2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

# Text-Independent Voice Conversion Using Deep Neural Network Based Phonetic Level Features

Huadi Zheng[‡], Weicheng Cai[*], Tianyan Zhou[*]
[*]SYSU-CMU Joint Institute of Eng.,Sun Yat-sen University
[†]SYSU-CMU Shunde International Joint Research Institute
liming46@mail.sysu.edu.cn

Shilei Zhang[§], Ming Li[*†]
[‡]Dept. of EIE, Hong Kong Polytechnic University
[§]Speech technology and Solution Group, IBM China Research

***Abstract*—This paper presents a phonetically-aware joint density Gaussian mixture model (JD-GMM) framework for voice conversion that no longer requires parallel data from source speaker at the training stage. Considering that the phonetic level features contain text information which should be preserved in the conversion task, we propose a method that only concatenates phonetic discriminant features and spectral features extracted from the same target speakers speech to train a JD-GMM. After the mapping relationship of these two features is trained, we can use phonetic discriminant features from source speaker to estimate target speaker's spectral features at conversion stage. The phonetic discriminant features are extracted using PCA from the output layer of a deep neural network (DNN) in an automatic speaker recognition (ASR) system. It can be seen as a low dimensional representation of the senone posteriors. We compare the proposed phonetically-aware method with conventional JD-GMM method on the Voice Conversion Challenge 2016 training database. The experimental results show that our proposed phonetically-aware feature method can obtain similar performance compared to the conventional JD-GMM in the case of using only target speech as training data.**

***Index Terms*—Gaussian mixture model; phoneme posterior probability; voice conversion; deep neural network**

## I. Introduction

Speech signals usually contain not only linguistic content but also some explicit personal identity information to help associate the speech with a specific speaker. For human beings, these non-linguistic cues can be easily caught by hearing perception. Voice conversion (VC) is an effective approach to capture this non-linguistic information and utilize it to synthesize an intended voice. The speech signal produced by one person (source speaker) can be modified by various transformation and mapping techniques to generate speech signals that sounds like another person (target speaker) while the linguistic message is preserved. VC system can be applied to different areas like electronic larynx [1] and text-to-speech system [2]. It has been reported that spectral attributes are important to characterize the speaker individuality [3].Therefore, most of VC systems are based on spectral mapping technique. The related mapping approach and model have been intensively studied over the past several years.

To conduct a typical parallel or text-dependent VC process, both paired data training and runtime conversion are usually required. During the data preparation stage, the parallel data, an utterance set containing speeches from both source speaker and target speaker on the same content, has to be prepared and aligned. Spectrum components separated from the paired data are further passed to a feature extraction module to extract spectral features such as Mel-cepstral coefficient (MCCs) [4], line spectral frequency (LSF) [2], line spectrum pair (LSP)[5] [6] and other types of acoustic feature. These features usually have a good representation of spectrum on low-resolution space, which provides convenience for computation. And spectrum can be easily reconstructed from these features for converted voice synthesis. Time alignment is employed on the parallel features for modifying the speech duration between the utterance pairs, such as using dynamic time warping (DTW) technique.

At the offline training stage, the spectral features will be used to estimate the parameters of the mapping function. A great number of statistical parametric approaches for VC have managed to transform these spectral features between speakers by implementing a robust feature mapping function, such as vector quantization (VQ) mapping codebooks [3], Gaussian mixture model (GMM) [2][4][7], artificial neural networks (ANN) [8], partial least squares regression (PLS) [9] and non-negative matrix factorization (NMF) [10]. In the GMM based approaches, joint density estimation technique has been proved to be robust for even a small amount of training data with a better perceptual test result [2]. The source features and target features are concatenated to train a joint density distribution Gaussian mixture model (JD-GMM) after time alignment. When it comes to runtime conversion, the spectral features can be estimated from the model and reversed back to spectrum component.

