Interspeech 2018
2-6 September 2018, Hyderabad

Wavelet Analysis of Speaker Dependent and Independent Prosody for Voice Conversion

Berrak Sisman, Haizhou Li

National University of Singapore, Singapore
berraksisman@u.nus.edu, haizhou.li@nus.edu.sg

Abstract

Thus far, voice conversion studies are mainly focused on the conversion of spectrum. However, speaker identity is also characterized by its prosody features, such as fundamental frequency (F0) and energy contour. We believe that with a better understanding of speaker dependent/independent prosody features, we can devise an analytic approach that addresses voice conversion in a better way. We consider that speaker dependent features reflect speaker's individuality, while speaker independent features reflect the expression of linguistic content. Therefore, the former is to be converted while the latter is to be carried over from source to target during the conversion. To achieve this, we provide an analysis of speaker dependent and speaker independent prosody patterns in different temporal scales by using wavelet transform. The centrepiece of this paper is based on the understanding that a speech utterance can be characterized by speaker dependent and independent features in its prosodic manifestations. Experiments show that the proposed prosody analysis scheme improves the prosody conversion performance consistently under the sparse representation framework.
Index Terms: Wavelet transform, prosody analysis, voice conversion

1. Introduction

The goal of voice conversion (VC) is to convert one speaker's voice to sound like that of another [1, 2]. Speaker identity is characterized by 1) linguistic factors that are reflected in sentence structure and lexical choice; 2) supra-segmental factors, or prosodic features, such as stress, tone, or word juncture that extend over syllables, words, or phrases; and 3) segmental factors that are related to short term features, such as short-time spectrum and formants [3, 4]. When the language content is fixed, the supra-segment and the segmental factors are related to speaker individuality. Ideally, the voice conversion technology is able to convert both the supra-segment and the segmental factors. Unfortunately, many voice conversion frameworks are mainly focusing on spectral conversion [5, 6, 7, 8, 9, 10, 11]. We also note that some have ventured into the idea of speaker independent representation such as conditional restricted boltzmann machines with speaker independent pre-training [12] and DNN-based voice conversion framework in speaker independent space [13], where a database from a large speaker population is required.

Computational modeling of prosody has been a challenging task for many reasons. For example, prosody is described at supra-segmental level while spectrum is at short-time frame; prosody consists of F0 and energy among others that can vary highly. Prosody can be used contrastively to express 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. We call them speaker dependent prosody [14].

Prosody is also hierarchical in nature [15][16] and it can be affected by both short term as well as long term dependencies [17]. Fundamental frequency (F0) is a crucial prosodic feature in speech, hence previous studies of prosody conversion mainly focus on transformation of F0 [18]. The continuous wavelet transform (CWT) has been used for the analysis and modeling of F0 within an hidden Markov model (HMM) framework [19][20]. With this motivation, voice conversion frameworks such as DKPLS [17] and exemplar-based prosody conversion [21][22] use CWT for F0 decomposition. More recently, we find that CWT decomposition for F0 and energy contour are effective in emotional voice conversion [23] and phonetically aware prosody conversion [15]. Unfortunately, the prior work hasn't provided a statistical analysis over the CWT decompositions from the viewpoint of speaker dependent and speaker independent prosody.

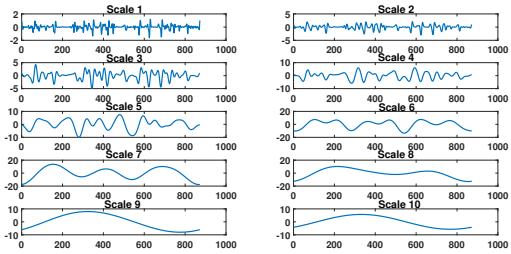
In this paper, we propose comprehensive frameworks, that are based on Pearson Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE), to understand the speaker dependent and independent characteristics of prosody. We report our findings from the CWT decompositions of F0 and energy contours at different scale through a statistical analysis. The proposed prosody analysis also improves the state-of-the-art voice conversion frameworks by carrying over the speaker independent (SI) prosody features from source to target, while transforming the speaker dependent (SD) prosody features of source to those of target. For the first time, we present the sparse representation frameworks that handles speaker dependent and speaker independent prosody differently during the transfer.

The main contributions of this paper include, 1) we provide an understanding from the perspective of speaker dependent and independent prosody features; 2) we devise a statistical analysis framework to assess the speaker dependent and independent prosody, that uses PCC and RMSE; 3) we incorporate the proposed prosody analysis with voice conversion, that consistently outperforms the state-of-the-art baseline.

