Many-to-Many Voice Conversion based on Bottleneck Features with Variational Autoencoder for Non-parallel Training Data

Yanping Li*§ and Kong Aik Lee† and Yougen Yuan‡ and Haizhou Li§ and Zhen Yang*
* College of Telecommunication & Information Engineering, Nanjing University of Posts and Telecommunications, China
† Data Science Research Laboratories, NEC Corporation, Japan
‡ Northwestern Polytechnical University, Xi'an, China
§ Department of Electronic and Computer Engineering, National University of Singapore, Singapore
E-mail: {liyp,yangz}@njupt.edu.cn, kongaik.lee@gmail.com, ygyuan@nwpu-aslp.org, haizhou.li@nus.edu.sg

***Abstract*—This paper proposes a novel approach to many-to-many (M2M) voice conversion for non-parallel training data. In the proposed approach, we first obtain bottleneck features (BNFs) as speaker representations from a deep neural network (DNN). Then, a variational autoencoder (VAE) implements the mapping function (i.e., a reconstruction process) using both the latent semantic information and the speaker representations. Furthermore, we propose an adaptive scheme by intervening the training process of the DNN, which can enrich the target speaker's personality feature space in the case of limited training data. Our approach has three advantages: 1) neither parallel training data nor explicit frame alignment process is required; 2) consolidates multiple pair-wise systems into a single M2M model (many-source speakers to many-target speakers); 3) expands M2M conversion task from closed set to open set when the training data of target speaker is very limited. The objective and subjective evaluations show that our proposed approach outperforms the baseline system.**

The aim of voice conversion (VC) is to modify one's voice to sound like that of another while preserving the language content [1], [2]. Most of the current developed methods require parallel training data, which require frame level alignment between speakers to establish voice conversion function. As speaker characteristics can be different, that may lead to undesired alignment errors, thus voice conversion quality. From the viewpoint of practical applications, the requirement of having large parallel training data is unrealistic, for example, in cross-language conversion or speaking aid for vocally handicapped people [2], [3]. Therefore, it is desirable to have voice conversion system that doesn't require parallel training data.

The study of voice conversion for non-parallel training data can be classified into three main categories in general. The first category is to create pseudo-parallel data from non-parallel data [4], [5]. There are two representative techniques. One is to mark phonemes by means of a speaker-independent automatic speech recognition system (ASR), the other is to splice small speech unit to form a parallel data by means of a text-to-speech conversion system (TTS). This category also includes frame selection, phoneme clustering [6] and INCA [7]. The advantage of these methods is that it is simple in principle and easy to implement, however, such methods rely on the quality of ASR or TTS systems.

The second category is to apply the model adaptation technique to update the existing parallel conversion model using the information of speaker's background set as the prior knowledge, which includes speaker adaptation [8], [9], speaker normalization, eigenvoices [10], [11]. Such methods usually require the assistance of background speakers with parallel training data, which not only fail to entirely get rid of parallel training data constraints but also increase the complexity of the system.

The third category is to establish the relationship between the two sets of data without any alignment. One of the examples is to obtain the linear or non-linear mapping relationship between source and target speech data, including KL divergence [12], manifold, local nonlinear principal component analysis [13] and i-vector PLDA [14]. Another example is to establish the relationship between semantic space and phonetic space, and render conversion process as a controlled version of self-reconstruction, which includes mixture of factor analyzers [15], non-negative matrix factorization [3], variational autoencoder [16], [17] and recurrent neural network [18]. The advantage of such method is that it can make full use of available speech data, and reduce the dependence on the amount of training data.

Recently, deep probabilistic generative models such as variational autoencoder (VAE) has achieved tremendous success in modeling natural images, speech, handwritten digits and segmentation [19], [20]. We propose to use bottleneck features (BNFs) extracted from a deep neural network (DNN), instead of simple one-hot vector as speaker representations [16]. The BNFs have been proven to be an effective speaker features in speaker recognition and speaker clustering [21].Then we would like to study how to corporate speaker representation (e.g BNFs) in a variational autoencoder (VAE) model for many-to-many (M2M) spectral conversion. We will also study an adaptive BNF strategy by intervening

the training process of the DNN, which allows us to expand the M2M problem from closed set to open set.

