LETTER

Statistical Approaches to Excitation Modeling in HMM-Based Speech Synthesis

June Sig SUNG[†a)], **Doo Hwa HONG**[†b)], **Hyun Woo KOO**[†c)], *Nonmembers*, *and* **Nam Soo KIM**[†d)], *Member*

SUMMARY In our previous study, we proposed the waveform interpolation (WI) approach to model the excitation signals for hidden Markov model (HMM)-based speech synthesis. This letter presents several techniques to improve excitation modeling within the WI framework. We propose both the time domain and frequency domain zero padding techniques to reduce the spectral distortion inherent in the synthesized excitation signal. Furthermore, we apply non-negative matrix factorization (NMF) to obtain a low-dimensional representation of the excitation signals. From a number of experiments, including a subjective listening test, the proposed method has been found to enhance the performance of the conventional excitation modeling techniques.

key words: *HMM-based speech synthesis, waveform interpolation, principal component analysis, non-negative matrix factorization*

1. Introduction

In previous studies on hidden Markov model (HMM)-based speech synthesis, emphasis has been placed on how to generate a natural trajectory of the spectrum parameters which account for the vocal tract characteristics. However, the issue of generating more realistic excitation signals is also very important to achieve naturally sounding speech. One of the simplest techniques for excitation modeling is to switch between the pulse train and noise depending on the voicing property, which usually turns out to provide poor speech quality. A number of attempts have been made to enhance excitation modeling. Yoshimura et al. apply a mixed excitation model to an HMM-based synthesizer [1] where the excitation signal is created by combining both the periodic impulse train and random noise with appropriate weights. A more sophisticated excitation model is proposed in [2], where the periodic impulse train and random noise are mixed after being passed through separate filters. As an enhanced mixed excitation method, STRAIGHT was developed by Kawahara [3]. In STRAIGHT, a mixed excitation is given as a weighted sum of a phase-manipulated pulse train and Gaussian noise.

In our previous work, we applied the waveform interpolation (WI) technique to an HMM-based speech synthesis system [4]. The WI framework enables us to analyze the excitation signals in the form of characteristic waveforms (CW's), which are described in the frequency domain as spectral coefficients. Each CW is given by a fixed-dimensional vector where the dimension equals the maximum pitch length, and its statistical distribution is approximated by HMM's. In order to reduce the dimension of each CW hence to improve robustness in training, we apply principal component analysis (PCA) to the extracted CW's.

In this letter, we propose an alternative representation of a CW to improve the performance of statistical handling. One of the significant drawbacks of the frequency domain representation of the CW in the previous work is that each element of the CW corresponds to a different frequency component if the dimension is adjusted to a fixed pitch length. For the purpose of alleviating this problem, we employ a time domain representation. In addition, we apply not only PCA but also non-negative matrix factorization (NMF) to obtain a compact representation of the CW's. The experimental results demonstrate that the proposed technique enhances the statistical representation of the CW's resulting in improved speech quality compared with the conventional excitation modeling techniques for HMM-based speech synthesis.

2. HMM-Based Speech Synthesis System with WI

The procedures in this work for modeling and synthesizing the spectrum and pitch parameter trajectories are identical to those adopted in a conventional HMM-based speech synthesis system [5]. In this section, we briefly describe the analysis and synthesis of excitation based on the WI technique [4].

Each CW is equivalent to a single pitch cycle of the excitation signal. A method to extract the CW from the given linear prediction residual is well explained in [6]. Let $s(n,m)$ denote the m-th sample of the CW extracted at the n-th frame. For convenience we let each CW be centered at $m = 0$. Then,

$$A_k(n) = \frac{2}{P(n)} \sum_{m=-\lfloor P(n)/2 \rfloor}^{\lfloor P(n)/2 \rfloor} \left[s(n,m) \cos\left(\frac{2km\pi}{P(n)} \right) \right] \quad (1)$$

$$B_k(n) = \frac{2}{P(n)} \sum_{m=-\lfloor P(n)/2 \rfloor}^{\lfloor P(n)/2 \rfloor} \left[s(n,m) \sin\left(\frac{2km\pi}{P(n)} \right) \right] \quad (2)$$

$$k = 1, 2, \ldots, (P(n) - 1)/2$$

where $P(n)$ is the pitch, which is given as a positive inte-

Manuscript received September 5, 2012.
Manuscript revised October 25, 2012.

