}_t$ is the target spectral feature vector at frame t, the total number of frames is T, $\boldsymbol{x}_{t,s}$ is the spectral feature vector of the s-th source speaker at frame t, S is the number of pre-stored source speakers in the training set. $f(\boldsymbol{x})$ denotes the mapping function built by ST-DNN, which is written as

$$f(\boldsymbol{x}_{t,s}, \theta) = \boldsymbol{W}^{(L+1)^\top}[\boldsymbol{h}^L \circ \cdots \circ \boldsymbol{h}^2 \circ \boldsymbol{h}^1(\boldsymbol{x}_{t,s})] + \boldsymbol{b}^{(L+1)}, \tag{5}$$

where L is the number of hidden layers, $\boldsymbol{h}^L \circ \cdots \circ \boldsymbol{h}^2 \circ \boldsymbol{h}^1(\boldsymbol{x})$ defines a combination of functions $\boldsymbol{h}^L, \cdots, \boldsymbol{h}^1$:

$$\boldsymbol{h}^i(\boldsymbol{x}) = 1./(1 + exp(-\boldsymbol{W}^{(i)^\top}\boldsymbol{x} - \boldsymbol{b}^{(i)})). \quad (6)$$

The ST-DNN can be applied to perform spectral conversion directly from arbitrary source speaker to the target speaker after this fine-tuning process. However, when parallel training utterances of the new source speaker and the target speaker are available, ST-DNN can be used as an initial model for further fine-tuning of the source-speaker-dependent DNN training.

3.4. DNN Trained with Multi-source and Multi-target

For ST-DNNs with different target speakers, what is learned for one ST-DNN could be beneficial to the others. The bottom part of these ST-DNNs would have lots in common as they all attempt to learn the source-speaker-independent mapping function using the same multiple source speakers. According to the multi-task learning theory [14], the learning processes of these ST-DNNs will be promoted if they are trained in parallel by sharing the information learned in the bottom parts of networks. Therefore, we merge these ST-DNNs together by sharing weights of the first L layers in order to improve the performance of each ST-DNN. Weights of the last layer are kept for each target speaker. We denote this new model as multi-target DNN (MT-DNN).

Similar to ST-DNN, the training process of MT-DNN also includes two steps. The pre-training process is exactly the same as that of the ST-DNN. At the fine-tuning stage, supposing there are S pre-stored source speakers and N pre-stored target speakers, feature sequence of each source speaker is aligned to that of each target speaker. The obtained $S \times N$ parallel data sets are used for MT-DNN training.

The mapping function from a given input spectral feature $\boldsymbol{x}_{t,s}$ to the n-th target speaker described by the MT-DNN is written as follows,

$$g(\boldsymbol{x}_{t,s}, \lambda_n) = \boldsymbol{W}_n^{(L+1)^\top}[\boldsymbol{h}^L \circ \cdots \circ \boldsymbol{h}^2 \circ \boldsymbol{h}^1(\boldsymbol{x}_{t,s})] + \boldsymbol{b}_n^{(L+1)}, \quad (7)$$

where $\lambda = \{\{\boldsymbol{W}^{(l)}, \boldsymbol{b}^{(l)}\}_{l=1}^{L}, \{\boldsymbol{W}_n^{(L+1)}, \boldsymbol{b}_n^{(L+1)}\}_{n=1}^{n=N}\}$ is the parameter set of MT-DNN, $\boldsymbol{W}_n^{(L)}, \boldsymbol{b}_n^{(L)}$ are the weight matrix and bias in the last layer corresponding to the n-th target.

The parameters of MT-DNN are learned by minimizing the summation of mean squared error of all the conversion pairs, the objective function for training the MT-DNN is defined as follows

$$L(\lambda) = \sum_{n=1}^{N}\sum_{s=1}^{S}\sum_{t=1}^{T_n} ||\boldsymbol{y}_{t,n} - g(\boldsymbol{x}_{t,s}, \lambda_n)||^2, \quad (8)$$

where T_n is the total frame number of the n-th target. Similarly, we use BP algorithm with MBGD to estimate the parameters.

The spectral conversion with MT-DNN model can be conducted using (7) directly. If parallel data set of the new source speaker and the desired target speaker are available, parameters in MT-DNN corresponding to the desired target speaker can serve as an initial model for further DNN training.

4. EXPERIMENTAL EVALUATIONS

4.1. Experimental Conditions

Our experiments were conducted on a Mandarin parallel speech data set. This data set includes 90 speakers. Each speaker utters the same 100 sentences. Waveforms are recorded in 16kHz/16bit format. We randomly chose 80 speakers in the corpus as the training speaker set and the remaining 10 speakers as the test speaker set. The number of sentences for training, validation and testing were 70, 15, 15, respectively. 24-order mel-cepstral coefficients were used as spectral features. Transcripts of the utterances were used to segment and align the feature sequences.

