ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/332412507

# Group Sparse Representation With WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion

**Article** *in* IEEE/ACM Transactions on Audio Speech and Language Processing · April 2019

DOI: 10.1109/TASLP.2019.2910637

CITATIONS
64

READS
677

**3 authors:**

Berrak Sisman
The University of Texas at Dallas
**116** PUBLICATIONS **2,436** CITATIONS
SEE PROFILE

Mingyang Zhang
National University of Singapore
**38** PUBLICATIONS **705** CITATIONS
SEE PROFILE

Haizhou Li
National University of Singapore
**1,078** PUBLICATIONS **28,287** CITATIONS
SEE PROFILE

All content following this page was uploaded by Berrak Sisman on 15 April 2019.

The user has requested enhancement of the downloaded file.

PGR2026-00032
Krisp Technologies EX1023 Page 1

# Group Sparse Representation with WaveNet Vocoder Adaptation for Spectrum and Prosody Conversion

Berrak Sisman, *Student Member, IEEE,* Mingyang Zhang, *Student Member, IEEE,* and Haizhou Li, *Fellow, IEEE*

***Abstract*—The statistical approach to voice conversion typically consists of a feature conversion module followed by a vocoder. So far, the feature conversion studies are mainly focused on the conversion of spectrum. However, speaker identity is also characterized by prosodic features, such as fundamental frequency (F0) and energy contour among others. In this paper, we study the transformation of speaker characteristics both in terms of spectrum and prosody. We propose two novel techniques that effectively use a limited amount of source-target training data and leverage a large general speech corpus to improve the voice conversion quality. Firstly, we study the phonetic sparse representation under the group sparsity mathematical formulation. We use Phonetic PosteriorGrams (PPGs) together with spectral and prosody features to form tandem feature in the phonetic dictionary. The tandem feature allow us to estimate an activation matrix that is less dependent on source speakers, thus providing a better voice conversion quality. Secondly, we study the use of WaveNet vocoder that can be trained on general speech corpus from multiple speakers and adapted on target speaker data to improve the vocoding quality. We benefit from the large general speech databases that are used to train the PPG generator, and the WaveNet vocoder. The experiments show that the proposed conversion framework outperforms the traditional spectrum and prosody conversion techniques in both objective and subjective evaluations.**

***Index Terms*—Voice conversion, prosody conversion, Phonetic PosteriorGrams (PPGs), Tandem Feature, WaveNet Vocoder.**

## I. Introduction

Human voice carries the information of speaker identity. The goal of voice conversion (VC) is to convert one speaker's voice to sound like that of another. Voice conversion has become an enabling technology for many new applications such as personalized speech synthesis, speaker de-identification, and voice dubbing of movies, etc.

Generally speaking, speaker identity can be characterized by 1) linguistic factors that are reflected in sentence structure and lexical choice; 2) supra-segmental factors such as the prosodic characteristics of a speech signal, and 3) segmental factors that are related to short term features, such as spectrum and formants. When the language content is fixed, the supra-segment and the segmental factors are the relevant factors to describe speaker individuality. An effective voice conversion technique is expected to convert both the supra-segment and the segmental factors.

The early studies of voice conversion were focused on spectrum mapping, that include vector quantization (VQ) [1] and fuzzy vector quantization [2]. While the statistical parametric approaches such as Gaussian mixture model [3]–[6], partial least square regression [7] and dynamic kernel partial least squares regression (DKPLS) [8] greatly improved the performance, the resulting speech quality remains much to be desired, particularly when the amount of training data are highly limited.

As a solution to the limited training data problem, non-negative matrix factorization (NMF) and sparse representation techniques were proposed [9]. Following the encouraging results in noise robust automatic speech recognition [10], the exemplar-based sparse representation techniques were studied [9], [11]–[13] to address the over-smoothing problem in voice conversion. Some more recent studies show that, by taking phonetic information into consideration [14], [15], the phonetic sparse representation technique outperforms the traditional sparse representation technique, that doesn't impose a phonetic sparse constraint on the activation matrix. In this paper, we formulate the phonetic sparse representation under the mathematical framework of group sparsity, or group sparse representation.

We note that both traditional sparse representation [9], [11], and phonetic sparse representation [14], [15] are under the assumption that the target speaker can share the activation matrix derived from source speech. However, such activation matrix is highly dependent on the source speaker. As a result, the activation matrix is affected by unwanted information such as speaker characteristics of source speakers, noise and channel distortions, that we don't want to carry over to the target voice. To alleviate this problem, we further the idea of phonetic sparse representation [14], [15] by taking into account phone-level as well as frame-level phonetic information. We construct the phonetic dictionaries by using Phonetic PosteriorGrams together with spectrum and prosody features, that we call Tandem Feature (TF). Phonetic PosteriorGrams represent the posterior probability of each phonetic class for each speech frame [16], [17] that are believed to be speaker independent [18].

The statistical approach to voice conversion typically consists of a feature conversion module that is followed by a traditional parametric vocoder. The speech synthesized by traditional parametric vocoders lacks naturalness due to the over-simplified assumptions in the signal processing process. WaveNet vocoder [19], that directly estimates

This research is supported by Ministry of Education, Singapore AcRF Tier1 NUS Start-up Grant FY2016. Berrak Sisman is also funded by SINGA Scholarship under A*STAR Graduate Academy.

PGR2026-00032
Krisp Technologies EX1023 Page 2

waveform samples from the input feature vectors, potentially addresses the problem. There have been studies of speaker dependent and independent WaveNet vocoders [20], [21] that generate natural sounding synthetic voices. The WaveNet approach transforms the vocoder design into a data-driven learnable process. 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 study the use of both speaker independent and speaker adapted WaveNet vocoders for voice generation.

There has been reported work on GMM-based voice conversion followed by WaveNet vocoder [22]. For the first time, this paper studies the interaction between the WaveNet vocoder and the sparse representation technique. We note that the sparse representation is known for producing high similarity voice, while WaveNet vocoder offers natural voice quality. We aim to benefit from the best of the two.

The main contributions of this paper are in three areas: a) we formulate group sparse representation as the theoretical foundation of phonetic sparse representation for the integrated solution to spectrum and prosody conversion; b) we study the speaker dependent and independent characteristics to motivate the design of sparse representation framework, namely phonetic activation matrix, that is based on Tandem Feature and differentiated multi-scale prosody transfer; c) we implement a voice conversion framework with phonetic sparse representation and WaveNet vocoder that provides superior performance for limited training data.

This paper is organized as follows: In Section II, we present the related work. In Section III, we study the modeling of prosody features with wavelet transform. In Section IV, we propose group sparse representation for voice conversion. In Section V, we propose the use of Tandem Feature in a joint spectrum and prosody conversion framework. In Section VI, we discuss the use of WaveNet vocoder in the sparse representation voice conversion. We report the experiments in Section VII and conclude in Section VIII.

## II. Related Work

This work is built on several previous studies on sparse representation voice conversion, and phonetic and phonological prosody. Here we briefly summarize the previous work in the context of voice conversion to set the stage for our study.

### A. Exemplar-based Sparse Representation

Non-negative matrix factorization (NMF) [23] is an effective data mining tool that has been widely used, especially when reconstruction of high quality signals is required, such as in speech enhancement [24], [25], speech de-noising [26], [27], noise and speech estimation [28]. It factorizes a matrix into two matrices (a dictionary and an activation matrix) with the property that all three matrices have no negative elements. The NMF-based techniques were shown effective in voice conversion with very limited training data. We just name a few as examples, non-negative spectrogram deconvolution [29], locally linear embedding (LLE) [30], and unit selection [13]. In NMF-based approaches, a target spectrogram is constructed as a linear combination of exemplars. Therefore, over-smoothing problem can arise.

To alleviate such over-smoothing problem, one effective solution is to apply sparsity constraint to the activation matrix [14], [15], that we call exemplar-based sparse representation. The recent works on sparsity constraint includes discriminative graph-embedded NMF approach [12], phonetic sparse representation for spectrum conversion [15], and conversion of timbre and prosody [31], [32].