However, the statistical property of GMM requires relatively large amounts of parallel training data to increase the mapping accuracy. The requirement for large amounts of parallel spectral features is not always feasible in practical application and impossible in cross-linguistic conversion. To utilize the non-parallel data set, text-independent method has been proposed, such as vocal tract length normalization (VTLN) [11], unit selection [12] and more. Though some of the mapping techniques have been proved to be useful in non-parallel training, these approaches still need to align the source and target data on frame or phoneme level and the model lacks of generalization with one-one mapping. To reduce the dependence on source data in the training stage

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

of voice conversion and increase the generalization of the model, we come up with a new framework that only considers target data when training a JD-GMM. The context-dependent phoneme posterior probability (PPP) based feature has been applied in speaker verification task with DNN method [13]. And using the features that combine MFCC and PPP at the feature level also reduces the equal error rate significantly [14]. We believe that the PPP features from the source utterance are similar to the target speakers as long as the phonetic content is the same. This motivates us to propose our framework with phonetic discriminant features as the input features for voice conversion.

The rest of the paper is organized as follows. In Section II, we introduce the parallel data and conventional JD-GMM voice conversion approach. In Section III, we describe our proposed phonetic discriminant features and text-independent conversion framework without any data from source speaker during training process. Section IV presents the experimental result achieved by conventional JD-GMM framework and proposed phonetically-aware framework. Finally, conclusion is provided in Section V.

## II. Conventional JD-GMM Method

The conventional text-dependent JD-GMM method contains two stages, namely the training stage and the runtime conversion stage. The process overview is shown in Figure 1.

### A. Spectral Feature Preparation

Conventional voice conversion research believes that there exists a latent relationship on how two persons speak. This relationship can be uncovered by fitting the spectral features into a model. A robust mapping function can be trained from the data to estimate the target spectral features from the source spectral features. Once a parallel data set is ready, the spectral features will be extracted from both source and target speeches at frame level. As we have collected the source spectral features and target spectral features, we can denote the utterance features as

$$X = [x_1^\top, x_2^\top, x_3^\top, ..., x_t^\top, ..., x_{L_x}^\top]^\top \quad (1)$$

$$Y = [y_1^\top, y_2^\top, y_3^\top, ..., y_t^\top, ..., y_{L_y}^\top]^\top \quad (2)$$

where the subscript denotes the frame index, $L$ is the total number of frames and $\top$ denotes the transpose operation.

After the frame alignment, a spectral feature vector from the source speaker will be concatenated with the aligned target spectral feature vector as a joint feature vector $[x_t^\top, y_t^\top]^\top$ to train the joint density distribution model.

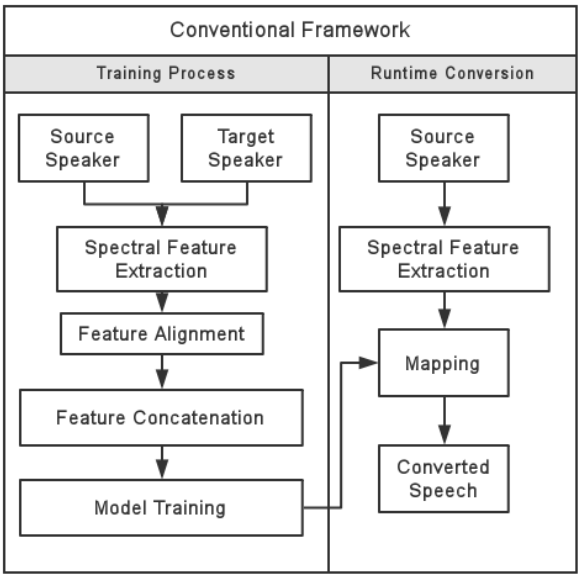


Fig. 1. Conventional method

### B. Probability Density Function of JD-GMM

JD-GMM has been recognized as one of the most effective voice conversion methods that fit the data with a probabilistic model. We employ JD-GMM as the mapping method for our framework as well. The model can be expressed as

$$P(z_t|\lambda^{(z)}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(z_k; \mu_k^{(z)}, \Sigma_k^{(z)}) \quad (3)$$

