Non-parallel Voice Conversion with Generative Attentional Networks

Tse Wei Chiu, You Sheng Guo, and Pao-Chi Chang
Department of Communication Engineering, National Central University, Taoyuan City, Taiwan
E-mail: twchiu.vaplab@gmail.com, ysguo.vaplab@gmail.com, pcchang.vaplab@gmail.com

***Abstract*— Non-parallel voice conversion is a challenging task because the converted speech contents are not included in the target speaker dataset. Given the good performance of generator adversarial network in non-parallel voice conversion, CycleGAN-VC is used as the baseline system here. This work proposes the CycleGAN-CAM architecture with additional class activation map (CAM) to convert spectra. The attention is achieved by giving more weight to features with greater differences between speakers so that it can focus on the regions that can distinguish the source and the target features. In the loss function calculation, we incorporated the CAM loss function so that the architecture can adjust the weights automatically. The experimental results show that we have achieved better results than the baseline system in Mel-cepstrum distortion (MCD) and Mean Opinion Score (MOS).**

***Index Terms*—Voice conversion, Generative Adversarial Network, Attention, Non-parallel data**

I. INTRODUCTION

Voice conversion technology aims to convert the acoustic features of the source speaker to the acoustic features of the target speaker, and to preserve the language content, so that the converted voice sounds like the voice of the target speaker. In voice conversion, many neural network algorithms have very good performance in parallel corpus, in which the source speaker and the target speaker have the same sentence content, and can be converted according to the correspondence between each other after time alignment. Architectures such as GMM [1], DNN [2], [3], LSTM [4], ..., *etc*. are good for mapping the relationship between source and target features for voice conversion. However, in the parallel corpus, the alignment will cause distortion and greatly limit the feasibility of expanding the speaker group.

In recent years, many studies gradually used non-parallel corpus, in which the contents of sentences between the source speakers and target speakers do not need to be the same, that can easily expand the group of speakers. However, it increases a lot of difficulty in conversion technology. At present, there exist many conversion technologies for non-parallel corpus, such as Variational Autoencoder (VAE) [5], [6], Generative Adversarial Network (GAN) [7], [8], PPG [9], ..., *etc*. These methods have exhibited good performance in non-parallel voice conversion.

In view of the fact that the GAN architecture has a very good effect on the image style transfer technology, the CycleGAN architecture is adopted in this work. The CycleGAN is a particular configuration of GAN, it has been applied to voice conversion and has good results, in which CycleGAN-VC [8], [10] can be regarded as a representative research. In this paper, we use an architecture similar to CycleGAN-VC2 [10] as a baseline system and further improve it.

The attention-based method is prominent in many neural network algorithms. In terms of image style transfer, [11], [12] adds attention module to the generator adversarial network and has a very good performance. This research refers to [12] using the class activation map method [13] to modify the architecture and loss function. The generator gives greater weights to feature regions that may be significantly different between speakers and forces the model to focus on these regions for transformation, thus makes the voice conversion better.

In this paper, the architecture of CycleGAN-CAM (Class Activation Map) is proposed to transform the speech spectrum. In the loss function part, we include the CAM loss function to make the network weights adaptive. It is worth noting that, unlike many other research studies, the approach we have proposed does not rely on any additional training data and external modules.

The main contributions of this paper include: 1) We add the attention-based method of class activation map to the generative adversarial network, and propose the CycleGAN-CAM architecture to convert the spectrum. 2) We include a CAM loss function to allow the network to adjust the weights adaptively.

This paper is organized as follows. In Section II, we review the research of CycleGAN voice conversion. In Section III, the proposed method is explained. In Section IV, experimental results are reported. Conclusions are drawn in Section V.

II. RELATED WORK

A. CycleGAN for Voice Conversion

Generative adversarial networks have shown very good performances on non-parallel data conversion in many fields such as computer vision and voice conversion. In this paper, we use an architecture similar to CycleGAN-VC2 [10] as a baseline system for voice conversion.

There are two major parts in CycleGAN, the generator and the discriminator. The purpose of the generator is to generate fake samples that are close to the reality, and the purpose of the discriminator is to identify the authenticity of the samples. During training, the discriminator assists the generator by identifying the fake samples which are produced by the generator. Four architectures are used during CycleGAN

training, two generators ($G_{x\to y}$ 、$G_{y\to x}$) and two discriminators (D_x 、D_y). During voice conversion, the generator $G_{x\to y}$ can map the acoustic features of the source speaker $x \in X$ to the acoustic features of the target speaker $y \in Y$. The discriminator D_x is to identify the true similarity between the real sample (x) and the fake sample ($G_{y\to x}(y)$). Four loss functions are used to update the network in CycleGAN-VC2 [10].

