

# A trainable excitation model for HMM-based speech synthesis

R. Maia<sup>†</sup>, T. Toda<sup>†,‡</sup>, H. Zen<sup>††</sup>, Y. Nankaku<sup>††</sup>, K. Tokuda<sup>†,††</sup>

<sup>†</sup>National Inst. of Inf. and Comm. Tech. (NiCT) / ATR Spoken Language Comm. Labs, Japan

<sup>‡</sup>Nara Institute of Science and Technology, Japan

<sup>††</sup>Nagoya Institute of Technology, Japan

ranniery.maia@atr.jp, tomoki@is.naist.jp, {zen,nankaku,tokuda}@sp.nitech.ac.jp

## Abstract

This paper introduces a novel excitation approach for speech synthesizers in which the final waveform is generated through parameters directly obtained from Hidden Markov Models (HMMs). Despite the attractiveness of the HMM-based speech synthesis technique, namely utilization of small corpora and flexibility concerning the achievement of different voice styles, synthesized speech presents a characteristic *buzziness* caused by the simple excitation model which is employed during the speech production. This paper presents an innovative scheme where mixed excitation is modeled through closed-loop training of a set of state-dependent filters and pulse trains, with minimization of the error between excitation and residual sequences. The proposed method shows effectiveness, yielding synthesized speech with quality far superior to the simple excitation baseline and comparable to the best excitation schemes thus far reported for HMM-based speech synthesis.

**Index Terms:** Speech Processing, Speech Synthesis, HMM.

## 1. Introduction

In the last years the speech synthesis technique in which the final waveform is generated by parameters directly obtained from Hidden Markov Models (HMMs) [1] has emerged as a good choice for synthesizing speech with different voice styles and characteristics, e.g. [2]. Nevertheless, synthesized speech for this technique presents a certain unnaturalness degree due to the waveform generation part, that consists in a source-filter model wherein the excitation is assumed to be either a periodic pulse train or a white noise sequence. Consequently, the same kind of artifacts that are usually observed in linear prediction (LP) vocoding [3] are noticed in the synthesized speech.

Although many attempts have been made to solve the unnaturalness problem of LP vocoders, the most successful reported approach has been the one in [4], which eventually became the Mixed Excitation Linear Prediction (MELP) algorithm [3]. Focusing on this idea, an excitation scheme for HMM-based synthesis was proposed in [5], which concerned the modeling of MELP parameters jointly with mel-cepstral coefficients and  $F_0$ . Following the same direction, in order to utilize the high-quality vocoding technique described in [6], the modeling of aperiodicity components with consequent application to HMM-based synthesis was performed by Zen et al. [7]. Approaches related to harmonic plus noise and sinusoidal models have also been reported, e.g. [8]. The common aspect among these methods is the fact of being based on the implementation of an excitation model through the utilization of some *special parameters* modeled by HMMs. Ideas related to the minimization of the distortion between artificial excitation and speech residual have

not been exploited until now.

In the method described in this paper, mixed excitation is produced by inputting pulse train and white noise into two state-dependent filters. These specific states may be represented, for instance, by leaves of phonetic decision-trees. The filters are derived so as to maximize the likelihood of residual sequences over the corresponding states through an iterative process. Aside from filter determination, the amplitudes and positions of the pulse trains are also optimized in the sense of residual likelihood maximization during the referred closed-loop training. Although some analysis-by-synthesis methods (similar to Code-Excited Linear Prediction (CELP) speech coding algorithms [3]) have already been proposed for unit concatenation-based speech synthesis, e.g. [9], the present approach targets residual instead of speech and considers the error of the system as the unvoiced component for the waveform generation part.

This paper is organized as follows: Section 2 outlines the proposed method; Section 3 presents some experiments; and the conclusions are in Section 4.

## 2. Mixed excitation by residual modeling

### 2.1. The idea

The proposed excitation model is depicted in Figure 1. The input pulse train,  $t(n)$ , and white noise sequence,  $w(n)$ , are filtered through  $H_v(z)$  and  $H_u(z)$ , respectively, and added together to result in the excitation signal  $e(n)$ . The voiced and unvoiced filters,  $H_v(z)$  and  $H_u(z)$ , respectively, are associated with each HMM state  $s = \{1, \dots, S'\}$ , as depicted in Figure 1, and their transfer functions are given by

$$H_v^s(z) = \sum_{l=-M/2}^{M/2} h_s(l)z^{-l}, \quad (1)$$

$$H_u^s(z) = \frac{K_s}{1 - \sum_{l=1}^L g_s(l)z^{-l}}, \quad (2)$$

where  $M$  and  $L$  are the respective orders.

