PARALLEL-DATA-FREE DICTIONARY LEARNING FOR VOICE CONVERSION USING NON-NEGATIVE TUCKER DECOMPOSITION

Yuki Takashima[1], *Hajime Yano*[1], *Toru Nakashika*[2], *Tetsuya Takiguchi*[1], *Yasuo Ariki*[1]

[1]Graduate School of System Informatics, Kobe University, Japan
[2]Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan

ABSTRACT

Voice conversion (VC) is a technique where only speaker-specific information in source speech is converted while preserving the associated phonological information. Non-negative Matrix Factorization (NMF)-based VC has been researched because of the natural-sounding voice it produces compared with conventional Gaussian Mixture Model-based VC. In conventional NMF-VC, parallel data are used to train the models; therefore, unnatural pre-processing of speech data to make parallel data is needed. NMF-VC also tends to be a large model because this method has many parallel exemplars for the dictionary matrix; therefore, the computational cost is high. In this paper, we propose a novel parallel dictionary learning method using non-negative Tucker decomposition (NTD) which uses tensor decomposition and decomposes an input observation into a set of mode matrices and one core tensor. Our proposed NTD-based dictionary learning method estimates the dictionary matrix for NMF-VC without using parallel data. Experimental results show that our proposed method outperforms conventional non-parallel VC methods.

Index Terms— Voice conversion, non-negative Tucker decomposition, non-negative matrix factorization, non-parallel training

1. INTRODUCTION

Recently, voice conversion (VC), which is a technique used to change speaker-specific information in the speech of a source speaker into that of a target speaker while retaining linguistic information, has been garnering much attention since the VC techniques can be applied to various tasks [1, 2] Various statistical approaches to VC have been studied so far as discussed in [3]. Among these approaches, the Gaussian mixture model (GMM)-based mapping method [4] is most widely used, and a number of improvements have been proposed [5, 6]. Other VC methods, such as approaches based on non-negative matrix factorization (NMF) [7, 8], neural networks (NNs) [9], and restricted Boltzmann machines (RBMs) [10, 11], have been also proposed.

NMF [12] is one of the most popular sparse representation methods. NMF decomposes the input observation into two matrices — the basis matrix and weight matrix. The goal of NMF is to estimate these two matrices from the input observation. In this paper, we refer to the basis matrix as the "dictionary" and the weight matrix as the "activity". The NMF-based method can be classified into two approaches: the dictionary-learning approach [8] and exemplar-based approach [13]. In the dictionary-learning approach, the dictionary and the activity are estimated simultaneously. On the other hand, in the exemplar-based approach, only the activity becomes sparse because the dictionary is determined using exemplars and the activity is estimated using NMF. In VC tasks, by using the basis matrices instead of the exemplars, the VC is performed with lower computation times than with the exemplar-based method.

However, the conventional NMF-based approach requires parallel data (aligned speech data from the source and the target speakers so that each frame of the source speaker's data corresponds to that of the target speaker) for training the models, which leads to several problems. First, the data is limited to pre-defined articles (both speakers must utter the same articles). Second, the training data (the parallel data) is not the original speech data anymore because the speech data is stretched and modified in the time axis when aligned, and it is not guaranteed that each frame is aligned perfectly. Because the dictionary is constructed from parallel data, the error of alignment in parallel data might adversely affect VC performance. Several other approaches have been proposed that do not use (or minimally use) parallel data of the source and the target speakers [14, 15]. In [14], for example, they model the spectral relationships between two arbitrary speakers (reference speakers) using GMMs, and convert the source speaker' s speech using the matrix that projects the feature space of the source speaker into that of the target speaker through that of reference speakers. In this paper, we expand a conventional NMF-based VC method into a non-parallel VC method.

