The Speaker and Language Recognition Workshop (Odyssey 2018)
26-29 June 2018, Les Sables d'Olonne, France

Phonetically Aware Exemplar-Based Prosody Transformation

Berrak Sisman, Grandee Lee, Haizhou Li

Department of Electrical and Computer Engineering
National University of Singapore

berraksisman@u.nus.edu, grandee.lee@u.nus.edu, haizhou.li@nus.edu.sg

Abstract

In this paper, we propose a novel prosody transformation framework for voice conversion by making use of phonetic information. The proposed framework is motivated by two observations. Firstly, the phonetic prosody is an important aspect of speech prosody, that is influenced by the phonetic content of utterances. We propose the use of phone-dependent dictionaries, or phonetic dictionary, that allows for effective phonetic prosody conversion. Secondly, in the traditional exemplar-based sparse representation frameworks, the estimated activation matrix highly depends on the source speech that is not the best for generating target speech. We propose to incorporate Phonetic PosteriorGrams (PPGs), that represent frame-level phonetic information, as part of the exemplars of the dictionaries. As the exemplars now consist of PPGs that are expected to be speaker-independent, the resulting activation matrix depends less on the source speaker, thus represents a better transformation function for prosody transformation. The experiments show that the proposed prosody transformation framework outperforms the traditional frameworks in both objective and subjective evaluations.

1. Introduction

We know that an individual can be uniquely identified by his/her voice. The fundamental objective of voice conversion (VC) is to modify the voice characteristics of one speaker to sound like those of another without changing the linguistic or phonetic content. Voice conversion is the enabling technology for a number of innovative applications such as personalized speech synthesis, speaking assistance, voice morphing, and dubbing of movies.

Generally speaking, a speaker can be characterized by both spectral features and prosody. Typically prosodic features include F0, energy and duration. Most of the well-known voice conversion frameworks focus only on spectral conversion. Some of the early voice conversion systems include vector quantization (VQ) [1] and fuzzy vector quantization [2]. The statistical voice conversion tenchniques include Gaussian mixture model [3, 4] and partial least square regression [5]. For applications with limited training data, the exemplar-based sparse representation framework, that is based on nonnegative matrix factorization (NMF) [6], was studied intensively [7, 8, 9, 10, 11, 12].

Recently, deep learning approach has significantly improved voice conversion performance [13, 14, 15, 16, 17]. However, such approach faces the curse of dimensionality and hinges on the size of training data, entailing an application scenario with moderate dimension and sufficient data. In addition, most of the deep learning-based VC frameworks focus on the conversion of spectral features, while carrying over the prosodic features of source speaker directly to the target with a simple F0 shifting. In this paper, we would like to address an equally important topic, that is to convert prosody of source speech to that of target in a systematic manner.

The studies on prosody and phonology [18, 19] point out the two aspects of prosody: 1) the phonetically-caused aspects of prosody, referred to phonetic prosody, that arise from purely physical phonetic factors and are not reflected in the mental lexicon; and 2) the phonologically-maintained aspects of prosody, that is related to the mental lexicon, phrase, and sentence, that is called phonological prosody. Moreover, prosody is influenced by short term as well as long term dependencies [20] as it is hierarchical in nature [21, 22]. We believe that it makes sense to convert phonetic prosody with a phone-dependent transformation scheme, and to convert phonological prosody with a suprasegmental scheme. In this paper, we will focus on the phone-dependent transformation scheme for phonetic prosody conversion.

The continuous wavelet transform (CWT) models F0 in different temporal scales that has been used to characterize F0 in hidden Markov model (HMM) [23, 24]. CWT has also been used for voice conversion such as DKPLS [20], exemplar-based approaches [25, 26] and emotional voice conversion with neural networks [27, 28]. The CWT decomposition provides us an effective tool to deal with phonetic prosody and phonological prosody analytically.

The main contributions of this paper include, 1) we propose a novel prosody conversion framework by building phonetic dictionaries for phone-dependent prosody conversion; 2) we propose to incorporate both frame-level and phone-level phonetic information to the sparse

10.21437/Odyssey.2018-38

representation to achieve a better activation matrix estimation, that depends less on source speech; 3) we validate that such activation matrix derived from Tandem Features achieves a better prosody conversion; and 4) we propose a back-off scheme by incorporating the frame-level phonetic information into the single dictionary sparse representation as a solution to insufficient training data.