$$\lambda^{(z)}) = \{\mu_1^{(z)}, \Sigma_1^{(z)}; \mu_2^{(z)}, \Sigma_2^{(z)}; ...; \mu_K^{(z)}, \Sigma_K^{(z)}\} \quad (4)$$

where $z_t = [x_t^\top, y_t^\top]^\top$ is the aligned joint feature vector concatenated from the parallel data. $\lambda^{(z)}$ is the set of parameters. This JD-GMM contains K components of multivariate Gaussian distribution. The mean vector $\mu_k^{(z)}$, the covariance matrix $\sigma_k^{(z)}$ and the component weight $\omega_k$ are parameters estimated from the training features. The notation of $\mathcal{N}(z_k; \mu_k^{(z)}, \Sigma_k^{(z)})$ represents the Gaussian distribution on the $k^{th}$ component with parameters of mean vector $\mu_k^{(z)}$ and covariance matrix $\Sigma_k^{(z)}$.

$$\mathcal{N}(z_k; \mu_k^{(z)}, \Sigma_k^{(z)}) = \quad (5)$$
$$\frac{1}{(2\pi)^{d/2}|\Sigma_k^{(z)}|^{1/2}} exp\left[-\frac{1}{2}(z_t - \mu_k^{(z)})^\top {\Sigma_k^{(z)}}^{-1}(z_t - \mu_k^{(z)})\right]$$

In each of the multivariate component, we can denote the mean vector $\mu_k^{(z)}$ and the covariance matrix $\Sigma_k^{(z)}$ with parameters from another perspective.

$$\mu_k^{(z)} = \begin{bmatrix} \mu_k^{(x)} \\ \mu_k^{(y)} \end{bmatrix} \quad \Sigma_k^{(z)} = \begin{bmatrix} \Sigma_k^{(xx)} & \Sigma_k^{(xy)} \\ \Sigma_k^{(yx)} & \Sigma_k^{(yy)} \end{bmatrix} \quad (6)$$

Here, $\mu_k^{(x)}$ and $\mu_k^{(y)}$ are mean vectors of $k^{th}$ multivariate Gaussian component for the source and target spectral features. Likewise, $\Sigma_k^{(xx)}$ and $\Sigma_k^{(yy)}$ are the covariance matrix of $k^{th}$ component for the source and target spectral features. $\Sigma_k^{(yx)}$ and $\Sigma_k^{(xy)}$ are the cross-covariance matrix of $k^{th}$ component.

The classic Expectation Maximization (EM) algorithm can be used to train a JD-GMM with joint feature vectors. Mixture parameters will be estimated through multiple iterations until the error converges stably.

## III. PROPOSED PHONETICALLY-AWARE FRAMEWORK

Shown in Figure 2, this novel framework uses two kinds of features extracted from the same target speech instead of using source spectral feature.

### A. Senone Posteriors

Senones are defined as tied triphone states which map multiple logical triphones into one identical physical triphone. Inside each physical triphone, there are several states tied and represented by senones and a typical choice for the number of states is 3 [15]. A series of senones can represent the pronunciation of all different words[16].

Final senones are obtained from the leaves of a decision tree since the senone set is usually defined by decision tree mechanism [16]. Senone posteriors are the computed observation probabilities for each leaf unit [17].

### B. DNN Senone Posteriors

In the recent studies, DNNs have shown significant success when replacing the traditional GMM to calculate the senone posteriors at the frame level[18]. The system is based on the multi-splice time delay DNN (TDNN) described in [19]. The labels for the DNNs are obtained from a standard tied-state triphone GMM-HMM system trained with maximum likelihood. The input features to the DNNs are 40 dimensional vectors obtained from an LDA+MLLT projection of 7 spliced frames of 13 MFCCs, and fMLLR transformation is not used. Cepstral mean subs traction is performed over a window of 6 seconds.

The TDNN has six layers, the hidden layers have an input dimension of 350 and an output dimension 3500. In the multi-splice system, a narrow temporal context is provided to the first layer and increasingly wide contexts are available to the subsequent hidden layers. The softmax output layer computes posteriors for 5621 senones.

