Interspeech 2018
2-6 September 2018, Hyderabad

A Voice Conversion Framework with Tandem Feature Sparse Representation and Speaker-Adapted WaveNet Vocoder

Berrak Sisman, Mingyang Zhang, Haizhou Li

National University of Singapore, Singapore

berraksisman@u.nus.edu, mingyang.zhang@u.nus.edu, haizhou.li@nus.edu.sg

Abstract

A voice conversion system typically consists of two modules, the feature conversion module that is followed by a vocoder. The exemplar-based sparse representation marks a success in feature conversion when we only have a very limited amount of training data. While parametric vocoder is generally designed to simulate the mechanics of the human speech generation process under certain simplification assumptions, it doesn't work consistently well for all target applications. In this paper, we study two effective ways to make use of the limited amount of training data for voice conversion. Firstly, we study a novel technique for sparse representation that augments the spectral features with phonetic information, or Tandem Feature. Secondly, we study the use of WaveNet vocoder that can be trained on multi-speaker and target speaker data to improve the vocoding quality. We evaluate that the proposed strategy with Tandem Feature and WaveNet vocoder, and show that it provides performance improvement consistently over the traditional sparse representations framework in objective and subjective evaluations.

Index Terms: Phonetic Sparse Representation, WaveNet Vocoder, Voice Conversion

1. Introduction

Voice conversion (VC) converts one speaker's voice to sound like that of another. With the advancement of the technology, voice conversion has enabled many applications such as personalized speech synthesis, spoofing attacks, and dubbing of movies.

The early studies of voice conversion were focused on spectrum mapping between source and target speakers [1, 2]. The statistical parametric approaches, such as Gaussian mixture model (GMM) [3], partial least square regression [4] and dynamic kernel partial least squares regression (DKPLS) [5] marked a success in spectrum conversion.

As a solution to the limited training data problem, non-negative matrix (NMF) based voice conversion frameworks [6] were proposed. With the NMF technique [6, 7], a group of exemplar-based sparse representation schemes were studied [8, 9, 10] to address the over-smoothing problem in voice conversion. More recently, the idea of phonetically aware multiple dictionaries [11, 12] was proposed to take into account the phonetic information in the speech content, that provided superior voice conversion quality. The traditional exemplar-based voice conversion frameworks [6, 8, 14] and phonetic sparse representation [11, 12, 13] work under the assumption that source and target speakers can share the same activation matrix. However, the activation matrix is highly dependent on the source speaker as only source spectral features have been used in the estimation process without considering the underlying phonetic state sequence. As a result, the activation matrix is affected by unwanted distortions such as speaker characteristics that we don't want to carry over to the target voice.

The phonetic posteriorgram (PPG) [15] represents the posterior probability of the each phonetic class for of a speech signal. Therefore, PPGs are supposed to be speaker independent [16]. Recently, PPG has been used in voice conversion [17, 12] to represent the underlying phonetic information to transfer across speakers. In this paper, we study voice conversion with a limited amount of parallel training data. By augmenting spectral features with PPG phonetic features to represent speech exemplars, that we call Tandem Feature, we explicitly incorporate frame-level phonetic information into the dictionaries. In this way, we believe that we can improve the estimation of activation matrix in phonetic sparse representation framework, therefore, the quality of converted speech.

It is noted that speech synthesized by traditional parametric vocoders lacks naturalness due to the over-simplified assumptions in signal processing. WaveNet vocoder [18], that directly estimates waveform samples from the input feature vectors, potentially addresses the problem. Speaker dependent and independent WaveNet vocoders [19, 20] have been proposed to make it possible to generate natural sounding synthetic voices. The WaveNet approach transforms the vocoder design into a learnable process based on the data. Through the learning, the network is expected to capture the dynamics of the complex mechanics of the human speech generation process. In this paper, we propose the use of both speaker independent and speaker adapted WaveNet vocoders to generate natural sounding speech for voice conversion.