The purpose of the adversarial loss function is to make the generator and the discriminator to form a state of adversarial. In order to make $G_{x\to y}$ better convert the acoustic features of x-speaker into the acoustic features of y-speaker, the adversarial loss is defined as (1)

$$L_{adv}(G_{X\to Y}, D_Y) = E_{y\sim P_Y(y)}[\log D_Y(y)] + E_{x\sim P_X(x)}[\log(1 - D_Y(G_{X\to Y}(x)))] \quad (1)$$

where D_Y is the discriminator for discriminating y-speaker acoustic features and $G_{X\to Y}(x)$ is the converted y-speaker pseudo-acoustic feature. In this formula, the discriminator expects this loss function to be maximized and to achieve the goal of capturing a converted fake sample, while the generator expects this loss function to be minimized and to achieve the goal of deceiving the discriminator.

Due to the use of a non-parallel corpus, the x-speaker and y-speaker sample contents are different in the training set. Therefore, it is not possible to compare the fake sample $G_{X\to Y}(x)$ with the real sample y. The cycle-consistency loss function produces a fake sample of the same content by allowing the real sample to go through two generators, and it calculates the L1 distance from the real sample. The cycle-consistency loss is defined as in (2)

$$L_{cyc}(G_{X\to Y}, G_{Y\to X}) = E_{x\sim P_X(x)}\left[\left\|G_{Y\to X}\left(G_{X\to Y}(x)\right) - x\right\|_1\right] + E_{y\sim P_Y(y)}\left[\left\|G_{X\to Y}\left(G_{Y\to X}(y)\right) - y\right\|_1\right] \quad (2)$$

where $G_{Y\to X}\left(G_{X\to Y}(x)\right)$ and $G_{X\to Y}\left(G_{Y\to X}(y)\right)$ are fake samples that have been converted twice to the same content as the input sample.

To save consistency in linguistic content, identity-mapping loss are used, which is defined as (3)

$$L_{id}(G_{X\to Y}, G_{Y\to X}) = E_{x\sim P_X(x)}[\|G_{Y\to X}(x) - x\|_1] + E_{y\sim P_Y(y)}[\|G_{X\to Y}(y) - y\|_1] \quad (3)$$

where $G_{Y\to X}(x)$ uses $G_{Y\to X}$ to transform the real samples of x speakers, expecting that the fake samples of x speakers after conversion retain the original linguistic content, and calculate the L1 distance from the real samples.

The second adversarial loss is similar to the adversarial loss. The difference is in the use of a discriminator for features that have been converted twice. It is defined as (4).

$$L_{adv2}(G_{X\to Y}, G_{Y\to X}, D_X) = E_{x\sim P_X(x)}[\log D_X(x)] + E_{x\sim P_X(x)}[\log(1 - D_X(G_{Y\to X}(G_{X\to Y}(x))))] \quad (4)$$

In the CycleGAN-VC2 [10] architecture, the generator uses Gate Linear Unit (GLU) [14] as the activation function, Pixel Shuffler [15] for Upsample. The 2D CNN is used to extract feature during Upsample and Downsample. The 1D CNN is used to extract the time relationship during conversion. The architecture of PatchGAN [16] is used in the discriminator to differentiate acoustic features.

III. Proposed Method

A. CycleGAN-CAM

This work proposes the CycleGAN-CAM architecture that is based on the CycleGAN-VC2 [10] architecture with additional class activation map for non-parallel voice conversion. The CycleGAN-CAM architecture is mainly used to convert spectrum features. It uses 36-dimensional mel-cepstrum (MCEPs) to represent spectrum features. During training, in each iteration it will randomly extract 128*36 MCEPs as the input.

The proposed system is an attention-based CycleGAN. The attention block is shown as in Fig 1. We use the global average pooling to quantify the features to obtain the feature vector representing the feature of each channel, and use the fully connected layer (only weight) to multiply it by the channel feature weight. The output CAM loss coefficient is used to calculate the CAM loss function so that the network can adjust the weight adaptively. Finally, the feature weight vector in the fully connected layer is multiplied by the features, which are originally input to the attention block, and the weighted features are generated as the output. Our goal is to allow the architecture to weight the channel features with large differences between speakers and to transform the different feature regions that can distinguish the source speaker from the target speaker.