This paper is organized as follows: In Section 2, we present the need for Tandem Features in the concept of sparse representation. In Section 3, we propose the novel prosody transformation framework that is based on Tandem Features. The experimental results and conclusion are given in Section 4 and 5.

2. Estimating Activation Matrix with Phonetic Information

It was shown that prosody conversion can be implemented under the traditional exemplar-based sparse representation framework [25]. Moreover, the phonetic sparse representation [26] improves the traditional framework [25] by using phone-dependent dictionaries. However, both of these frameworks are under the assumption that source and target speakers can share the same activation matrix. We argue that such sharing is not well grounded on theoretical and practical basis [29]. Ideally, the activation matrix should capture the information that represents the underlying phonetic and prosodic patterns, and work equally well for both source and target speakers. Unfortunately, the activation matrix in the frameworks above [25, 26] is optimized for source speech, thus, highly biased towards source speaker.

To alleviate this problem, we take the idea of phonetic sparse representation a step forward by incorporating frame-level phonetic information. Specifically, we propose to include the frame-aligned Phonetic Posteriograms (PPG) features, spectral features, and prosodic features (F0 and energy contours) in a single exemplar, that we call Tandem Features hereafter. PPG features represent the posterior probabilities of phonetic classes given a speech frame [30, 31] that are supposed to be speaker independent [32]. As the exemplar includes speaker independent PPG feature, we expect to derive an activation matrix that is more speaker independent, thus, effectively convert underlying phonetic and prosodic features.

In the exemplar-based sparse representation approach to spectral and prosody conversion [25], a pair of dictionaries, denoted as $\mathbf{A}$ and $\mathbf{B}$, each consists of spectrum, aperiodicity component, energy contour and 5-scale CWT representation of F0 is constructed from a parallel corpus. At run-time, the activation matrix is estimated with source spectral and prosody features, that we call the source activation matrix $\mathbf{H}$ as given below:

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

where λ is the sparsity penalty factor. A generalised Kullback-Leibler (KL) divergence is used to estimate activation matrix $\mathbf{H}$. As the target activation matrix does not exist at run-time, the traditional frameworks use the source activation matrix for the target speaker to convert the spectral and prosody features, that yields to a lower voice conversion accuracy.

We propose to build dictionaries using Tandem Features to estimate a more phonetically informed activation matrix, that is learnt from the source utterance, and to be used for the target at run-time. We note that PPGs are estimated with a large amount of temporal context, that are only weakly correlated with spectral features, but highly independent of speakers. They were used successfully as the intermediate representations of speech in LSTM-based voice conversion [33] and PSR-PPG [12].

Figure 1 shows the relationship between the proposed method and two other reference models. A group of studies [7, 8, 9, 34] under the framework of exemplar-based sparse representation (SR) provide the basic formulation, called model A, that works well for very limited training data. Phonetic sparse representation (PSR) [12, 26] marks an important progress by improving the exemplar dictionary, that is called model B. In this paper, we would like to propose a novel model, model C, by making use of frame-level phonetic information (Tandem Features) and phone-level information (phonetic dictionary) for the first time.

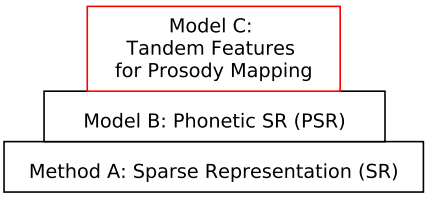

Figure 1: The relationship between the proposed method and the two other reference models.

3. Phonetic Sparse Representation with Tandem Features (PSR-TF)

We now propose a novel prosody conversion framework that uses phone-dependent dictionaries with Tandem Features as the exemplars.

We believe that a better way to convert phonetic prosody (F0, energy) is to have phone dependent dictionaries, or phonetic dictionaries. Furthermore, phonetic dictionaries with Tandem Features allow for a better estimation of activation matrix that is less dependent on