In exemplar-based sparse representation, a pair of dictionaries, denoted as $\mathbf{A}$ and $\mathbf{B}$, consists of aligned exemplars from source and target. [$\mathbf{A}$; $\mathbf{B}$] is called the coupled dictionary. At run-time, the spectrogram of a source utterance $\mathbf{X}$ can be represented as

$$\mathbf{X} \approx \mathbf{A}\hat{\mathbf{H}} \tag{1}$$

Due to the nonnegative nature of spectrogram, NMF technique is employed to estimate the source activation matrix $\hat{\mathbf{H}}$, which is constrained to be sparse. Mathematically, we estimate $\hat{\mathbf{H}}$ by minimizing an objective function,

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

where $\lambda$ is the sparsity penalty factor. To estimate activation matrix $\hat{\mathbf{H}}$, a generalised Kullback-Leibler (KL) divergence is used. It is assumed that source and target dictionaries $\mathbf{A}$ and $\mathbf{B}$ can share the same source activation matrix $\hat{\mathbf{H}}$ from Eq.(2). Therefore, the converted spectrogram can be written as

$$\hat{\mathbf{Y}} = \mathbf{B}\hat{\mathbf{H}}. \tag{3}$$

We note that the source activation matrix $\hat{\mathbf{H}}$ is estimated from the source speech, therefore, is biased very much towards the source signals. In this paper, we further the idea of exemplar-based sparse representation by studying how to reduce the bias, and the way to improve the sparse representation through phonetic group sparsity.

### B. Spectrum Conversion with Phonetic Sparse Representation

It is known that a compact and effective dictionary has an important role in achieving a good voice conversion quality [33]. In addition, multiple dictionaries with a dictionary selection strategy [14] also help to improve the conversion performance of sparse representation. Our recent work in phonetic sparse representation [15] voice conversion is another example that employs a phonetic dictionary with phonetically labeled exemplars, or phonetic sub-dictionaries. It has been shown that multiple sub-dictionaries outperform single dictionary in exemplar-based sparse representation voice conversion [14], [15].

The studies in [14], [15] haven't formulated the problem of multiple sub-dictionaries under a theoretical framework. They assume that source and target speakers can share the same activation matrix, without looking into how to estimate such a shareable activation matrix. Moreover, they

PGR2026-00032
Krisp Technologies EX1023 Page 3

focus on spectrum conversion without studying the prosody transfer between the speakers. To gain insights into the previous work, we introduce group sparse representation in this paper to describe both exemplar-based sparse representation, that is referred to as traditional sparse representation, and phonetic sparse representation under an unified mathematical framework. We further study the use of Tandem Feature for phonetic dictionary to obtain speaker invariant activation matrix for spectrum and prosody conversion.

### C. Phonetic and Phonological Prosody

The prosody of speech manifests itself prominently at two different levels [34], [35]: 1) the phonetically-caused aspects of prosody, referred to phonetic prosody, that arise from purely physical phonetic factors. Phonetic prosody are associated with phones and not reflected in the mental lexicon; and 2) the phonologically-maintained aspects of prosody, that is related to the mental lexicon, phrase, and sentence, referred to as phonological prosody. It is observed that prosody pattern is influenced by short term as well as long term dependencies [36] as it is hierarchical in nature [37], [38].

Prosody conveys linguistic, para-linguistic and various types of non-linguistic information, such as speaker identity, intention, attitude and mood [39], [40]. It is inherently supra-segmental [41], [42] due to the fact that the variations of prosody cannot be derived solely from the segmental sequence [37]. Speech prosody is affected by long-term dependencies at different levels such as word, phrase and utterance. At the same time, it is also affected by segmental differences [37]. For example, voiceless segments lack explicit F0 values and high vowels generally have higher F0 than low vowels. Unfortunately, there have been a few studies on how voice conversion can benefit from the findings in prosodic analysis.

The continuous wavelet transform (CWT) was used for the analysis and modeling of F0 within an hidden Markov model (HMM) framework [43] [44]. The CWT decomposition of F0 and energy contours was also used in voice conversion framework such as DKPLS [36], exemplar-based prosody conversion [31], [32], [45], and emotional voice conversion [46].

Motivated by the previous work, we would like to further the study of CWT decomposition for prosody modeling under the sparse representation voice conversion framework. We also believe that the phonetic sparse representation is a logical choice of platform for phonetic-phonological prosody conversion.

## III. Prosody Modeling with CWT

Prosody is hierarchical in nature [32] [37] [38] and it can be affected by both short term as well as long term dependencies [36]. Fundamental frequency (F0) is a crucial prosodic feature in speech, hence previous studies of prosody conversion mainly are focused on transformation of F0. We will study a technique to convert both F0 and energy contours next.

One of the challenges is that prosody is highly dynamic. The limited amount of training utterances are generally not enough to characterize the prosody patterns of a speaker. The traditional technique for F0 conversion usually transforms the mean and variance of source speaker's F0 to those of the target [11]. One of the disadvantages of this technique is that it is based on a frame-level operation. We note that human manipulates F0 in a segmental manner, including phone, syllable, word, phrase or sentence level. GMM-based mapping [47], piecewise linear mapping [48] represent some promising extensions of this traditional approach [11].

Recently, it has been shown that CWT can effectively model F0 in different temporal scales that improves the speech synthesis performance [42]–[44]. Inspired by this, CWT is studied in F0 modeling within sparse representation [31], DKPLS for voice conversion [36], emotion conversion with arbitrary F0 scales [49] and adaptive scales for emotional voice conversion [50]. It was also studied that CWT decompositions of F0 and energy contours provide a way to manipulate the prosody of utterances [46].

The continuous wavelet transform of an input signal $f_0$ can be written as

$$W(\tau, t) = \tau^{-1/2} \int_{-\infty}^{\infty} f_0(x) \psi \left( \frac{x-t}{\tau} \right) dx, \tag{4}$$

where $\psi$ is the Mexican hat mother wavelet. If we fix the analysis at 10 discrete scales, $f_0$ can be represented as follows [31]

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

where $i = 1, ..., 10$ and $\tau_0 = 5ms$. These timing scales were originally proposed in [44] and in hierarchical prosody model [43], that were further used in voice conversion [31], [32], [45], [46]. The use of these particular scales is motivated by the attempt to relate the scales to different levels of linguistic structure. In this representation, lower scales capture short-term variations and higher scales capture the long-term variations associated with the utterance. The reconstruction formula is given as follows:

$$f_0(t) = \sum_{i=1}^{10} W_i(f_0)(t)(i+2.5)^{-5/2}. \tag{6}$$

In this study, we adopt CWT to decompose both F0 and energy contour into 10 temporal scales, that can be used to model different prosodic levels ranging from micro-prosody to sentence levels. As CWT is sensitive to discontinuties in the prosody features, the following pre-processing steps are needed: 1) transformation of F0 and energy values from linear to logarithmic scale, 2) smoothing F0 and energy contour by using 3-point mean filter 3) linear interpolation over unvoiced regions, 4) normalizing the resulting F0 and energy contour to zero mean and unit variance.

Prosody typically reflects a combination of features from both the speaker and the linguistic content of the utterance.

PGR2026-00032
Krisp Technologies EX1023 Page 4

The examples of speaker features include the dialectal background, the lexical choice in the expression, the speaking rate, the emotional state of the speaker etc; while the examples of linguistic content, referred to as utterance features, include the form of the utterance (statement, question, or command); the presence of irony or sarcasm; emphasis, contrast, and focus etc. In general, we consider speaker features are speaker dependent, and utterance features are speaker independent because they are manifested in the same way by all speakers. It is generally believed that the elements of prosody can be identified as either speaker dependent and independent. In voice conversion [51], [52], we would like to carry over the speaker independent prosody from the source to the target, but to replace the speaker dependent prosody of source speaker with that of target speaker.

We believe that a better way to convert phonetic-phonological prosody (F0, energy) is to associate the phonetic prosody with the phonetic dictionary. The 10-scale decomposition of F0 and energy contour provides a way to model 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, studies have shown that scales 4-8 in a 10-scale decomposition represent the phonetic prosody that are more speaker dependent than other scales [32], [53], [54].

## IV. Group Sparse Representation for Voice Conversion

### A. Group Structure and Group Sparsity

Many real world data are inherently organized in groups with group prior information. For example, phoneme in a speech signal, speaker identity in a telephone recording, and topic in document archive form natural groupings. NMF is an effective data mining tool that has been widely used in many pattern recognition tasks. The idea of group sparsity is to incorporate group prior information into a mixed norm regularization process for non-negative matrix factorization, that constrains the activation matrix to the relevant data group in the optimization process. NMF with group sparsity has been used to improve speaker identity discrimination [55] and speaker identification [56] where speaker identity is used to induce group sparsity. In this paper, we would like to show that, given the phonetic class as the group prior information, sparse representation actually leads to phonetically aware sparse representation [15].

Let's now describe the phonetic group structure that we consider in Fig. 1. Suppose that we would like to convert a word *bed* with 3 phonemes from the source to the target speaker. The spectrum features are characterized by 3 prevalent phonetic groups */b/*, */e/* and */d/*. The motivation of our work is that the spectrum features belonging to a specific phoneme may share the same sparsity pattern, thus the activation matrix should be constrained to be sparse similarly within the the relevant phonetic group. We propose an extension of NMF that incorporates phonetic identity as the prior information to show the theoretical intuition of phonetic sparse representation.