[†]The authors are with the School of Electrical Engineering and the Institute of New Media and Communications, Seoul National University, Seoul 151–742, Korea.
a) E-mail: jssung@hi.snu.ac.kr
b) E-mail: dhhong@hi.snu.ac.kr
c) E-mail: hwkoo@hi.snu.ac.kr
d) E-mail: nkim@snu.ac.kr

DOI: 10.1587/transinf.E96.D.379

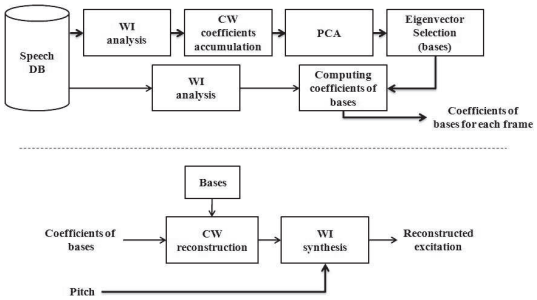

Fig. 1 The procedures for parameter extraction (top) and excitation generation (bottom) of the WI-based approach.

ger in this work, and $A_k(n)$ and $B_k(n)$ are the k-th discrete time Fourier series (DTFS) coefficients computed at frame n. In (1) and (2), $P(n)$ is assumed to be an odd integer. For the case when $P(n)$ is an even integer, a slight modification is required. Each CW is described in terms of the derived DTFS coefficients, $\{A_k(n), B_k(n)\}$, which can be considered a frequency-domain representation of the CW. Considering that each CW has a different dimension depending on the corresponding pitch period, zeros are appended so that all the coefficients of CW's can be described in the same fixed dimensional space. After the zeros are padded, each CW coefficient is converted to the magnitude CW where each element represents the magnitude of the corresponding CW coefficient. PCA is then applied to the covariance matrix of the magnitude CW's. As a result of the PCA, the eigenvector matrix is acquired, and M dominant eigenvectors are chosen as the basis for approximating each magnitude CW. By taking the inner product of a magnitude CW with these basis vectors, a reduced dimensional representation is obtained. During the synthesis process, the magnitude CW at a specific frame is reconstructed by a linear combination of the basis vectors with the corresponding coefficients which are the elements of the reduced-dimensional representation. The CW coefficients are computed by applying a default or a random phases to the reconstructed magnitude CW depending on the voicing type. The reconstructed CW, $s_R(n,m)$ is then generated by following

$$s_R(n,m) = \sum_{k=1}^{\lfloor P(n)/2 \rfloor} \left[A_k(n) \cos\left(\frac{2\pi km}{P(n)}\right) + B_k(n) \sin\left(\frac{2\pi km}{P(n)}\right) \right] \\ -(P(n)-1)/2 \le m \le (P(n)-1)/2. \quad (3)$$

Finally, the excitation signal is generated from the reconstructed CW's by following a continuous pitch track. Figure 1 describes the procedures for excitation parameter extraction and reconstruction under the WI framework.

As mentioned above, the frequency to which a specific element of a CW coefficient originally corresponds will be altered if zeros are padded in order to adjust the dimension to a fixed value. Since it is known that human auditory perception is more sensitive to spectral distortion than the temporal mismatch, this frequency domain representation of the CW's is likely to deteriorate the synthesized speech quality with the WI scheme. In order to alleviate this problem, we propose an alternative approach in this section. Basically, the proposed approach employs a time domain representation.