#### 2.1.1. Function of $H_v(z)$

The function of the voiced filter  $H_v(z)$  is to transform the input pulse train  $t(n)$ , yielding the signal  $v(n)$  which is intended to be as similar as possible to the residual sequence  $e(n)$ . Because pulses are mostly considered in the voiced regions,  $v(n)$  is referred to as the voiced excitation component. The property of having finite impulse response leads to stability and phase information retention. Further, since the final waveform is synthesized offline, an anti-causal structure is advantageous in terms of resolution.

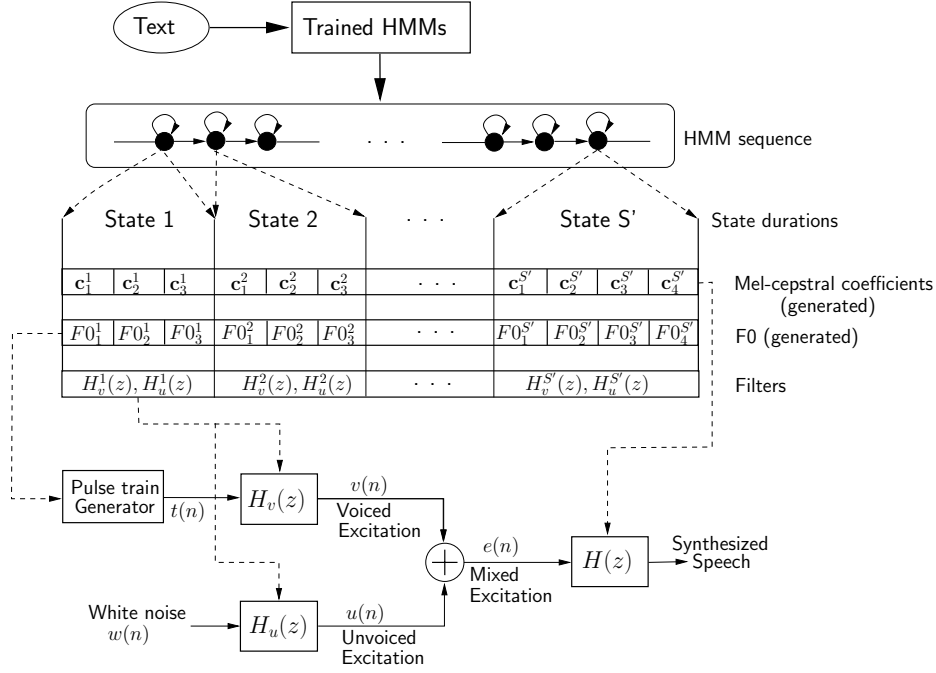


Figure 1: Proposed excitation scheme for HMM-based speech synthesis: filters  $H_v(z)$  and  $H_u(z)$  are associated with each state  $s$ .

### 2.1.2. Function of $H_u(z)$

Because white noise is assumed to be the input of the unvoiced filter, the function of  $H_u(z)$  is thus to weight the noise - in terms of spectral shape and power - which is eventually added to the voiced excitation  $v(n)$ . The reason for the all-pole structure is elucidated in Section 2.3.

## 2.2. Problem formulation

Through the re-arrangement of the excitation construction block of Figure 1, Figure 2 can be obtained. In this case pulse train and speech residual correspond to the input of the system whereas white noise is the output, as a result of the filtering of  $u(n)$  through the inverse unvoiced filter  $G(z)$ .

By observing the system shown in Figure 2, an analogy with analysis-by-synthesis speech coders [3] can be made as follows. The target signal is represented by the residual  $e(n)$ , the error of the system is  $w(n)$ , and the terms whose incremental modification can minimize the power of  $w(n)$  are the filters and pulse train. Therefore, according to this interpretation, the problem of achieving an excitation signal whose waveform can be as close as possible to the residual will consist in the design of  $H_v(z)$  and  $H_u(z)$ , and optimization of the positions,  $\{p_1, \dots, p_Z\}$ , and amplitudes,  $\{a_1, \dots, a_Z\}$ , of  $t(n)$ .

