Transformation of Prosody in Voice Conversion

Berrak Şişman *, Haizhou Li *, Kay Chen Tan ‡
* National University of Singapore, Singapore
‡ City University of Hong Kong, Hong Kong SAR, China
berraksisman@u.nus.edu, haizhou.li@nus.edu.sg, kaytan@cityu.edu.hk

***Abstract*—Voice Conversion (VC) aims to convert one's voice to sound like that of another. So far, most of the voice conversion frameworks mainly focus only on the conversion of spectrum. We note that speaker identity is also characterized by the prosody features such as fundamental frequency (F0), energy contour and duration. Motivated by this, we propose a framework that can perform F0, energy contour and duration conversion. In the traditional exemplar-based sparse representation approach to voice conversion, a general source-target dictionary of exemplars is constructed to establish the correspondence between source and target speakers. In this work, we propose a Phonetically Aware Sparse Representation of fundamental frequency and energy contour by using Continuous Wavelet Transform (CWT). Our idea is motivated by the facts that CWT decompositions of F0 and energy contours describe prosody patterns in different temporal scales and allow for effective prosody manipulation in speech synthesis. Furthermore, phonetically aware exemplars lead to better estimation of activation matrix, therefore, possibly better conversion of prosody. We also propose a phonetically aware duration conversion framework which takes into account both phone-level and sentence-level speaking rates. We report that the proposed prosody conversion outperforms the traditional prosody conversion techniques in both objective and subjective evaluations.**

I. Introduction

Human voice carries unique speaker identity. Voice conversion refers to a process of modifying the characteristics of one speaker such as spectrum or/and prosody, to sound as if it was spoken by another speaker. Over the last few decades, there has been immense research in voice conversion technology with the applications such as personalized speech synthesis, disguising one's voice, and dubbing of movies, etc.

Speaker characteristics are carried by multiple speech features including spectrum, F0, energy and duration. Besides spectrum, transformation of other features is also reqiured to build a comprehensive voice conversion framework. On the other hand, most of well-known voice conversion frameworks focus only on spectral conversion. Vector quantization (VQ) [1] and fuzzy vector quantization [2] were used to establish the mapping between source and target speakers in early studies. The voice conversion frameworks such as Gaussian mixture model [3] [4], partial least square regression [5] and dynamic kernel partial least squares regression (DKPLS) [6] are the statistical parametric approaches, which marked a success of converting speaker identity. Even though the statistical parametric approaches convert speaker identity better than the frequency warping approaches including bilinear frequency warping [7] and correlation-based frequency warping [8], the resulting speech quality remains to be improved, especially when we have highly limited training data. To alleviate this problem, nonnegative matrix factorization (NMF) [9] based voice conversion frameworks such as exemplar-based voice conversion with non-negative spectrogram deconvolution [10], locally linear embedding (LLE) for exemplar-based voice conversion [11], and an exemplar-based unit selection framework called Cute [12] have been proposed. To address the over-smoothing problem arising from linear combination of exemplars, exemplar-based sparse representation technique [13], discriminative graph-embedded NMF approach [14] and multiple dictionaries in an NMF-based framework [15] suggest to constrain the activation vector to be sparse.

So far, spectral mapping mechanism is central to the study of voice conversion. On the other hand, how to effectively generate the prosody in the target voice remains a challenge. In case the listener is familiar with the speaker and/or his speaking style, source related cues such as fundamental frequency, energy contour and duration, play a crucial role in transmitting the speaker identity [16].

It is generally agreed that prosody is inherently suprasegmental and hierarchical in nature [17] [18] and it can be affected by both short term as well as long term dependencies [19]. Previous studies of prosody conversion mainly focus on fundamental frequency (F0) which is an essential prosodic factor in speech. F0 modeling and generation continue to be a a challenging in voice conversion, due to the fact that the amount of training data is limited with tens to a few hundreds of utterances. One of the widely-used technique for F0 conversion is to transform the mean and variance of source speaker's F0 to that of the target [13]. This method is based on a frame-level operation. However, human manipulates F0 in a segmental manner at phone, syllable, word, phrase or sentence level. There have been some extensions of this widely-used technique such as GMM-based mapping [20] and a piecewise linear mapping for transformation of F0 contour by using a small linguistically motivated parameter set [21].

The continuous wavelet transform (CWT) models F0 in different temporal scales that has been used to characterize F0 within an hidden Markov model (HMM) framework [22]

[23]. Moreover, voice conversion frameworks such as DKPLS [19] and exemplar-based prosody conversion [24] [25] show that CWT decomposition of F0 contour works well in voice conversion. A recently proposed emotional voice conversion framework [26] also motivates the use of CWT decomposition for F0 and energy contour in voice conversion. In a study [15] on exemplar-based sparse representation for voice conversion, the strategy of multiple dictionaries is shown to be more effective than single dictionary. In this paper, we further the idea of exemplar-based sparse representation by incorporating phonetic information. We believe that phonetically aware prosody dictionaries allow us to provide a better estimation of the activation matrix, therefore, yielding a better conversion of prosody.

Duration transformation has not been considered in many of the well-known voice conversion frameworks. One of the widely used technique is to keep the duration of target speaker same as that of the source speaker. On the other hand, it is important to mention that speakers tend to speak at different rates and with different rhythms. In such cases, transformation of duration becomes crucial in voice conversion. There have been few attemps to incorporate durational modification in voice conversion such as GMM based durational modification [27], Gaussian normalized transformation of phone durations in ANN based voice conversion [28] and prosody conversion using DNN segmental models [29]. Considering that phone duration is regulated both locally at phone-level and globally at sentence-level, we propose a phone-dependent duration conversion framework, which takes into account both phone-level and sentence-level speaking rates between the source and target speakers, that we call phonetically aware duration conversion.

