Excitation Modeling Based on Waveform Interpolation for HMM-based Speech Synthesis

June Sig Sung, Doo Hwa Hong, Kyung Hwan Oh and Nam Soo Kim

Institute of New Media and Communications
School of Electrical Engineering and Computer Science
Seoul National University
{jssung, dhhong, khoh}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

It is generally known that a well-designed excitation produces high quality signals in hidden Markov model (HMM)-based speech synthesis systems. This paper proposes a novel techniques for generating excitation based on the waveform interpolation (WI). For modeling WI parameters, we implemented statistical method like principal component analysis (PCA). The parameters of the proposed excitation modeling techniques can be easily combined with the conventional speech synthesis system under the HMM framework. From a number of experiments, the proposed method has been found to generate more naturally sounding speech.

Index Terms: HMM-based speech synthesis, Waveform Interpolation, Principal Component Analysis

1. Introduction

Even though hidden Markov model (HMM) based speech synthesis shows a high performance for generating various speech styles using only a small size of parameters, the quality of the synthesized speech is still rated lower than that of the unit selection-based systems due to the inadequate modeling capability of the vocoder. A traditional vocoder employed in speech synthesis consists of two sets of parameters, where one is devoted to characterize the resonance structure of the vocal tract and the other is used to generate artificial excitation signals. Linear prediction (LP) analysis has been the predominant approach to approximating the vocal tract frequency response. As for the excitation generation, the most popular technique has been the pulse or noise (PoN) model in which a periodic impulse train and random noise are exclusively used for voiced and unvoiced parts, respectively. Even though the PoN model can be easily implemented with a very few number of parameters, the quality of the synthesized speech is usually poor and we can perceive unpleasant artifacts such as buzziness.

In order to enhance the poor quality of PoN model, many attempts have been made. Yoshimura et al. applied the mixed excitation (ME) model to the HMM-based synthesizer [1]. In this method the excitation signal is created by combining both the periodic impluse train and the random noise with an appropriate weight. In [2], a two-band excitation (TBE) model was proposed. TBE model separates the whole frequency range into high and low bands with respect to a cut-off frequency, and the periodic impulse train and the random noise are respectively applied to constant the low band and high band components of the excitation signal. An excitation generation model more sophisticated than ME and TBE was proposed in [3] where the periodic impulse train and the random noise are mixed after passing through separate filters. The parameters of these two separate filters are optimized so as to minimize the difference between the original and synthesized speech. The data-driven method was also employed to model excitation [4]. In here, principal component analysis (PCA) was applied to time-domain signal for getting excitation parameter which is used for generating excitation.

In this paper, we propose a novel approach to excitation modeling under the waveform interpolation (WI) framework. For parameterizing the excitation generation model, a characteristic waveform (CW) is extracted from each frame of LP residual signals. To derive a compact representation of each CW, we apply principal component analysis (PCA) to a collection of the extracted CW's. Once PCA is done, each CW can be compactly approximated as a linear combination of a few PCA basis vectors. The statistical distribution of the linear combination coefficients and their dynamics can be efficiently described by means of HMM's for which the relevant parameters are estimated by following the conventional HMM training procedure. Given a sentence we want to synthesize, the sequence of CW's can be generated from the trained HMM's according to the maximum likelihood (ML) criterion. The WI algorithm enables a smooth transition between adjacent CW's resulting in a more natural excitation signal. The major advantages of the proposed technique are twofold. First, instead of using a fixed set of waveforms such as the impulse train and the random noise, the proposed method finds CWs which represents the excitation waveforms from the various kinds of modeling in frequency domain. Second, the WI approach lets the excitation signal evolve smoothly, which may reduce the audible artifacts of the synthesized speech. From a number of experiments on speech synthesis, it has been demonstrated that the propose technique enhances the quality of the synthesized speech.

2. Waveform Interpolation

WI was first introduced by Kleijn with the name Prototype Waveform Interpolation (PWI) [5], and has been further refined and extended since then. In the WI framework, each cycle of the excitation signal is represented by a CW. There are a variety of ways to describe a CW, and we use the method based on the discrete time Fourier series (DTFS) analysis in this work.