Recently, GMM-based voice conversion, that is followed by a WaveNet vocoder [21], has been proposed and shown to achieve a good conversion performance. To our best knowledge, this paper is the first attempt to study the interaction between the speaker independent and speaker adapted WaveNet vocoder and the phonetic sparse representation technique for voice conversion with small training data. It is important to mention that the sparse representation is known for producing high similarity voice, while WaveNet vocoder offers natural sounding voice. In this paper, we aim to benefit from the best of the two techniques.

The main contributions of this paper include, 1) we propose a conversion framework by building sparse representation dictionaries based on PPG Tandem Feature (TF), that we call TF-dictionaries; 2) we propose to incorporate both frame-level and phone-level phonetic information to the sparse representation to improve the activation matrix, 3) we propose a back-off scheme as a solution to insufficient training data; 4) we propose a voice conversion framework that consists of feature conversion and a speaker independent WaveNet vocoder for voice conversion

This research is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016, Non-parametric approach to voice morphing. Berrak Sisman is also funded by SINGA Scholarship under A*STAR Graduate Academy.

10.21437/Interspeech.2018-1131

with small training data.

This paper is organized as follows: In Section 2, we explain the role of activation matrix estimation in sparse representation. In Section 3, we present the novel idea of TF exemplars, and formulate the training and run-time processes. In Section 4, we study the interaction between the WaveNet vocoder and sparse representation for voice conversion. We report the objective and subjective test results in Section 5 and conclude in Section 6.

2. Activation Matrix: A Bridge Between Speakers

In the traditional sparse representation frameworks [6][8], we construct a pair of dictionaries, denoted as $\mathbf{A}$ and $\mathbf{B}$, that consists of aligned exemplars between source and target. Due to the nonnegative nature of spectrogram, nonnegative matrix factorization (NMF) technique is employed to estimate the activation matrix $\mathbf{H}$, which is constrained to be sparse. Mathematically, the objective function is written as

$$\mathbf{H} = \underset{\mathbf{H}\geq 0}{\operatorname{argmin}}\, d\left(\mathbf{X}, \mathbf{AH}\right) + \lambda||\mathbf{H}|| \tag{1}$$

where λ is the sparsity penalty factor and $\mathbf{X}$ is the spectrogram of a source utterance. Estimating the activation matrix $\mathbf{H}$, a generalised Kullback-Leibler (KL) divergence [22] is used. It is assumed that source and target dictionaries $\mathbf{A}$ and $\mathbf{B}$ can share the same activation matrix $\mathbf{H}$. Therefore, the converted spectrogram can be written as $\hat{\mathbf{Y}} = \mathbf{BH}$.

As discussed in [23], the sharing of source and target activation matrix in [6, 8, 11, 12] is grounded neither well in theory nor in practice. Such sharing is based on the assumption that the source activation matrix $\mathbf{H}$ mostly captures the phonetic content. However, the source activation matrix in reality carries information such speaker characteristics from the input source utterance, that we don't want to carry over to the target utterance.

Given a pair of dictionaries, and a pair of parallel utterances, we can derive an activation matrix from the source utterance following Eq.(1), that we call the source activation matrix $\mathbf{H}$; If we replace $\mathbf{X}$ with $\mathbf{Y}$, and $\mathbf{A}$ with $\mathbf{B}$ in Eq.(1), we can derive an activation matrix for the target utterance, that we call target activation matrix. In the traditional approaches [6, 8, 11, 12], we typically assume that we can use the source activation matrix $\mathbf{H}$ for the target speaker at run-time because the target activation matrix is not available.

To improve the estimation of activation matrix so that it is more sharable with the target speaker, we propose to build dictionaries using Tandem Feature that consists of spectral features and PPGs. As PPGs are estimated with a large amount of temporal context, they represent the phonetic information independent of speakers.

3. Activation Matrix with Tandem Feature

We now study a novel technique for activation matrix estimation in conjunction with phonetic sparse representation, such that the activation matrix depends less on the source speaker.

3.1. Tandem Feature

A recent study shows that phonetic sparse representation [11, 12] achieves better voice conversion quality than the traditional sparse representation by using phonetic dictionaries. In this paper, we further the idea of phonetic sparse representation by augmenting spectral feature with PPG feature, that we call Tandem Feature. We believe that the phonetic dictionaries with Tandem Feature allow the activation matrix to capture the speaker independent phonetic content, thus, making the source activation matrix more shareable with the target speaker.