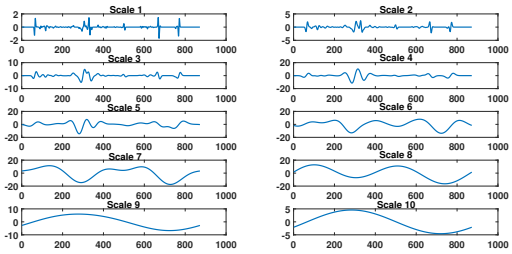
This paper is organized as follows: In Section 2, we describe the details of CWT decomposition of F0 and energy contour. In Section 3, we explain the analysis of speaker dependent and independent prosody features. In Section 4, we describe the proposed prosody transfer. Section 5 reports the experimental results. Finally, we conclude in Section 6.

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

10.21437/Interspeech.2018-1499



(a) *Decomposition of energy contour*



(b) *Decomposition of F0 contour*

Figure 1: *CWT analysis of prosody features of an utterance*

2. Prosody Modelling in Voice Conversion

Prosody conveys linguistic, para-linguistic and various types of non-linguistic information, such as speaker identity, intention, attitude and mood [24, 25]. It is inherently supra-segmental [26, 27] due to the reason that the variations of prosody cannot be derived from the segmental sequence [16]. It 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 [16]. For example, voiceless segments lack explicit F0 values and high vowels generally have higher F0 than low vowels.

The fundamental frequency (F0) of speech is one of the most important prosodic features that should be taken into account in a comprehensive voice conversion framework. F0 features extracted from STRAIGHT vocoder are low dimensional features that have lower data complexity than spectral features. A simple approach to F0 conversion is what we call Gaussian normalized transformation [28], a linear transformation. It is clear that it is not adequate to use a linear model to represent all variations in different temporal scales.

Recently, CWT was shown to effectively model F0 in different temporal scales that improves the speech synthesis performance [19, 27, 20]. It was also introduced to sparse representation [22], DKPLS for voice conversion [17], emotion conversion with arbitrary F0 scales [29] and adaptive scales for emotional voice conversion [30]. It was also studied that CWT decompositions of F0 and energy contours provide a way to manipulate the prosody of utterances [23] .

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

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

where ψ is the Mexican hat mother wavelet. If we fix the analysis at 10 discrete scales, f_0 can be represented as [20]

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

where $i = 1, ..., 10$ and $\tau_0 = 5ms$. These time scales were originally proposed in [22] and in a hierarchical prosody model [19], that were further used in some voice conversion applications [23, 21, 22, 15]. 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:

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

We adopted CWT to decompose the 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 discontinuities 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. Figure 1 illustrates the CWT decompositions of F0 and energy contour at 10 different scales for the same utterance. Different from all the previous studies, our main focus here is to analyze the speaker dependent and speaker independent elements through CWT decompositions in different temporal scales. We hope to associate the temporal scales with speaker dependency through a statistical analysis.

3. Analysis of Prosody Features for Voice Conversion

Prosody typically reflects a combination of features from both the speaker and the utterance. 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 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 that speaker features are speaker dependent, and utterance features are speaker independent because they are manifested in the same way by all speakers.

The prosody mapping approaches [18, 17, 21, 22, 15] that have been proposed so far, don't study the problem from the view point of speaker dependent and independent features in speech utterances. Recently proposed deep learning approaches for spectral mapping [31, 32] do not handle F0 and energy contours either due to lack of appropriate modeling mechanism.

In voice conversion, 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. To do this, we need to be able to identify speaker dependent and speaker independent prosodic elements in an utterance. We consider that the CWT decompositions of F0 and energy contour at different temporal scales represent the speaker dependent and speaker independent prosodic elements, that we would like to empirically prove in this paper. To our best knowledge, this paper is the first to perform such a statistical analysis to benefit the design of voice conversion system.