Let $s(n, m)$ denote the m-th sample of the CW extracted at the n-th frame. Then,

$$s(n,m) = \sum_{k=0}^{P(n)/2} [A_k(n)\cos(\tfrac{2\pi km}{P(n)}) + B_k(n)\sin(\tfrac{2\pi km}{P(n)})], \quad 0 \le m < P(n) \tag{1}$$

$$A_k(n) = \frac{2}{P(n)} \sum_{m=0}^{P(n)-1} \left[s(n,m)\cos\left(\frac{2\pi km}{P(n)}\right)\right]$$
$$B_k(n) = \frac{2}{P(n)} \sum_{m=0}^{P(n)-1} \left[s(n,m)\sin\left(\frac{2\pi km}{P(n)}\right)\right]$$
$$k = 1, 2, ..., (P(n)-1)/2 \tag{2}$$

where $P(n)$ is the pitch period and $A_k(n)$ and $B_k(n)$ are the k-th DTFS coefficients computed at frame n. When deriving (1) and (2), we have assumed that the pitch period $P(n)$ is odd. A slight modification is required if $P(n)$ is even [6]. For the convenience of generating the phase track, it is usually better to modify (1) into

$$s(n,\phi) = \sum_{k=0}^{P(n)/2} [A_k(n)cos(k\phi) + B_k(n)sin(k\phi)] \tag{3}$$

by which, all the CW's have the same length of 2π. After applying some approximations, the phase track is obtained as follows:

$$\phi(n) \simeq \phi(n-1) + \pi\left(\frac{1}{P(n-1)} + \frac{1}{P(n)}\right) \tag{4}$$

where $\phi(n)$ denotes the instantaneous phase at the n-th frame. Fig. 1 shows the whole analysis and synthesis procedures of WI.

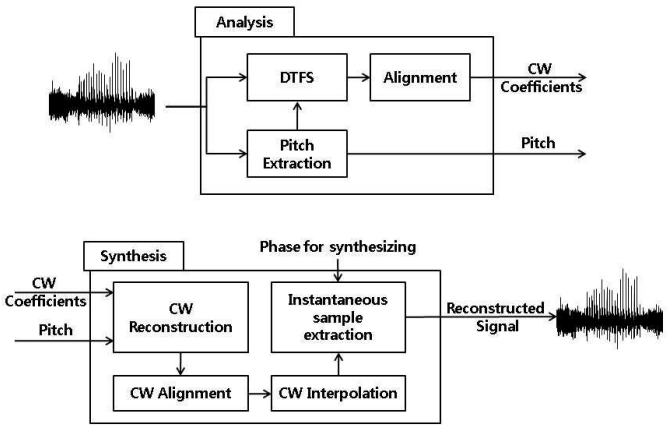

Figure 1: WI analysis (top) and synthesis (bottom) procedures.

In the WI technique, CW is extracted from the LP residual at a fixed rate, and the sampled CW's are interpolated to fill up the remaining intervals. Care must be taken when we interpolate adjacent CW's. First, the CW's should be modified such that they can be described in the same dimensional space. This operation is needed when the pitch periods of the CW's are different from each other, and pitch doubling or halving occurs. The next step is to align the CW's such that they have the same phase offset. Once these preprocessing stages are completed, interpolated waveform is generated from the interpolated DTFS coefficients of the adjusted CW's. For a detailed implementation of the WI algorithm and its wide-band version, the readers are referred to [6] [7].

3. Excitation Modeling based on PCA

In this section, we propose a technique to approximate a CW based on the PCA approach. PCA is very popular in various applications such as dimensionality reduction, lossy data compression and feature extraction [8]. For PCA, a covariance matrix $\mathbf{C}$ is constructed from the statistics of the given data vectors. Let $\mathbf{x}_n$ be the n-th data vector of dimension D. Then, the covariance matrix is computed as

$$\mathbf{C} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T \tag{5}$$

where N denotes the total number of samples, $\bar{\mathbf{x}}$ is the sample mean and T indicates the transpose of a matrix. The PCA method leads us to the following matrix factorization:

$$\mathbf{U}^{-1}\mathbf{C}\mathbf{U} = \mathbf{D} \tag{6}$$

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:

$$\tilde{\mathbf{x}} = \sum_{i=1}^{M}\alpha_i\mathbf{u}_i + \bar{\mathbf{x}} \tag{7}$$

in which $M \ll D$, $\mathbf{u}_i$ is the i-th column of $\mathbf{U}$, and α_i is the coefficient associated to $\mathbf{u}_i$. Since $\{\mathbf{u}_i\}$ forms an orthonormal basis, α_i can be easily obtained by taking the inner product between $(\mathbf{x} - \bar{\mathbf{x}})$ and $\mathbf{u}_i$. If the covariance matrix $\mathbf{C}$ has few dominant eigenvalues, (7) results in a very small approximation error and the original data can be represented in a more compact way.

Based on PCA, the excitation modeling procedures are given as follows: First, a large number of CW's are extracted from the LP residuals of the training data. Then, the conventional PCA is performed to the covariance matrix of the accumulated CW. A proper number of eigenvectors, denoted by M, is chosen considering the eigenvalue distribution. The corresponding eigenvectors are treated as the basis for the subspace onto which each CW is projected. By (7), each D-dimensional CW is described in terms of a M-dimensional coefficients vector, $[\alpha_1, \alpha_2, \cdots \alpha_M]^T$.

Once the PCA basis, $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_M\}$ is determined, the CW extracted from each frame of the training data is projected onto the subspace spanned by this basis, and the corresponding coefficients vector $[\alpha_1, \alpha_2, \cdots \alpha_M]^T$ is calculated. The coefficients vector obtained at each frame serves an additional feature vector to the conventional HMM-based speech synthesis system. Each state of the HMM has a distribution of the coefficients vector, and the parameters of this distribution are estimated through a usual HMM training procedure. For an efficient modeling of the dynamic variation of the excitation signal,

the coefficients vector can be used in conjunction with its delta and delta-delta parameters. Fig. 2 describes the parameter extraction from CW and reconstruction procedures.

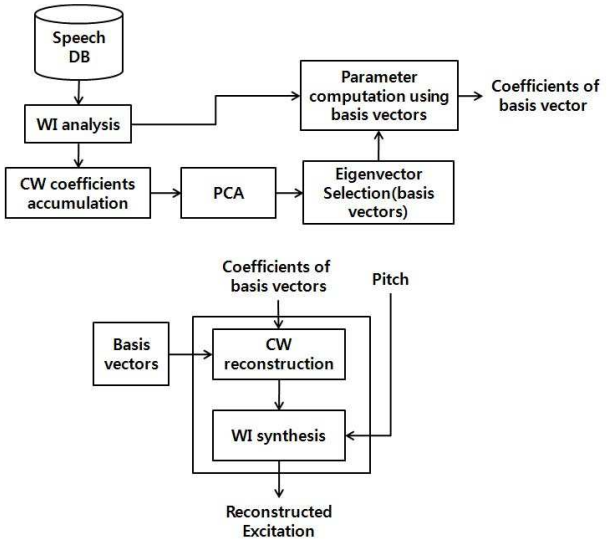

Figure 2: Parameter extraction from CW (top) and CW reconstruction from parameter (bottom) procedures using PCA.

4. Experiments

To show the effectiveness of the proposed method, speech data uttered by a Korean male speaker was applied in the experiments. The speech material consisted of 3200 phonetically balanced utterances in which extracted by similar method in [9]. Speech signals were sampled at 16kHz and quantized in 16 bits, with information such as the phoneme segmentation and context dependency. We used 43 phonemes including silence, and the following contextual factors were taken into account:

- preceding, current, succeeding phonemes
- preceding, current, succeeding syllable types
- syllable position in current word from beginning
- syllable position in current phrase from beginning
- word position in current phrase from beginning
- number of syllables in current word and phrase
- number of words in current phrase
- symbol in current word

For feature extraction, speech signals were windowed by a 20ms Hamming window with a 5ms frame shift. At each frame, we extracted 25 mel-cepstral coefficients and pitch for the characterization of the spectrum and periodicity. A 78-dimensional feature vector was obtained by appending the delta and delta-delta parameters. We applied a 5 state left-to-right HMM with no skip for each context dependent phone. The HMM's were trained by following the procedure given in [10]. As a result of training, we obtained decision trees for spectrum, pitch and duration for each phone-sized unit.