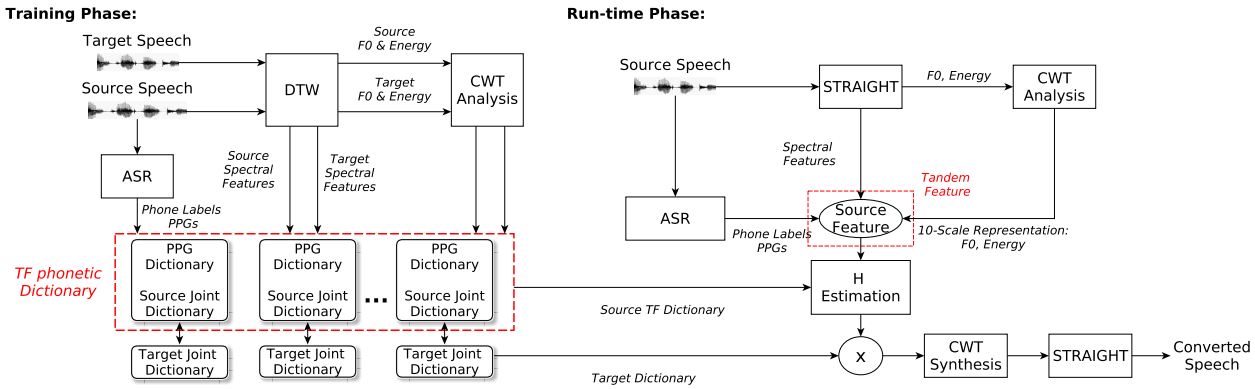

Figure 2: Training and run-time phases of the proposed prosody conversion framework, that is called Phonetic Sparse Representation with Tandem Features (PSR-TF).

source speaker. We consider that the phonetic sparse representation framework will benefit from better activation matrix with Tandem Features. To our best knowledge, this paper is the first to use PPG-spectral-prosodic Tandem Features in sparse representation for prosody conversion.

The proposed idea shares similar motivation with [26] as far as phonetic dictionary is concerned. But it differs from [26] in many ways: 1) we motivate the use of monophone and biphone phonetic dictionaries from prosody and phonological study for effective prosody conversion; 2) we propose the use of Tandem Features (PPGs, spectral and prosody features) as exemplars in the phonetic dictionary for a more balanced estimation of activation matrix; 3) we propose a back-off scheme that uses frame-level phonetic information, as an integrated solution to the case where we do not have the phonetic dictionary for a particular phone.

Figure 2 shows the training and run-time phases of the proposed prosody conversion framework. During training, we use a DNN-HMM based Automatic Speech Recognizer (ASR) to find the phone labels, boundaries and PPGs for each utterance. Instead of using a single coupled dictionary as in Section 2, we now have 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, 26], our source dictionary here consists of PPGs, source spectral features and CWT-based prosody features, that include F0 and energy contour.

The continuous wavelet transform of an input signal $f_0(t)$ can be written as:

$$W(\tau, t) = \tau^{-1/2} \int_{-\infty}^{\infty} f_0(x) \psi \left(\frac{x - t}{\tau} \right) dx, \quad (2)$$

where ψ is the Maxican hat mother wavelet. If we fix the analysis at 10 discrete scales, f_0 can be represented as

$$W_i(f_0)(t) = W_i(f_0)(2^{i+1}\tau_0, t)(i + 2.5)^{-5/2}, \quad (3)$$

where $i = 1, ..., 10$ and $\tau_0 = 5ms$. These timing scales were used in many prosody modelling [23] and voice conversion frameworks such as [25, 35, 36]. After performing wavelet analysis, the original signal can be approximated by the following reconstruction formula:

$$f_0(t) = \sum_{i=1}^{10} W_i(f_0)(t)(i + 2.5)^{-5/2} \quad (4)$$

At run-time conversion, we obtain the spectral and prosody 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 system that was used in the training phase. We then estimate the activation matrix for each phone by using KL-divergence. For phone $i = k$, the objective function can be formulated as

$$\mathbf{H}_k = \underset{\mathbf{H}_k \geq 0}{\operatorname{argmin}}\, d\left([\mathbf{X}_k; \mathbf{P}_k], \mathbf{A}_k \mathbf{H}_k\right) + \lambda ||\mathbf{H}_k|| \quad (5)$$

The converted spectral and prosody features for phone k can be written as

$$\hat{\mathbf{Y}}_k = \mathbf{B}_k \mathbf{H}_k. \quad (6)$$

The proposed phonetic sparse representation with Tandem Features is called *PSR-TF* hereafter.

We note that the 10-scale decomposition of F0 and energy contour represent the short term as well as long term dependencies of prosody. With wavelets of different time spans, some scales are more language dependent, and others are more speaker dependent. For example, scales 3-8 represent the phonetic prosody that are shown to be speaker dependent [26]. We transform scales 3-8 and carry over scales 1, 2, 9 and 10 from source to target

speakers because we consider they are speaker independent. For example, scales 1 and 2 represent long term phonological dependencies of prosody, that don't vary much from speaker to speaker.