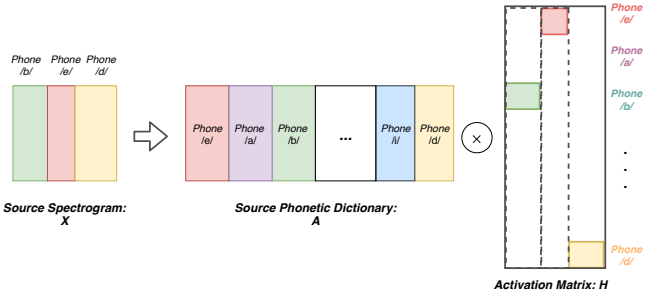



Fig. 1: The group sparse representation of a source spectrogram **X** for the word *bed* that has 3 phonetic groups highlighted in 3 colors. **X** is factorized with group sparsity into the product of the source phonetic dictionary **A** and the activation matrix **H**. The colored rows of **H** represent nonzero sections of the activation matrix.

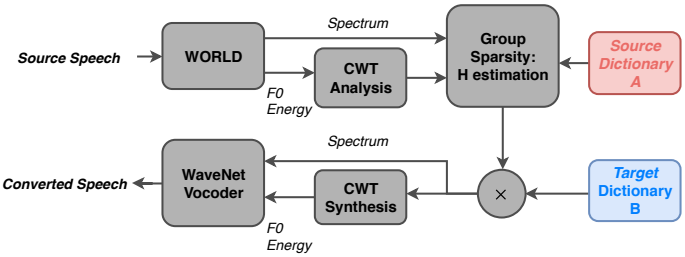



Fig. 2: The activation matrix, H, estimation and run-time conversion of the sparse representation framework under the group sparsity constraint. Exemplars in source and target dictionaries are with the phonetic labels.

### B. Formulation of Group Sparsity for Voice Conversion

The traditional sparse representation framework for voice conversion constrains the activation matrix to be sparse without taking into account the group membership information. By defining phoneme class as the group membership, we formulate the problem of sparse representation. Next we show how the proposed group sparse representation works in voice conversion.

During training, we first use a vocoder, for example, WORLD [57], to analyze the spectrum and prosody features. Then, we decompose F0 and energy into 10 temporal scales by using Continuous Wavelet Transform. At the same time, we use an Automatic Speech Recognizer (ASR) to find the phoneme labels and boundaries for each pair of source-target parallel utterances. We construct a coupled dictionary $[\mathbf{A};\mathbf{B}]$, of which the spectrum and prosody exemplars are phonetically aligned, referred to as phonetic dictionary. The phonetic dictionaries **A** and **B** are divided into sections that are also called $\mathbf{A}_i$ and $\mathbf{B}_i$ sub-dictionaries, where $i = 1, ..., N$ with $N$ being the number of phone entries in the phonetic dictionary.

The run-time conversion of group sparse representation is given in Fig. 2. We note that ASR is only used during training when we construct the phonetic dictionary. The

| Framework | ASR | Phonetic Dict. | H Estimation |
|---|---|---|---|
| Group Sparse Representation | Training only | Yes | Eq. (7) |
| Phonetic Sparse Representation [15] | Training & Testing | Yes | Eq. (8) |
| Sparse Representation [11] | N. A. | No | Eq. (2) |

TABLE I: The comparison of sparse representation approaches in terms of the usage of ASR and phonetic dictionary. We consider group sparse representation is a constrained version of traditional sparse representation, and phonetic sparse representation assigns the phonetic class as the group in group sparse representation.

group sparsity criterion for sparse representation [58] can be written as follows:

$$\hat{\mathbf{H}} = \underset{\mathbf{H}\geq 0}{\operatorname{argmin}}\ d\left(\mathbf{X}, \sum_{i=1}^{N} \mathbf{A}_i\mathbf{H}_i\right) + \lambda \sum_{i=1}^{N} \sqrt{p_i}||\mathbf{H}_i|| \tag{7}$$

where $A_i$ is the subdictionary for phone $i$ and $p_i$ is the length of $\mathbf{H}_i$. This criterion exploits the non-differentiability of $||\mathbf{H}_i||$ at $\mathbf{H}_i = 0$; setting some groups of coefficients to exactly 0. In this way, the group sparsity leads to a sparse set of groups. If the group sparsity decides to activate a group in the model, all coefficients in the group carry non-zero values.

To understand the difference between the group sparse representation and the traditional sparse representation, let's compare Eq. (2) and Eq. (7). We note that Eq. (2) decomposes an source spectrogram into $\mathbf{A}\times\hat{\mathbf{H}}$ where $\hat{\mathbf{H}}$ is sparse over the speech exemplars in the dictionary $\mathbf{A}$. Eq.(7) decomposes an source spectrogram into $\sum\mathbf{A}_i \times \hat{\mathbf{H}}_i$ where $\hat{\mathbf{H}}_i$ is sparse over the phonetic group of speech exemplars. In the latter case, only a few phonetic groups in the phonetic dictionary $\mathbf{A}_i$ are activated in the factorization. During training, the group sparse representation uses an ASR to derive a phonetic dictionary where each exemplar is phonetically labeled, while the traditional sparse representation puts all exemplars into a single dictionary indiscriminatively.

With the group sparsity regularization, only the phonetic sub-dictionary that is relevant to the input features is likely to be activated at run-time. As the activated sub-dictionary only contains the phonetic exemplars that are similar to the input features, we can expect the group sparse representation offers a better activation matrix, thus better converted voice than the traditional sparse representation [11].

We consider that the proposed formulation of group sparse representation provides the theoretical intuition to the phonetic sparse representation that was reported in [15]. To summarize, we compare the family of sparse representation frameworks in Table I. Group sparse representation and phonetic sparse presentation both use ASR to derive phonetic dictionary during training. At run-time, phonetic sparse representation estimates activation matrix using phonetic labels of source features from ASR [15], while group sparse representation doesn't use phonetic labels from ASR. In phonetic sparse representation, the objection function for estimating the activation matrix $\hat{\mathbf{H}}_k$ for phone $i = k$ is formulated [15] by selecting the corresponding phonetic sub-dictionary $\mathbf{A}_k$,

$$\hat{\mathbf{H}}_k = \underset{\mathbf{H}_k\geq 0}{\operatorname{argmin}}\ d\left(\mathbf{X}_k, \mathbf{A}_k\mathbf{H}_k\right) + \lambda||\mathbf{H}_k|| \tag{8}$$

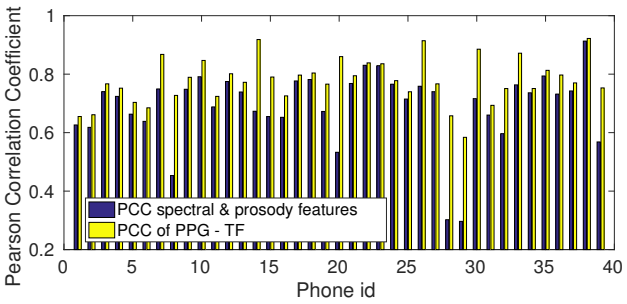



Fig. 3: Pearson Correlation Coefficient of spectrum & prosody features vs. Tandem Feature. As Tandem Feature shows consistently high PCCs, thus is more speaker invariant than the spectrum and prosody features.

which is similar to Eq.(2) except that $\hat{\mathbf{H}}_k$ is now phone dependent [15] and the overall $\hat{\mathbf{H}}$ can be written as $[\hat{\mathbf{H}}_1, \hat{\mathbf{H}}_2, ..., \hat{\mathbf{H}}_k, ..., \hat{\mathbf{H}}_M]$, where $M$ is the number of phones in a given utterance. The phonetic labels come from either an ASR or the transcript of the source speech at run-time. The estimation of activation matrix for group sparse representation, or Eq. (7), is different from Eq. (8) in that it doesn't explicitly require the phonetic labels of source speech. It optimizes Eq. (7) numerically to find $\hat{\mathbf{H}}_k$. In the ideal case, we end up with $\hat{\mathbf{H}}_k$ that corresponds to the phone i section of in the source speech features through Eq. (7). In this way, we can expect that Eq. (7) and Eq. (8) lead to almost the same matrix decomposition. As the traditional sparse representation does not take into account the phonetic information, it doesn't involve ASR either during training or testing.

## V. Joint Spectrum and Prosody Conversion with Tandem Features

In this section, we propose an integrated voice conversion framework that jointly transforms both spectrum and prosody under the formulation of group sparse representation. As shown in Table I, phonetic sparse representation is the most constrained version of sparse representation among the three variants, and offers the best performance. Next we study the use of speaker invariant features under the phonetic sparse representation framework. Without loss of generality, the same feature scheme will be applicable to the implementation of group sparse representation and traditional sparse representation frameworks as well.