## 2.3. Filter determination

### 2.3.1. Likelihood of $e(n)$ given the excitation model

The likelihood of the residual vector  $\mathbf{e} = [e(0) \dots e(N-1)]^T$ , with  $[\cdot]^T$  meaning transposition and  $N$  being the whole database length in number of samples<sup>1</sup>, given the voiced excitation vector  $\mathbf{v} = [v(0) \dots v(N-1)]$  and  $\mathbf{G}$ , is

$$P[\mathbf{e}|\mathbf{v}, \mathbf{G}] = \frac{1}{\sqrt{(2\pi)^N (|\mathbf{G}^T \mathbf{G}|)^{-1}}} e^{-\frac{1}{2} [\mathbf{e} - \mathbf{v}]^T \mathbf{G}^T \mathbf{G} [\mathbf{e} - \mathbf{v}]}, \quad (3)$$

<sup>1</sup>The entire database is considered to be contained in a single vector.

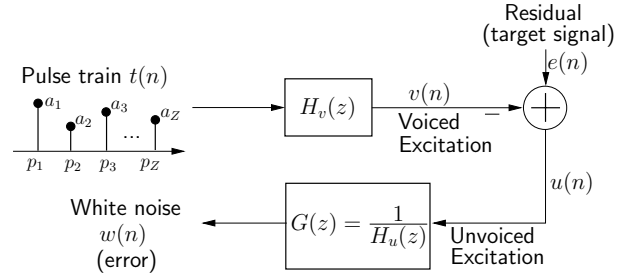


Figure 2: Re-arrangement of the excitation part: pulse train and residual are the input while white noise is the output.

where  $\mathbf{G}$  is an  $N \times N$  matrix containing the overall impulse response of the inverse unvoiced filter  $G(z)$  including all the states  $\{1, \dots, S\}$ , with  $S$  being the number of states considering the entire database. Since the filters are state-dependent,  $\mathbf{v}$  can be written as

$$\mathbf{v} = \sum_{s=1}^S \mathbf{A}_s \mathbf{h}_s = \mathbf{A}_1 \mathbf{h}_1 + \dots + \mathbf{A}_S \mathbf{h}_S, \quad (4)$$

where  $\mathbf{h}_s = [h_s(-M/2) \dots h_s(M/2)]^T$  is the impulse response vector of the voiced filter for state  $s$ , and the term  $\mathbf{A}_s$  is the overall pulse train matrix where only pulse positions belonging to state  $s$  are non-zero.

After substituting (4) into (3), and taking the logarithm, the following expression can be obtained for the log likelihood of  $\mathbf{e}$ , given the filters  $H_v(z)$  and  $H_u(z)$  and pulse train  $t(n)$ ,

$$\log P[\mathbf{e}|H_v(z), H_u(z), t(n)] = -\frac{N}{2} \log 2\pi + \frac{1}{2} \log |\mathbf{G}^T \mathbf{G}| - \frac{1}{2} \left[ \mathbf{e} - \sum_{s=1}^S \mathbf{A}_s \mathbf{h}_s \right]^T \mathbf{G}^T \mathbf{G} \left[ \mathbf{e} - \sum_{s=1}^S \mathbf{A}_s \mathbf{h}_s \right]. \quad (5)$$

### 2.3.2. Determination of $H_v(z)$

For a given state  $s$ , the corresponding vector of coefficients  $\mathbf{h}_s$  which maximizes (5) is determined from

$$\frac{\partial \log P[\mathbf{e}|H_v(z), H_u(z), t(n)]}{\partial \mathbf{h}_s} = 0. \quad (6)$$

The expression above results in

$$\mathbf{h}_s = \left[ \mathbf{A}_s^T \mathbf{G}^T \mathbf{G} \mathbf{A}_s \right]^{-1} \mathbf{A}_s^T \mathbf{G}^T \mathbf{G} \left[ \mathbf{e} - \sum_{\substack{l=1 \\ l \neq s}}^S \mathbf{A}_l \mathbf{h}_l \right], \quad (7)$$

that corresponds to the least-squares formulation for filter design by the solution of an over-determined linear system [10].