The main contribution of this paper is a novel prosody conversion framework that focuses on F0, energy contour and duration transformation. The proposed framework uses 10-scale representation of F0 and energy contour while constructing the phone-dependent prosody dictionary. We have good reasons to believe that phonetic exemplars allow us to achieve a better estimation of activation matrix, hence a better conversion performance. The 10-scale representation of F0 and energy contour allows us to capture the temporal changes at different levels. The proposed Phonetically Aware Sparse Representation framework differs from the previous studies [13] [15] [25] in the following ways:

- We make use of an automatic speech recognition (ASR) system to extract the phone labels and their boundaries to obtain phone-dependent CWT decompositions of F0 and energy contours.
- We construct phone-dependent dictionaries that include spectral and prosody features.
- To include the segmental information in speech, we use biphone exemplars together with monophone exemplars to construct the dictionary.

Moreover, we propose a phone-dependent duration conversion framework that estimates the converted duration for each phone by interpolating phone-level and sentence-level speaking rates. We obtain the phone-dependent optimal weight by performing convex optimization. Overall, we show that the proposed prosody conversion framework for F0, energy and duration can work in conjunction with a spectrum conversion technique, and outperforms the state-of-the-art techniques both in objective and subjective evaluations.

This paper is organized as follows: In section II, we describe the details of phone-dependent CWT decomposition of F0 and energy conversion. In section III, we propose a framework for F0 and energy conversion. Section IV describes the proposed duration transformation framework. The experimental results and conclusion are given in Section V and VI.

II. Phone-dependent F0 and Energy Representation

Speaker identity is characterized by various speech features, that include spectrum and prosody. A comprehensive voice conversion should include the transformation of both spectrum and prosody among others. However, prosody conversion is a challenging task for many reasons. For example, prosody can be used contrastively to communicate meaning such as emotions (angry or joyful), lexical stress, or speech acts in a dialogue (statement or question) that we call speaker independent prosody, it also carries personal, dialectal, and other background characteristics that belong to an individual (speaker dependent prosody) [30]. The individual's characteristics are not linguistically significant. However, they are significant in speaker characterization, that is needed in voice conversion. In voice conversion, we would like to carry over the speaker independent prosody, both F0 and energy contours, from the source to the target, but to replace the speaker dependent prosody of source speaker with that of target speaker. Prosody is demonstrated in a hierarchical structure [17] which is affected by short term as well as long term dependencies [18]. Therefore, it is not adequate to use a linear model to represent all variations in different temporal scales.

Recently, CWT has been proposed for the analysis and and modeling of F0 in speech synthesis [22] [23] and voice conversion [19] [25]. It was also shown that CWT decompositions of F0 can be used effectively as the exemplars in the exemplar-based prosody conversion [25]. In addition, a recently proposed emotional voice conversion framework [26] shows that CWT can also be used to model energy contour. We consider that the CWT decomposition of a F0 or energy contour represents the speaker- dependent and speaker-independent prosody in different scales, that provides an useful tool for prosody transformation. We also believe that prosody patterns, either speaker dependent or speaker independent, are phone dependent. Therefore, we propose a phone-dependent prosody conversion framework

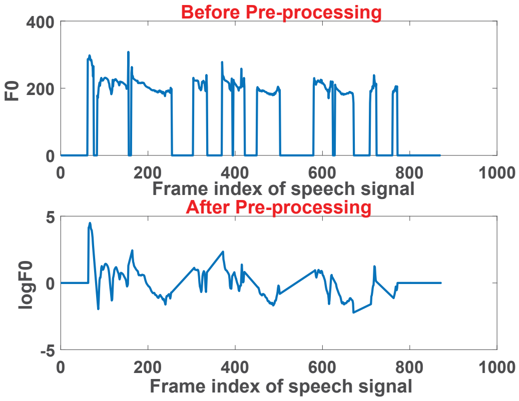


Fig. 1: F0 contours from a female speaker, uttering the sentence ”And even beyond, where the paddocks were, and the berry patches." before and after applying pre-processing steps.

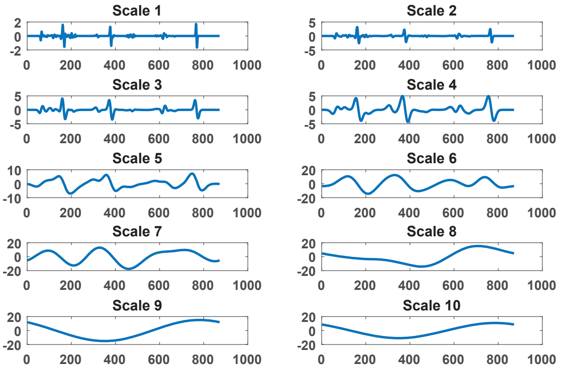


Fig. 2: Wavelet transform of the F0 contour from a female speaker, uttering the sentence ”And even beyond, where the paddocks were, and the berry patches."

that decomposes F0 and energy contours into CWT decompositions in different scales to facilitate the conversion.

To start with, the continuous wavelet transform of an input signal $f_0(t)$ can be written as

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

where ψ is the Maxican hat mother wavelet. If we fix the analysis at 10 discrete scales, f_0 can be represented as [25]

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

where $i=1,...,10$ and $\tau_0=5ms$. These timing scales were originally proposed by [23] in a hierarchic prosody model [22], and then used in some voice conversion frameworks such as [24]–[26].

After performing wavelet analysis, the original signal can be approximated by the following formula:

$$f_0(t)=\sum_{i=1}^{10} W_i(f_0)(t)(i+2.5)^{-5/2}. \qquad (3)$$

In the proposed framework, we use Eq. (1) and (2) to decompose F0 and energy contour into 10 temporal scales. It is important to mention that the wavelet transform is sensitive to the abrupt F0 changes due to unvoiced frames but having little to do with prosody patterns, therefore as shown in Figure 1, some pre-processing steps such as transformation to logarithmic scale, smoothing F0 contour by using 3-point mean filter, filling the gaps produced by unvoiced frames, and normalizing the resulting F0 contour to zero mean and unit variance are needed [19]. Figure 2 is an example to illustrate the corresponding F0 scales of one speech signal.