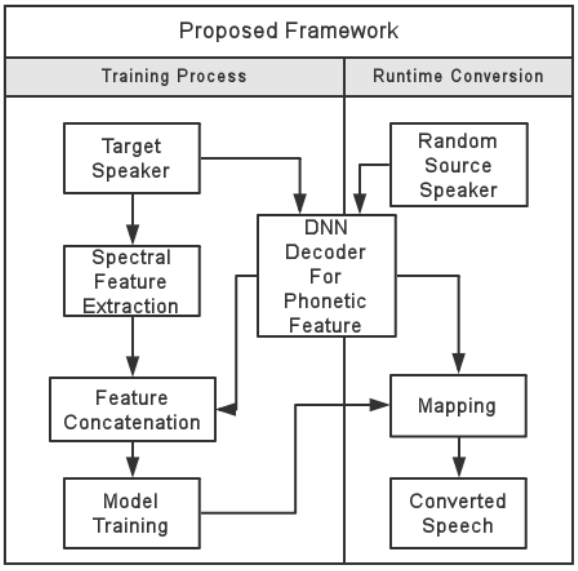


Fig. 2. Proposed method

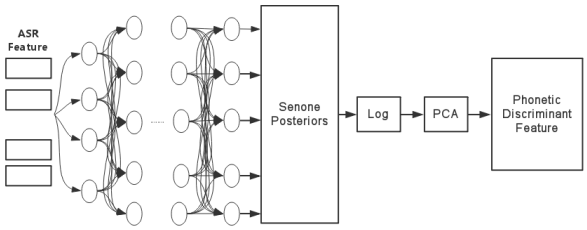


Fig. 3. Phonetic discriminant features extraction

### C. Low Resolution Representation of DNN Senone Posteriors

The resulting vector has the same dimension as the size of senone set and may be much larger than the typical spectral feature. Moreover, for each frame, a large number of senone posteriors are close to 0. So we should find a low resolution representation of these senone posteriors.

Log transform, principal component analysis (PCA) and mean variance normalization (MVN) are applied to the process that converts the senone posteriors into phonetic discriminant features[20], [21], as shown in Figure 3.

### D. Proposed text-indepentdent framework

As we obtain the model after the training stage, we can input phonetic discriminant features extracted from the new source data to estimate the corresponding target spectral features. The JD-GMM is trained with the joint vector $z_t = [x_t^\top, y_t^\top]^\top$ where $x$ is replaced by the proposed target phonetic discriminant features and remains as the target spectral features. The conversion can be achieved by the following method.

$$P(k|x_t, \lambda^{(z)}) = \frac{\omega_k \mathcal{N}(x_k; \mu_k^{(x)}, \Sigma_k^{(xx)})}{\sum_{k=1}^{K} \omega_k \mathcal{N}(x_k; \mu_k^{(x)}, \Sigma_k^{(xx)})} \quad (7)$$

$P(k|x_t, \lambda^{(z)})$ is the posterior probability for frame x in the $k^{th}$ Gaussian component from the trained model. By applying the minimum mean-square error method (MMSE), we can estimate the converted spectral feature $y$ by

$$y_t = \sum_{k=1}^{K} P(k|x_t, \lambda^{(z)})(\mu_k^{(y)} + \Sigma_k^{(yx)} {\Sigma_k^{(xx)}}^{-1}(x_t - \mu_k^{(x)})) \quad (8)$$

The phonetic discriminant feature of each source speech frame is converted with the same model and mapping technique.Therefore, we will obtain the same numbers of estimated spectral feature.

## IV. EXPERIMENTS

We conduct experiments on the Fisher English Training Speech and Voice Conversion Challenge 2016 training data to evaluate our proposed phonetically-aware voice conversion framework using DNN based phonetic discriminant features.