Our work has three main advantages: 1) we do not require parallel training data and avoid the possible alignment errors; 2) a trained model can be applied to M2M conversion, which is different from traditional conversion model that only applies to a specific speaker pair(as in one-to-one, O2O); 3) it expands the M2M problem from closed set to open set when the training data of target speaker is very limited, which is an important step towards practical applications.

I. Voice conversion using VAE

Given spectral frames $X_s = \{X_{s,n}\}_{n=1}^{N_s}$ from the source speaker and $X_t = \{X_{t,n'}\}_{n'=1}^{N_t}$ from the target, the process of voice conversion using VAE can be decomposed into two stages. In the first stage, a speaker independent encoder f_ϕ infers a latent content z_n from x_n, which is similar to a phone recognizer. In the second stage, a speaker dependent decoder f_θ mixes z_n with a target speaker representation y_n to reconstruct a speaker dependent frame $\hat{x}_n$ ($\hat{x}_{s,n}$ or $\hat{x}_{t,n}$, depending on y_n), which operates as a synthesizer, thus the traditional conversion function f can be reformulated as:

$$\hat{x}_n = \hat{f}(x_n, y_n) = f_\theta(z_n, y_n) = f_\theta(f_\phi(x_n), y_n) \quad (1)$$

Alignment plays no roles in the formulation, because the frame feature x, speaker representation y and latent content z are on a point-wise basis instead of the pair-wise basis of traditional conversion function. We will drop the frame and speaker indices whenever readability is unharmed. The final approximated objective function is to maximize a variational lower bound of the log-likelihood:

$$logp_\theta(x|y) \leqslant -J_{vae}(x|y) = -(J_{obs}(x|y) + J_{lat}(x)) \quad (2)$$

$$J_{lat}(\phi; x) = D_{KL}(q_\phi(z|x)||p_\theta(z)) \quad (3)$$

$$J_{obs}(\phi, \theta; x, y) = -E_{q_\phi(z|x)}[logp_\theta(x|z, y)] \quad (4)$$

where $x \in X_s \cup X_t$, D_{KL} is the Kullback-Leibler divergence, $p_\theta(z)$ is the prior distribution model of z, which is chosen to be a standard normal distribution, $p_\theta(x|z, y)$ is the decoder model, and $q_\phi(z|x)$ is the encoder model. More details can be found in [16].

From the above analysis, VAE-based VC can be considered as a reconstruction process with latent content z and a replacing speaker representation y, which is suitable for non-parallel corpora. However, the speaker representation y is as simple as a one-hot vector, pre-defined for each speaker, which will lead to two problems in terms of system performance and practicability. First, in the decoding stage of VAE model, speaker representation y is not fully utilized, which only contains speaker identity, but does not contain specific and sufficient personality characteristics. Therefore, it is expected to get a better performance by introducing an improved speaker representation containing rich speaker information. Second, the baseline method can't complete the conversion task under limited training data, which is an inevitable problem in practical application.

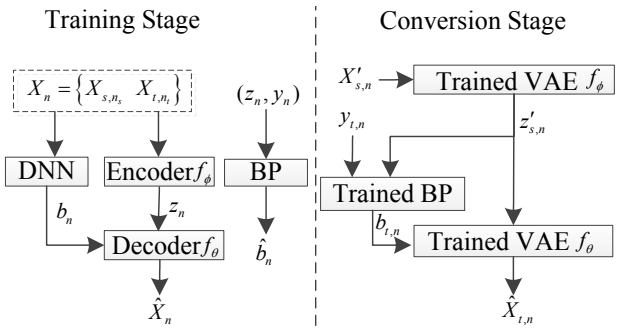

Fig. 1. Schematic diagram of VC using BNF-VAE with non-parallel training data.