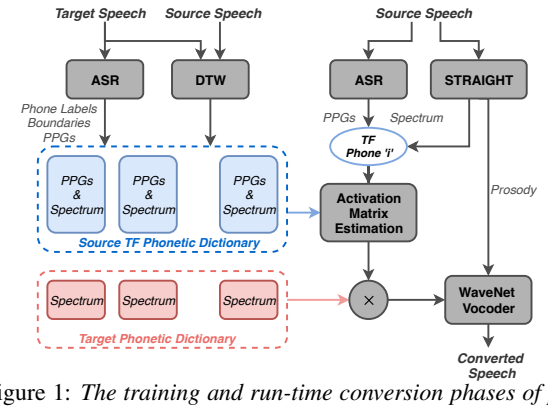


Figure 1: *The training and run-time conversion phases of proposed phonetic sparse representation framework with PPG Tandem Feature and WaveNet vocoder.*

While our proposal shares similar motivation with [11, 12], it differs from [11, 12] in many ways, for example: 1) for the first time, we incorporate both frame-level (PPGs) and phone-level (phonetic dictionary) phonetic information into the sparse representation framework, while [11, 12] only uses phone level information; 2) we propose a backoff scheme by using PPG Tandem Feature as a solution to insufficient training data; 3) for the first time, we study the interaction between exemplar-based sparse representation and WaveNet vocoder to achieve high voice quality.

3.2. Phonetic Sparse Representation with Tandem Feature

Figure 1 shows the training and run-time phases of the proposed voice conversion framework. During training, we first use a DNN-HMM based Automatic Speech Recognizer (ASR) to find the phone labels, boundaries and PPGs for each training utterance. We propose to construct multiple coupled dictionaries $[\mathbf{A}_i; \mathbf{B}_i]$, one for each phone i, where $i = 1, ..., n$, $\mathbf{A}_i$ is the source phonetic dictionary, and $\mathbf{B}_i$ the target phonetic dictionary. Different from the previous studies [11, 12], our source phonetic dictionary here consists of both source spectral features and PPGs, called *Source TF Phonetic Dictionary*, while the target phonetic dictionary $\mathbf{B}_i$ only includes spectral features.

At run-time conversion, we obtain the spectral features, denoted as $\mathbf{X}_i$ and its corresponding PPGs denoted as $\mathbf{P}_i$, for each phone of the source speaker with the same ASR in the training phase. For phone $i = k$, the objective function for estimating the activation matrix can be formulated as:

$$\mathbf{H}_k = \underset{\mathbf{H}_k\geq 0}{\operatorname{argmin}}\, d\left([\mathbf{X}_k; \mathbf{P}_k], \mathbf{A}_k\mathbf{H}_k\right) + \lambda||\mathbf{H}_k|| \tag{2}$$

The activation matrix is applied to the target phonetic dictionary to perform conversion. The converted spectrogram for phone k can be generated as $\hat{\mathbf{Y}}_k = \mathbf{B}_k\mathbf{H}_k$. The use of Tandem Feature is applicable to any exemplar-based sparse representation schemes. In this paper, we study the use of Tandem Feature in the phonetic sparse representation framework, that we call PSR-TF hereafter.

So far, no contextual information is taken into consideration. In other words, each frame is converted independently. This may lead to sharp changes across frames. By considering contextual information, one can expect a smoother output during conversion [8]. We implement exemplars which span multiple consecutive frames in phonetic dictionary to achieve a more reliable activation matrix estimation. Moreover, to account for

phone transition, we use both monophone and biphone exemplars in the TF-dictionary. We don't use a higher order of phone segments than biphone because biphone segments are enough to cover intended phone transition.

3.3. Back-off Scheme

As we have a limited amount of training data, it is not guaranteed to have an adequately constructed phonetic sub-dictionary $\mathbf{A}_k$ that corresponds to a phone $i = k$ in the input speech at run-time. In such case, a backoff Tandem Feature dictionary will be used to estimate the activation matrix, and to perform spectral mapping. In the extreme case when a acoustic dictionary for all phones is used, the proposed framework will be reduced to the traditional sparse representation [8] with Tandem Feature.