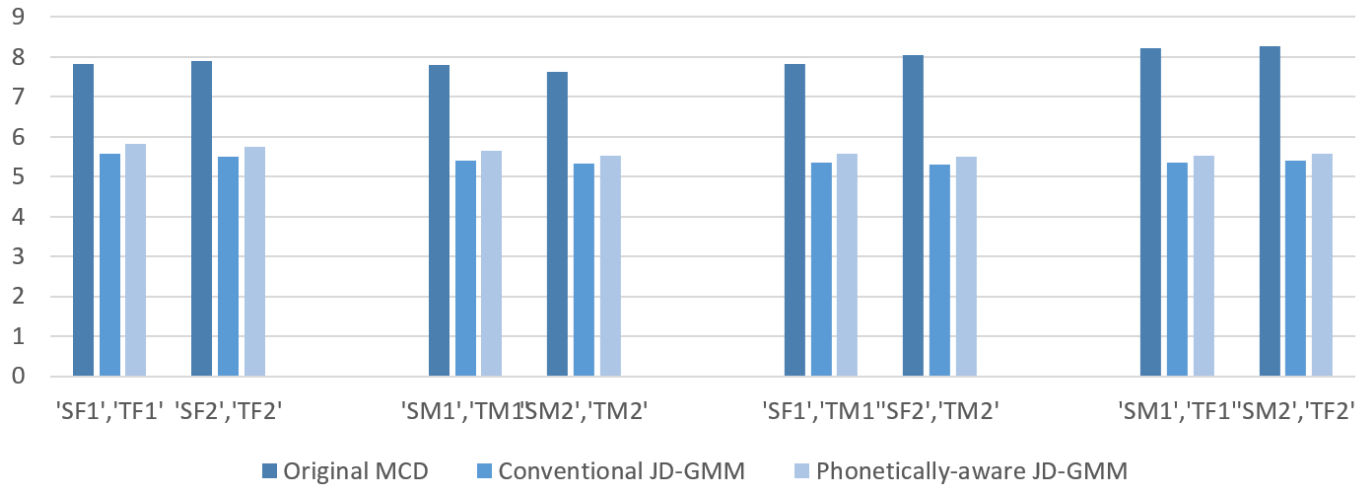


Fig. 4. Objective evaluation

### A. Setups

First, we employ voice conversion on speech data from four males (SM1, SM2, TM1, TM2) and four females (SF1, SF2, TF1, TF2). SF means source female speaker while TM means target male speaker. 8 speaker pairs with 4 intra-gender and 4 inter-gender conversion types are used. 142 utterances from no.21 to no.162 are used as the training data and the rest of 20 utterances from no.1 to no.20 are used as evaluation data. The speech is sampled at 16 kHz. The STRAIGHT [22] approach is applied to spectral extraction and converted speech synthesis.

Since our implementation is based on the conventional JD-GMM mapping technique, we compare the performance between conventional parallel spectral features mapping and proposed phonetic discriminant features mapping. The configurations for the experiments are as follows.

- ***Conventional JD-GMM:*** 24-dimensional MCCs are extracted from both source data and target data. Joint-vectors are concatenated by the aligned MCCs from the source speaker and the target speaker. A JD-GMM with 64 components is trained. MCCs extracted from the evaluation source data are used as the input features for mapping function.
- ***Phonetically-aware JD-GMM:*** 24-dimensional phonetic discriminant features and 24-dimensional MCCs are extracted only from target data. Joint-vectors are concatenated by phonetic discriminant features and MCCs on the corresponding frame. No alignment required. A JD-GMM with 64 components is trained. Phonetic discriminant features extracted from the source evaluation data are used as the input features for the mapping function.

It should be noted that 24 order MCCs actually have 25 parameters. The $0^{th}$ order coefficient is usually considered as the power information of the frame. We ignore $0^{th}$ order coefficient since it is not directly related to speaker identity, so we only convert the $1^{th}$ through $24^{th}$ coefficients. The converted 24-dimensional MCCs will use the source $0^{th}$ order coefficients to produce the final converted MCCs.