In this paper, we propose a non-negative Tucker decomposition (NTD) [16, 17]-based dictionary learning method for a NMF-based VC. NTD is a non-negative extension of Tucker decomposition that decomposes the input observation into a set of matrices and one core tensor. Because we use spectral features as the input observation, a set of matrices consists of two mode matrices for frequency and time, and a core tensor corresponding to a core matrix. We assume that these matrices correspond to the frequency basis matrix, the phonemic information, and the codebook between the frequency basis and each phone, respectively. We assume that the activity matrix in NMF is decomposed into the codebook and the phonemic information. When learning the dictionaries, although the activity matrix is shared between speakers using parallel data in the conventional NMF, in our proposed method the codebook is shared between speakers and the phonemic information is dependent on a speaker. Hence, the time-varying phonemic information can be captured for each speaker. When converting, we estimate only the phonemic information matrix as the activity matrix. Our proposed method is able to have time-dependent factors for each speaker; therefore, parallel data is not necessary.

The rest of this paper is organized as follows: In Section 2, VC using exemplar-based NMF is described. In Section 3, our proposed method is described. In Section 4, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. NMF-BASED VOICE CONVERSION

2.1. Basic idea

Fig. 1 shows the basic approach of the dictionary-learning NMF-based VC [8], where F, T, and K represent the numbers of dimensions, frames, and bases, respectively. This VC method needs two dictionaries that are phonemically parallel. $\mathbf{A}^s$ represents a source dictionary, and $\mathbf{A}^t$ represents a target dictionary. In exemplar-based VC, these two dictionaries consist of the same words and are aligned with dynamic time warping (DTW) just as conventional GMM-based VC is. In dictionary-learning VC, these two dictionaries are estimated simultaneously. Hence, these dictionaries have the same number of bases. In this paper, we employ dictionary-learning VC.

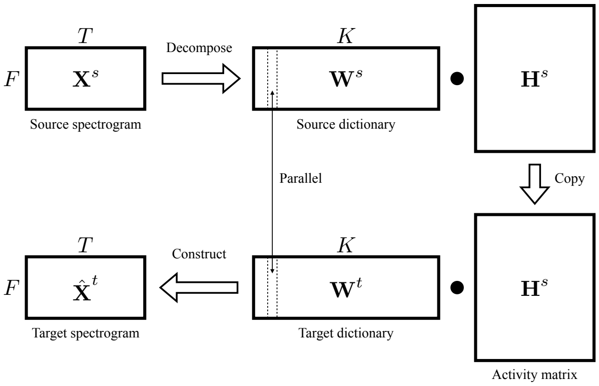

Fig. 1. Basic approach of NMF-based voice conversion

At first, for the training source data $\mathbf{X}^s$, $\mathbf{A}^s$ and the source speaker's activity $\mathbf{H}^s$ are estimated using NMF. The cost function of NMF is defined as follows:

$$d_{KL}(\mathbf{X}^s, \mathbf{A}^s\mathbf{H}^s) + \lambda||\mathbf{H}^s||_1 \; s.t. \; \mathbf{A}^s, \mathbf{H}^s \geq 0 \qquad (1)$$

In Eq. (1), the first term is the Kullback-Leibler (KL) divergence between $\mathbf{X}^s$ and $\mathbf{A}^s\mathbf{H}^s$ and the second term is the sparsity constraint with the L1-norm regularization term that causes the activity matrix to be sparse. λ represents the weight of the sparsity constraint. This function is minimized by iteratively updating.

Next, using the activity matrix $\mathbf{H}^s$ obtained by Eq. (1), the target basis matrix $\mathbf{A}^t$ of the training target data $\mathbf{X}^t$ is optimized. Then, $\mathbf{A}^t$ is optimized so that the activity matrix is equivalent to $\mathbf{H}^s$; i.e., $\mathbf{A}^t$ is optimized to minimize the following cost function:

$$d_{KL}(\mathbf{X}^t, \mathbf{A}^t\mathbf{H}^s) \; s.t. \; \mathbf{A}^t \geq 0 \qquad (2)$$

In this optimization, the activity matrix is fixed to $\mathbf{H}^s$, and only $\mathbf{A}^t$ is updated.