Overall, the proposed back-off scheme is an extension to the traditional sparse representation (SR) framework by incorporating PPG features. Therefore, we call it sparse representation with Tandem Feature, or SR-TF. As the proposed back-off scheme incorporates PPGs as frame-level phonetic information, we expect that it outperforms the traditional sparse representation counterpart. Last but not least, with the TF dictionary, SR-TF can also be used for spectrum conversion by itself.

4. Speaker-Adapted WaveNet Vocoder

The state-of-the-art voice conversion frameworks [1, 2, 3, 24, 25, 26, 27, 28] including sparse representation [6, 11, 12], typically use a statistical parametric vocoder. Traditional parametric vocoder is generally designed to simulate the complex mechanics of the human speech generation process under certain simple assumptions, for example, the interaction between F0 and formant structure is ignored, the phase information is discarded [29], the assumption of stationary process in the short-time window, a time-invariant linear filter. As a result, the traditional vocoding voice lacks naturalness in general. Such a problem becomes more serious in voice conversion where the feature conversion changes both F0 and the formant structure of speech among others. We expect that a good vocoder can help reconstruct the speech by harmonizing various changes.

WaveNet [18] is a well-known deep neural network that can generates raw audio waveforms. Recently proposed WaveNet vocoder [21, 20] achieves remarkable sound quality improvement over the traditional vocoders. WaveNet vocoder is able to learn the relationship between input features and output waveforms, and also able to learn the interaction among the input features. Moreover, it is shown to be successful in speech synthesis [30] and in GMM-based voice conversion [21] by improving the naturalness of the synthetic voice. Recently, a speaker independent WaveNet vocoder [20] is studied by utilizing the acoustic features such as F0, aperiodicity, and spectrum as the additional inputs of WaveNet. In doing so, WaveNet learns a sample-by-sample correspondence between the time-domain waveform and the corresponding acoustic features. To make use of human speech of a larger speaker population that is publicly available, we train a speaker independent WaveNet vocoder [20] to generate speech waveforms. Furthermore, to take into account the training samples from the target speaker, we also propose to adapt the speaker independent WaveNet vocoder towards the target speaker.

In this paper, we propose a framework where we only have a limited number of speech samples from the source and target speakers. We believe that the WaveNet vocoder can benefit from the limited samples. We train the speaker independent WaveNet vocoder in a similar way that is described in [20]. During training phase, we do not use the speech data from any of the source or target speakers. To adapt the WaveNet vocoder, we use the same target utterances that are used in the training of sparse representation. At run-time conversion, we use the adapted WaveNet vocoder to generate the converted speech waveforms.

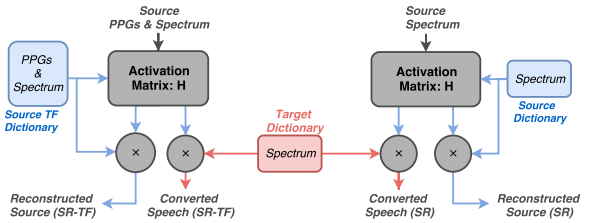


Figure 2: *SR vs SR-TF: The experiment setups for MCD study reported in Table 1.*

5. Experiments

We conduct the experiments on VCC 2016 database [31, 32] to assess the performance of spectrum conversion. For fundamental frequency (F0), we perform linear conversion, that is to normalize the mean and variance of the source speech to those of target. In all experiments, 30 source-target utterance pairs are used during training. We use a DNN-HMM based ASR [33] to obtain phone labels, phone boundaries and PPGs. The ASR is reported with 18.0% word error rate (WER) on WSJ Eval92 database. We adopt the Mel ceptral distortion (MCD) [34] between converted speech and target speech as the objective evaluation measure.

We train the speaker-independent WaveNet Vocoder with 5 hours of data from CMU Arctic and Voice Conversion Challenge (VCC) 2016 datasets. We note that the batch size is 20,000 and the iteration number is 200,000. Then, we use this speaker-independent WaveNet Vocoder as the initialized network for speaker adaptation. During the adaptation, we use about 3 minutes of speech from the target speaker, with a batch size of 20,000 over 100,000 iterations.