II. Voice conversion using BNF-VAE

A. Improving speech models with BNFs

We are interested in whether an effective and informative speaker representation feature can improve the performance and circumvent the problem of limited training data. Recent works have proven the effectiveness of BNFs as a speaker representation in speaker recognition and speaker clustering [21]. BNFs are generated by a kind of the DNN which has a hidden layer with less neurons than other layers. Once trained, the network model is truncated at the bottleneck layer. The BNFs provide good speaker discriminative information, which are extracted from a DNN-based speaker recognition system. In this case, the expectation in (4) of the objective function can be rewritten as:

$$J_{obs}(\phi, \theta; x, b) = -E_{q_\phi(z|x)}[logp_\theta(x|z, b)] \quad (5)$$

During the conversion stage, we only have the random testing utterance of the source speaker. In order to obtain the BNFs b_n of desired target speaker during decoding in the conversion stage, we need to establish the mapping between the joint vector (z_n, y_n) and BNFs b_n in our proposed framework. We train a Back-propagation (BP) network that takes the joint vector as the input, and outputs the corresponding BNFs b_n [22]. In this way, it is possible to obtain the BNFs b_n of desired target speaker from any testing data in the conversion stage.

Fig. 1 depicts the graphical models of our proposed BNF-VAE voice conversion framework which can consolidate many pair-wise systems into one. For example, when we choose speech from five speakers as training data, BNF-VAE can convert any of the 20 permutations. In other words, it can consolidate 20 systems into one. In this way, the BNF-VAE model can accommodate M2M task under a closed set condition.

B. Open set problem with limited training data

From the view of practical application, an M2M system under an open set condition has two further requirements. First, it must be capable of converting an arbitrary, even unseen

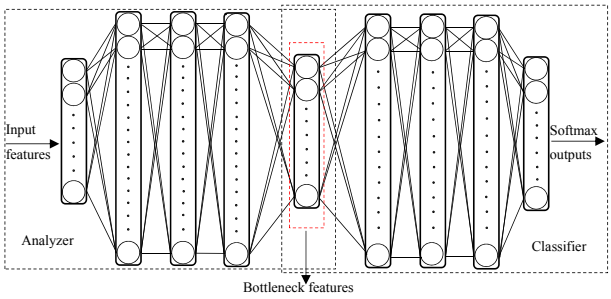

Fig. 2. BNF extraction in a speaker recognition system

source to a given target. Second, it must be able to convert a source to an arbitrary target which only has limited training speech. To answer the first requirement, the source speaker's speech is only used to extract semantic information, so the effect on system performance is very small when the source speaker's training speech is limited or even does not appear in the training stage. To answer the second requirement, we further study an adaptive scheme by intervening the training process of the DNN to learn and supplement a new space from other speakers' sufficient feature space.

Considering the speaker representative in a speaker recognition task, we can divide the DNN model into two modules, namely the analyzer and classifier. As illustrated in Fig 2, the bottleneck layer and the previous network can be considered as an analyzer, and the bottleneck layer and the beyond can be considered as a classifier. Once trained, it can be simply interpreted that analyzer acts as a feature extraction function, which can obtain BNFs from the original speech spectral feature, and classifier acts as a traditional classifier based on BNFs. It can be considered that the trained DNN not only gets the optimal classification boundary, but also forms a suitable feature space distribution in speaker recognition system at the bottleneck layer.

The steps in detail are as follows: 1) Prepare the data for training DNN, including limited data of expected target speaker and sufficient data of other speakers. The softmax nodes are equal to the total number of speaker's participating in the training; 2) After pre-training the DNN layer-by-layer, the whole DNN is optimized and trained, and the error rate of each mini-batch is monitored. When the error rate is less than a threshold (e. g. 30%), we need to suspend the training process temporarily and save the current network structure and parameters; 3) Analysis the DNN-based classification results to filter out all the wrongly classified frames into the target speaker, and then assign the target speaker's label to those frames; 4) Continue the training of the DNN until gradient convergence.