3.1. Contextual Information

So far, each frame is converted independently, in other words, contextual information is not taken into account. This may lead to sharp changes across frames. By considering contextual information, one can expect a better conversion performance [7, 12, 26]. To achieve a more reliable activation matrix estimation, we implement exemplars that span multiple consecutive frames in TF phonetic dictionary.

In unit selection approach to speech synthesis [10, 37], we favor speech units that share similar phonetic context as the intended context by using bi-phones or triphone context. In a similar way, we use biphones together with monophones in TF phonetic dictionary to achieve a smoother phone transition. As far as prosody conversion is concerned, the studies on prosody and phonology [18, 19] reveal that the phonetic prosody is highly influenced by the transition between phonemes such as F0 perturbation caused by pre-vocalic voiced and voiceless consonants. We consider that bi-phone phonetic dictionary with Tandem Features captures the prosodic patterns arising from specific phonetic transitions.

3.2. Back-off Scheme

As we have limited training data, it is not guaranteed that there are always enough exemplars for each phonetic dictionary. In such cases, we would like to propose a back-off scheme for PSR-TF that includes exemplars for all phones.

During the training phase of PSR-TF, we also construct a back-off dictionary that includes all exemplars from source and target speakers. We incorporate PPGs, spectral features and CWT-based prosodic features to form Tandem Features in the source dictionary, called *TF back-off dictionary*, while we only include spectral and prosody features in the target back-off dictionary. In practice, the TF back-off dictionary is the union of all phonetic dictionaries.

Figure 3 shows the run-time conversion process. We note that ASR is still used to obtain Phonetic PosteriorGrams (PPGs), denoted as $\mathbf{P}$ for each test utterance of the source speaker. We augment PPGs with spectral and prosody features to obtain the source feature matrix $[\mathbf{X}; \mathbf{P}]$. With the TF back-off dictionary, we estimate the activation matrix and perform the prosody conversion. The objective function for estimating the activation matrix $\mathbf{H}$ can be formulated as follows:

$$\mathbf{H} = \underset{\mathbf{H} \geq 0}{\operatorname{argmin}}\, d\left([\mathbf{X}; \mathbf{P}], \mathbf{A}\mathbf{H}\right) + \lambda ||\mathbf{H}|| \qquad (7)$$

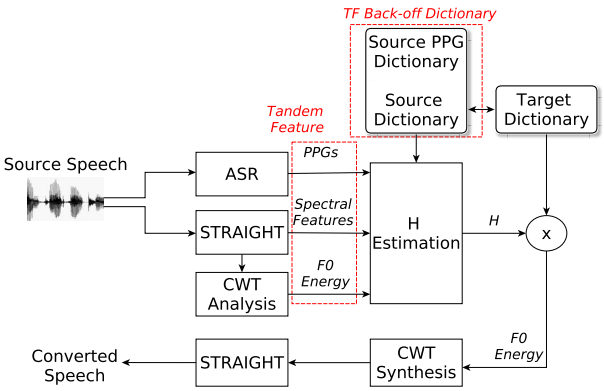

Figure 3: The run-time workflow of the proposed back-off scheme, that is Sparse Representation with Tandem Features (SR-TF).

where $\mathbf{H}$ is estimated by KL divergence.

Overall, the proposed back-off scheme is an extension to the traditional sparse representation (SR) framework [25] by incorporating PPG features. Therefore, we call it *SR-TF* for prosody conversion. We note that SR can perform prosody conversion under very limited training data. As the proposed back-off scheme incorporates PPGs as frame-level phonetic information, we expect that it outperforms the traditional SR framework. Last but not least, with the TF back-off dictionary, SR-TF can also be used for prosody conversion by itself.

4. Experiments

We conducted the experiments on the Voice Conversion Challange (VCC) 2016 dataset [38, 39] to assess the performance of the proposed prosody conversion framework for F0 and energy contour under the assumption of parallel training data. In experiments, we use a DNN-HMM based ASR [40] to obtain phone labels, phone boundaries and PPGs. The ASR is reported with 18.0% WER on WSJ Eval92 database.