SR		SR-TF	
MCD_s	MCD	MCD_s	MCD
3.66	6.02	4.03	5.81

Table 1: *Comparison of spectral distortions of the re-estimated source and converted target with and without Tandem Feature. We use exemplars, that span over 3 consecutive frames.*

5.1. Objective Evaluation

To establish the baseline, we implemented some of the well established voice conversion schemes, such as traditional sparse representation approach (SR) [8] and Phonetic Sparse Representation (PSR) [12]. We compare the proposed PSR-TF scheme against the baselines.

First of all, we would like to validate that the TF-dictionary leads to an activation matrix that is more shareable between the source and target speaker, although it is estimated only from the source speech. We devise an experiment as illustrated in Figure 2, and report the MCD values of 4 different settings in Table 1. Given a test set of parallel utterances, we estimate the source activation matrix $\mathbf{H}$ with and without using Tandem Feature, denoted as SR-TF and SR [8]. We then estimate the MCD between the actual source speech $\mathbf{X}$ and the re-estimated source speech $\hat{\mathbf{X}}$ using the activation matrix $\mathbf{H}$, that we call MCD_s; and the MCD between the actual target Y and the converted target $\hat{\mathbf{Y}}$ using the same activation matrix $\mathbf{H}$, that we call MCD. In practice, $\hat{\mathbf{X}}$ is not needed, we only need $\hat{\mathbf{Y}}$. Here, we estimated $\hat{\mathbf{X}}$ just to examine whether $\mathbf{H}$ is fair to both source or target. If the activation $\mathbf{H}$ is biased to the source, we will see a low MCD_s and a high MCD. It is logical that MCD_s is

Dictionary	Monophone			Monophone+Biphone		
# Frames	1			3		
Script: Training	yes	yes	no	yes	yes	no
Script: Testing	yes	no	no	yes	no	no
MCD: PSR [12]	5.28	5.33	5.48	5.16	5.23	5.44
MCD: PSR-TF	5.22	5.26	5.39	5.08	5.19	5.37

Table 2: *Comparison of spectral distortions between the proposed phonetic sparse representation with PPG Tandem Feature (PSR-TF) and the baseline without PPG Tandem Feature (PSR [12]). We also compare the effect of contextual information, i.e., the number of consecutive frames (# Frames) as an exemplar entry in sparse representation.*

lower than MCD. By introducing TF-dictionary, we attempt to find $\mathbf{H}$ that reduces the MCD between the actual target and the converted target.

We observe that the SR-TF framework reduce the target MCD of the traditional sparse representation framework [8], from 6.02 to 5.81, with a slight increase of MCD_s from 3.66 to 4.03. This experiment not only ascertains the fact that source activation matrix is not as sharable as we expected for the target speaker, but also shows that the activation matrix obtained from PPG Tandem Feature lowers the MCD to the target, in other words, depends less on the source. In general, as SR-TF uses the frame-level phonetic information, its activation matrix depends less on the source speaker than that of SR approach, hence yields a better conversion. It is important to mention that PSR-TF uses SR-TF as the back-off scheme, just like SR being the back-off of PSR.

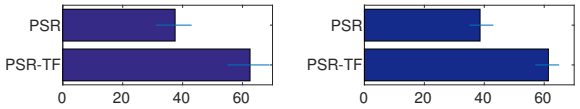
Table 2 reports the MCD values for a number of settings in a comparative study for PSR and PSR-TF. By taking into account the phonetic information at phoneme level, or segmental level, PSR has proven effective [12] to outperform the baseline sparse representation (SR) framework [8]. The proposed PSR-TF approach is an extension to the PSR by taking into account segmental level as well as frame-level phonetic information. We observe that all PSR-TF settings consistently outperform the phonetic sparse representation frameworks. In addition, we observed that multiple-frame exemplars is apparently helpful to avoid sharp changes across frames. We also observe that when scripts are available for training and/or test utterances, we obtain better phone segmentation, therefore, lower MCD values. Last but not least, we observe that both monophones and biphones in TF-Dictionaries enhances the conversion performance.