This method assumes that when the source signal and the target signal (which are the same words but spoken by different speakers) are expressed with sparse representations of the source dictionary and the target dictionary, respectively, the obtained activity matrices are approximately equivalent. The estimated source activity $\mathbf{H}^s$ is multiplied to the target dictionary $\mathbf{W}^t$ and the target spectra $\mathbf{X}^t$ are constructed.

$$\hat{\mathbf{X}}^t = \mathbf{W}^t\mathbf{H}^s \qquad (3)$$

2.2. Problems

NMF-based VC has several problems. First, if the source and target utterances are aligned using DTW in advance, the estimated parameters are affected by the quality of the alignment. There still seems to be a mismatch of alignment. This mismatch degrades the performance of exemplar-based VC [18]. Second, it seems that the activity matrix contains not only phonetic information but also another information. In [13, 19], Aihara *et al.* assumed that the activity matrix contains the phonetic information and speaker information, proposed some frameworks for dealing with this effect, and was able to improve the performance of NMF-based VC. In this paper, we propose another approach. We decompose the activity matrix into the speaker-shared matrix and the speaker-dependent phonetic information matrix. Then, estimating only the phonetic information matrix as the activity matrix when converting, we expect to improve the activity estimation accuracy.

3. DICTIONARY LEARNING USING NTD

3.1. Non-negative Tucker decomposition

Given a non-negative N-way tensor, non-negative Tucker decomposition (NTD) [20] decomposes the input tensor into a core tensor and a set of mode matrices that are restricted to have only non-negative elements. In this paper, because we use spectral features as the input observation, a core tensor is represented as a matrix, and the number of mode matrices is two. Under these conditions, NTD is simply defined as follows:

$$\mathbf{X} \approx \mathbf{U}\mathbf{G}\mathbf{V}^{\top} \; s.t. \; \mathbf{U} \geq 0, \mathbf{G} \geq 0, \mathbf{V} \geq 0 \qquad (4)$$

where $\mathbf{X} \in \mathbb{R}^{F \times T}$, $\mathbf{U} \in \mathbb{R}^{F \times M}$, $\mathbf{V} \in \mathbb{R}^{T \times L}$, $\mathbf{G} \in \mathbb{R}^{M \times L}$ are an input spectrogram, mode matrices along the frequency and time axes and a core matrix, respectively. F, T, M, and L indicate the number of frequency bins and frames, and the frequency and time basis, respectively. The cost function of NTD is defined as follows:

$$||\mathbf{X} - \mathbf{U}\mathbf{G}\mathbf{V}^{\top}||_F^2, \qquad (5)$$

where $|| \cdot ||_F$ indicates the Frobenius norm. NTD provides a general form of the non-negative tensor factorization including special case NMF, and updating algorithms have been proposed in [20]. These updating algorithms are based on that of NMF.

3.2. Parallel dictionary learning using NTD

In this section, we describe how a parallel dictionary between the source and target speakers is estimated by NTD. The objective function is represented as follows:

$$\begin{aligned} &||\mathbf{X}^s - \mathbf{U}^s\mathbf{G}\mathbf{V}^{s\top}||_F^2 + ||\mathbf{X}^t - \mathbf{U}^t\mathbf{G}\mathbf{V}^{t\top}||_F^2 \\ &s.t. \; \mathbf{U}^s \geq 0, \mathbf{U}^t \geq 0, \mathbf{G} \geq 0, \mathbf{V}^s \geq 0, \mathbf{V}^t \geq 0 \qquad (6) \end{aligned}$$

where $\mathbf{X}^s \in \mathbb{R}^{F \times T_s}$, $\mathbf{X}^t \in \mathbb{R}^{F \times T_t}$, $\mathbf{U}^s \in \mathbb{R}^{F \times M}$, $\mathbf{U}^t \in \mathbb{R}^{F \times M}$, $\mathbf{V}^s \in \mathbb{R}^{T_s \times L}$, $\mathbf{V}^t \in \mathbb{R}^{T_t \times L}$, and $\mathbf{G} \in \mathbb{R}^{M \times L}$ are the source and target spectrograms, the source and target frequency basis matrices, the source and target time basis matrices, and a core matrix, respectively. F, T_s, T_t, M, and L indicate the number of frequency bins, the source and target frames, and the frequency and time basis, respectively. This function is minimized by iteratively updating each parameter in the same manner as NTD. The core matrix G is estimated by iteratively updating the following equation.