The stress of the syllables, the stress of the words, and the intonation patterns are important criteria among speakers [19]. By using CWT, we hope to separately represent the speaker dependent and speaker independent prosody patterns in different temporal scales, for example scale 1 represents microprosody events and scale 3 and 4 represent stress of the syllables. By converting the scales 3 to 8 out of the 10 scales, we hope to convert the prosody at syllable, word and sentence levels, that are speaker dependent. We will directly carry over the less speaker dependent scales, i.e. 1, 2, 9, 10 from the source speech to the target speech [19].

In the traditional framework for spectrum and prosody conversion [25], a randomly selected subset of paired joint exemplars is used to construct the coupled joint dictionary. In the proposed framework for prosody conversion, we first obtain phonetic labels and their boundaries by using an ASR system. Then, the coupled joint dictionary, that consist of spectrum, 10-scale representation of F0 and energy contour, are organized according to the phonetic labels. We believe that rather than selecting the exemplars randomly to construct the joint dictionary, phonetically aware CWT representation of F0 and energy may yield a better estimation of activation matrix, hence better prosody conversion.

III. F0 and Energy Conversion

A. Traditional NMF Framework for Spectrum and Prosody Conversion

The traditional exemplar-based sparse representation [13], performs spectrum conversion by describing a magnitude spectrum as a linear combination of exemplars. Recently, this framework was extended to perform both spectrum and prosody conversion [25]. To convert spectral and prosody features simultaneously, a pair of dictionaries, denoted as **A** and **B**, each consists of spectrum, aperiodicity component, energy contour and 5-scale CWT representation of F0 is constructed. **A** and **B** are derived from a parallel database, where exemplars are aligned frame-by-frame.

At run-time, the matrix denoted as **X** that consists of both spectral and prosody features of a source utterance can be

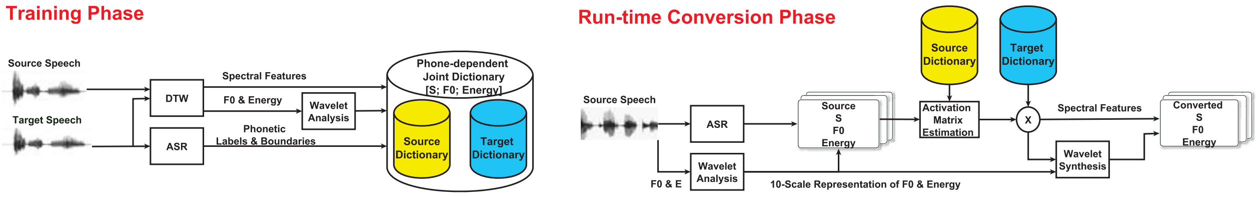


Fig. 3: Training and run-time phases of F0 and energy conversion using Phonetically Aware Sparse Representation.

represented as

$$\mathbf{X} \approx \mathbf{AH} \tag{4}$$

Nonnegative matrix factorization (NMF) technique is employed to estimate the activation matrix H, which is constrained to be sparse. Mathematically, the objection function is written as

$$\mathbf{H} = \underset{\mathbf{H} \geq 0}{\text{argmin}}\ d\left(\mathbf{X}, \mathbf{AH}\right) + \lambda ||\mathbf{H}|| \tag{5}$$

where λ is the sparsity penalty factor. A generalised Kullback-Leibler (KL) divergence is used to estimate activation matrix **H**. It has been showed in the previous studies [13] [14] that as long as the dictionaries **A** and **B** are aligned, the source and target speakers can share the same activation matrix **H**. Therefore, the converted spectral and prosody features can be written as

$$\hat{\mathbf{Y}} = \mathbf{BH}. \tag{6}$$

Both subjective and objective evaluations show the effectiveness of prosody transformation in voice conversion.

B. Phonetically Aware Sparse Representation

In the traditional NMF framework for spectrum or/and prosody conversion, the frame alignment between source and target is used as a collection of exemplars without referring to the phonetic labels of the data [10], [13], [24], [25]. It is important to note that with the scripts of the training data and a general purpose ASR, we easily obtain valuable phonetic information [31], [32], for example phonetic labels and their boundaries.

In this paper, we propose an approach for prosody conversion which uses such phonetic information to construct the coupled joint dictionary, that we will call *phone-dependent joint dictionary.* We propose replacing the acoustic dictionary in the traditional NMF framework [13], [25] with phone-dependent dictionary. Instead of having a single coupled joint dictionary [**A**; **B**], we construct multiple phone-dependent joint dictionaries, denoted as $\mathbf{A}_i$ and $\mathbf{B}_i$ for phone i where $i = 1, ..., n$. We call such phone-dependent joint dictionaries the sub-dictionaries.

We propose to construct the phone-dependent joint dictionary that consists of spectral features, and CWT representation of F0 and energy contour as well. Figure 3 presents the training and run-time conversion phases of the proposed F0 and energy conversion framework.

At training phase, we construct phone-dependent joint dictionary, that includes aligned spectral and prosody features of both source and target speaker. Given the parallel source and target utterances, we use STRAIGHT to extract spectrum and fundamental frequency that are denoted as **S** and **F0**, respectively. To obtain the energy contour of the speech signal, we first find the energy e_m of a speech frame m, where $m = 1, ..., M$ with M being the number of frames in the speech signal. With the energy of every speech frame, we obtain the energy contour of the speech signal $[e_1, ...e_m, ...e_M]$.