As the same source-target utterances are used for training, we can compare Table 1 and Table 2, where we use 3 consecutive frames. We observe that PSR-TF consistently achieves lower MCD values than SR-TF, as PSR-TF takes into account both phoneme and frame level phonetic information, while SR-TF only benefits from frame level phonetic information. By comparing SR with SR-TF, and PSR with PSR-TF, we find the use of phonetic information in both segmental level and frame level is rewarding.

5.2. Subjective Evaluation

We conduct four listening experiments to assess the performance of Tandem Feature, and the effect of WaveNet vocoder in PSR frameworks, in terms of voice quality and speaker similarity. 10 subjects participated in all the listening tests. Each listener listens to 30 converted utterances from 2 target speakers.

We conduct the first two listening experiments, as reported in Fig. 3a and 3b, to examine the effect of Tandem Feature in terms of voice quality and speaker similarity. We note that the original PSR framework [12] uses the STRAIGHT vocoder to generate speech waveform. To assess the effect of Tandem Feature, without changing the vocoding process, we use STRAIGHT in both PSR and PSR-TF in these experiments. Each listener is asked to decide the better sample in terms of voice quality and speaker similarity. We observe that PSR-TF outperforms the baseline PSR consistently in both voice quality and speaker similarity.



(a) *Voice quality* (b) *Speaker similarity*

Figure 3: *The preference percentage tests with 95 % confidence interval for PSR and PSR-TF.*

We further conduct a listening experiment, that is reported in Table 3, to study the listener preference of WaveNet vocoders. We perform synthesis by the speaker independent WaveNet [20] and the WaveNet that is adapted to the target speaker by using 30 utterances. We observe that the Adapted WaveNet outperforms the speaker independent WaveNet in terms of voice quality, that validates our proposed idea.

Motivated by the success of Tandem Feature and the adapted WaveNet, we now move on to the fourth listening experiment, as reported in Table 4, to assess the performance of the proposed PSR-TF with adapted WaveNet vocoder. We perform adaptation by using the same 30 utterances from the target speaker, that have been used for TF-dictionary construction. We evaluate the sound quality of the converted voices by using the mean opinion score (MOS). The listeners rate the quality of the converted voice using a 5-point scale: "5" for excellent, "4" for good, "3" for fair, "2" for poor, and "1" for bad. Table 4 shows that the proposed voice conversion framework PSR-TF with WaveNet vocoder significantly outperforms the traditional frameworks SR and PSR.

Speaker independent WaveNet [20]	**Adapted WaveNet**
$(41.0 \pm 3.2)\,\%$	$(59.0 \pm 3.7)\,\%$

Table 3: *The preference test between the speaker independent WaveNet vocoder [20] and the adapted WaveNet vocoder.*

SR	**PSR**	**PSR-TF**
2.78±0.12	3.02 ±0.15	3.41 ±0.11

Table 4: *Comparison of evaluated MOS for SR [8], PSR[12] and PSR-TF with adapted WaveNet vocoder.*

6. Conclusion

We have studied two effective ways to improve the conversion performance under limited training data. We first propose the Tandem Feature sparse representation strategy. We implement the TF-dictionary in two sparse representation frameworks as a solution to the very limited parallel training data problem. We show that the proposed strategy effectively improve the voice quality. We also propose the use of speaker adapted WaveNet vocoder. Experiment results show that the proposed framework makes good use of the limited training data and outperforms the baselines in both objective and subjective evaluations.

7. References

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

[2] Kiyohiro Shikano, Satoshi Nakamura, and Masanobu Abe, "Speaker Adaptation and Voice Conversion by Codebook Mapping," *IEEE International Sympoisum on Circuits and Systems*, pp. 594–597, 1991.

[3] Tomoki Toda, Alan W. Black, and Keiichi 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.

[4] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[5] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, "Voice conversion using dynamic kernel partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 806–817, 2012.

[6] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion in noisy environment," *In IEEE SLT*, pp. 313–317, 2012.

[7] Yi Luan, Daisuke Saito, Yosuke Kashiwagi, Nobuaki Minematsu, and Keikichi Hirose, "Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition ," *In ICASSP*, pp. 1764–1767, 2014.

[8] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou 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.