Pearson correlation coefficient (PCC) [20, 36] between the converted and reference target prosody features were employed as an objective evaluation measure. Pearson correlation coefficient of two signals is a measure of their linear dependence. Mathematically, the PCC can be defined as:

$$p(S,T) = \frac{covariance(S,T)}{\sigma_S \sigma_T} \qquad (8)$$

where σ_S and σ_T are the standard deviations of signals S and T, respectively.

Besides PCC, we also examine the Frame Disturbance between the converted prosody and the reference [41, 42] . We first perform dynamic programming (DTW) to obtain the frame alignment between the original target and converted F0 contour, and calculate the number

F0 conversion	# Frames	Phonetic Dict.	# Tr. Pairs	PCC	Frame Disturbance
PSR-TF	1	Monophone	20	0.835	24.72
	1	Monophone	30	0.847	22.18
	3	Monophone+Biphone	20	0.889	20.23
	3	**Monophone+Biphone**	**30**	**0.898**	**18.47**
PSR [26]	1	Monophone	20	0.825	26.32
	1	Monophone	30	0.836	23.61
	3	Monophone+Biphone	20	0.876	21.67
	3	Monophone+ Biphone	30	0.891	19.62
SR-TF	3	-	20	0.809	31.92
	3	-	30	0.817	30.02
SR [25]	3	-	20	0.793	35.36
	3	-	30	0.801	33.31
Linear conversion (Eq. 9)	-	-	20	0.703	42.61
	-	-	30	0.721	40.55

Table 1: Comparison of the proposed phonetic sparse representation with PPG Tandem Features (PSR-TF), sparse representation with PPG Tandem Features (SR-TF), the baseline frameworks without PPG Tandem Features (SR [25] and PSR [26]) and the traditional linear F0 conversion.

of frame deviations between the target and converted F0 contour. It is noted that a large Frame Disturbance indicates poor prosody conversion performance.

4.1. Objective Evaluation

We report the experiments for F0 and energy conversion as presented in Section 3, with 20 and 30 source-target utterance pairs in training phase. We first report the experiments for F0 conversion by implementing the phonetic sparse representation with Tandem Features (PSR-TF). We use the linear F0 conversion [43], traditional sparse (SR [25]) and phonetic sparse representation (PSR [26]) frameworks as the reference baselines. Moreover, we also added the proposed back-off scheme denoted as SR-TF to show its performance compared to the baseline frameworks. The formula for linear conversion of F0 is given as follows:

$$\hat{y} = \frac{\sigma_y}{\sigma_x}(x_t - \mu_x) + \mu_y \tag{9}$$

where x_t and $\hat{y}$ are log-scaled F0 of the run-time source speech, and converted one at frame t. The parameters μ_x and σ_x are the mean and the standard deviaton of log-scaled F0 calculated from training data of source speaker, and μ_y and σ_y are the mean and the standard deviaton of log-scaled F0 calculated from training data of target speaker.

Table 1 shows the PCC and Frame Disturbance values of F0 conversion for a number of settings in a comparative study. To start with, we observed that the proposed prosody conversion framework, that is Phonetic Sparse Representation with Tandem Features (PSR-TF), outperforms all traditional approaches; PSR [26], SR [25], and the linear conversion [43]. 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. Therefore, we expect PSR-TF to outperform SR-TF as well. Moreover, the proposed back-off scheme SR-TF uses the frame-level phonetic information. As a result, the activation matrix of SR-TF depends less on the source speaker than that of SR approach, hence yields a better F0 conversion. Last but not least, by comparing SR with SR-TF, and PSR with PSR-TF, we show the effect of Tandem Features in achieving a better conversion.

We further report the experiments for energy contour conversion as presented in Table 2. As expected, we observed that all Phonetic Sparse Representation with Tandem Features settings for energy contour conversion outperform the traditional frameworks PSR and SR as well as the baseline system where we use source speaker's energy contour directly. In addition, we observed that the proposed back-off scheme consistently outperforms the traditional sparse representation.

Overall, Table 1 and 2 confirm the effectiveness of the novel idea to incorporate PPGs as phonetic features to achieve a more speaker independent activation matrix estimation. PSR has proven effective [26] to outperform the baseline sparse representation (SR) framework [25] by taking into account the phonetic information at phone level, or segmental level. However, in this paper, we propose a novel prosody conversion framework PSR-TF, that takes into account frame-level as well as phone-level phonetic information, hence achieve a better estimation of activation matrix. Lastly, we find that the use of phonetic information in both segmental level and frame level is rewarding.