Previously, we note that prosody is influenced both at a supra-segmental level and at a segmental level. To represent all temporal scales of F0 and energy contour, we perform wavelet analysis by using Eq. (1) and (2). We first use an ASR to obtain phonetic labels and their boundaries. Then, the phone-dependent joint dictionaries, that consist of spectrum, 10-scale representation of F0 and energy contour, are organized according to the phonetic labels.

At run-time conversion, we estimate the phone-dependent activation matrices. For phone $i = k$, a source speaker 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

$$\mathbf{X}_k \approx \mathbf{A}_k \mathbf{H}_k \tag{7}$$

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

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

A generalised Kullback Leibler divergence [33] is used to estimate the activation matrix $\mathbf{H}_k$. The activation matrix is applied to the target phone-dependent joint dictionary to perform conversion. The converted speech, represented by the matrix $\mathbf{Y}_k$, consists of converted spectral and prosody features, can be written as:

$$\mathbf{Y}_k = \mathbf{B}_k \mathbf{H}_k \tag{9}$$

Once the prosody conversion is achieved, we transform the syllable, word and sentence levels (scale 3-8) and copy the rest of the decompositions in other scales from the source speaker. By converting the scales from 3 to 8, we aim to convert the prosody at syllable, word and sentence levels. To reconstruct the converted F0 and energy contour, the

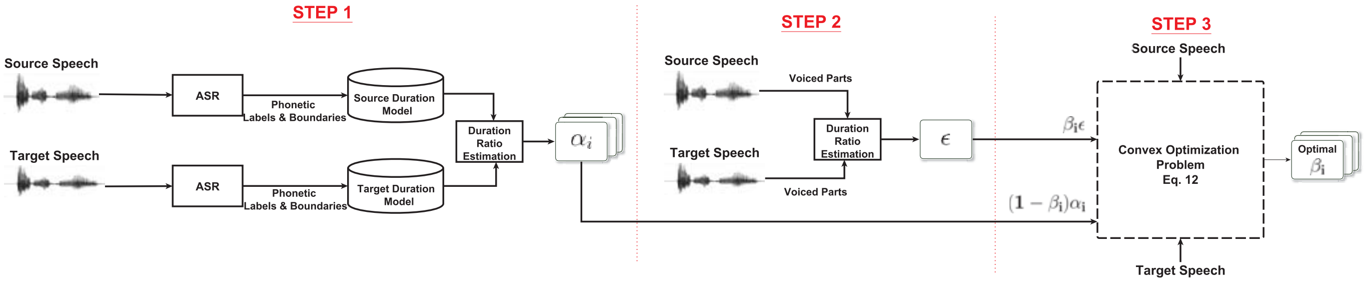


Fig. 4: Training phase of the proposed framework for duration modeling where we estimate the parameters for the conversion function Eq. (13).

reconstruction formula given in Eq. (3) is used. As a post-processing step to energy conversion, we perform energy contour improvement to obtain an energy contour of the converted spectrum which is more close to that of the target [25].

1) Phone-dependent Dictionary with Contextual Information: So far, we haven't taken into account contextual information, which means that each frame is converted independently. To alleviate the sharp changes across frames and achieve a more reliable activation matrix estimation, we use exemplars which span multiple consecutive frames in phone-dependent joint dictionary [13].

To achieve a smooth phone transition, we use biphone exemplars together with monophone exemplars while constructing the phone-dependent dictionary. Our idea is motivated by unit selection approach to speech synthesis [12], [34] where we favor speech units that share similar phonetic context as the intended context by using bi-phone or tri-phone context. We note that in the non-negative matrix factorization and signal reconstruction, the sequence of frames in the dictionaries are not particularly informative. By using the biphone exemplars, we would like to make sure that the intended phone transition frames are captured in the dictionary.

It is important to mention that, under the constraint of limited training data, it is not guaranteed that there is always a sub-dictionary for each phone in the test utterance. In this case, a backoff scheme will be helpful. For instance we can use all voiced exemplars to form a sub-dictionary. In the extreme case where use all phonetic exemplars are used in the backoff scheme, our proposed framework is reduced to the traditional NMF framework [25].

IV. Duration Conversion

So far, we perform conversion on spectrum, fundamental frequency and energy contour. We now move on to study the duration transformation, which is a part of the prosody description of speaker identity.

A. Prior work in Phone-dependent Duration Transformation

Duration transformation has not been considered in many well-known voice conversion frameworks. In [28], it was proposed to incorporate duration transformation into the Artificial Neural Network (ANN) for voice conversion.

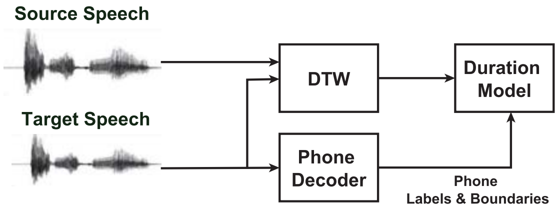


Fig. 5: Training phase of the duration transformation framework [28].

The prior work [28] studied a training process in which we find the duration of each phone from all utterances and estimate the mean and variance of both source and target speakers. At run-time, it is proposed to have a duration formulation via a Gaussian normalized transformation, that is given as:

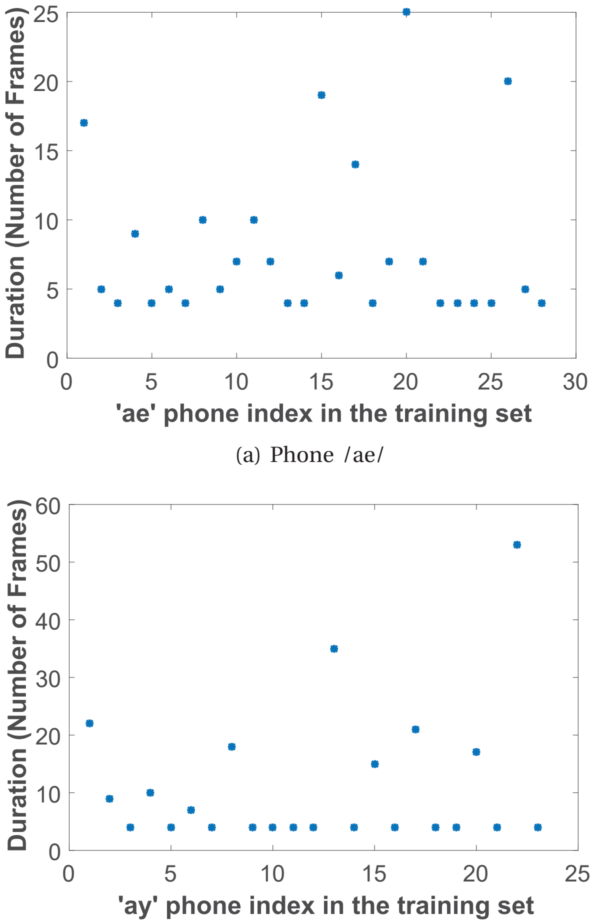
$$d_{t,i} = \mu_{t,i} + \frac{\sigma_{t,i}}{\sigma_{s,i}}\left(d_{s,i} - \mu_{s,i}\right) \qquad (10)$$

where $\mu_{t,i}, \sigma_{t,i}$ are the mean and variance of target speaker's duration, and $\mu_{s,i}, \sigma_{s,i}$ are the mean and variance of source speaker's duration for phone i.

The experimental results show that segmental duration transform can be done in the baseline voice conversion framework, and yield a significant improvement on conversion performance.

B. Proposed Back-off Scheme for Duration Conversion

In a voice conversion framework, duration transformation is an essential prosody feature that should be taken care of. In some cases, duration transformation becomes more vital, for example the scenario where speaking rate of source speaker differs from target speaker by a large margin and depends on the phonetic composition of the sentence. In such case, a sentence level duration conversion doesn't pay sufficient attention to the details. Therefore, either phone or syllable level duration conversion is necessary for a better conversion performance. With this motivation, the traditional duration transformation approach [28] has



(a) Phone /ae/

(b) Phone /ay/

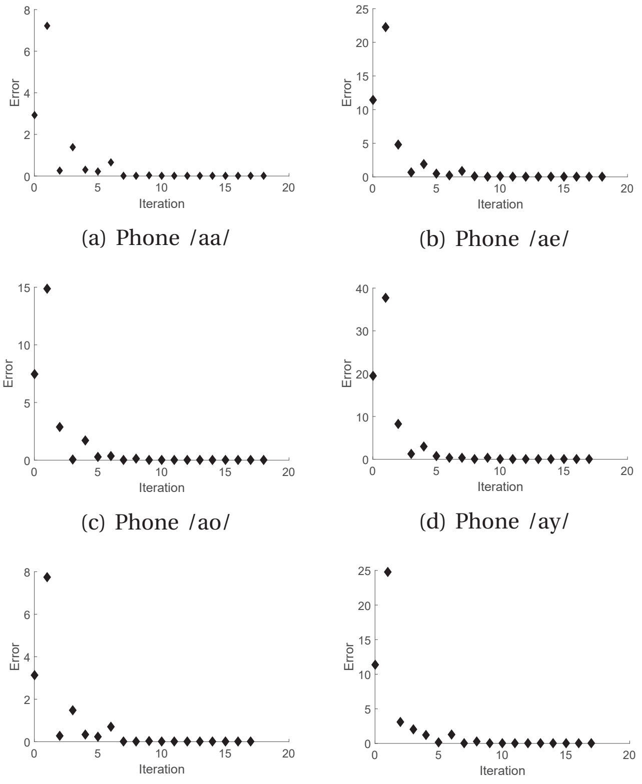
Fig. 6: The distribution of phones /ae/ and /ay/ has a high variance, hence performing transformation of duration, that is phone-dependent only [28], may not be reliable.

been proposed to perform segmental duration conversion. On the other hand, it highly depends on how reliable the data are for each phone. In the experiments that we conducted using VCC dataset [35], [36], a specific phone might have very different duration values, which is shown in Figure 6. As a result, it may not be very reliable to use only phone-level duration information of source and target speaker. To avoid this problem, we need a better duration transformation scheme, which takes into account phone-level as well as sentence-level speaking rates.

Representative works in duration modification also include a modification based on HMM model called DeBi-HMM [37], which proposes a voice conversion model as the post processing of a text-to-speech (TTS) system for speech synthesis. Another duration conversion framework [38] suggests to attach duration models to statistical models.

In this paper, we propose a back-off scheme by taking into account phone level speaking rates as well as sentence level speaking rates, which is more reliable than [28] when only limited parallel data are available. Furthermore, this approach is seen as converting speaker-dependent duration information at both phone level and sentence level. As shown in Figure 8, we estimate the duration of target speech before performing spectrum, F0 and energy contour conversion. Traditionally [10], [12], [13], [25], the duration patterns of the source speech are directly carried to the target speech. Such technique assumes that the target speaker and source speaker share the same duration patterns, which cannot be true in general. With our proposed duration conversion technique, duration of the converted speech will be closer to that of target.

We summarize the proposed framework in Figure 4. The training process involves three steps: 1) to estimate phone-level duration ratio, 2) to estimate sentence-level duration ratio, and 3) to find the optimal weigths for phone and sentence level duration ratios by performing convex optimization.



(a) Phone /aa/

(b) Phone /ae/

(c) Phone /ao/

(d) Phone /ay/

(e) Phone /ey/

(f) Phone /iy/

Fig. 7: Illustration of Step 3 in Training Phase. For each phone, number of iteration needed for convergence is given in x-axis, and the error in terms of number of frames is given in y-axis. We perform this optimization by using a development set.