Figure 2 shows the proposed system diagram for the analysis of speaker dependent and speaker independent prosody patterns. Pearson correlation coefficient (PCC) and root mean square error (RMSE) between the respective wavelet decompo-

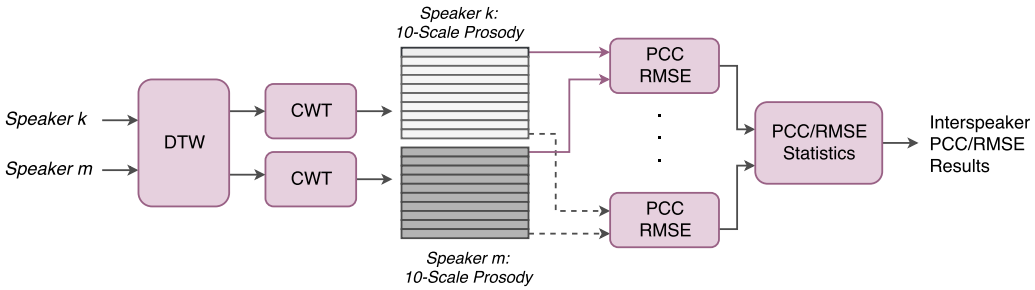
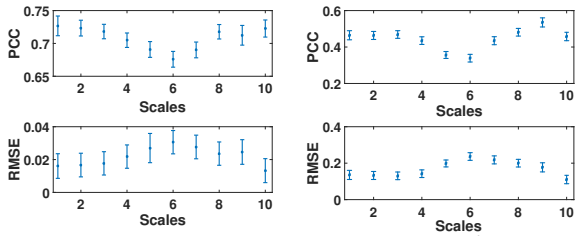


Figure 2: *The proposed workflow to assess the inter-speaker correlation of prosodic patterns at different temporal scales, where* $k, m = 1, 2, ..., 10$.



(a) *Energy* (b) *F0*

Figure 3: *Comparison of inter-speaker prosody contours in different temporal scales using Pearson correlation coefficient (PCC) and root mean square error (RMSE), with 95 % confidence interval.*

sitions of all speaker pairs (k, m) are employed as the objective measures to compare the speakers. Suppose that we have two signals K, and M. The Pearson correlation coefficient of two signals is a measure of their linear dependence that is calculated as follows:

$$p(K, M) = \frac{cov(K, M)}{\sigma_K \sigma_M} \tag{4}$$

where σ_K and σ_M are the standard deviation of K and M, respectively. The proposed framework includes the following steps: 1) As needed for PCC and RMSE, we perform frame alignment by using DTW, 2) using Eq. (1) and (2), we perform CWT to decompose the F0 and energy contour into different temporal scales, 3) we calculate the inter-speaker linear dependency between the same temporal scales by using PCC. We also calculate the inter-speaker distance between the same temporal scales by RMSE, 4) finally, we report the overall mean and variance of all speaker pair combinations at different scales for F0 and energy contour. We conduct the experiments on Voice Conversion Challenge (VCC) 2016 database [33] that is recorded from 5 female and 5 male US English speakers. In our experiments, we use data from all 10 speakers. STRAIGHT is used as the vocoder for both analysis and synthesis.

As can be seen from Fig. 3, middle scales (scales 4-8) provides lower correlation and higher RMSE values, that means that they carry more speaker dependent information. However, the rest of the scales carry less speaker dependent information, in other words, more linguistic information. We are glad that we are now able to identify the relationship between temporal scales and speaker dependency.

We can interpret the findings as follows, the low temporal scales (scales 1-3) represent short-term/micro-level prosodic elements and the high temporal scales (scales 9-10) represent long-term ones. We note that prosodic features are the properties of speech units larger than the individual segments. Short-term prosodic elements describe what is within the individual segments, therefore, don't vary much from speaker to speaker. Long-term prosodic elements describe the sentence level prosody, such as the form of the utterance (statement, question, or command), by following certain prosodic templates, therefore, don't vary with the speaker either. The middle temporal scales (scales 4-8) capture supra-segmental information such as intonation, tone, stress, and rhythm over words and phrases, therefore, we observe a large speaker variation. In the next section, we will propose a voice conversion framework that carry over low and high scales prosodic elements from source to target, while transforming the middle scales prosodic elements to the target.

4. Prosody Conversion

In exemplar-based sparse representation (SR)[22], spectral and prosody features are converted via a pair of coupled dictionaries, denoted as $\mathbf{A}$ and $\mathbf{B}$, each consists of spectrum, aperiodicity component and the CWT representation of F0. At run-time, both the spectral and prosody features of a source utterance $\mathbf{X}$ can be represented as $\mathbf{X} \approx \mathbf{AH}$. Non-negative factorization (NMF) technique is employed to estimate the activation matrix $\mathbf{H}$, which is constrained to be sparse. The converted spectral and prosody features can be written as $\hat{\mathbf{Y}} = \mathbf{BH}$.