### A. Tandem Feature

The sparse representation marks a success in voice conversion when only a very limited amount of training data is available [11], [15]. In the sparse representation framework,

PGR2026-00032
Krisp Technologies EX1023 Page 6

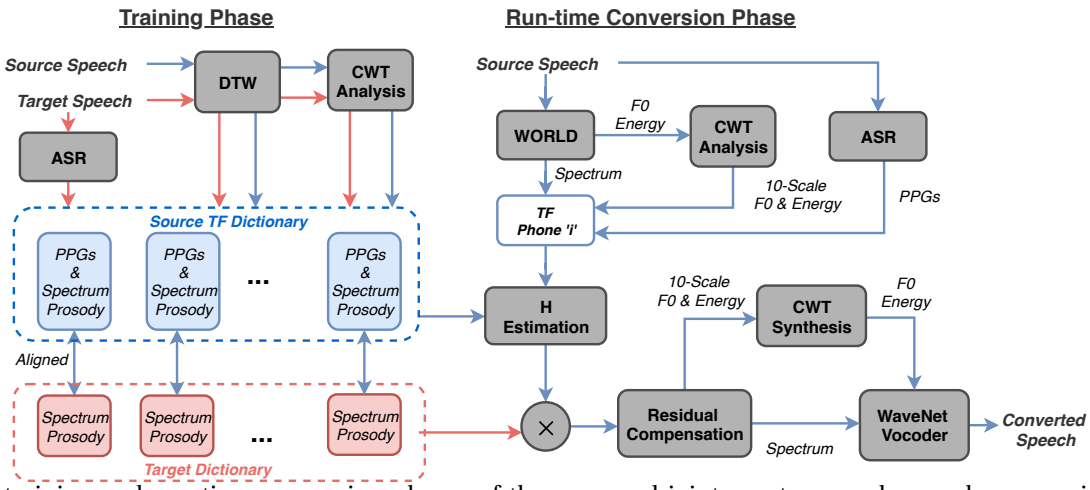



Fig. 4: The training and run-time conversion phases of the proposed joint spectrum and prosody conversion framework with Tandem Feature.

the activation matrix is shared between the source and target speakers. Ideally, the activation matrix should capture the speaker invariant phonetic information [59]. We now argue that such sharing might not be grounded well either in theory or in practice.

Suppose that we have a coupled dictionary and a pair of parallel utterances. For ease of discussion, we use the traditional sparse representation framework here as an example to discuss the sharing of activation matrix. The same argument is valid for both group sparse representation and phonetic sparse representation.

Using Eq.(2), we can derive an activation matrix for the source utterance, that we call source activation matrix $\hat{\mathbf{H}}$; 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 $\hat{\mathbf{H}}^t$, as follows:

$$\hat{\mathbf{H}}^t = \underset{\mathbf{H}^t \geq 0}{\operatorname{argmin}}\ d\left(\mathbf{Y}, \mathbf{B}\mathbf{H}^t\right) + \lambda ||\mathbf{H}^t|| \tag{9}$$

In [9] [11] and [15], it is assumed that the source activation matrix $\hat{\mathbf{H}}$ and target activation matrix $\hat{\mathbf{H}}^t$ are similar enough, therefore the source activation matrix is used for the target speaker. Such sharing is based on the assumption that the source activation matrix $\hat{\mathbf{H}}$ mostly captures the phonetic content. However, the source activation matrix in reality carries speaker characteristics from the input source utterance, that we don't want to carry over to the target utterance. The same problem exists in to Eq. (7) and Eq. (8).

To alleviate this problem, we propose to use Tandem Feature that consists of Phonetic PosteriorGrams (PPGs), spectrum and prosody features to represent the exemplars, thus, the dictionaries. Phonetic PosteriorGrams represents the posterior probability of each phonetic class for each speech frame [16]–[18], [60]. We propose to convert all speaker's voice into PPGs, and construct dictionaries that consist of source spectrum and prosody features as well as PPGs. As the dictionaries now consist of PPGs, that are considered speaker-independent phonetic features, we believe that the resulting activation matrix depends less on the source speaker, thus represents a better transformation function for voice conversion.

In Fig. 3, we compare the two types of features by examining their inter-speaker Pearson's Correlation Coefficients (PCCs): a) spectrum and prosody features; b) Tandem Feature: PPGs, spectral features, and prosodic features. We report the phone-dependent PCCs across all the speaker pairs on Voice Conversion Challenge (VCC) 2016 database. As a pair of speakers share the same linguistic content, the feature of a higher PCC suggests that it is more speaker invariant.

### *B. Joint Spectrum and Prosody Conversion*

Recent studies show that phonetic sparse representation [14], [15], [32] achieves better voice conversion quality by using phonetic transcription of the speech both during training and at run-time. We now extend the idea of Tandem Feature to the phonetic sparse representation framework to perform joint spectrum and prosody conversion. We call the proposed framework as PSR-TF. When the idea of Tandem Feature is extended to group sparse representation, we call the framework as GSR-TF. For simplicity, we only discuss PSR-TF in detail.

We believe that a better way to convert the spectrum and prosody (F0, energy) features is to use phonetic dictionary [32]. Furthermore, phonetic dictionary with Tandem Feature allows for a better estimation of activation matrix that is less dependent on source speaker. We consider that the phonetic sparse representation framework will benefit from better activation matrix with Tandem Feature.

Fig. 4 shows the training and run-time phases of the proposed joint spectrum and prosody conversion framework. During training, we use WORLD to extract spectrum and fundamental frequency. To obtain the energy contour of the speech signal, we first find the energy of each speech frame. With the energy of every speech frame, we obtain the energy contour of the speech signal. We then use CWT to obtain 10-scale decomposition of F0 and energy contour. In addition, we use a DNN-HMM based Automatic Speech

PGR2026-00032
Krisp Technologies EX1023 Page 7

Recognizer (ASR) to find the phone labels, boundaries and PPGs for each utterance. Instead of using a single coupled dictionary, 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. The source phonetic dictionary here consists of both PPGs, source spectrum features as well as CWT-based prosody features, called *source TF-dictionary* while the target phonetic dictionary $\mathbf{B}_i$ only includes spectrum features and prosody features, called *target dictionary*.

At run-time conversion, we use CWT to obtain the 10-scale representations of F0 and energy contour, and the spectrum features 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. For phone $i = k$, a source speech is represented by a matrix, denoted as $\mathbf{X}_k$, that consists of spectrogram, and 10-scale representations of F0 and energy contour, can be written as

$$\mathbf{X}_k \approx \mathbf{A}_k \hat{\mathbf{H}}_k \tag{10}$$

The objective function for estimating the activation matrix $\mathbf{H}_k$ can be formulated as follows.

$$\hat{\mathbf{H}}_k = \underset{\mathbf{H}_k \geq 0}{\operatorname{argmin}}\ d\left([\mathbf{X}_k;\mathbf{P}_k], \mathbf{A}_k\mathbf{H}_k\right) + \lambda||\mathbf{H}_k|| \tag{11}$$

Finally, the activation matrix is applied to the target phonetic dictionary to perform conversion. The converted spectrum and prosody features for phone $k$ can be generated as

$$\hat{\mathbf{Y}}_k = \mathbf{B}_k \hat{\mathbf{H}}_k. \tag{12}$$

We note that the 10-scale decomposition of F0 and energy contour represents 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. We transform scales 3-8 and carry over scales 1, 2, 9 and 10 from source to target speakers because we consider they are more speaker independent [32].

In phonetic sparse representation, we use the phonetic labels of the run-time input speech from either an ASR or the given scripts. Next we discuss three ways to take advantage of such phonetic labels, the phone-dependent residual compensation, context information, and back-off scheme.

### C. Phone-dependent Residual Compensation

It is well-known that residual compensation (RC) reduces spectral distortion arising from reconstruction of speech features [11]. If we have the phonetic label of the speech at run-time, either from ASR or given the scripts, we can introduce phone dependent residual estimation and compensation to phonetic sparse representation for the first time, which is different from [11] [14] where residual compensation is only introduced to spectrum conversion. Here we use dynamic partial least square (DPLS) regression [8] for spectrum and prosody features, that we call *PSR-TF-RC*. During training, we model the residual mapping from source to target speaker for both spectrum and prosody features. At run-time, we use the estimated residual mapping function to compensate the conversion error.