In order to evaluate the general performance of our proposed methods, conversions from the test speakers to 6 target speakers were conducted and performed. The 6 target speakers, including 3 male and 3 female, were randomly selected from the training speaker set. Four systems were compared in our experiments:

a) *JDGMM*: Static mel-cepstral feature, together with dynamic and acceleration components were used as the spectral feature. The mixture number for all conversions in the experiments were set as 256.

b) *DNN*: Conventional DNN system trained with single source and single target speaker, DNNs were initialized with DBNs;

c) *ST-DNN*: The proposed source-speaker-independent DNN trained with multi-source speakers and single target speaker. The target speaker was randomly selected from the training speaker set and all the 80 speakers in the training speaker set were adopted as the source speakers.

d) *MT-DNN*: The proposed source-speaker-independent DNN trained with multi-source speakers and multi-target speakers. All the 80 speakers in the training set were used as the source speakers as well as the target.

In addition, conversions conducted by DNNs initialized with ST-DNN and MT-DNN were also built in our experiments in order to evaluate the effectiveness of applying ST-DNN and MT-DNN as initial models for source-speaker-dependent DNNs training.

3 concatenated mel-cepstral feature vectors were used as the inputs of the networks, while the networks output static feature vectors. The training data were normalized to zero mean and unit variance before training. Architectures of DNN, ST-DNN and MT-DNN were set to the same. The number of hidden layers is 4. Each layer has 512 hidden nodes. Learning rates for all the listed networks are 0.002, 0.01 and 0.01 while the sizes of mini-batch of the corresponding systems are set as 10, 10 and 100, separately. The learning rate and mini-batch size for fine-tuning of DNNs with the initialization of ST-DNN or MT-DNN are 0.001 and 10. L_2 regularization items are employed during the training processes in order to prevent overfitting, the value of regularization coefficients for the training of all networks is 2×10^{-5}.

In this paper, pitch conversion was conducted using the traditional linear transformation in log-scale.

4.2. Experimental Results

4.2.1. Source-speaker-independent Mapping

The performance of ST-DNN and MT-DNN for directly conversions of new speakers are evaluated in Table 1. Mel-cepstral distortions (MCDs) between the converted mel-cepstra and the target were calculated as the objective measurement. For each target speaker, the average MCD of conversions from the 10 test speakers was calculated. We can see that DNN shows less conversion accuracy than JDGMM in this Mandarin database in general while in conversions to some target speakers it still gets smaller average MCDs, e.g. m_1, m_3. As a whole, both JDGMM and DNN demonstrate more accurate conversion performance than ST-DNN and MT-DNN, which is

Table 1. Average mel-cepstral distortions (dB) for conversions to 6 target speakers.

target	JDGMM	DNN	ST-DNN	MT-DNN
m_1	4.86	4.69	4.85	4.81
f_1	4.91	5.33	5.44	5.38
m_2	4.82	5.17	5.31	5.22
f_2	5.02	5.17	5.24	5.22
m_3	4.92	4.80	4.88	4.83
f_3	4.86	4.95	5.04	5.01
average	4.90	5.02	5.13	5.08

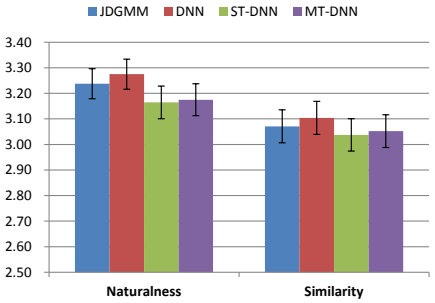

Fig. 2. Average mean opinion scores of JDGMM, DNN, ST-DNN and MT-DNN on speech naturalness and similarity.

reasonable as training data of new speakers are used in construction of JDGMM and DNN. While using no training data, ST-DNN and MT-DNN obtain acceptable spectral distortions. Besides, MT-DNN outperforms ST-DNN due to the effect of multi-task learning.

To evaluate the speeches converted by the source-speaker-independent DNNs, we carried out mean opinion score (MOS) tests on speech naturalness and similarity. 40 utterances were set as the listening set, which were randomly selected from the test utterances of all the 60 conversion pairs. All the conversion types, i.e., male-to-male, male-to-female, female-to-male, female-to-female, were included in this set. Six expert listeners took part in the tests. The results are given in Figure 2. As DNN conducts non-linear conversion while without statistical average effect of statistical modelling, waveforms generated by DNN are perceived to be better than those generated by JDGMM. MT-DNN gets almost the same performance of ST-DNN in both speech naturalness and similarity. It is possible that the inadequate number of layers in the top of MT-DNN for different target speakers weaken the multi-task learning effect. Speeches generated by ST-DNN and MT-DNN obtain comparable auditory performance to those by JDGMM, while perform slightly worse than DNN. Nevertheless, the proposed source-speaker-independent DNNs enjoy the advantage that no utterances of new source speaker are needed for training compared with both JDGMM and DNN. It would be more convenient to perform VCs with the use of ST-DNN and MT-DNN.