$$\begin{aligned} \mathbf{G} \leftarrow \; &\mathbf{G}. * (\mathbf{U}^{s\top}\mathbf{X}^s\mathbf{V}^s + \mathbf{U}^{t\top}\mathbf{X}^t\mathbf{V}^t) \\ &./(\mathbf{U}^{s\top}\mathbf{U}^s\mathbf{G}\mathbf{V}^{s\top}\mathbf{V}^s + \mathbf{U}^{t\top}\mathbf{U}^t\mathbf{G}\mathbf{V}^{t\top}\mathbf{V}^t), \qquad (7) \end{aligned}$$

where $.*$ and $./$ denote element-wise multiplication and division, respectively.

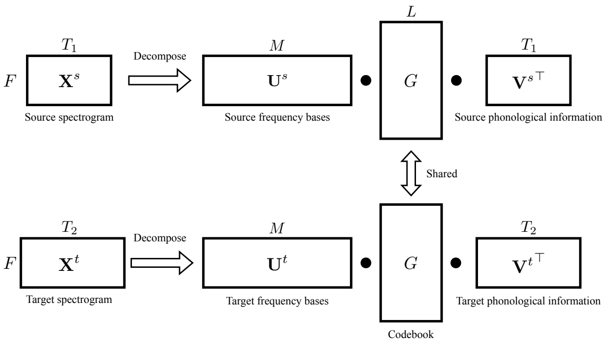

Fig. 2. Parallel dictionary learning using NTD

We assume that $\mathbf{U}^s$ and $\mathbf{U}^t$ represent the frequency basis matrices, and $\mathbf{V}^s$ and $\mathbf{V}^t$ represent the phonemic information. Because the core matrix is not dependent on both the frequency and the time, we assume that this matrix represents the codebook between the frequency bases and the phones. According to this assumption, the core matrix makes a correspondence between frequency bases and phones. Specifically, there are L phones, and the spectrum of each phone is constructed using M frequency bases. Although the information contained in the activity matrix is not just the phonological information, in conventional NMF-based approaches, the activity matrix is estimated as only the phonological information. Therefore, the estimated activity is degraded. In contrast, our proposed NTD-based approach expressly decomposes the activity matrix into the speaker-shared information and the speaker-dependent phonemic information. Therefore, it is expected that the performance of the activity estimation will be improved when converting.

After each matrix in the model is estimated, parallel dictionaries are calculated as follows:

$$\mathbf{W}^s = \mathbf{U}^s\mathbf{G} \qquad (8)$$

$$\mathbf{W}^t = \mathbf{U}^t\mathbf{G}. \qquad (9)$$

Then the input source speaker's spectra are converted using the dictionaries in the same manner described in Section 2.

4. EXPERIMENTAL RESULTS

4.1. Conditions

The proposed VC technique was evaluated by comparing it with the conventional GMM-based method [4], the conventional dictionary-learning NMF-based method [8], and an adaptive restricted Boltzmann machine (ARBM)-based method [11] which do not use parallel data, in a speaker-conversion task using clean speech data. The source speaker and target speaker were one male and one female speaker, respectively, whose speech is stored in the ATR Japanese speech database [21]. The sampling rate was 12 kHz. Fifty sentences were used for training and another 10 sentences