The training process of the DNN is intervened in step 2 and 3 for two assumptions. Considering that the classification boundary has been initially formed when the error rate is less than 30%, this training process needs to be intervened. The error rate of 30% is empirical and self-defined, which can be set according to the change of each mini-batch's error rate by fully training the DNN in advance. Second, it can be considered that the wrongly assigned frames to the target speaker containing the characteristic information of the target speaker to some extend. In order to enrich and supply the feature space of the target speaker, we try to change the label of wrongly classified frames. As training continues, the analyzer module is expected to reinforce and retrieve the information of target speaker from these frames changed the label.

III. Experiments and Discussions

A. The dataset and feature set

The proposed VC system is evaluated on the CMU ARCTIC dataset [23] including 7 speakers (5 male and 2 female). The signals are sampled at 16 kHz with mono channel, windowed with 25ms and shifted every 5ms using Hamming window. Acoustic features including Mel Frequency Cepstral Coefficients (MFCC)(19 dimension), F0 (1 dimension) and maximum voiced frequency (1 dimension) are extracted by Ahocoder [24]. Delta and delta-delta are appended giving rise to a 57-dimensional spectral features per frame, and then two adjacent frames are spliced together to obtain 171-dimensional spectral features. We normalize the spectral features, and the normalization factor is considered as an independent feature without any change. Then the normalized spectral feature is used in the experiments. We use the traditional Gaussian normalization method to convert the fundamental frequency parameters F0, and keep maximum voiced frequency unchanged. After the converted spectral and fundamental frequency are obtained, the normalization factor is compensated to the converted spectral, and finally the speech synthesis is performed by Ahocoder.

B. Configurations and hyper-parameters

Our DNN is a multi-layer stacked and fully connected artificial neural network, including 7 hidden layers, and the bottleneck layer in the middle has 57 nodes, while the other hidden layers have 1,200 nodes respectively. 171-dimensional spectral features are chosen as input and the output is the softmax classification of all training speakers. For the VAE model, the encoder and the decoder are fully connected. The encoder has two hidden layers with 500 nodes and 64 nodes respectively. The decoder has one hidden layer with 500 nodes. The latent layer has 32 nodes. Rectifier linear units (ReLU) is applied to each layer to provide non-linearity (except for latent layer and output layer , which are linear). The size of a mini-batch is the spectral feature of 30 frames. The optimizer is ADAM. The hidden layer of BP network has 1,200 nodes. The input layer had 37 nodes (32+5), of which 5 is one-hot vector for 5 speakers. The output layer has 57 nodes using softmax activation function.

C. Performance comparison with sufficient training data

To compare the performance of the baseline system VAE and our proposed BNF-VAE in the case of sufficient training

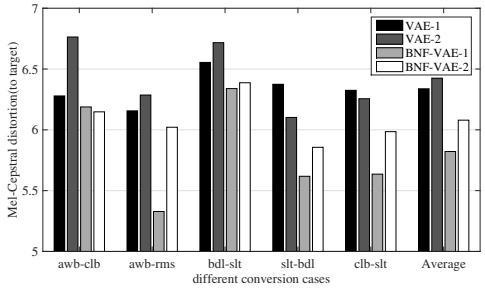

Fig. 3. Average MCD of baseline VAE and proposed BNF-VAE for different conversion cases in two training situations.

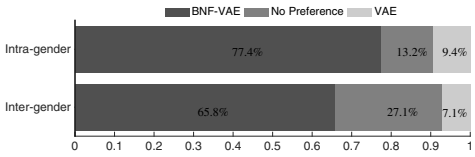

Fig. 4. ABX preference test results of baseline VAE and proposed BNF-VAE for inter-gender and intra-gender.

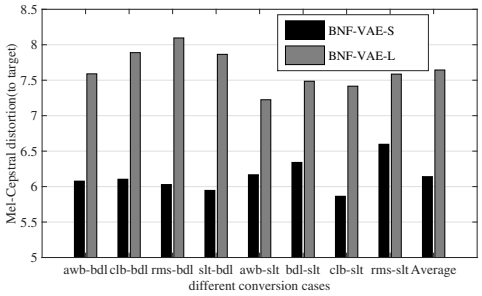

Fig. 5. Average MCD of BNF-VAE-S and BNF-VAE-L for different conversion cases.