[9] Ryo Aihara, Kenta Masaka, Tetsuya Takiguchi, and Yasuo Ariki, "Parallel dictionary learning for multimodal voice conversion using matrix factorization," *In INTERSPEECH*, pp. 27–40, 2016.

[10] Zeyu Jin, Adam Finkelstein, Stephen Di Verdi, Jingwan Lu, and Gautham J Mysore, "Cute: a concatenative method for voice conversion using exemplar- based unit selection," *In ICASSP*, 2016.

[11] Ryo Aihara, Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, "Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," *In ICASSP*, 2014.

[12] Berrak Sisman, Haizhou Li, and Kay Chen Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," *IEEE ASRU*, 2017.

[13] Berrak Sisman, Haizhou Li, and Kay Chen Tan, "Transformation of Prosody in voice conversion," *APSIPA ASC. accepted for publication*, 2017.

[14] Berrak Sisman, Grandee Lee, Haizhou Li, and Kay Chen Tan, "On the analysis and evaluation of prosody conversion techniques," *IALP*, 2017.

[15] Keith Kintzley, Aren Jansen, and Hynek Hermansky, "Event selection from phone posteriorgrams using matched filters," *In INTERSPEECH*, pp. 1905–1908, 2011.

[16] Lifa Sun, Hao Wang, Shiyin Kang, Kun Li, and Helen Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," *In INTERSPEECH*, pp. 322–326, 2016.

[17] Lifa Sun, Kun Li, Hao Wang, Shiyin Kang, and Helen Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *In IEEE ICME*, 2016.

[18] Aaron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[19] Akira Tamamori, Tomoki Hayashi, and Kazuhiro Kobayashi, "Speaker-dependent wavenet vocoder," *INTERSPEECH*, 2017.

[20] Tomoki Hayashi, Akira Tamamori, Kazuhiro Kobayashi, Kazuya Takeda, and Tomoki Toda, "An investigation of multi-speaker training for wavenet vocoder," *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 712–718, 2017.

[21] Kazuhiro Kobayashi, Tomoki Hayashi, Akira Tamamori, and Tomoki Toda, "Statistical voice conversion with wavenet-based waveform generation," *INTERSPEECH*, 2017.

[22] Jort F. Gemmeke, Tuomas Virtanen, and Antti Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[23] Ryo Aihara, Tetsuya Takiguchi, and Ariki Yasuo, "Activity-mapping non-negative matrix factorization for exemplar-based voice conversion," *In ICASSP*, 2015.

[24] Heiga Zen, Yoshihiko Nankaku, and Keiichi Tokuda, "Probabilistic feature mapping based on trajectory HMMs," *In INTERSPEECH*, pp. 1068–1071, 2008.

[25] Takuhiro Kaneko and Hirokazu Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv*, 2017.

[26] Wei-Ning Hsu, Yu Zhang, and James Glass, "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data," *arXiv*, 2017.

[27] Wei-Ning Hsu, Yu Zhang, and James Glass, "Learning Latent Representations for Speech Generation and Transformation," *arXiv*, 2017.

[28] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice Conversion from Unaligned Corpora using Variational Autoencoding Wasserstein Generative Adversarial Networks," *arXiv*, 2017.

[29] Sadaoki Furui, "Digital speech processing, synthesis, and recognition(revised and expanded)," *Digital Speech Processing, Synthesis, and Recognition*, 2000.

[30] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv:1712.05884*, 2018.

[31] Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "Multidimensional scaling of systems in the Voice Conversion Challenge 2016," *In INTERSPEECH*, pp. 40–45, 2016.

[32] Tomoki Toda, Ling-Hui Chen, Daisuke Saito, Fernando Villavicencio, Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi, "The Voice Conversion Challenge 2016," *In INTERSPEECH*, pp. 1632–1636, 2016.

[33] Daniel Povey, Arnab Ghoshal, Nagendra Goel, Mirko Hannemann, Yanmin Qian, Petr Schwarz, Jan Silovsk, and Petr Motl, "The Kaldi Speech Recognition Toolkit," *In IEEE ASRU*, 2011.

[34] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Communications, Computers and Signal Processing*, pp. 125–128, 1993.