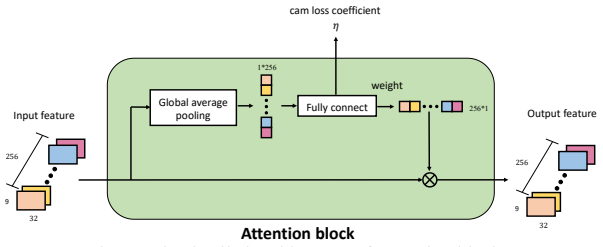

Fig. 1 The detailed architecture of Attention block

The generator architecture and network parameters are shown as in the Fig 2, where the GLU [14] is used as the activation function, and the Pixel Shuffler [15] is used for the upsample part. We add an attention block after the downsample part and use the residual block to convert feature.

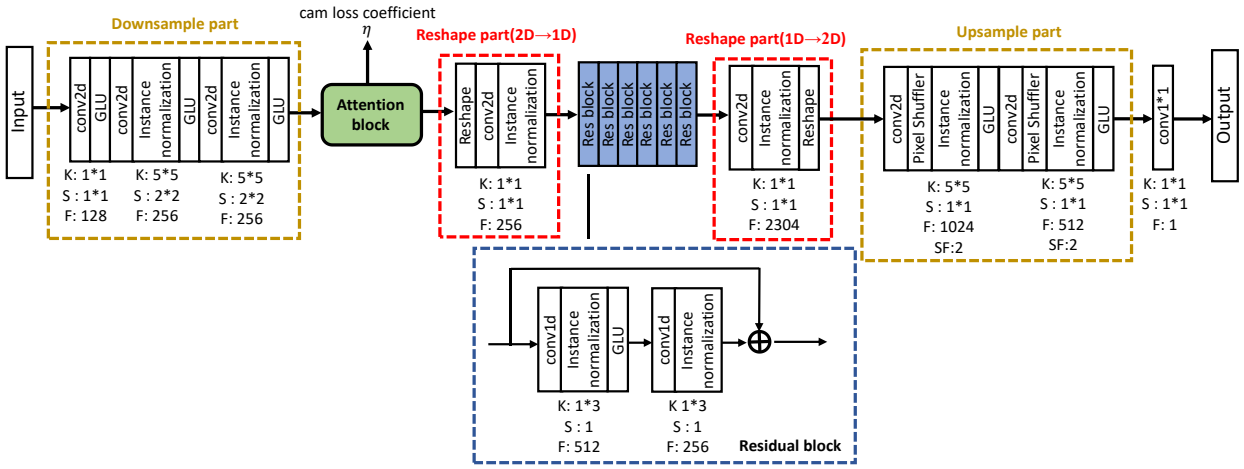

Fig. 2 Network architecture of the generator. Input and output are MCEPs. In the architecture, K, S, F denote kernel size, stride, and filter number of convolution layer. In Upsample part, SF denote scale factor of pixel shuffler.

The discriminator architecture and network parameters are shown as in the Fig 3. We add an attention block at the corresponding position of the generator. The detailed structure is the same as the attention block in the generator, and the CAM loss coefficient is also generated to calculate the loss function. It makes it more accurate for the discriminator to distinguish the authenticity of the weighted features.

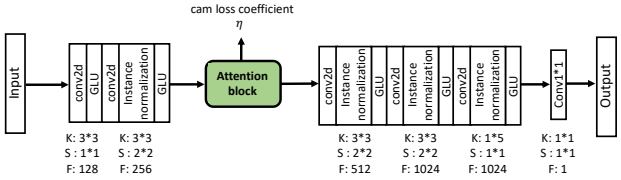

Fig. 3 Network architectures of discriminator. Input is MCEPs Output is similarity matrix. In the architecture, K, S, F denote kernel size, stride, and filter number of convolution layer.

B. CAM Loss Function

In order to allow the attention block in the architecture to adaptively adjust the weights, we use the CAM loss coefficient η to adjust the parameters in the network when calculating the loss function. It gives large weights to the feature region that differ significantly between speakers. This loss function is divided into two parts. In the generator part, take $G_{x\to y}$ as an example, the CAM loss function $L_{cam}^{x\to y}$ can be defined as (5).

$$L_{cam}^{x\to y} = -(E_{x\sim P_X(x)}\left[log\left(\eta_{x\to y}(x)\right)\right] + E_{y\sim P_Y(y)}\left[log\left(1-\eta_{x\to y}(y)\right)\right]) \quad (5)$$

When input the features of different speakers, we expect that the weights added in the same region will make η produce different extreme values. Through $L_{cam}^{x\to y}$, we can increase the weights of the region features that differ greatly from source and target speaker. For $G_{x\to y}$, it should be expected that the output value of (5) is as small as possible.