For excitation modeling, we applied 25th-order LP analysis to the speech data and generated LP residuals. A CW was extracted at each frame from the LP residual based on the technique proposed in [6]. Since each CW had different dimension depending on the corresponding pitch period, we appended zeros so that all the CW's could be described in the same dimensional space. If P_{max} denotes the maximal pitch period, each CW becomes a vector in the P_{max}-dimensional space. Even though the PCA can be applied directly to the original CW's, we found that it was more useful to apply it to the magnitude spectra. For this reason, we converted each CW to the magnitude CW where each element represented the magnitude of the corresponding CW components. If A_k and B_k are the k-th components of a CW as given in (3), then the k-th component of the magnitude CW, γ_k is given by $\gamma_k = \sqrt{A_k^2 + B_k^2}$. Generally, a phase alignment process is important in WI. but we did not consider it because we converted CW to magnitude CW which neglected a phase effect.

PCA was applied to the covariance matrix of the magnitude CW's. We chose 8 eigenvectors as the basis for approximating each CW. Once the 8 projection coefficients were computed, a 24-dimensional feature vector was obtained by appending the delta and delta-delta parameters. HMM's for the coefficients of the magnitude CW were trained according to the conventional ML framework.

When a text is given, appropriate models for the spectrum, pitch, duration and coefficients of excitation basis should be selected through text analysis. Then, according to the ML criterion, the trajectory of each feature vector is determined [11]. As for the excitation signal, only the magnitude CW could be reconstructed by a linear combination of the 8 basis vectors with the corresponding coefficients. However, to generate an excitation waveform, the phase of each CW component should be also available. For this, we applied a default phase when the phone unit was decided to be voiced and a random phase if it was unvoiced. The default phase was extracted from a representative example of a voiced sound, and we found that the speech quality did not make a great difference if the default phase was extracted from other voiced sounds. Finally, the speech was synthesized from the generated excitations and spectrum using the mel-log spectral approximation (MLSA) filter. Fig. 3 shows all the subblocks of training and synthesis.

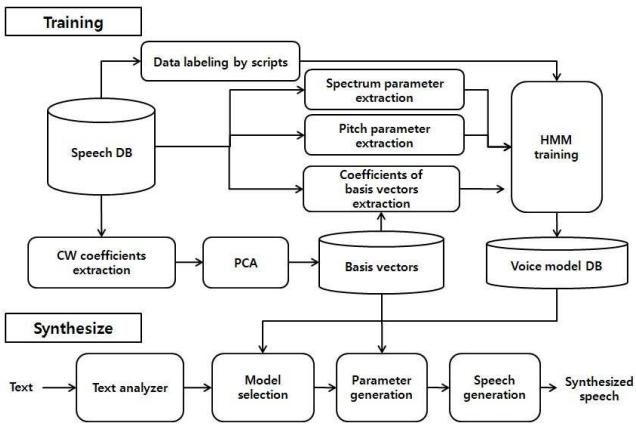

Figure 3: HMM training and synthesis procedures including excitation modeling.

Fig. 4 shows an example of the generated excitation from a Korean voiced syllable /a/ sampled from the sentence which is not included in the training. It can be seen that the excitation generated from proposed method are getting closer to those from the real speech. Note that because a real pitch and generated pitch to be synthesized are different, it makes a different