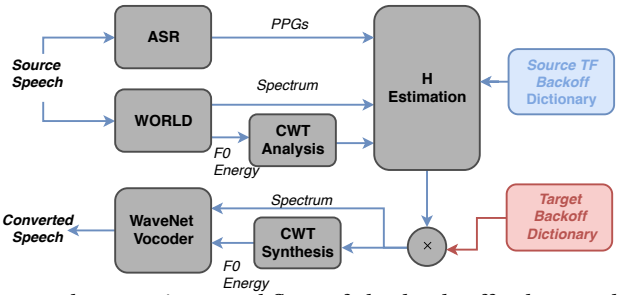



Fig. 5: The run-time workflow of the back-off scheme, that is Sparse Representation with Tandem Feature (SR-TF) for the joint spectrum and prosody conversion.

For phone $i = k$, the modeling error between the observed source features $\mathbf{X}_k$ and the reconstructed source features $\hat{\mathbf{X}}_k$ is called source residual, and the modeling error between the observed target features $\mathbf{Y}_k$ and the converted target features $\hat{\mathbf{Y}}_k$ is called target residual. The process of finding source and target residuals can be formulated as:

$$\mathbf{R}_{S,k} = \log(\mathbf{X}_k) - \log(\hat{\mathbf{X}}_k) \tag{13}$$

$$\mathbf{R}_{T,k} = \log(\mathbf{Y}_k) - \log(\hat{\mathbf{Y}}_k) \tag{14}$$

where $\mathbf{R}_{S,k}$ and $\mathbf{R}_{T,k}$ are source and target residuals, respectively. As given below, we can train a DPLS to estimate a residual mapping function denoted as $F_k$, from each phone of source speaker to target speaker by using the paired source-target residuals.

$$\mathbf{R}_{T,k} = F_k\left(\mathbf{R}_{S,k}\right) \tag{15}$$

### D. 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 [11], [15], [32]. 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 [13], [61], we favor speech units that share similar phonetic context as the intended context by using bi-phones or tri-phone context. In a similar way, if we know the phonetic labels of the input speech at run-time, we can 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 [34], [35] reveal that the phonetic prosody is highly influenced by the transition between phonemes. We consider that bi-phone phonetic dictionary with Tandem Feature captures the prosodic patterns arising from specific phonetic transitions.

PGR2026-00032
Krisp Technologies EX1023 Page 8

### E. Back-off Scheme

As we have limited training data, it is not guaranteed that there are always enough exemplars for every phone in the phonetic dictionary that is needed at run-time. In such cases, we propose a back-off scheme that uses exemplars from all phones.

During training phase, we construct a back-off dictionary that includes all exemplars from source and target speakers. We incorporate PPGs, spectrum features and CWT-based prosodic features to form Tandem Feature in the source dictionary, while we only include spectrum and prosody features in the target dictionary. In practice, such a back-off dictionary is the union of all phonetic sub-dictionaries. The run-time conversion phase of the proposed backoff scheme is given in Fig. 5.

The proposed back-off scheme is an extension to the traditional sparse representation (SR) framework [31] by incorporating PPG features. Therefore, we call it *SR-TF* for spectrum and 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 proposed back-off dictionary, SR-TF can also be used alone as a system for voice conversion.

## VI. WaveNet as Vocoder

Many well-known voice conversion frameworks [1]–[4], [62]–[65] including sparse representation [9], [14], [15], 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 [66], the assumption of stationary process in the short-time window, a time-invariant linear filter. Therefore, the speech that is generated from the traditional vocoders, lacks naturalness in general. This problem is even more serious in voice conversion as the feature conversion changes both F0 and the formant structure of speech among others. We believe that a good vocoder can help reconstruct the speech by harmonizing various changes.

### A. WaveNet

WaveNet [19] is a deep neural network that can generate high quality raw audio waveforms. The joint probability of a waveform $x = x_1, x_2, ..., x_N$ can be factorized as a product of conditional probabilities.

$$p(x) = \prod_{n=1}^{N} p(x_n | x_1, x_2, ..., x_{n-1}) \tag{16}$$

WaveNet approximates the conditional probability by canceling the effect of past samples. The structure of WaveNet consists of many residual blocks. Each of these residual blocks consist of $2 \times 1$ dilated causal convolutions, a gated activation function and $1 \times 1$ convolutions. Using additional

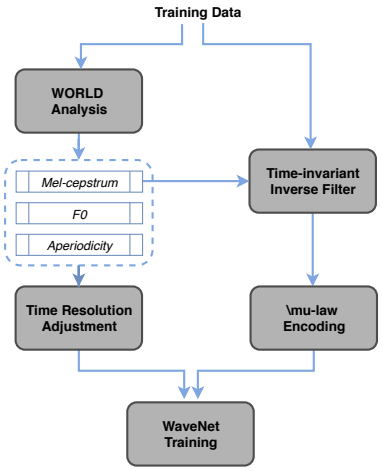



Fig. 6: The training phase of speaker independent WaveNet vocoder [21].

auxiliary features $h$, WaveNet can also model conditional distribution $p(x|h)$. Eq. (16) can then be written as follows:

$$p(x|h) = \prod_{t=1}^{T} p(x_t | x_1, x_2, ..., x_{t-1}, h) \tag{17}$$

We note that we can control the characteristics of generated samples by conditioning the network with auxiliary features. In the original WaveNet [19], linguistic features and/or speaker codes are conditioned to generate speech samples according to the given text input while keeping specific speaker characteristics.

### B. WaveNet Vocoder

We note that WaveNet vocoder [21] achieves remarkable sound quality improvement over the traditional vocoders. Not only can it learn the relationship between input features and output waveforms, but also it learns the interaction among the input features. Moreover, it has been sucessfully used for speech synthesis [67] and in GMM-based voice conversion [22] to improve the naturalness of the synthetic voice. Recently, a speaker independent WaveNet vocoder [21] 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 leverage from human speech of a larger speaker population, we apply the recently proposed speaker independent WaveNet vocoder [21] to generate the speech waveforms. We also propose to adapt the speaker independent WaveNet vocoder towards the target speaker to make full use of the training samples from the target speaker.

### C. Adaptation of WaveNet Vocoder

In this paper, we study the case where we only have a limited number of speech samples from the target speaker, which is a challenging task in voice conversion. While WaveNet vocoder is a deep neural network that typically

PGR2026-00032
Krisp Technologies EX1023 Page 9

| Spectrum Conversion | Features | # Frames | Phonetic Dict. | Script: Training | Script: Testing | MCD-35 | MCD-25 |
|---|---|---|---|---|---|---|---|
| PSR-TF | Tandem Feature | 1 | M | yes | yes | 5.09 | 5.21 |
| | Tandem Feature | 3 | M+B | yes | yes | 4.98 | 5.15 |
| | Tandem Feature | 3 | M+B | yes | no | 5.06 | 5.19 |
| | Tandem Feature | 3 | M+B | no | no | 5.12 | 5.25 |
| PSR [15] | spectrum+prosody | 1 | M | yes | yes | 5.18 | 5.31 |
| | spectrum+prosody | 3 | M+B | yes | yes | 5.09 | 5.23 |
| | spectrum+prosody | 3 | M+B | yes | no | 5.16 | 5.29 |
| | spectrum+prosody | 3 | M+B | no | no | 5.22 | 5.34 |
| GSR-TF | Tandem Feature | 3 | M | yes | no | 5.31 | 5.43 |
| SR-TF | Tandem Feature | 3 | - | - | - | 5.58 | 5.72 |
| SR [11] | spectrum+prosody | 3 | - | - | - | 5.78 | 5.92 |

TABLE II: A comparison of spectral distortion (MCD) [68] between the baseline exemplar-based sparse representation [9] and a variant of group sparse representation. We compare the use of monophone (M) and the combination of monophone and biphone (M+B) in the phonetic dictionary that are derived from 25 and 35 parallel utterances for each speaker pair.

requires a large amount of training data, we believe that it can benefit from the limited speech sample through adaptation process. We first train the speaker independent WaveNet vocoder with CMU Arctic and VCC 2016 datasets, in a similar way that is described in [21] and given in Fig. 6. During training, we exclude the speech data from any of the source or target speakers to make sure we have a speaker-independent WaveNet Vocoder to start with. Then, we propose to adapt the WaveNet vocoder for a target speaker by using the same target utterances that are used in the training of sparse representation. The adaptation brings the speaker independent WaveNet vocoder close to the target speaker.

At run-time conversion, we use the proposed adapted WaveNet vocoder to generate the target speech waveforms. It is not straightforward to report the performance of the WaveNet vocoder using objective metrics such as Mel-cepstral distortion [68]. We report the listening test of the WaveNet vocoder under various conditions in Section VII.