In Step 1, we use an ASR to find the phone labels and boundaries for each training utterance. Then we estimate the duration ratio between source and target phone i denoted as α_i where $i = 1, 2, ..., N$ and N is the total number of phones in the training data. In Step 2, we estimate the

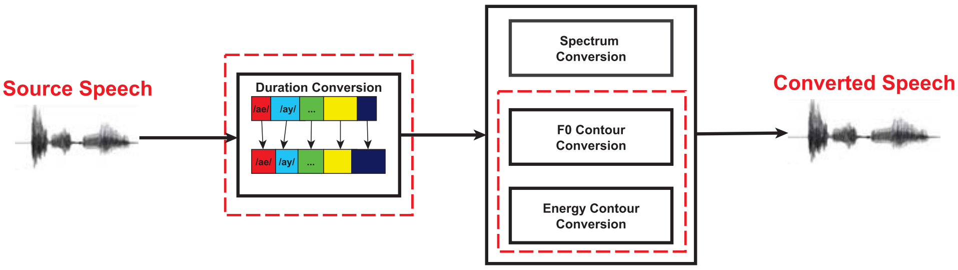


Fig. 8: The work flow of run-time spectrum and prosody conversion. The dotted red boxes are prosody conversion modules discussed in Section III (F0 and energy conversion) and Section IV (Duration conversion).

sentence level duration ratio, denoted as ϵ. In step 3, we assign the weights, that are denoted as $(1-\beta_i)$ and β_i, to interpolate the duration estimates between phone-level and sentence-level duration values. Now we would like to estimate the weight β_i between α_i and ϵ using the parallel training data. We note that α_i,ϵ,β_i characterize a speaker pair involved in the conversion. In other words, they will apply to the duration conversion of all utterances between a pair of given speakers at run-time.

The formulation can be written as follows:

$$d_i^t = \left((1-\beta_i)\alpha_i + \beta_i\epsilon\right) d_i^s \tag{11}$$

where d_i^t is the original target duration and d_i^s is the original source duration of phone i in the parallel data. After Step 1 and 2, the only unknown in Eq. (11) is the weight value β_i for $i = 1,...,N$. Then, we solve the problem of finding optimal β_i for phone i by performing convex optimization. The objection function for estimating the optimal weights for each phone can be formulated as follows.

$$\hat{\beta}_i = \min_{0\le\beta_i\le 1} \left(\left((1-\beta_i)\alpha_i + \beta_i\epsilon\right) d_i^s - d_i^t\right)^2 \tag{12}$$

where $i = 1,...,N$ and N is the total number of phones. We can consider that this is the phone duration ratio estimate with sentence duration ratio as the back-off. Figure 7 shows that this optimization problem is actually convex and can be optimized by simply varying β_i from 0 to 1.

At run-time conversion phase, the converted duration of phone i can be written as:

$$d_i^c = \left((1-\hat{\beta}_i)\alpha_i + \hat{\beta}_i\epsilon\right) d_i^s \tag{13}$$

where d_i^c is the converted duration, and d_i^s is the phone duration of the source speech. After estimation of converted duration for each phone in the testing utterance, we use a time scaling modification algorithm called SOLAFS [39], that is similar to Synchronized Overlapp-Add Algorithm with reduced computational requirements.

V. Experiments

We conducted the experiments on the Voice Conversion Challange (VCC) 2016 dataset [35], [36] to assess the performance of the proposed prosody conversion framework for F0, energy contour and duration with parallel training data. The VCC 2016 dataset, that is recorded by professional US English speakers, includes 5 female and 5 male speakers. However, we only considered speakers TF1, SF2, and SM1 for simplicity. In experiments, we use a DNN-HMM based ASR [40] to obtain phone labels and phone boundaries.

A. Objective Evaluation

1) Conversion of F0 and Energy Contour: We first report the experiments for F0 and energy conversion as presented in Section II, with 10, 20 and 30 source-target utterance pairs in training phase. The correlation coefficient of two signals is a measure of their linear dependence. As an objective evaluation for F0 and energy contour conversion, we first calculated the correlation coefficient between the converted and the reference target F0, then the correlation coefficient between the converted and the reference target contour.

The Pearson correlation coefficient(CC) can be defined as:

$$p(S,T) = \frac{cov(S,T)}{\sigma_S\sigma_T} \tag{14}$$

where σ_S and σ_T are the standard deviations of signals S and T, respectively. It is important to mention that the correlation coefficients for both F0 and energy are calculated between the frames aligned by dynamic time warping.

Table I compares the proposed Phonetically Aware Sparse Representation approach with two reference approaches, namely the traditional NMF-based approach [25] that we call Single Dictionary Sparse Representation, and linear conversion of F0 where F0 is linearly converted by normalizing the mean and variance of source speaker to target speaker. The formula for linear conversion of F0 is given as follows:

$$\hat{y} = \frac{\sigma_y}{\sigma_x}\left(x_t - \mu_x\right) + \mu_y \tag{15}$$

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.

F0 conversion framework	# Frames	Phonetic Dict.	# Training Pairs	Pearson CC for F0
Phonetically Aware Sparse Representation	1	Monophone	10	0.817
	1	Monophone	20	0.825
	1	Monophone	30	0.836
	3	Monophone+Biphone	10	0.852
	3	Monophone+Biphone	20	0.876
	3	Monophone+Biphone	30	0.891
Single Dictionary Sparse Representation [25]	3	-	20	0.793
	3	-	30	0.801
Baseline: Linear F0 conversion (Eq. 15)	-	-	20	0.703
	-	-	30	0.721

TABLE I: Comparison of correlation coefficients of the proposed Phonetically Aware Sparse Representation for F0 conversion, the traditional exemplar-based sparse representation [25], and the traditional approach to convert F0 linearly given in Eq. (15). The number of frames indicates the window size for activation matrix computation, and '# Training Pairs' indicates the number of parallel utterances used in the training phase.