Phonetic sparse representation (PSR) [15] is an extension to sparse representation [22] by replacing the coupled dictionary [$\mathbf{A}$;$\mathbf{B}$] with multiple phonetic sub-dictionaries that consist of both spectrum and prosody features. Phonetic sparse representation makes use of phonetic information to achieve better activation matrix estimation, thus, better voice conversion.

In Section 3, we provide an analysis of speaker dependent and independent prosody features through CWT decomposition. We show that the middle scales of both F0 and energy contour contain more speaker dependent information while the rest of the scales carry less speaker dependent information. With this findings, we would like to see if it helps in voice conversion experiments to only convert the speaker dependent prosodic elements (scales 4-8).

5. Experiments

We conduct experiments using VCC 2016 database [33] to assess the performance of the proposed speaker dependent and independent prosody features in voice conversion. We would like to validate the idea that the CWT decomposition of F0 and energy contour represents the speaker dependent and speaker independent prosody in different scales. We choose exemplar-

Framework	# Scales	# Training Pairs	PCC
SR: F0	1-10	20	0.753
	4-8	20	0.771
	1-10	30	0.768
	4-8	30	0.794
SR: E	1-10	20	0.796
	4-8	20	0.807
	1-10	30	0.812
	4-8	30	0.821
PSR: F0	1-10	20	0.811
	4-8	20	0.824
	1-10	30	0.828
	4-8	30	0.837
PSR: E	1-10	20	0.842
	4-8	20	0.854
	1-10	30	0.851
	4-8	30	0.862

Table 1: *The PCC between the target prosodic contour and the converted one in sparse and phonetic sparse representation experiments. '# Scales' represents the CWT scales of F0 and energy contour that are converted. SR: F0 is for the F0 contour, while SR: E is for the energy contour*

based sparse representation [22, 15] for prosody conversion experiments. Pearson correlation coefficient is used as an evaluation for F0 and energy contour conversion. It is important to mention that the correlation coefficients for both F0 and energy are calculated between the frames aligned by dynamic time warping.

5.1. Objective Evaluations

We first conduct experiments for the wavelet analysis of prosody features in sparse representation framework with 20, 30 source-target utterance pairs in training phase. We use 3 consecutive frames to achieve a more reliable activation matrix estimation. To observe the relationship between the wavelet scales and speaker dependency, we conduct experiments with 2 different settings: 1) all scales of F0 and energy contour are converted, and 2) middle scales (4-8) are converted from source speaker to target, while the rest is directly copied from source speaker.

Table 1 reports the PCC values for a number of settings in a comparative study. We examine the Pearson correlation coefficients (PCC) between the target prosody contour and the converted one in sparse and phonetic sparse representation (PSR) frameworks. In all experiments, we use exemplars that span 3 consecutive frames. Moreover, in PSR experiments, we use a DNN-HMM based ASR [36] to obtain phone labels and phone boundaries. To capture the phone transition, we use biphone exemplars together with monophone exemplars while constructing the phonetic dictionary. In both F0 and energy contour conversion, we observed that we get better PCC results by converting scales 4-8 than converting all scales. These results validate our findings in Section 3 that the middle scales carry more speaker dependent information, while the other scales are relatively speaker independent. We also observed that increasing the number of training data improves the prosody conversion performance.

As PSR [15] takes into account phonetic information while estimating the activation matrix, it is not suprising that PSR consistently outperforms SR [22] framework for both F0 and energy conversion. PSR achieves higher PCC values than SR for

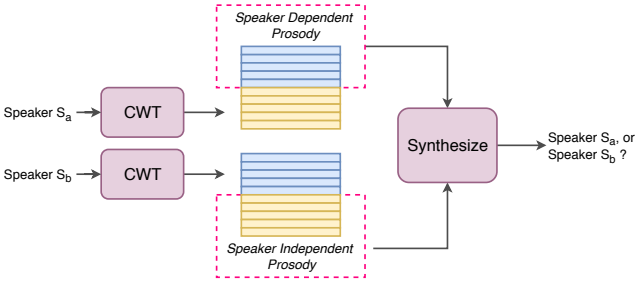


Figure 4: *The listening experiment setup for speaker dependent independent prosody study reported in Table 2.*

both F0 and energy contour conversion. Overall, these results show that the CWT decomposition of F0 and energy contours can represent the speaker dependent and speaker independent prosody in different scales, that improves the performance of prosody transformation.