## VII. Experiments

We report the voice conversion experiments on VCC 2016 database [69]. The VCC 2016 dataset is recorded from 5 female and 5 male US English speakers. The spectrum compression factor is set to 0.4 in all experiments [11]. In addition, we use a DNN-HMM based ASR [70] to obtain phone labels, phone boundaries and PPGs. The ASR is reported with 18.0% WER on WSJ Eval92 database. While we conduct joint spectrum and prosody (F0 and energy) conversion, we report objective evaluation on spectrum and prosody conversion separately. We conduct experiments with 25 and 35 pairs of parallel training utterances.

Mel-cepstral distortion (MCD) [68] between the converted and reference Mel-cepstra was employed as an objective evaluation measure for spectrum conversion.

We use Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE) as the evaluation metrics for prosody conversion. The PCC measures the linear dependence of two signals, that can be defined as:

$$p(S,T) = \frac{cov(S,T)}{\sigma_S \sigma_T} \tag{18}$$

where $\sigma_S$ and $\sigma_T$ are the standard deviations of signals $S$ and $T$, respectively. We report the PCCs for F0 and energy contour between the converted and the reference target speech.

### A. Evaluation of WaveNet Vocoder

We use MCD to assess the spectrum conversion, and PCC to assess the prosody conversion. Both are calculated before the vocoding. As a result, it is not straightforward to compare the vocoding effect between WaveNet vocoder and traditional vocoders using objective measures. Therefore, we only evaluate the performance of WaveNet through listening test in the subjective evaluation.

### B. Reference Methods and Setups

To validate our proposals, we implement the state-of-the-art methods for joint spectrum and prosody conversion. In addition, we also implement a family of sparse representation schemes. The setups are summarized as follows:

- Reference baseline: linear F0 conversion.
- SR: Exemplar-based sparse representation [31].
- PSR: Phonetic sparse representation [15], [32].
- GSR: Group sparse representation, as described in Section IV *(proposed)*.
- PSR-TF: Phonetic sparse representation with Tandem Feature, as described in Section V *(proposed)*.
- PSR-TF with Speaker Independent WaveNet Vocoder *(proposed)*.
- PSR-TF with Speaker Adapted WaveNet Vocoder *(proposed)*.
- PSR-TF with Residual Compensation, denoted as *PSR-TF-RC*. *(proposed)*.

We conduct both objective and subjective evaluations to assess their performance in a comparative study.

The GMM-based approaches [3], [6], such as ML-GMM, represent a group of successful voice conversion frameworks. The dynamic kernel partial least squares (DKPLS) regression method has been shown to be effective for implementing a nonlinear conversion function [8]. The studies [11] have shown that the sparse representation approach in general outperforms ML-GMM [3] and DKPLS [8]. The recent work on phonetic sparse representation [14], [15] shows that it further improves the sparse representation framework by incorporating phonetic information. As the phonetic sparse representation (PSR) provides a competitive performance among the state-of-the-art techniques for the limited training data problem, we use PSR as the baseline in the experiments.

PGR2026-00032
Krisp Technologies EX1023 Page 10

### C. Objective Evaluation

*1) Spectrum Conversion Performance:* Table II reports the MCD values for a number of settings in a comparative study. Firstly, we observed that all PSR-TF settings consistently outperform those of other sparse representation frameworks both with 25 and 35 training utterances. Secondly, when scripts are available for training and/or test utterances, we perform forced alignment to obtain better phone segmentation, therefore, we get lower MCD values. Thirdly, we observed that multiple-frame exemplars is apparently helpful to avoid sharp changes across frames. Furthermore, we observed that using monophones together with biphones enhances the conversion performance. Lastly, by comparing PSR with PSR-TF, we show the effect of Tandem Feature in achieving a better conversion.

*2) Prosody Conversion Performance:* Table V reports the PCC and RMSE values of F0 and energy contour conversion for a number of settings in a comparative study. To start with, we observe that the proposed prosody conversion, based on phonetic sparse representation with Tandem Feature (PSR-TF), outperforms the PSR [32] that doesn't use Tandem Feature. The proposed PSR-TF uses the frame-level phonetic information. As a result, the activation matrix of PSR-TF depends less on the source speaker than that of PSR approach, hence yields a better prosody conversion. We also note that group sparse representation (GSR) outperforms the linear conversion [71] and the exemplar-based sparse representation (SR) [31] that doesn't use phonetic dictionary.

*3) The Effect of Tandem Feature:* Table II and V confirm the effectiveness of the novel idea to incorporate PPGs as phonetic features. The phonetic features allow for an improved activation matrix in Eq. (2), Eq. (7), and Eq. (11), that depends less on the source speaker. If we consider PSR [15], [32] performs well because of the use of phonetic information through a phonetic dictionary, the proposed PSR-TF benefits from both segmental level (phonetic dictionary) and frame level (PPGs) phonetic information. We observe that the use of phonetic information in both segmental level and frame level is rewarding for spectrum as well as prosody conversion.

*4) The Effect of Residual Compensation:* We report a phone dependent residual estimation and compensation by using dynamic partial least square (DPLS) regression for spectrum and prosody features under the PSR framework. Table III reports the results in a comperative study for spectrum conversion. We observed that all PSR-TF-RC settings outperform the PSR-TF settings by yielding lower MCD values. Moreover, we have conducted some experiments to see the effect of residual compensation in prosody conversion, as reported in Table IV. As expected, with residual compensation, we achieve higher correlation, hence a better prosody conversion.

### D. Subjective Evaluation

We further conduct listening experiments to assess the performance of Tandem Feature for joint spectrum and prosody conversion, and the effect of WaveNet vocoder in terms of voice quality and speaker similarity. In all listening experiments, we use the scripts of the training utterances to perform forced alignment. It should be noted that at run-time conversion, we do not perform forced alignment. 15 subjects participated in all the listening tests, each of them listens to 30 converted utterances. The p-values are calculated in a similar way that is reported in [72].

| Spectrum Conversion | Script: Training | Script: Testing | MCD |
|---|---|---|---|
| PSR-TF | yes | yes | 4.98 |
| | yes | no | 5.06 |
| | no | no | 5.12 |
| PSR-TF-RC | yes | yes | 4.85 |
| | yes | no | 4.90 |
| | no | no | 4.99 |

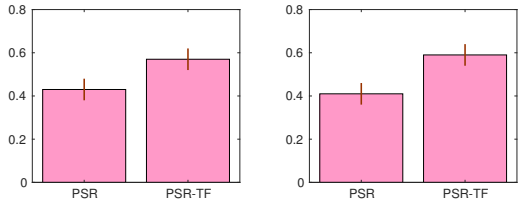
TABLE III: The effect of residual compensation for spectrum conversion. 35 parallel utterances are used during training. Monophones together with biphones are used in phonetic dictionary.

| F0 Conversion | Script: Training | Script: Testing | PCC |
|---|---|---|---|
| PSR-TF | yes | yes | 0.908 |
| | yes | no | 0.899 |
| | no | no | 0.891 |
| PSR-TF-RC | yes | yes | 0.914 |
| | yes | no | 0.906 |
| | no | no | 0.901 |

TABLE IV: The effect of residual compensation in prosody conversion. 35 parallel utterances are used during training. Monophones together with biphones are used in phonetic dictionary.

*1) Tandem Feature:* We first conduct listening experiments, as reported in Fig. 7.a and 7.b, to examine the effect of Tandem Feature in terms of voice quality and speaker similarity for both spectrum and prosody conversion. We note that the original PSR framework [15] does not use the WaveNet vocoder to generate speech waveform. To assess the effect of Tandem Feature, without changing the vocoding process, we use WORLD 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, that validates the proposal of the use of Tandem Feature in sparse representation voice conversion.

*2) Phone-dependent Residual Compensation:* We also conduct listening experiments, as reported in Fig. 8.a and 8.b, to examine the effect of phone-dependent residual



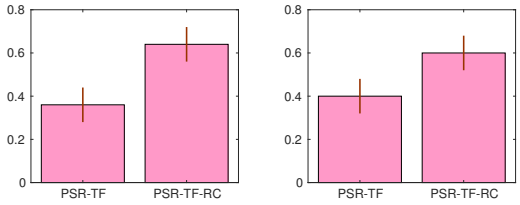

(a) Voice quality (b) Speaker similarity

Fig. 7: The preference percentage tests with 95 % confidence interval between PSR and PSR-TF for joint spectrum and prosody conversion. The p-values [72] are (a) $7.0382e^{-21}$ and (b) $2.0564e^{-21}$ respectively.