To adapt the prosodic feature from the source speaker to the target speaker, fundamental frequency is converted linearly by

$$log(F_0^y) = \frac{\sigma^{(y)}}{\sigma^{(x)}}(log(F_0^x - \mu^{(x)}) + \mu^{(y)} \tag{9}$$

where $F_0^x$ is the $F_0$ of the source frame while $F_0^y$ is the converted result. $\sigma^{(x)}$ and $\mu^{(x)}$ are the log-scale mean and standard deviation of in source data. $\sigma^{(y)}$ and $\mu^{(y)}$ are the log-scale mean and standard deviation $F_0$ in target data.

### B. Objective Evaluations

To evaluate the performance of our system, we calculate the Mel-cepstral distortion [4] between the target speech and the converted speech by the following method

$$Mel - CD[dB] = \frac{10}{ln10}\sqrt{2\sum_{d=1}^{24}(mc_d^{(y)} - mc_d^{(\tilde{y})})^2} \tag{10}$$

where $mc_d^{(y)}$ and $mc_d^{(\tilde{y})}$ are the $d^{th}$ dimensional MCCs.

The distortion indicates dissimilarity of two speeches on the MCCs representation. If MCCs are extracted from an identical audio, the distortion is supposed to be zero. In Figure.4, the result shows that the mel-cepstral distortion is both reduced for the two conversion systems on all four types of conversion. The conventional text-dependent JD-GMM achieve the lowest Mel-cepstral distortion. Table I shows that under the circumstances of no source data at training stage at all, the overall performance of our proposed system achieves a similar result with an approximate 0.2dB distortion difference from the conventional JD-GMM framework.

### C. Subjective Evaluation

Listeners tests are conducted on the converted speech to compare the performance of the two systems. First, An XAB test is conducted with respect to the individuality of the target speaker. 5 volunteers are presented with target speech X and

TABLE I
OVERALL AVERAGE PERFORMANCE

| Total Average MCD | | | |
|---|---|---|---|
| Original MCD | Conventional JD-GMM | Phonetically-aware JD-GMM | Difference |
| 7.93779 | 5.40354 | 5.61196 | -0.20842 |

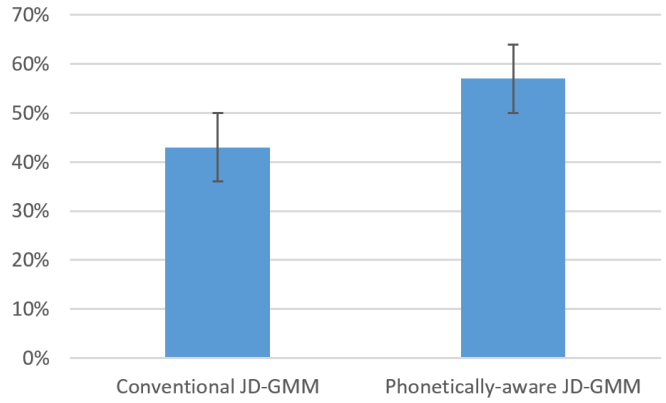


Fig. 5. XAB preference results

converted speech from two methods randomly as A or B. After they listened to the speech audio, they are asked to make a preference choice between A and B according to similarity to the identity of X.

In Figure 5, volunteers show preferences for the audio generated from our phonetically-aware JD-GMM system, which indicates that our system reaches a fairly satisfying performance in speaker identity conversion. In practice, most of the volunteers show hesitation and difficulty to make a decision when they are asked to make a preference choice because the two audio files are similar to the target speech in a very close degree for human hearing, while the distortion results show a slightly higher value for our proposed framework.

Secondly, an opinion test is conducted to assess the speech quality of the converted audio. In this test, volunteers are given a scale of 5 points including: 1-bad, 2-poor, 3-fair, 4-good and 5-excellent. The audio of original target speaker, the converted speech from both conventional JD-GMM and phonetically-aware JD-GMM are presented one at a time to the volunteers. They will provide a score that best describes the quality of each speech audio if the original target speech is assumed to have the score of 5.