Let the maximum pitch length be denoted by D, which is assumed to be an odd integer. The basic idea of our approach is to perform zero padding in the time domain before computing the DTFS coefficients. Let $s'_D(n,m)$ denote the zero padded CW at frame n. Then,

$$s'_D(n,m) = \begin{cases} s(n,m) & \text{if } |m| \le P(n)/2 \\ 0 & otherwise. \end{cases} \quad (4)$$

The DTFS coefficients of $s'_D(n,m)$ is given by

$$A'_k(n) = \frac{2}{D} \sum_{m=-\lfloor D/2 \rfloor}^{\lfloor D/2 \rfloor} \left[s'_D(n,m) \cos\left(\frac{2\pi km}{D}\right) \right] \quad (5)$$

$$B'_k(n) = \frac{2}{D} \sum_{m=-\lfloor D/2 \rfloor}^{\lfloor D/2 \rfloor} \left[s'_D(n,m) \sin\left(\frac{2\pi km}{D}\right) \right] \quad (6)$$

$$k = 1, 2, \ldots, (D-1)/2.$$

Now, each CW is given by a $(D-1)/2$ dimensional vector consisting of $\{A'_k(n), B'_k(n)\}$. The reconstructed signal $s'_{DR}(n,m)$ is obtained straightforwardly as

$$s'_{DR}(n,m) = \sum_{k=1}^{\lfloor D/2 \rfloor} \left[A'_k(n) \cos\left(\frac{2\pi km}{D}\right) + B'_k(n) \sin\left(\frac{2\pi km}{D}\right) \right] \\ -(D-1)/2 \le m \le (D-1)/2. \quad (7)$$

We refer to the proposed approach as time domain zero padding (TDZ) in contrast to the previous technique which is called frequency domain zero padding (FDZ).

3. Low Dimensional Representation of CW's

3.1 Principal Component Analysis

For PCA, a covariance matrix $\mathbf{C}$ is constructed from the statistics of the given data vectors. The PCA method leads us to the following matrix factorization [7]:

$$\mathbf{U}^{-1}\mathbf{C}\mathbf{U} = \mathbf{D} \quad (8)$$

where $\mathbf{U}$ is a unitary matrix whose columns are eigenvectors of $\mathbf{C}$ and $\mathbf{D}$ is a diagonal matrix consisting of the corresponding eigenvalues arranged in a descending order.

Any vector can be uniquely expressed as a linear combination of the columns of $\mathbf{U}$. The usefulness of PCA lies on the fact that there exists a compact representation approximating the given data. Let $\mathbf{x}$ be an arbitrary D dimensional vector and M be the number of eigenvectors for compact representation. Then, it can be approximated as follows:

$$\tilde{\mathbf{x}} \simeq \sum_{i=1}^{M} \alpha_i \mathbf{u}_i + \bar{\mathbf{x}} \quad (9)$$

in which $M \ll D$, $\mathbf{u}_i$ is the i-th column of $\mathbf{U}$, α_i is the coefficient associated to $\mathbf{u}_i$, and $\bar{\mathbf{x}}$ is the mean of the data vectors. Since $\{\mathbf{u}_i\}$ forms an orthonormal basis, α_i, denoting a weight for the i-th basis vector, can be easily obtained by taking the inner product between $(\mathbf{x} - \bar{\mathbf{x}})$ and $\mathbf{u}_i$.

3.2 Non-negative Matrix Factorization

NMF is a signal analysis method in which the data matrix is factorized into two constrained matrices of non-negative elements [8]. When a collection of D-dimensional positive valued input data is represented by a $D \times q$ matrix $\mathbf{V}$ with q denoting the number of samples, it can be approximately factorized into two matrices $\mathbf{W}$ and $\mathbf{H}$ with dimensions $D \times M$ and $M \times q$, respectively, i.e.,:

$$\mathbf{V} \simeq \mathbf{WH}. \tag{10}$$

The way of factorizing a certain matrix is generally non-unique and a lot of methods have been developed with different constraints. NMF is different from the other methods in that it has the constraint that all the factors of $\mathbf{W}$ and $\mathbf{H}$ must be non-negative. For the factorization of an input data matrix with this constraint, we can apply the multiplicative update rules which find a suboptimal solution iteratively. In this work, we apply the algorithm presented in [8] where Euclidean distance is used as the measure which results in the following update rules:

$$H_{kj} \leftarrow H_{kj} \frac{(\mathbf{W}^T\mathbf{V})_{kj}}{(\mathbf{W}^T\mathbf{WH})_{kj}}, W_{ik} \leftarrow W_{ik} \frac{(\mathbf{VH}^T)_{ik}}{(\mathbf{WWH}^T)_{ik}}. \tag{11}$$