In the discriminator part, we take D_y as an example. The CAM loss function $L_{cam}^{D_y}$ can be defined as (6)

$$L_{cam}^{D_y} = E_{y\sim P_Y(y)}\left[\left(\eta_{D_y}(y)\right)^2\right] + E_{x\sim P_X(x)}[(1-\eta_{D_y}\left(G_{x\to y}(x)\right))^2] \quad (6)$$

where CAM loss coefficient η can be regarded as the preliminary identification result after weighting the features. The structure of this loss function is similar to the adversarial loss, also input real sample y and fake sample $G_{x\to y}(x)$ for identification. In this formula, the generator expects $L_{cam}^{D_y}$ to be minimized and achieves the goal of deceiving the discriminator. On the contrary, the discriminator expects $L_{cam}^{D_y}$ to be maximized and is capable of capturing the converted fake samples.

Combining the original loss function of CycleGAN-VC [10] and the CAM loss function proposed in this work, the overall loss function objective can be defined as (7)

$$\begin{aligned} L_{full}(G_{X\to Y}, G_{Y\to X}, D_X, D_Y) = \\ \lambda_{adv}(L_{adv}(G_{X\to Y}, D_Y) + L_{adv}(G_{Y\to X}, D_X) \\ +L_{adv2}(G_{X\to Y}, G_{Y\to X}, D_X') + L_{adv2}(G_{Y\to X}, G_{X\to Y}, D_Y')) \\ +\lambda_{cyc}L_{cyc}(G_{X\to Y}, G_{Y\to X}) \\ +\lambda_{id}L_{id}(G_{X\to Y}, G_{Y\to X}) \\ +\lambda_{cam}L_{cam}\left(\eta_x, \eta_y, \eta_{D_x}, \eta_{D_y}\right) \end{aligned} \quad (7)$$

where λ_{adv}, λ_{cyc}, λ_{id}, and λ_{cam} are the weights of each loss function.

After incorporating the CAM loss function, the generator can convert the voice better, and the relative discriminator can identify the real and fake samples more easily.

C. Total Conversion Model

The framework conversion process is shown in Fig 4, WORLD vocoder is used to extract Fundamental frequency (F0), Spectrum (SP) feature, and Aperiodic parameter (AP) [17]. Spectrum part uses 36-dimensional MCEPs representation and utilize the CycleGAN-CAM architecture conversion. The F0 part uses Logarithm Gaussian normalization transformation [18] to convert. Finally, we use the WORLD vocoder to reconstruct the voice waveform by Converted F0, Converted SP, and unconverted AP.

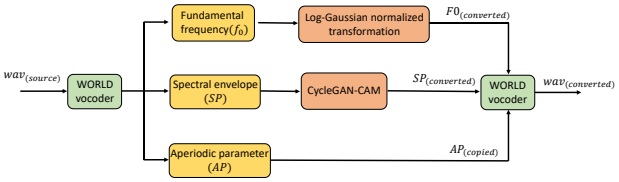

Fig. 4 The total conversion phase of the proposed voice conversion framework

IV. Experiment Result

In this section, we describe the experiments and evaluate the effectiveness of our proposed voice conversion framework in terms of spectrum.

We use the non-parallel corpus (spoke task) part of the VCC2018 dataset, which is composed of English data. We have selected 2 source speakers (SM1, SF1) and 2 target speakers (TM1, SM1) to form 4 pairs for experiments, denoted by SM→TM, SM→TF, SF→TF, and SF→TM, respectively. They are divided into Intra-gender (SF→TF, SM→TM) and Inter-gender(SM→TF, SF→TM) groups. Each speaker in the training set has 81 sentences, each sentence is about 2-7 seconds, and the total length is about 5 minutes. No additional data is used for training. Each speaker in the evaluation set has 35 sentences, and they are only used when calculating objective and subjective evaluation values.

A. Experimental Setup

For training data, we use a sampling rate of 22.05kHz, and use the WORLD vocoder to extract 36-dimensional mel cepstrum coefficients (MCEP), fundamental frequency (F0) and aperiodic parameter (AP) every 5ms. The conversion is mainly for MCEPs and F0, the AP part is directly copied without conversion.