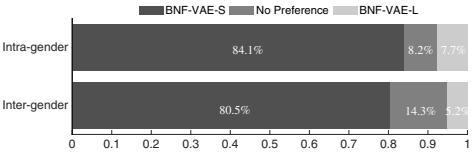

Fig. 6. ABX preference test results of BNF-VAE-S and BNF-VAE-L for inter-gender and intra-gender.

data, we choose five speaker's speech (awb, clb, rms, slt and bdl) to establish five conversion cases with intra-gender and inter-gender. The training data consists of two situations, in the first situation, we choose the same 100 sentences from each speaker, but we don't pair them with alignment. In the second situation, we randomly choose 100 sentences from each speaker. The testing data has 10 additional sentences in each conversion case.

Mel-cepstral distortion (MCD) is used to measure how close the converted is to the target speech [1], [2]. Fig. 3 shows the performance of VAE and BNF-VAE in two training data situations. For example, VAE-1 means VAE approach in the first training situation. As shown, the performance of the two approaches in the first situation is better than the second situation, but the average difference between two situations is not very great. We analyze that maybe the same utterances will be beneficial for the model learning, which inspire us that the models might not have been fully exploited. We will investigate the cause more profoundly in the future. The performance of BNF-VAE is better than VAE in different conversion cases consistently, with an average improvement of 6.75%.

We randomly chose 10 sentences from intra-gender and inter-gender conversion cases respectively for subjective listening tests. In the mean opinion scores (MOS) test, 30 listeners are asked to rate the naturalness and clearness of the converted speech on a 5-point scale. BNF-VAE achieved 3.106 MOS, with 95% confidence interval (2.925-3.287), while VAE achieved 2.258 MOS, with 95% confidence interval (2.057-2.459), which indicates that the speech quality is consistent with MCD performance.

For the ABX preference test, listeners are asked to choose which of the converted sentences A and B (generated by the two approaches) sound more like the target speaker's recording X or no preference. Each pair of A and B are shuffled to avoid preferential bias. As shown in Fig. 4, BNF-VAE is obviously preferred over the baseline approach.

D. Performance comparison with limited training data

We compare the performance of BNF-VAE in the situation of sufficient and limited training data, denoted as BNF-VAE-S and BNF-VAE-L respectively. In the sufficient data situation, the settings are similar to Sec. III-C. In the limited data situation, we choose two speakers of bdl and slt as the target speaker with only 5 sentences for training respectively, while other four speaker's training data consists of 100 sentences. Then we conduct eight conversion cases to compare the performance.

Fig. 5 shows the performance of BNF-VAE-S and BNF-VAE-L. As shown, the performance of BNF-VAE-L has directly decreased, but the semantics of synthesized speech can be discerned to a certain extent.

BNF-VAE-S achieved 3.053 MOS, with 95% confidence interval (2.878-3.228), while BNF-VAE-L achieved 2.070 MOS, with 95% confidence interval (1.885-2.255), indicating the quality of converted speech has a certain degree of decline with limited training data.

From the perspective of speaker similarity, as shown in Fig. 6, the performance of BNF-VAE-S is dominantly better than that of BNF-VAE-L. In the case of BNF-VAE-L, although the adaptation of BNFs improves the reconstruction process and the desired conversion is made possible by enriching and supplementing target speaker's feature space, the mix of other speakers' information is inevitable. From the application

point of view, our work takes an important step forward and represents a new perspective. In the next step, we will try to use i-vector [25] and x-vector [26] features as speaker representations, hoping to improve the performance of the system.

IV. Conclusion

In this paper, we proposed a BNF-VAE framework on M2M voice conversion for non-parallel training data. Objective and subjective evaluations verified that the performance of the proposed method outperforms the baseline system. Furthermore, the proposed adaptive BNF can extend M2M problem from closed set to open set, although there are still problems in speaker similarity. After all, acquiring enough personality information from limited data is still a big challenge. We will continue to improve the conversion performance, especially under limited training data.