PGR2026-00032
Krisp Technologies EX1023 Page 11

| Prosody Conversion | Features | Number of Frames | Phonetic Dict. | Script: Training | Script: Testing | PCC: F0 | PCC: E |
|---|---|---|---|---|---|---|---|
| PSR-TF | Tandem Feature | 1 | M | yes | yes | 0.861 | 0.831 |
| | Tandem Feature | 3 | M+B | yes | yes | 0.901 | 0.850 |
| | Tandem Feature | 3 | M+B | yes | no | 0.892 | 0.843 |
| | Tandem Feature | 3 | M+B | no | no | 0.881 | 0.837 |
| PSR [15] | spectrum+prosody | 1 | M | yes | yes | 0.847 | 0.819 |
| | spectrum+prosody | 3 | M+B | yes | yes | 0.893 | 0.836 |
| | spectrum+prosody | 3 | M+B | yes | no | 0.879 | 0.831 |
| | spectrum+prosody | 3 | M+B | no | no | 0.865 | 0.825 |
| GSR-TF | Tandem Feature | 3 | M | yes | no | 0.839 | 0.805 |
| | Tandem Feature | 3 | M | no | no | 0.830 | 0.793 |
| SR-TF | spectrum+prosody | 3 | - | - | - | 0.827 | 0.779 |
| SR [11] | spectrum+prosody | 3 | - | - | - | 0.810 | 0.765 |
| Linear Conversion | - | 3 | - | - | - | 0.780 | N. A. |

TABLE V: A comparison of PCC values of converted F0 and energy, denoted as E, across several sparse representation implementations. We also report linear F0 conversion result as the F0 conversion baseline. PSR-TF and GSR-TF are proposed in this paper, while linear conversion, PSR [15] and SR [11] were studied previously.




(a) Voice quality (b) Speaker similarity

Fig. 8: The preference percentage tests with 95 % confidence interval between PSR-TF and PSR-TF-RC for joint spectrum and prosody conversion. The p-values [72] are (a) $4.2350e^{-23}$ and (b) $2.0564e^{-21}$ respectively.

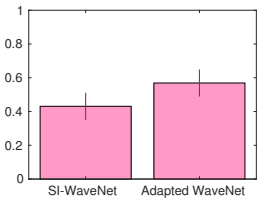



Fig. 9: The preference test between the speaker independent WaveNet vocoder [21] and the adapted WaveNet vocoder in terms of voice quality. The p-value [72] is $2.3393e^{-20}$.

compensation. We observe that PSR-TF with residual compensation outperforms PSR-TF consistently in both voice quality and speaker similarity.

*3) Speaker Independent and Adapted WaveNet Vocoder:* We futher conduct a listening experiment, that is reported in Fig. 9, to study the listener preference of WaveNet vocoders. We train the speaker independent WaveNet vocoder with CMU Arctic and VCC 2016 datasets, in a similar way that is described in [21]. It is important to note that the speech data from any of the source or target speakers in the experiment are excluded during training. To perform WaveNet adaptation, we use the same 35 target utterances that are used in the training of sparse representation. At run-time conversion, we use the speaker independent WaveNet [21] and the adapted WaveNet vocoder to generate the target speech waveforms. We observe that Adapted WaveNet outperforms the speaker independent WaveNet in terms of voice quality, that validates our proposed idea. Motivated by the encouraging results of Tandem Feature and the adapted WaveNet for joint spectrum and prosody conversion, we now move on to the last listening experiment, that is reported in Fig. 10, to assess the performance of the proposed PSR-TF-RC with adapted WaveNet vocoder. We perform adaptation by using the same 35 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. Fig. 10 shows that the voice conversion framework PSR-TF-RC with adapted WaveNet vocoder outperforms others that include SR, GSR, PSR, PSR-TF and PSR-TF-RC.

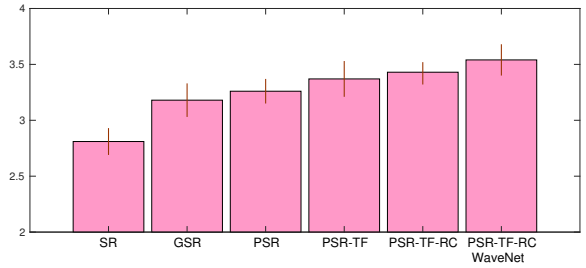



Fig. 10: A comparison of MOS for joint spectrum and prosody conversion over the sparse representation schemes discussed in this paper, that include SR [11], PSR [15], GSR, PSR-TF, PSR-TF-RC and PSR-TF-RC with adapted WaveNet vocoder (PSR-TF-RC-WaveNet).

## VIII. Conclusion

We have proposed a novel voice conversion framework that includes spectrum, F0 and energy contour transformation. We provide a theoretical formulation for phonetic sparse representation by proposing group sparse representation approach for voice conversion. Moreover, we study the speaker dependent and independent characteristics (phonetic activation matrix and differentiated multi-scale prosody transfer) to motivate the design of sparse representation framework. We show two effective ways to improve the conversion performance under limited training data. Firstly, we propose the Tandem Feature sparse representation strategy for spectrum and prosody features. We implement the TF-dictionary in phonetic sparse representation frameworks as a solution to improve the con-

PGR2026-00032
Krisp Technologies EX1023 Page 12

version performance under limited training data. Secondly, we propose the use of speaker adapted WaveNet vocoder together with a conversion framework to achieve higher voice quality. 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.

## 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)*, vol. 2, pp. 655–658, 1988.

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

[3] 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.

[4] H. Zen, Y. Nankaku, and K. Tokuda, "Probabilistic feature mapping based on trajectory HMMs," *In INTERSPEECH*, pp. 1068–1071, 2008.

[5] T. Toda, Y. Ohtani, and K. Shikano, "ONE-TO-MANY AND MANY-TO-ONE VOICE CONVERSION BASED ON EIGENVOICES," *ICASSP*, 2007.

[6] K. Kobayashi, S. Takamichi, S. Nakamura, and T. Toda, "The NU-NAIST voice conversion system for the Voice Conversion Challenge 2016," *in INTERSPEECH*, 2016.

[7] E. Helander, T. Virtanen, J. Nurminen, and M. Gabbouj, "Voice conversion using partial least squares regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[8] 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.

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

[10] Y. Luan, D. Saito, Y. Kashiwagi, N. Minematsu, and K. Hirose, "Semi-supervised noise dictionary adaptation for exemplar-based noise robust speech recognition ," *In ICASSP*, pp. 1764–1767, 2014.

[11] Z. Wu, T. Virtanen, E. S. Chng, and H. 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.

[12] R. Aihara, K. Masaka, T. Takiguchi, and Y. Ariki, "Parallel Dictionary Learning for Multimodal Voice Conversion Using Matrix Factorization," *INTERSPEECH*, pp. 27–40, 2016. [Online]. Available: http://link.springer.com/10.1007/978-3-319-40171-3_3

[13] Z. Jin, A. Finkelstein, S. Di Verdi, J. Lu, and G. J. Mysore, "Cute: a concatenative method for voice conversion using exemplar- based unit selection," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2–6, 2016.

[14] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, "voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," *ICASSP*, pp. 7894–7898, 2014.

[15] B. Sisman, H. Li, and K. C. Tan, "Sparse representation of phonetic features for voice conversion with and without parallel data," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). accepted for publication*, 2017.

[16] T. J. Hazen, W. Shen, and C. White, "Query-By-Example Spoken Term Detection Using Phonetic Posteriorgram Templates," *In IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 421–426, 2009.

[17] K. Kintzley, A. Jansen, and H. 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.

[18] L. Sun, H. Wang, S. Kang, K. Li, and H. Meng, "Personalized, cross-lingual TTS using phonetic posteriorgrams," *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 322–326, 2016.

[19] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[20] A. Tamamori, T. Hayashi, and K. Kobayashi, "Speaker-dependent wavenet vocoder," *INTERSPEECH*, 2017.

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

[22] K. Kobayashi, T. Hayashi, A. Tamamori, and T. Toda, "Statistical voice conversion with wavenet-based waveform generation," *INTERSPEECH*, 2017.

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

[24] S.-S. Wang, A. Chern, Y. Tsao, J.-W. Hung, X. Lu, Y.-H. Lai, and B. Su, "Wavelet speech enhancement based on nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 23, 2016.

[25] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[26] K. A. Akarsh, "Speech Enhancement using Non negative Matrix Factorization and Enhanced NMF," *International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 2015.

[27] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.

[28] M. Sun, Y. Li, J. F. Gemmeke, and X. Zhang, "Speech Enhancement Under Low SNR Conditions Via Noise Estimation Using Sparse and Low-Rank NMF with Kullback – Leibler Divergence," vol. 23, no. 7, pp. 1233–1242, 2015.

[29] Z. Wu, T. Virtanen, T. Kinnunen, E. S. Chng, and H. Li, "Exemplar-Based Voice Conversion Using Non-Negative Spectrogram Deconvolution," *8th ISCA Speech Synthesis Workshop*, 2013.