Energy conversion	# Frames	Phonetic Dict.	# Tr. Pairs	PCC	Frame Disturbance
PSR-TF	1	Monophone	20	0.820	28.07
	1	Monophone	30	0.834	24.76
	3	Monophone+Biphone	20	0.833	24.80
	3	**Monophone+Biphone**	**30**	**0.845**	**22.39**
PSR [26]	1	Monophone	20	0.812	31.23
	1	Monophone	30	0.828	26.02
	3	Monophone+Biphone	20	0.826	25.87
	3	Monophone+Biphone	30	0.835	23.69
SR-TF	3	-	20	0.743	37.36
	3	-	30	0.762	36.12
SR [25]	3	-	20	0.732	38.19
	3	-	30	0.754	37.25
Direct Transfer	-	-	-	0.724	41.06

Table 2: Comparison of the proposed phonetic sparse representation with PPG Tandem Features (PSR-TF), sparse representation with PPG Tandem Features (SR-TF), the baseline without PPG Tandem Features (SR [25] and PSR [26]), and the direct transfer method.

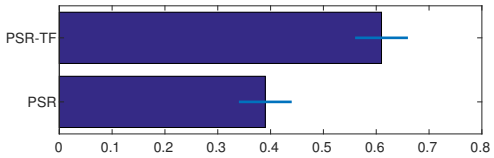

Figure 4: The preference test results of prosody similarity for F0 conversion.

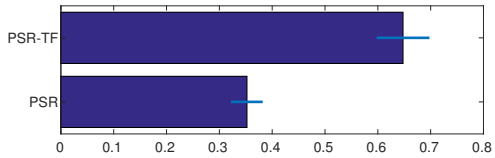

Figure 5: The preference test results of prosody similarity for F0 and energy contour conversion.

4.2. Subjective Evaluation

We further conducted the listening tests to assess the performance of the proposed prosody conversion framework PSR-TF in terms of speaker similarity. We use 30 utterance pairs in PSR and PSR-TF frameworks. 15 subjects participated in all the listening tests. 3-frame exemplars with monophone+biphone are used in all PSR and PSR-TF setups. We conducted the following 2 listening experiments to assess the conversion performance of PSR-TF:

- F0 conversion, and
- both F0 and energy contour conversion.

In these experiments, each listener was asked to listen both the converted samples and the original target samples. Then, each listener choses the sample that is closest to the target in terms of prosody similarity.

The first listening experiment that is given in Fig. 4, assesses the performance of F0 conversion. In the second listening experiment, we evaluate the performance of F0 conversion together with energy conversion, as reported in Fig. 5. We observed that PSR-TF outperforms the baseline PSR framework consistently in both F0 and energy contour conversion, that confirms the effectiveness of the proposed prosody conversion framework.

5. Conclusions

We have proposed a novel prosody conversion framework that includes F0 and energy contour. In the proposed framework, by augmenting spectral and prosody features with PPG phonetic features to represent speech exemplars, we explicitly incorporate frame-level phonetic information into the dictionaries. Moreover, we propose a back-off scheme, that is shown to be more effective than the traditional single dictionary sparse representation. The activation matrix derived from TF-dictionaries depends less on source speaker, therefore, improves the quality of converted speech. Both subjective and objective experiment results show that PSR-TF marks a success by outperforming the baseline frameworks.

6. Acknowledgment

This research is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016. Berrak Sisman is also funded by SINGA Scholarship under A*STAR Graduate Academy. The authors would like to thank Dr. Tong Rong for her support and valuable comments.

7. References

[1] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara, "Voice conversion through vector quantization," *In ICASSP*, vol. 2, 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] Heiga Zen, Yoshihiko Nankaku, and Keiichi Tokuda, "Probabilistic feature mapping based on trajectory HMMs," *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1068–1071, 2008.

[5] 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.

[6] Dd Lee and Hs Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, , no. 1, pp. 556–562, 2001.

[7] Zhizheng Wu, Tuomas Virtanen, Eng Siong Chng, and Haizhou Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[8] Zhizheng Wu, Tuomas Virtanen, Tomi Kinnunen, Eng Siong Chng, and Haizhou Li, "Exemplar-Based Voice Conversion Using Non-Negative Spectrogram Deconvolution," *8th ISCA Speech Synthesis Workshop*, 2013.