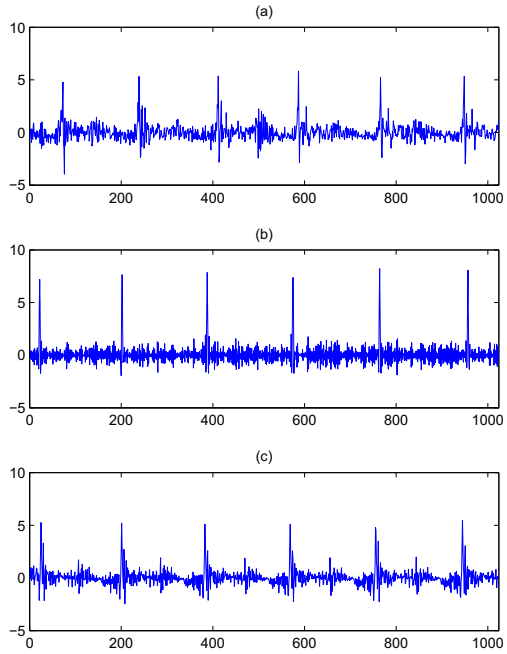

Figure 4: Excitation comparison between real and generated ones. (a) excitation from real speech. (b) generated excitation by TBE. (c) generated excitation by the proposed method.

interval of impulses between the real and generated excitation in the Fig. 4.

In order to evaluate the performance of the proposed technique, subjective listening test was carried out. For the test, twenty-five sentences were synthesized with different excitation modeling schemes. We compared the proposed method with PoN and TBE [2]. The speech quality was measured in terms of paired comparison where for each test a pair of two speech files were given and the subject provided the relative quality of the latter file compared to the former in five scales: 2 (much better), 1 (better), 0 (about the same), -1 (worse), -2 (much worse). Fourteen listeners participated in the tests and the result is shown in Fig. 5. From the result, we can see that the proposed method generated a high quality speech compared with PoN and was better than TBE.

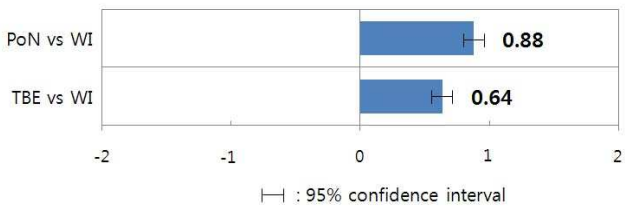

Figure 5: Results of the subjective listening test.

5. Conclusions

In this paper, we have proposed an approach to excitation modeling based on WI framework for speech synthesis. In our proposed method, the magnitude CW of the excitation is compactly expressed as a linear combination of a few basis vectors which are extracted from the training data using PCA. The coefficients of basis vectors are incorporated to the HMM system as additional features. The subjective quality tests have shown that the proposed method generates a good synthesized speech quality.

For future work, it can be applied to other statistical methods for modeling CW that show a better representation. Furthermore, more natural speech will be achieved if an appropriate phase modeling can be employed to modeling for phase of CW.

6. Acknowledgements

This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2009-0083044) and by Samsung Electronics Co. LTD.

7. References

[1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Mixed Excitation for HMM-based Speech Synthesis", In Eurospeech 2001, 2263-2266, 2001.

[2] S.-J. Kim, M.-S. Hahn, "Two-band excitation for HMM-based speech synthesis", IEICE Trans. Inf. & Syst., vol.E90-D, no.1, pp.378-381, Jan. 2007.

[3] R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling", Proc. ISCA SSW6, Aug. 2007.

[4] T.Drugman, G.Wilfart, T.Dutoit, "A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis", Interspeech2009, Brighton, U.K, 2009.

[5] W.B. Kleijn, "Continuous representations in linear predictive coding", Proc. IEEE Int. Conf. on Acoustics, Speech, signal Processing(Toronto), pp.201-204, May 1991.

[6] Eddie 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.H. Ritz, I.S. Burnett, J. Lukasiak, "Extending waveform interpolation to wideband speech coding", Speech coding, IEEE Workshop Proceedings, pp.32-34, Oct 2002.

[8] Christopher M. Bishop, *Pattern Recognition and Machine Learning.* pp.559-586, Springer, 2006

[9] A.W. Black, K.H. Lenzo, "Optimal data selection for unit selection synthesis", 4th ISCA Tutorial and Research Workshop (ITRW), 2001.

[10] K. Tokuda, H. Zen, J. Yamagishi, T. Masuko, S. Sako, A.W. Black, T. Nose, "The HMM-based speech synthesis system (HTS)", http://hts.ics.nitech/ac.jp.

[11] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis", *Proc. of ICASSP*, pp.1315-1318, June 2000.