According to the Figure 6, proposed system is able to generate a speech file with a close quality to conventional result. The phonetically-aware JD-GMM is more likely to provide converted speech with equal or even better quality in the inter-gender conversion while conventional JD-GMM has a slightly better result in the intra-gender conversion.

## V. CONCLUSION

We propose a phonetically-aware framework which converts the source speech to the designated target speech. This novel

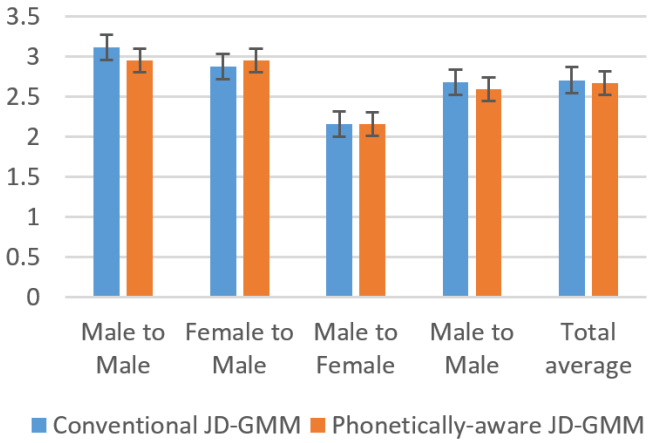


Fig. 6. Mean opinion score results

text-independent voice conversion method is full of potential. By using the JD-GMM, we can estimate the target spectral feature with phonetic discriminant feature extracted from DNN decoder without any consideration of source speaker data. Our text-independent framework not only reduces the dependence for parallel data but also increases the generalization of a trained model. It enables a new approach to achieve many-to-one conversion.

The performance of our system has been investigated on both objective and subjective evaluation. The results show that under the circumstance of no source training data, the proposed method still achieves a similar performance compared to conventional method. After some investigations on phonetic feature used in speaker verification, we believe that this system can be further improved by decomposing the current phonetic discriminant feature to a better content-dependent component, which will be our future work.

## ACKNOWLEDGMENT

This research was funded in part by the National Natural Science Foundation of China (61401524), Natural Science Foundation of Guangdong Province (2014A030313123), the Fundamental Research Funds for the Central Universities(15lgjc10) and National Key Research and Development Program (2016YFC0103905)

## REFERENCES

[1] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[2] 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.

[3] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*. IEEE, 1988, pp. 655–658.

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

[5] F. Soong and B. Juang, "Line spectrum pair (lsp) and speech data compression," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.*, vol. 9. IEEE, 1984, pp. 37–40.

[6] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[7] T. Toda, A. W. Black, and K. Tokuda, "Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter." in *ICASSP (1)*, 2005, pp. 9–12.

[8] 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.

[9] 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.

[10] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[11] D. Sundermann, H. Ney, and H. Hoge, "Vtln-based cross-language voice conversion," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 676–681.

[12] D. Sundermann, H. Hoge, A. Bonafonte, H. Ney, A. Black, and S. Narayanan, "Text-independent voice conversion based on unit selection," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[13] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1695–1699.

[14] M. Li, L. Liu, W. Cai, and W. Liu, "Generalized i-vector representation with phonetic tokenizations and tandem features for both text independent and text dependent speaker verification," *Journal of Signal Processing Systems*, vol. 82, no. 2, pp. 207–215, 2016.

[15] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated, 2014.

[16] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*. Association for Computational Linguistics, 1994, pp. 307–312.

[17] L. Ferrer, Y. Lei, M. McLaren, and N. Scheffer, "Study of senone-based deep neural network approaches for spoken language recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 105–116, 2016.

[18] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[19] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 92–97.

[20] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional hmm systems," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, vol. 3. IEEE, 2000, pp. 1635–1638.

[21] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "Shifted-delta mlp features for spoken language recognition," *IEEE Signal Processing Letters*, vol. 20, no. 1, pp. 15–18, 2013.

[22] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "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. 187–207, 1999.