Energy conversion framework	# Frames	Phonetic Dict.	# Training Pairs	Pearson CC for Energy
Phonetically Aware Sparse Representation	1	Monophone	10	0.804
	1	Monophone	20	0.812
	1	Monophone	30	0.828
	3	Monophone+Biphone	10	0.814
	3	Monophone+Biphone	20	0.826
	3	Monophone+Biphone	30	0.836
Baseline: Direct Transfer	-	-	-	0.802

TABLE II: Comparison of correlation coefficient of the proposed Phonetically Aware Sparse Representation for energy contour conversion and Direct Transfer that uses the energy contour of the source speech to reconstruct converted target utterance. As in Table I, the number of frames indicates the window size when computing activation matrix, and '# Training Pairs' indicates number of parallel utterances used in the training phase.

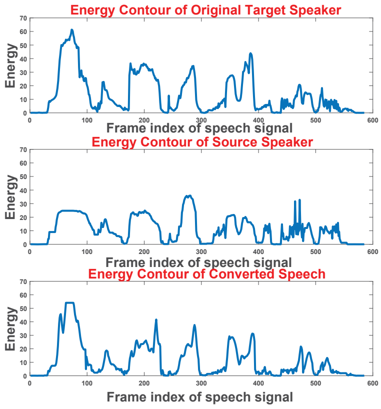


Fig. 9: An illustration of energy contours of target, source and converted speech. The energy contour of the converted speech is obtained by performing the proposed prosody conversion framework called Phonetically Aware Sparse Representation. In this experiment, 30 parallel utterances were used during training.

Firstly, in Table I, we observed that all Phonetically Aware Sparse Representation settings for F0 conversion outperform the traditional F0 conversion frameworks. Secondly, we observed that contextual information, e.g. multiple-frame exemplars together with biphones is effective to improve the conversion performance.

We further report the experiments for energy conversion as presented in Table II. We perform energy contour conversion by using Phonetically Aware Sparse Representation of different temporal scales by using CWT. As expected, we observed that all Phonetically Aware Sparse Representation settings for energy contour conversion outperform the baseline system where we use source speaker's energy contour directly. In addition, the experiment results show that modeling of contextual information is also crucial for conversion of energy contour.

2) *Conversion of Duration:* We further report the experiments for duration transformation framework as presented in Section III, with parallel training data. As in the Phonetically Aware Sparse Representation for F0 and energy conversion, we use the same ASR to get the phone labels and boundaries [40]. By taking into account both phone-level and sentence-level speaking rates, we expect to see a more reliable duration estimates. As an objective evaluation for the proposed framework, we calculated the distance between the converted and corresponding reference target durations in phone-level as well as sentence-level. Mathematically, the distance between the converted duration and the corresponding target duration denoted as δ_i for phone i can be written as:

$$\delta_i = \left| d_i^t - \left((1-\hat{\beta}_i)\alpha_i + \hat{\beta}_i \epsilon \right) d_i^s \right| \qquad (16)$$

where $i = 1, ..., M$ and M is the total number of phones in a test utterance. We calculated the mean of δ_i for $i = 1, ..., M$, and call it *Phone-level error*. Sentence-level error is calculated as the mean of all the utterances that occured in testing data. Table III shows the errors calculated in phone-level as well as sentence-level for a number of settings in a comparative study. One of the widely used techniques

Duration Conversion Framework	# Training Pairs	Phone-level Error	Sentence-level Error
Proposed Back-off Scheme	10	10.48	59.80
	20	10.29	57.30
	30	10.13	55.70
Traditional Framework [28]	20	10.67	62.21
	30	10.53	60.52
Baseline: Direct Transfer	-	11.36	78.34

TABLE III: Comparison of the proposed back-off scheme for duration transformation, the prior phone-dependent approach [28] for duration transformation and the baseline system that uses source speaker's duration directly. '# Training Pairs' indicates the number of parallel utterances used in the training phase.

in duration transformation is to keep the duration of target speaker same as that of the source speaker. For that reason, we use this approach as a baseline framework, that we call Direct Transfer. In addition, we implemented the traditional framework [28], and include the results to Table III as a reference. Firstly, we observed that our proposed scheme outperforms both reference approaches. Secondly, as expected, increasing the number of training utterances yields a better estimation of converted duration.

B. Subjective Evaluation

We further conducted listening tests to assess the performance of Phonetically Aware Sparse Representation for F0 and energy contour conversion in terms of prosody similarity. In all of the listening experiments, we use 30 utterance pairs from source and target speaker during training. In VC literature, XAB preference test is the evaluation technique which have been widely used [13]. For that reason, we prefer to use XAB preference test to evaluate our proposed framework. 12 subjects participated in all the listening tests. Overall, we conducted the following 2 listening experiments to assess the prosody conversion performance.

- F0 conversion
- F0 and energy contour conversion

The first listening experiment that is given in Figure 10, assesses the performance of F0 conversion. In this experiment, each listener was asked to listen both the converted samples and the original target samples. Then, each listener chose the sample that is closest to the target in terms of prosody similarity. We observe that Phonetically Aware Sparse Representation for F0 conversion outperforms the traditional framework where F0 is linearly converted by normalizing the mean and variance of source speaker to target speaker.

In Experiment 2, we evaluated the performance of F0 conversion together with energy conversion, as reported in Figure 11. In this experiment, Phonetically Aware Sparse Representation was used to perform both F0 and energy conversion. In the baseline system, the source speaker's energy contour is directly used, and F0 conversion is performed linearly as given in Eq. 15. The same listeners were asked to listen the target reference sample first, then the converted samples. Next, they decide which sample is the closest to the reference target in terms of prosody similarity. We observed that Phonetically Aware Sparse Representation for F0 and energy conversion outperforms the baseline framework.