4.2.2. Source-Speaker-Dependent DNN Initialized by ST-DNN/MT-DNN

In Table 2, we evaluate the effectiveness of setting the source-speaker-independent DNNs as initial models for DNNs training. The results show that the conversion accuracy of DNNs can be improved by using ST-DNN or MT-DNN instead of RBMs as initial

Table 2. Average mel-cepstral distortions (dB) of DNNs with different initializations for conversions to 6 target speakers. DNN-1 represents conventional DNN initialized with RBMs while DNN-2 and DNN-3 are source-speaker dependent DNNs initialized by ST-DNN and MT-DNN models respectively.

target	DNN-1	DNN-2	DNN-3
m_1	4.69	4.55	4.53
f_1	5.33	5.21	5.18
m_2	5.17	5.04	5.01
f_2	5.17	5.01	4.98
m_3	4.80	4.65	4.65
f_3	4.95	4.73	4.71
average	5.02	4.86	4.84

Table 3. Results of preference tests. DNNs initialized with RBMs, ST-DNN and MT-DNN were compared with each other. N/P represents for no preference, p is the p-value of a two-tail t-test.

	DNN-1	DNN-2	DNN-3	N/P	p
Naturalness	8.33	**18.75**	-	72.92	<0.01
	9.58	-	**19.17**	71.25	<0.01
	-	15.42	10	74.58	>0.05
Similarity	6.25	5.83	-	87.92	>0.05
	7.5	-	4.58	87.92	>0.05
	-	10.42	7.5	82.08	>0.05

models, which is owing to the prior conversion information contained in the source-speaker-independent-DNNs. MT-DNN works slightly better than ST-DNN all the same in this test.

Furthermore, conversion performances of DNNs initialized with RBMs, ST-DNNs and MT-DNNs were compared with each other using preference tests. The same six persons participated in these tests. Results are presented in Table 3 and show that the speech naturalness converted by DNN can be improved with the initialization of source-speaker-independent models while speech similarity gets no significant improvements. Though in objective tests, DNNs initialized with MT-DNNs works slightly better than those with ST-DNNs , subjective results show that there is no significant difference between them on both speech naturalness and similarity.

5. CONCLUSIONS

This paper describes a new training strategy for spectral mapping using deep neural network (DNN). Multiple source speakers are used during the training process for the purpose of constructing a source-speaker-independent mapping function. The source-speaker-independent DNN can perform conversion of arbitrary source speaker when no training data of this speaker is available. When parallel data set of new source speaker and the target speaker is available, it can serve as an initial model for source-speaker-dependent DNN training. Experimental results show that the source-speaker-independent DNNs can achieve comparable performance to the source-speaker-dependent approaches, such as JDGMM and DNN based approaches. And DNNs initialized with the source-speaker-independent models outperform those with RBMs. Due to the limitation of the number of available pre-stored speakers, the performance of ST-DNN trained with more source speakers are not investigated in this paper. This is one of our future works.

6. REFERENCES

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

[2] 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.

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

[4] M Narendranath, Hema 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] Naoto Iwahashi and Yoshinori Sagisaka, "Speech spectrum conversion based on speaker interpolation and multi-functional representation with weighting by radial basis function networks," *Speech Communication*, vol. 16, no. 2, pp. 139–151, 1995.

[6] Z.W. Shuang, R. Bakis, S. Shechtman, D. Chazan, and Y. Qin, "Frequency warping based on mapping formant parameters," in *Proc. Interspeech*. ISCA, 2006.

[7] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversion based on bilinear frequency warping plus amplitude scaling," *IEEE Trans. Audio, Speech, and Lang. Process*, vol. 21, no. 3, pp. 556C566, 2013.

[8] T. Toda, A.W. 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.

[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 *Proc. InterSpeech*, 2013, pp. 3052–3056.

[10] Zhizheng Wu, Eng Siong Chng, and Haizhou 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.

[11] 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 Transactions on*, vol. 22, no. 12, pp. 1859–1872, Dec 2014.

[12] Srinivas Desai, Alan W Black, B Yegnanarayana, and Kishore Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 5, pp. 954–964, 2010.

[13] Toru Nakashika, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Voice conversion in high-order eigen space using deep belief nets," in *INTERSPEECH'13*, 2013, pp. 369–372.

[14] Rich Caruana, *Multitask learning*, Springer, 1998.

[15] Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton, "Acoustic modeling using deep belief networks," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 14–22, 2012.

[16] Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.