[30] Y. C. Wu, H. T. Hwang, C. C. Hsu, Y. Tsao, and H. M. Wang, "Locally linear embedding for exemplar-based spectral conversion," *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-September-2016, no. 1, pp. 1652–1656, 2016.

[31] H. Ming, D. Huang, L. Xie, S. Zhang, M. Dong, and H. Li, "Exemplar-based sparse representation of timbre and prosody for voice conversion," *In IEEE ICASSP*, no. 61175018, pp. 5175–5179, 2016.

[32] B. Sisman, H. Li, and K. C. Tan, "Transformation of Prosody in voice conversion," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017.

[33] H. Chin-Cheng, H. Hsin-Te, W. Yi-Chiao, T. Yu, and W. Hsin-Min, "Dictionary update for nmf-based voice conversion using an encoder-decoder network," *10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, vol. 22, no. 3, pp. 293–297, 2016.

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

[35] J. J. Ohala, A. Dunn, and R. Sprouse, "Prosody and Phonology," *ISCA Speech Prosody*, 2004.

[36] G. Sanchez, H. Silen, J. Nurminen, and M. Gabbouj, "Hierarchical modeling of F0 contours for voice conversion," *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 2318–2321, 2014.

[37] "Speech prosody: A Methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2015.

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

[39] M. S. Ribeiro and R. A. J. Clark, "A multi-level representation of f0 using the continuous wavelet transform and the Discrete Cosine Transform," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[40] A. Wennerstrom, "the music of everyday speech prosody and discourse analysis," *Oxford University Press*, pp. 153–158, 2001.

[41] D. R. Ladd, "Intonational phonology," *Cambridge University Press*, pp. 153–158, 2008.

[42] B. K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech Synthesis Based on Hidden Markov Models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[43] M. Vainio, A. Suni, and D. Aalto, "Continuous wavelet transform for analysis of speech prosody," *In TRASP*, pp. 78–81, 2013.


PGR2026-00032
Krisp Technologies EX1023 Page 13

[44] A. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio, "Wavelets for intonation modeling in HMM speech synthesis," *In 8th ISCA Speech Synthesis Workshop*, no. 1, pp. 285–290, 2013.
[45] H. Ming, D. Huang, M. Dong, H. Li, L. Xei, and S. Zhang, "Fundamental Frequency Modeling Using Wavelets for Emotional Voice Conversion," *In ACII*, pp. 804–809, 2015.
[46] H. Ming, D. Huang, L. Xie, J. Wu, M. Dong, and H. Li, "Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion," *INTERSPEECH*, vol. 08-12-September-2016, pp. 2453–2457, 2016.
[47] Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, "Text-Independent F0 Transformation with Non-Parallel Data for Voice Conversion," *In INTERSPEECH*, pp. 1732–1735, 2010.
[48] B. Gillett, S. King, and U. Kingdom, "Transforming F0 Contours," *In INTERSPEECH*, no. 2, pp. 101–104, 2003.
[49] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using neural networks with arbitrary scales F0 based on wavelet transform," *EURASIP Journal on Audio, Speech, and Music Processing*, 2017.
[50] ——, "Emotional Voice Conversion with Adaptive Scales F0 based on Wavelet Transform using Limited Amount of Emotional Data," *INTERSPEECH*, pp. 3399–3403, 2017.
[51] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *arXiv*, 2018.
[52] ——, "StarGAN-VC: Non-parallel many-to-many voice conversion with star generative adversarial networks," *arXiv*, 2018.
[53] B. Sisman, G. Lee, and H. Li, "Phonetically aware exemplar-based prosody transformation," *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018.
[54] B. Sisman and H. Li, "Wavelet analysis of speaker dependent and independent prosody for voice conversion," *INTERSPEECH*, 2018.
[55] A. Hurmalainen, R. Saeidi, and T. Virtanen, "Group Sparsity for Speaker Identity Discrimination in Factorisation-based Speech Recognition," *In INTERSPEECH*, 2012.
[56] R. Serizel, S. Essid, and G. Richard, "Group non-negative matrix factorisation with speaker and session variability compensation for speaker identification," *In ICASSP*.
[57] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, 2016.
[58] J. Kim and H. Park, "Group Sparsity in Nonnegative Matrix Factorization ," *In proceedings of the SIAM International Conference on Data Mining*, 2012.
[59] R. Aihara, T. Takiguchi, and A. Yasuo, "Activity-mapping non-negative matrix factorization for exemplar-based voice conversion," *ICASSP*, 2015.
[60] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *In the Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, 2016.
[61] P. Taylor, "Text-to-Speech Synthesis," *Cambridge University Press*, 2009.
[62] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv*, 2017.
[63] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data," *arXiv*, 2017.
[64] ——, "Learning Latent Representations for Speech Generation and Transformation," *arXiv*, 2017.
[65] 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," *arXiv*, 2017.
[66] S. Furui, "Digital speech processing, synthesis, and recognition(revised and expanded)," *Digital Speech Processing, Synthesis, and Recognition*, 2000.
[67] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," *arXiv:1712.05884*, 2018.
[68] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," *Communications, Computers and Signal Processing*, pp. 125–128, 1993.
[69] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The Voice Conversion Challenge 2016," *In INTERSPEECH*, pp. 1632–1636, 2016.
[70] D. Povey, A. Ghoshal, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, J. Silovsk, and P. Motl, "The Kaldi Speech Recognition Toolkit," *In IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.
[71] G. K. Anumanchipalli, L. C. Oliveira, and A. W. Black, "A style capturing approach to F0 transformation in voice conversion," *ICASSP*, 2013.
[72] S. Lifa, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *2016 IEEE ICME*, 2016.

**Berrak Sisman** She pursues Ph.D. at the National University of Singapore. She received Singapore International Graduate Award (SINGA) for her Ph.D studies. She obtained 2 Bachelor degrees, that are electrical and electronic engineering, and computer engineering from Isik University, Istanbul, Turkey, in 2015. She obtained Master of Science degree in electronics engineering from Isik University Graduate School of Science and Engineering, in 2016. She was awarded APSIPA PhD Forum Best Presentation Award for her presentation titled "Limited Data Voice Conversion from Sparse Representation to GANs and WaveNet" in 2018, Hawaii, United States. Her current research interests include speech synthesis and voice conversion. She is the general coordinator of the Student Advisory Committee (SAC) of International Speech Communication (ISCA).

**Mingyang Zhang** received the B.Eng. degree in Electrical Information Engineering from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2014. He is currently working toward the Ph.D. degree in Key Laboratory of Underwater Acoustic Signal Processing of Ministry of Education, Southeast University, Nanjing, China. He has been with the Department of Electrical & Computer Engineering at the National University of Singapore. His current research interests include speech synthesis and voice conversion.

**Haizhou Li** Haizhou Li (M'91-SM'01-F'14) received the B.Sc., M.Sc., and Ph.D degree in electrical and electronic engineering from South China University of Technology, Guangzhou, China in 1984, 1987, and 1990 respectively. Dr Li is currently a Professor at the Department of Electrical and Computer Engineering, National University of Singapore (NUS). He is also a Conjoint Professor at the University of New South Wales, Australia. His research interests include automatic speech recognition, speaker and language recognition, and natural language processing. Prior to joining NUS, he taught in the University of Hong Kong (1988-1990) and South China University of Technology (1990-1994). He was a Visiting Professor at CRIN in France (1994-1995), Research Manager at the Apple-ISS Research Centre (1996-1998), Research Director in Lernout & Hauspie Asia Pacific (1999-2001), Vice President in InfoTalk Corp. Ltd. (2001-2003), and the Principal Scientist and Department Head of Human Language Technology in the Institute for Infocomm Research, Singapore (2003-2016). Dr Li is currently the Editor-in-Chief of IEEE/ACM Transactions on Audio, Speech and Language Processing (2015-2018), a Member of the Editorial Board of Computer Speech and Language (2012-2018). He was an elected Member of IEEE Speech and Language Processing Technical Committee (2013-2015), the President of the International Speech Communication Association (2015-2017), the President of Asia Pacific Signal and Information Processing Association (2015-2016), and the President of Asian Federation of Natural Language Processing (2017-2018). He was the General Chair of ACL 2012 and INTERSPEECH 2014. Dr Li is a Fellow of the IEEE and the ISCA. He was a recipient of the National Infocomm Award 2002 and the President's Technology Award 2013 in Singapore. He was named one of the two Nokia Visiting Professors in 2009 by the Nokia Foundation, and Bremen Excellence Chair Professor in 2019.

View publication stats

PGR2026-00032
Krisp Technologies EX1023 Page 14