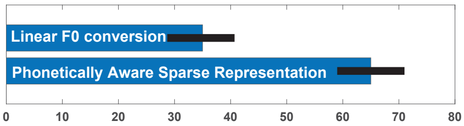


Fig. 10: Listening Experiment 1: Preference test results of prosody similarity for F0 conversion. The proposed framework Phonetically Aware Sparse Representation of F0 is compared with linear conversion of F0 given in Eq. 15. The results are provided with 95% confidence intervals.

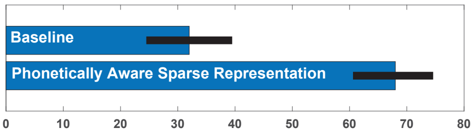


Fig. 11: Listening Experiment 2: Preference test results of prosody similarity for conversion of F0 and energy contour. The baseline system refers to the scenario where energy contour is directly transferred from source speaker, and F0 is converted linearly by using Eq. 15. The results are provided with 95% confidence intervals.

VI. Conclusion

We have proposed a novel prosody conversion framework that includes F0, energy contour and duration transformation. By using CWT, we modelled F0 and energy contour in different temporal scales effectively. In the proposed framework, we converted F0 and energy contour with Phonetically Aware Sparse Representation. In addition, we proposed a durational transformation approach that considers both phone-level and sentence-level speaking rates. We have validated that phonetically aware conversion of F0 and energy contour outperforms the traditional methods in both the objective and subjective evaluations. Moreover, the proposed back-off scheme for durational transformation marked a success in the estimation of converted duration. The proposed prosody conversion framework can work together with a spectrum conversion framework as an integrated solution to voice conversion which will be an interesting future work.

VII. Acknowledgment

This research is supported by Ministry of Education, Singapore AcRF Tier 1 NUS Start-up Grant FY2016. Berrak

Sisman is funded by SINGA Scholarship under A*STAR Graduate Academy.

The authors would like to thank Dr. Tong Rong for her support and invaluable comments.

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 Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1068–1071, 2008.

[5] 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.

[6] 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.

[7] X. Tian, Z. Wu, S. W. Lee, and E. S. Chng, “Correlation-based Frequency Warping for Voice Conversion,” *9th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 2–6, 2014.

[8] D. Erro, E. Navas, and I. Hernáez, “Parametric Voice Conversion Based on Bilinear Frequency Warping Plus Amplitude Scaling,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, no. 3, pp. 556–566, 2013.

[9] D. Lee and H. Seung, “Algorithms for non-negative matrix factorization,” *Advances in neural information processing systems*, no. 1, pp. 556–562, 2001. [Online]. Available: http://papers.nips.cc/paper/1861-algorithms-for-non-negative-matrix-factorization

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

[11] 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.

[12] 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.

[13] 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.

[14] R. Aihara, K. Masaka, T. Takiguchi, and Y. Ariki, “Parallel Dictionary Learning for Multimodal Voice Conversion Using Matrix Factorization,” *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 27–40, 2016.

[15] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, “voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary,” *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7894–7898, 2014.

[16] R. Patel, “Acoustic Characteristics of the Question-Statement Contrast in Severe Dysarthria Due to Cerebral ,” *Journal of Speech, Language, and Hearing Research*, pp. 1401–1415, 2003.

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

[18] J. Latorre, “Multilevel parametric-base F0 model for speech synthesis,” no. May, 2014.

[19] 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.

[20] Z. Wu, T. Kinnunen, E. S. Chng, and H. Li, “Text-Independent F0 Transformation with Non-Parallel Data for Voice Conversion,” *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1732–1735, 2010.

[21] B. Gillett, S. King, and U. Kingdom, “Transforming F0 Contours,” *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, no. 2, pp. 101–104, 2003.

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

[23] 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.

[24] 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.

[25] 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, no. 61175018, pp. 5175–5179, 2016.

[26] 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,” *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 08-12-September-2016, pp. 2453–2457, 2016.

[27] A. Toth, A. W. Black, and W. Em, “Incorporating durational modification in voice transformation,” *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1088–1091, 2008.

[28] S. Ronanki, B. Bajibabu, and K. Prahallad, “Duration Modeling in Voice Conversion Using Artificial Neural Networks,” *In IWSSIP*, pp. 11–13, 2012.

[29] H. Q. Nguyen, S. W. Lee, X. Tian, M. Dong, and E. S. Chng, “High quality voice conversion using prosodic and high-resolution spectral features,” *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5265–5285, 2016.

[30] D. Crystal and R. Quirk, “Systems of Prosodic and Paralinguistic Features in English,” *Mouton*, 1964.

[31] T. Kinnunen and H. Li, “An Overview of Text-Independent Speaker Recognition : from Features to Supervectors,” *Speech Communication*, vol. 52, pp. 12–40, 2010.

[32] H. Li and B. Ma, “Spoken Language Recognition: From Fundamentals to Practice,” *Proceedings of the IEEE*, vol. 101, no. 5, 2013.

[33] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, “Exemplar-based sparse representations for noise robust automatic speech recognition,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[34] P. Taylor, “Text-to-Speech Synthesis,” *Cambridge University Press*, 2009.

[35] M. Wester, Z. Wu, and J. Yamagishi, “Multidimensional scaling of systems in the Voice Conversion Challenge 2016,” *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 40–45, 2016.

[36] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, “The Voice Conversion Challenge 2016,” *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 1632–1636, 2016.

[37] C.-h. Wu, S. Member, C.-c. Hsia, T.-h. Liu, and J.-f. Wang, “Voice Conversion Using Duration-Embedded Bi-HMMs for Expressive Speech Synthesis,” vol. 14, no. 4, pp. 1109–1116, 2006.

[38] K. Yutani, Y. Nankakul, T. Toda, and K. Tokuda, “Simultaneous conversion of duration and spectrum based on statistical models including time-sequence matching,” *In Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2008.

[39] D. Hejna, B. R. Musicus, and B. Beranek, “The SOLAFS time-scale modification algorithm,” 1991.

[40] 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.