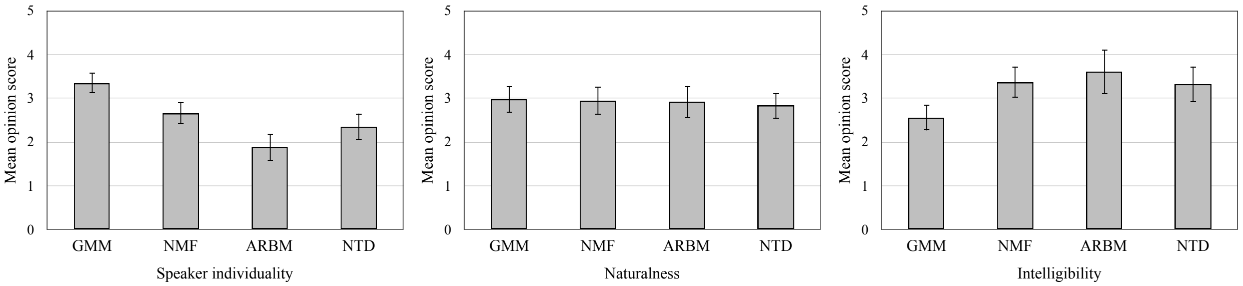

Fig. 3. Mean opinion scores (MOS) for each method

were used for testing. The GMM- and NMF-based methods were trained using parallel data that were aligned using dynamic programming matching (DPM). The maximum number of iterations is set to 300 for dictionary learning in NTD and 300 for conversion in NMF. Those parameters are chosen experimentally.

In the proposed method, a 513-dimensional STRAIGHT spectrum [22] is used as a spectral feature. We set the number of frequency bases M and time bases L to 1,000 and 200, respectively. In the conventional GMM-based method, mel-cepstrum is used as a spectral feature. Its number of dimensions is 24. The number of Gaussian mixtures in the GMM-based method was set to 64, which is experimentally selected. In the conventional dictionary-learning NMF-based method, the number of bases is 1,000. In the ARBM-based method, as an input vector, we used 32-dimensional mel-cepstrum that were calculated from the 513-dimensional STRAIGHT spectra. We set the number of hidden units as 128.

In this paper, F0 information is converted using a conventional linear regression based on the mean and standard deviation [5]. The other information, such as aperiodic components, is synthesized without any conversion.

4.2. Results and discussion

The subjective evaluation was conducted on “similarity to the target speaker (individuality)”, “naturalness” and “intelligibility”. For the subjective evaluation, 10 sentences were evaluated by 12 Japanese speakers. For the evaluation on speech quality, we performed a Mean Opinion Score (MOS) test. The opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad).

Fig. 3 shows the results of the subjective evaluation for each method. The error bars show 95% confidence intervals. In this figure, GMM and NMF are methods using parallel data, and ARBM and NTD are methods that do not use parallel data. As shown in this figure, the two methods that use parallel data obtained higher scores than those that do not use parallel data regarding the speaker individuality evaluation criteria. However, the MOS of the proposed method preserves speaker identity better than that of the conventional non-parallel method, ARBM. (The result is confirmed with a p-value test result of 0.05.) Regarding intelligibility evaluation criteria, the MOS of the proposed method is significant in the p-value test result of in 0.05 for a GMM-based method. The performances of the all methods were not so different regarding the naturalness evaluation criteria. These result shows that our proposed method effectively converts speaker individuality compared to conventional non-parallel VC methods, and has comparable performance to the parallel VC method. Although our proposed method is comparable to conventional NMF-based VC as shown in Fig. 3, our method slightly degraded in speaker individuality. For this reason, our proposed method is a non-parallel method. Moreover, for another reason, our proposed method does not have the sparse constraint that NMF-based methods have. NTD methods carry out more complex decomposition than NMF-based methods, so it seems that some constraints are required to obtain more stable performance. In this modeling, it is not guaranteed that the indices of the frequency bases and phones are matched between speakers. To further improve the speaker individuality, we will need to overcome this weakness.

5. CONCLUSIONS

In this paper, we proposed parallel dictionary learning for NMF-VC based on NTD that does not require parallel data during training. NTD decomposes an input observation into a set of mode matrices and one core tensor. In our proposed framework, the spectrogram is decomposed into the frequency basis matrix, the phonological information matrix, and the codebook matrix. In our experiments, we confirmed that our proposed method improves intelligibility compared to conventional GMM-based methods. Because the model parameters are estimated without any constraints in our proposed model, the corresponding relationship between the frequency bases and phones could not be obtained adequately. Nevertheless, we expanded NMF-based dictionary learning to a non-parallel method that achieved performance that is comparable to conventional NMF-based VC. For our future work, we will examine an apposite constraint and a model architecture to improve the performance.