In our proposed voice conversion framework, we use CycleGAN-CAM to convert spectrum features. In order to better maintain the feature structure, the CycleGAN-CAM generator uses 2D CNN in upsample part and downsample part, 1D CNN in the residual block, and the discriminator uses 2D CNN. In terms of loss function, we use adversarial loss, cycle-consistency loss, identity-mapping loss, two-step adversarial loss, and CAM loss. We set $\lambda_{adv} = 1$, $\lambda_{cyc} = 10$, $\lambda_{id} = 5$ and $\lambda_{cam} = 10$, and we only use L_{id} for the first 10^4 iterations. In the training phase, we use the Adam optimizer with a batch size of 1 to train the networks. We train the networks for $2 * 10^5$ iterations. We set the initial learning rate 0.0002 for the generator and 0.0001 for the discriminator and with momentum term β_1 of 0.5.

B. Evaluations

In the evaluation phase, we use Mel-Cepstrum Distortion (MCD) and Mean Opinion Score (MOS) to evaluate the voice quality and similarity after conversion.

MCD is a common objective evaluation value, which mainly evaluates the voice similarity after conversion. This evaluation uses 35 sentences of voice data in the evaluation set. We use Dynamic Time Warping (DTW) to align the converted voice length to the target speaker's voice length of the same content, and calculate MCD between each other. The results are shown in the Table I. We compare the baseline system (CycleGAN-VC) and our proposal attention-based system (CycleGAN-CAM). The table shows that CyclGAN-CAM architecture is more effective than the baseline system.

TABLE I
A Comparison of the MCD results between baseline and our proposed method

Method	Intra-gender		Inter-gender	
	SM-TM	SF-TF	SM-TF	SF-TM
CycleGAN-VC (baseline)	7.00	6.32	7.03	7.01
CycleGAN-CAM	**6.92**	**6.28**	**6.58**	**6.96**

MOS is a common subjective evaluation value. We randomly select 10 sentences of converted voice from different systems, and present them to the listeners in random order. We invite 10 listeners to listen to the converted sentences and score them according to the converted voice quality and naturalness (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The evaluation results are shown in the Table II. The results show that, in subjective hearing, the voice converted by CycleGAN-CAM is more clear than the baseline system. The results show that the CycleGAN-CAM proposed in this paper has a good performance compared with the baseline system.

TABLE II
A Comparison of the MOS results between baseline and our proposed method

Method	Intra-gender		Inter-gender	
	SM-TM	SF-TF	SM-TF	SF-TM
CycleGAN-VC (baseline)	3.19	2.32	2.14	2.87
CycleGAN-CAM	**3.37**	**2.61**	**2.30**	**3.05**

V. Conclusion

This paper proposes a new non-parallel voice conversion architecture, which is trained under limited non-parallel data and does not use external modules. We use the CycleGAN-CAM architecture with an attention-based method to convert the spectrum. The objective and subjective evaluation values indicate that our system has a good performance on the speaker similarity and voice quality.

REFERENCE

[1] 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.

[2] S. Desai, E. V. Raghavendra, B. Yegnanarayana, A. W. Black, and K. Prahallad, "Voice conversion using artificial neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 3893–3896, 2009.

[3] T. Nakashika, T. Takiguchi and Y. Ariki, "Voice Conversion Using RNN Pre-Trained by Recurrent Temporal Restricted Boltzmann Machines," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 580-587, 2015.

[4] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 4869-4873, 2015.

[5] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *Proc. IEEE, Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, pp. 1–6, 2016.

[6] J. Chorowski, R. Weiss, S. Bengio, and A. Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 12, pp. 2041–2053, Dec. 2019.

[7] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, pp. 5279–5283, 2018.

[8] T. Kaneko and H. Kameoka, "CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks," in *Proc. 26th IEEE Eur. Signal Process. Conf.*, pp. 2100–2104, 2018.

[9] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams formany-to-one voice conversion without parallel data training," in *Proc. IEEE Int. Conf. Multimedia Expo.*, pp. 1–6, 2016.

[10] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "Cycle GANVC2: Improved cycle GAN-based non-parallel voice conversion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, pp. 6820–6824, 2019.

[11] K. Junho, K. Minjae, K. Hyeonwoo, and L. Kwanghee, "U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation," *8th International Conference on Learning Representations (ICLR)*, 2020.

[12] H. Tang, H. Liu, D. Xu, P. H.S. Torr, and N. Sebe, "Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks," *arXiv:1911.11897*, 2019.

[13] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921-2929, 2016.

[14] Y. N. Dauphin, A. Fan, M. Auli, and D. Grngier, "Language modeling with gated convolutional networks," *Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, pp. 933-941, Aug. 2017.

[15] W. Shi *et al*., "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network,"*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1874-1883, Jun. 2016.

[16] P. Isla, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967-5976, 201

[17] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoderbased high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. 99, no. 7, pp. 1877–1884, 2016.

[18] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with straight for mandarin,"*in Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 4, pp. 410–414, 2007.