where T denotes the transpose of a matrix, and H_{kj} and W_{ik} are the (k, j)-th and (i, k)-th element of $\mathbf{H}$ and $\mathbf{W}$, respectively. After $\mathbf{W}$ and $\mathbf{H}$ are obtained from a set of training data, we use columns of $\mathbf{W}$ as the basis vectors. A reduced dimensional representation $\mathbf{a}$ for a vector $\mathbf{x}$ is derived according to

$$\mathbf{a} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{x} \tag{12}$$

where $\mathbf{a}$ denotes an M-dimensional weight vector and $\mathbf{x}$ is a D-dimensional input vector.

4. Experiments

To prove the effectiveness of the proposed method, several experiments on speech synthesis were conducted. A Korean speech database spoken by a male (HNC) and female (YMK) speakers was applied. For each speaker, 1,000 sentences were used for training the synthesizer and another 20 sentences were used for performance evaluation. Speech data was sampled at 16 kHz and quantized in 16 bits in conjunction with the information on the phone segmentation and context dependency. For feature extraction, speech waveforms were windowed by a 20 ms Hamming window with a 5 ms frame shift. The maximum length of CW was set to 320 in samples. At each frame, we extracted a CW and converted it to the fixed-dimensional magnitude CW vector with either the FDZ or TDZ technique. Examples of the excitation from a sample of the male speaker are compared in Fig. 2.

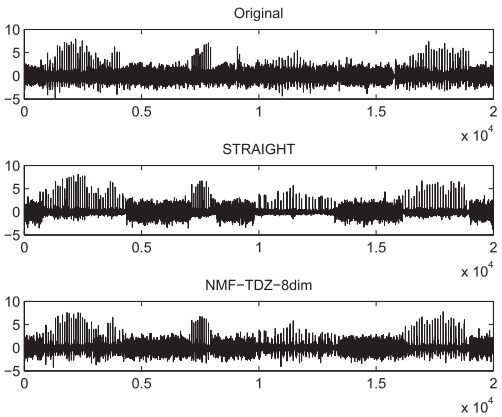

Fig. 2 Examples of excitation for male voice.

Table 1 Averaged approximation errors with low-dimensional representations.

		FDZ	TDZ
8 dim	PCA	0.4394	0.2245
	NMF	0.4376	0.2103
12 dim	PCA	0.3953	0.2033
	NMF	0.3987	0.1888
16 dim	PCA	0.3599	0.1865
	NMF	0.3586	0.1706
20 dim	PCA	0.3299	0.1705
	NMF	0.3302	0.1555

4.1 Performance of Low-Dimensional Representation

To investigate the usefulness of low dimensional representation via PCA and NMF, we measured the reconstruction error for the magnitude CW vectors. For this, we extracted 5,558 magnitude CW's from the actual excitations. Among them, 3,000 magnitude CW's were used to generate the basis, and the other 2,558 vectors were used to evaluate approximation errors. The number of basis vectors was set to 8, 12, 16, and 20 for both the PCA and NMF analyzes. Euclidian distance was used to compute mismatches between the original magnitude CW's and those reconstructed from the low-dimensional representation. Table 1 shows the averaged approximation errors.

From Table 1, we can see that TDZ reduces the approximation error more than FDZ, and NMF shows a slightly better performance than PCA.

4.2 Evaluation of Objective Measure

In the next experiment, we compared the original excitations with those synthesized by the proposed techniques. We measured the distance between the original magnitude CW's and those obtained through HMM-based synthesis. To do this, we extracted the magnitude CW's from the excitations of 20 test sentences which were not included in the training database. We carried out the experiment with different combinations of the low-dimensional representation (PCA and NMF) and the CW models (TDZ and FDZ). In addition, as

Table 2 Distance measurement between original magnitude CW's and those synthesized by proposed technique.