6. REFERENCES

[1] Alexander Kain and Michael W. Macon, “Spectral voice conversion for text-to-speech synthesis,” in *ICASSP*, 1998, pp. 285–288.

[2] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano, “Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech,” *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[3] Hlne Valbret, Eric Moulines, and Jean-Pierre Tubach, “Voice transformation using PSOLA technique,” *Speech Communication*, vol. 11, no. 2-3, pp. 175–187, 1992.

[4] Yannis Stylianou, Olivier Capp, and Eric Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[5] Tomoki Toda, Alan W. Black, and Keiichi Tokuda, “Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[6] Elina Helander, Tuomas Virtanen, Jani Nurminen, and Moncef Gabbouj, “Voice conversion using partial least squares regression,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 18, no. 5, pp. 912–921, 2010.

[7] Ryoichi Takashima, Tetsuya Takiguchi, and Yasuo Ariki, “Exemplar-based voice conversion in noisy environment,” in *IEEE Workshop on Spoken Language Technology*, 2012, pp. 313–317.

[8] Ryoichi Takashima, Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, “Noise-robust voice conversion based on spectral mapping on sparse space,” in *Speech Synthesis Workshop*, 2013, pp. 71–75.

[9] Srinivas Desai, E. Veera Raghavendra, B. Yegnanarayana, Alan W. Black, and Kishore Prahallad, “Voice conversion using artificial neural networks,” in *ICASSP*, 2009, pp. 3893–3896.

[10] Ling-Hui Chen, Zhen-Hua Ling, Yan Song, and Li-Rong Dai, “Joint spectral distribution modeling using restricted boltzmann machines for voice conversion,” in *INTERSPEECH*, 2013, pp. 3052–3056.

[11] Toru Nakashika, Tetsuya Takiguchi, and Yasuhiro Minami, “Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine,” *IEEE/ACM Trans. Audio, Speech & Language Processing*, vol. 24, no. 11, pp. 2032–2045, 2016.

[12] Daniel D. Lee and H. Sebastian Seung, “Algorithms for non-negative matrix factorization,” in *NIPS*, 2000, pp. 556–562.

[13] Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, “Activity-mapping non-negative matrix factorization for exemplar-based voice conversion,” in *ICASSP*, 2015, pp. 4899–4903.

[14] Athanasios Mouchtaris, Jan Van der Spiegel, and Paul Mueller, “Nonparallel training for voice conversion based on a parameter adaptation approach,” *IEEE Trans. Audio, Speech & Language Processing*, vol. 14, no. 3, pp. 952–963, 2006.

[15] Daisuke Saito, Keisuke Yamamoto, Nobuaki Minematsu, and Keikichi Hirose, “One-to-many voice conversion based on tensor representation of speaker space,” in *INTERSPEECH*, 2011, pp. 653–656.

[16] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle, “A multilinear singular value decomposition,” *SIAM J. Matrix Anal. Appl.*, vol. 21, no. 4, pp. 1253–1278, 2000.

[17] L. R. Tucker, “Some mathematical notes on three-mode factor analysis,” *Psychometrika*, vol. 31, pp. 279–311, 1966.

[18] Ryo Aihara, Toru Nakashika, Tetsuya Takiguchi, and Yasuo Ariki, “Voice conversion based on non-negative matrix factorization using phoneme-categorized dictionary,” in *ICASSP*, 2014, pp. 7894–7898.

[19] Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki, “Parallel dictionary learning for voice conversion using discriminative graph-embedded non-negative matrix factorization,” in *INTERSPEECH*, 2016, pp. 292–296.

[20] Yong-Deok Kim and Seungjin Choi, “Nonnegative tucker decomposition,” in *CVPR*, 2007.

[21] Akira Kurematsu, Kazuya Takeda, Yoshinori Sagisaka, Shigeru Katagiri, Hisao Kuwabara, and Kiyohiro Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Communication*, vol. 9, no. 4, pp. 357–363, 1990.

[22] Hideki Kawahara, Ikuyo Masuda-Katsuse, and Alain de Cheveign, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.