V. Acknowledgments

This work is supported by the National Nature Science Foundation of China under Grant No.61401227, No.61671252, the Neuromorphic Computing Programme under the RIE2020 Advanced Manufacturing and Engineering Programmatic Grant A1687b0033 in Singapore, Jiangsu Government Scholarship for Overseas Studies under Grant JS-2015-002 and the Postdoctoral Fund in Jiangsu Province under Grant 1402067B.

References

[1] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[2] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016." in *INTERSPEECH*, 2016, pp. 1632–1636.

[3] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders based on non-negative matrix factorization," in *Proc. ICASSP*, 2013, pp. 8037–8040.

[4] M. Dong, C. Yang, Y. Lu, J. W. Ehnes, D. Huang, H. Ming, R. Tong, S. W. Lee, and H. Li, "Mapping frames with dnn-hmm recognizer for non-parallel voice conversion," in *Proc. APSIPA*, 2015, pp. 488–494.

[5] M. Zhang, J. Tao, J. Tian, and X. Wang, "Text-independent voice conversion based on state mapped codebook," in *Proc. ICASSP*, 2008, pp. 4605–4608.

[6] 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.

[7] D. Erro, A. Moreno, and A. Bonafonte, "Inca algorithm for training voice conversion systems from nonparallel corpora," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 944–953, 2010.

[8] T. Nakashika, T. Takiguchi, and Y. Minami, "Non-parallel training in voice conversion using an adaptive restricted boltzmann machine," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[9] A. Mouchtaris, J. Van der Spiegel, and P. Mueller, "Nonparallel training for voice conversion based on a parameter adaptation approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 952–963, 2006.

[10] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," in *Proc. ICASSP*, 2007, pp. 1245–1249.

[11] T. Hashimoto, H. Uchida, D. Saito, and N. Minematsu, "Parallel-data-free many-to-many voice conversion based on dnn integrated with eigenspace using a non-parallel speech corpus," in *Proc. INTERSPEECH*, 2017, pp. 1278–1282.

[12] F.-L. Xie, F. K. Soong, and H. Li, "A kl divergence and dnn-based approach to voice conversion without parallel training sentences." in *Proc. INTERSPEECH*, 2016, pp. 287–291.

[13] B. Makki, M. N. Hosseini, S. A. Seyyedsalehi, and N. Sadati, "Unaligned training for voice conversion based on a local nonlinear principal component analysis approach," *Neural computing and applications*, vol. 19, no. 3, pp. 437–444, 2010.

[14] T. Kinnunen, L. Juvela, P. Alku, and J. Yamagishi, "Non-parallel voice conversion using i-vector plda: Towards unifying speaker verification and transformation," in *Proc. ICASSP*, 2017, pp. 5535–5539.

[15] Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, "Mixture of factor analyzers using priors from non-parallel speech for voice conversion," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 914–917, 2012.

[16] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," in *Proc. INTERSPEECH*, 2017, pp. 3364–3368.

[17] ——, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. APSIPA*, 2016, pp. 1–6.

[18] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," in *Proc. ICME*, 2016, pp. 1–6.

[19] M. Blaauw and J. Bonada, "Modeling and transforming speech using variational autoencoders." in *Proc. INTERSPEECH*, 2016, pp. 1770–1774.

[20] C. Doersch, "Tutorial on variational autoencoders," *arXiv preprint arXiv:1606.05908*, 2016.

[21] J. Jorrın, P. Garcıa, and L. Buera, "Dnn bottleneck features for speaker clustering," in *Proc. INTERSPEECH*, 2017, pp. 1024–1028.

[22] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 4869–4873.

[23] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[24] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.

[25] K. A. Lee and H. Li, "Gain compensation for fast i-vector extraction over short duration," in *Proc. INTERSPEECH*, 2017, pp. 1527–1531.

[26] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *Proc. ICASSP*, 2018.