method \ speaker	HNC	YMK
PCA-FDZ-8dim	*3.6720*	*2.0472*
PCA-FDZ-20dim	4.4499	2.0473
PCA-TDZ-8dim	1.3465	0.7585
PCA-TDZ-20dim	1.5195	**0.7523**
NMF-FDZ-8dim	4.0257	2.0309
NMF-FDZ-20dim	6.4102	2.0701
NMF-TDZ-8dim	**1.3136**	0.7613
NMF-TDZ-20dim	1.3700	0.7549

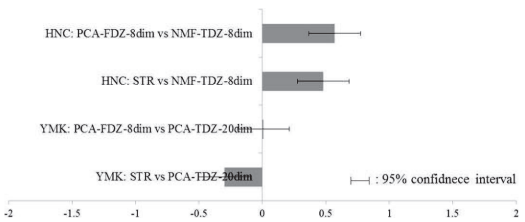

Fig. 3 Result of subjective listening tests.

for the number of basis vectors we tried $M = 8$ and 20. For each test condition, we computed averaged Euclidean distance between the actual magnitude CW and the corresponding synthesized magnitude CW.

The results are shown in Table 2 from which we can discover that TDZ much reduced distortion compared with FDZ in all the tested conditions. Note that more number of basis vectors does not always guarantee a better performance. This implies that some of the basis vectors of PCA and NMF did not help to improve the robustness of statistical modeling in the HMM-based framework.

4.3 Subjective Tests

We performed a set of subjective listening tests for which 11 listeners participated. In these tests, each listener was provided with two speeches synthesized by two different methods, and gave his/her preference as a score in the range $\{-2, 1, 0, 1, 2\}$.

To compare the performance between FDZ and TDZ, we applied PCA-FDZ-8dim for both HNC and YMK as the reference method, and NMF-TDZ-8dim for HNC and PCA-TDZ-20dim for YMK as the proposed method because these two methods produced the best results in the previous experiments. Furthermore, to compare the performance with that of other excitation generation technique, we synthesized excitation signals using the parameters provided by the STRAIGHT [5] method, which is denoted as STR in the result. The averaged listening preference scores are given in Fig. 3 where ′A vs B′ means that the techniques ′A′ and ′B′ were compared and a positive value indicates that ′B′ was preferred and vice versa.

From the results, we can see that TDZ produced better speech quality than FDZ for the speaker HNC while no significant difference was occured for the speaker YMK. In comparison with STRAIGHT, TDZ showed a better result for the male (HNC) voices while it produced slightly worse scores for the female voices. This phenomenon might be partly caused by the current implementation of the WI-based system in which the phases of the original CW's are ignored.

5. Conclusions

In this letter, we have proposed several approaches to improve the excitation modeling based on the WI framework for excitation generation in HMM-based speech synthesis. In the proposed method, we append zeros before extracting the DTFS coefficients of the CW's. In addition, NMF is applied to obtain a more efficient low-dimensional representation of the excitation signals. A number of experiments have proven that the proposed method improves the statistical representation of the CW's, resulting in enhanced speech quality compared with the conventional excitation modeling techniques.

For the future work, research on an appropriate phase modeling is required to further enhance the quality of the synthesized speech.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No.2012R1A2A2A01045874).

References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," Eurospeech2001, pp.2263–2266, 2001.
[2] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," ISCA SSW6, Aug. 2007.
[3] H. Kawahara and H. Matsui, "Auditory morphing based on an elastic perceptual distance metric in an interference-free time frequency representation," ICASSP2003, vol.I, pp.256–259, 2003.
[4] J.S. Sung, D.H. Hong, K.H. Oh, and N.S. Kim, "Excitation modeling based on waveform interpolation for HMM-based speech synthesis," Interspeech2010, pp.813–816, Sept. 2010.
[5] H. Zen, T. Toda, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard challenge 2006," Blizzard Challenge 2006 Workshop, Sept. 2006.
[6] E.L.T. Choy, Waveform interpolation speech coder at 4 kb/s. Master of Engineering Thesis, Department of Electrical Engineering, McGill University, Montreal, Canada, 1998.
[7] C.M. Bishop, Pattern Recognition and Machine Learning. pp.559–586, Springer, 2006.
[8] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," Advances in Neural Information Processing Systems, vol.13, pp.556–562, 2001.