[9] Yi Chiao Wu, Hsin Te Hwang, Chin Cheng Hsu, Yu Tsao, and Hsin Min Wang, "Locally linear embedding for exemplar-based spectral conversion," *In INTERSPEECH*, vol. 08-12-September-2016, no. 1, pp. 1652–1656, 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*, pp. 2–6, 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*, pp. 7894–7898, 2014.

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

[13] Lifa Sun, Shiyin Kang, Kun Li, and Helen Meng, "Voice Conversion Using Deep Bidirectional Long Short-Term Memory Based Recurrent Neural Networks," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, , no. 1, pp. 4869–4873, 2015.

[14] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang, "Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[15] Toru Nakashika, Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Voice conversion in high-order eigen space using deep belief nets," *In INTERSPEECH*, , no. August, pp. 369–372, 2013.

[16] Ling-hui Chen, Zhen-hua Ling, Li-juan Liu, and Li-rong Dai, "Voice Conversion Using Deep Neural Networks With Layer-Wise Generative Training," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1859–1872, 2014.

[17] Jie Wu, Zhizheng Wu, and Lei Xie, "On the Use of I-vectors and Average Voice Model for Voice Conversion without Parallel Data," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[18] D. L. Bolinger, "Intonation as a universal," *Int. COngr. Linguists, Cambridge*, 1962.

[19] John J. Ohala, Alexandra Dunn, and Ronald Sprouse, "Prosody and Phonology," *ISCA Speech Prosody*, 2004.

[20] Gerard Sanchez, Hanna Silen, Jani Nurminen, and Moncef Gabbouj, "Hierarchical modeling of F0 contours for voice conversion," *In INTERSPEECH*, pp. 2318–2321, 2014.

[21] "Speech prosody: A Methodological review," *Journal of Speech Sciences*, pp. 85–115, 2015.

[22] Javier Latorre, "Multilevel parametric-base F0 model for speech synthesis," , no. May, 2014.

[23] Martti Vainio, Antti Suni, and Daniel Aalto, "Continuous wavelet transform for analysis of speech prosody," *In TRASP*, pp. 78–81, 2013.

[24] Antti Suni, Daniel Aalto, Tuomo Raitio, Paavo Alku, and Martti Vainio, "Wavelets for intonation modeling in HMM speech synthesis," *In 8th ISCA Speech Synthesis Workshop*, , no. 1, pp. 285–290, 2013.

[25] Huaiping Ming, Dongyan Huang, Lei Xie, Shaofei Zhang, Minghui Dong, and Haizhou Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," *In ICASSP*, pp. 5175–5179, 2016.

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

[27] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki, "Emotional voice conversion using neural networks with arbitrary scales F0 based on wavelet transform," *EURASIP Journal on Audio, Speech, and Music Processing*, 2017.

[28] Zhaojie Luo, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki, "Emotional Voice Conversion with Adaptive Scales F0 based on Wavelet Transform using Limited Amount of Emotional Data," *In INTERSPEECH*, pp. 3399–3403, 2017.

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

[30] Timothy J Hazen, Wade Shen, and Christopher White, "Query-By-Example Spoken Term Detection Using Phonetic Posteriorgram Templates," *In IEEE ASRU*, pp. 421–426, 2009.

[31] Keith Kintzley, Aren Jansen, and Hynek Hermansky, "Event selection from phone posteriorgrams using matched filters," *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1905–1908, 2011.

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

[33] 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.

[34] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, "Exemplar-based voice conversion in noisy environment," *In IEEE Workshop on Spoken Language Technology (SLT)*, pp. 313–317, 2012.

[35] Huaiping Ming, Dongyan Huang, Lei Xie, Jie Wu, Minghui Dong, and Haizhou Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," *In INTERSPEECH*, vol. 08-12-September-2016, pp. 2453–2457, 2016.

[36] Huaiping Ming, Dongyan Huang, Minghui Dong, Haizhou Li, Lei Xei, and Shaofei Zhang, "Fundamental Frequency Modeling Using Wavelets for Emotional Voice Conversion," *In ACII*, pp. 804–809, 2015.

[37] Paul Taylor, "Text-to-Speech Synthesis," *Cambridge University Press*, 2009.

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

[39] 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.

[40] 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.

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

[42] Chitralekha Gupta, Haizhou Li, and Ye Wang, "Perceptual Evaluation of Singing Quality," *APSIPA ASC*, 2017.

[43] Gopala Krishna Anumanchipalli, Luis C. Oliveira, and Alan W. Black, "A style capturing approach to F0 transformation in voice conversion," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.