5.2. Subjective Evaluations

We further conduct listening experiment to assess the effect of speaker dependent and independent prosody features in speaker similarity. Our findings in Section 3 suggests that the middle scales carry more speaker dependent information, while the other scales are more speaker independent. We devise a subjective evaluation framework as given in Fig. 4. We first perform CWT on speaker S_a and speaker S_b, to decompose their F0 contours into 10 temporal scales. Then we synthesize the speech of speaker S_a with his speaker dependent prosody features (scales 4-8) together with the speaker independent prosody features (scales 1, 2, 3, 9, 10) of speaker S_b. Then we ask to 10 subjects to listen and choose the speaker identity that is closer to the synthesized speech. We use four different S_a-S_b combinations with each of them having 10 speech samples. In theory, we expect listeners to choose Speaker S_a, as we carry over only the speaker independent features, that we call linguistic information, from Speaker S_b.

Speaker S_a	Speaker S_b	No preference
(82.0±2.0) %	(6.0±1.8) %	(12.0 ±1.5) %

Table 2: *The preference tests with 95 % confidence interval for the experiments illustrated in Fig. 4.*

As given in Table 2, the speaker identity does not change as long as we preserve the speaker dependent scales in the wavelet analysis. This is also confirms our findings, that are reported in objective evaluations.

6. Conclusion

In this paper, we perform an analysis on prosodic features in terms of speaker characteristics. We have proposed a system model to assess the speaker dependent and independent prosody features and provide a better understanding of prosody patterns of different speakers. Then, we incorporate this knowledge to prosody conversion by carrying the speaker independent prosody features from source speaker to the target, and replace the speaker dependent prosody of source speaker to that of target speaker in sparse and phonetic sparse representation. We show that the CWT decomposition of a F0 and energy contour can offer the speaker dependent and speaker independent prosody in different scales, that can provide an useful tool for prosody transformation.

7. References

[1] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *In 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] R. Aihara, T. Nakashika, T. Takiguchi, and Y. Ariki, "voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary," *In ICASSP*, 2014.

[6] 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)*, 2017.

[7] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data," *arXiv*, 2017.

[8] ——, "Learning Latent Representations for Speech Generation and Transformation," *arXiv*, 2017.

[9] 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.

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

[11] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," *arXiv*, 2017.

[12] S. H. Mohammadi and A. Kain, "Voice conversion using deep neural networks with speaker-independent pre-training," *In IEEE Spoken Language Technology Workshop (SLT)*, 2014.

[13] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion in time-invariant speaker-independent space," *In ICASSP*, pp. 7889–7893, 2014.

[14] D. Crystal and R. Quirk, "Systems of Prosodic and Paralinguistic Features in English," *Mouton*, 1964.

[15] B. Sisman, H. Li, and K. C. Tan, "Transformation of Prosody in voice conversion," *APSIPA ASC*, 2017.

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

[17] G. Sanchez, H. Silen, J. Nurminen, and M. Gabbouj, "Hierarchical modeling of F0 contours for voice conversion," *In INTERSPEECH*, pp. 2318–2321, 2014.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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 ICASSP*, pp. 5175–5179, 2016.

[23] 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 INTERSPEECH*, vol. 08-12-September-2016, pp. 2453–2457, 2016.

[24] 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.

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

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

[27] 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.

[28] 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.

[29] 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.

[30] ——, "Emotional Voice Conversion with Adaptive Scales F0 based on Wavelet Transform using Limited Amount of Emotional Data," *In INTERSPEECH*, pp. 3399–3403, 2017.

[31] L. Sun, K. Li, H. Wang, S. Kang, and H. Meng, "Phonetic posteriorgrams for many-to-one voice conversion without parallel data training," *In IEEE ICME*, 2016.

[32] C.-C. Hsu, H.-T. Hwang, Y.-C. Wu, Y. Tsao, and H.-M. Wang, "Voice Conversion from Non-parallel Corpora Using Variational Auto-encoder," *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016. [Online]. Available: http://arxiv.org/abs/1610.04019

[33] 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.

[34] C. Gupta, H. Li, and Y. Wang, "Perceptual Evaluation of Singing Quality," *APSIPA ASC*, 2017.

[35] B. Sisman, G. Lee, H. Li, and K. C. Tan, "On the analysis and evaluation of prosody conversion techniques," *IALP*, 2017.

[36] D. Povey, A. Ghoshal, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, J. Silovsk, and P. Motl, "The Kaldi Speech Recognition Toolkit," *In IEEE ASRU*, 2011.