### 2.3.3. Determination of $H_u(z)$

Based on the assumption that  $w(n)$  is white noise, the function of  $H_u(z)$  is thus to remove long and short-term correlation from the unvoiced excitation  $u(n)$ . For such purpose the all-pole structure based on LP coefficients shown in (2) with large  $L$  seems to be a good choice, due to its simplicity in terms of computational complexity. Therefore, the coefficients of the unvoiced filter for state  $s$ ,  $\{g_s(1), \dots, g_s(L)\}$ , and related gain,  $K_s$ , are determined through an autoregressive spectral estimation [11] of  $u(n)$  over speech segments belonging to state  $s$ .

## 2.4. Pulse optimization

Aside from the determination of the filters, the positions and amplitudes of  $t(n)$ ,  $\{p_1, \dots, p_Z\}$  and  $\{a_1, \dots, a_Z\}$ , with  $Z$  being the number of pulses of the entire training database, are modified in the sense of minimizing the mean squared error of the system of Figure 2, given by

$$\varepsilon = \frac{1}{N} \mathbf{w}^T \mathbf{w} = \frac{1}{N} \left[ \mathbf{e} - \sum_{s=1}^S \mathbf{A}_s \mathbf{h}_s \right]^T \mathbf{G}^T \mathbf{G} \left[ \mathbf{e} - \sum_{s=1}^S \mathbf{A}_s \mathbf{h}_s \right]. \quad (8)$$

In fact, it can be verified that the minimization of (8) corresponds to the maximization of (5) when  $G(z)$  is minimum phase, which is assured by the assumption that  $H_u(z)$  is stable.

The procedure for calculating the positions and amplitudes resembles multipulse excitation linear prediction coding algorithms [3]. For the present case the search range is performed towards each pulse position  $\{p_1, \dots, p_Z\}$ .

## 2.5. Recursive algorithm

The overall procedure for the determination of the filters  $H_v(z)$  and  $H_u(z)$ , and optimization of the positions and amplitudes of  $t(n)$  is depicted in Figure 3. Pitch marks may represent the best choice to construct the initial pulse trains  $t(n)$ . Furthermore, initialization of the voiced filters as the unit sample sequence  $\delta(n)$  consists in a good approximation considering that the pitch marks may indicate the actual pitch pulses in  $e(n)$ . The variation of the voiced filter coefficients is taken into account as the convergence criterion.

## 3. Experiment

In order to verify the effectiveness of the proposed excitation approach, the English speech database CMU-ARCTIC [12], female speaker SLT, was utilized. For the present case, the states  $\{1, \dots, S\}$  were regarded as leaves of phonetic decision-trees generated for mel-cepstral coefficients. The choice of the states

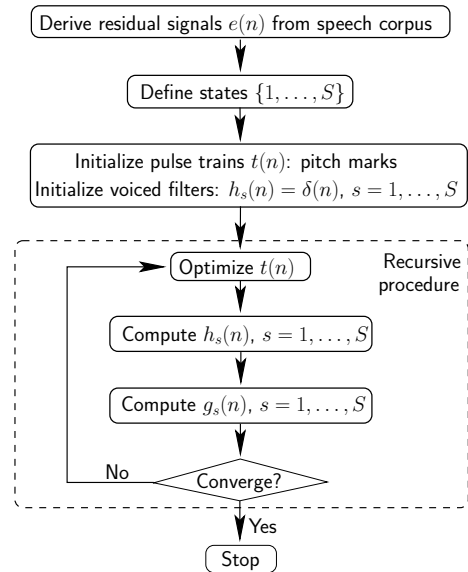


Figure 3: Closed-loop training for the determination of the excitation model.

was made upon the assumption that residual signals are highly correlated with their corresponding spectral parameters [13]. The trees were constructed with the utilization of phone-based questions and the clustering parameters were adjusted to derive small trees, in order to group general phonetic characteristics within each terminal node. Under these conditions, the number of states obtained in the end of the clustering process was a hundred and thirty-one, i.e.,  $S = 131$ .

Residual segmentation was performed as follows. First, Viterbi alignment of the database was done using the usual context-dependent models yielded by the training of the HMM-based synthesizer. After that, the contextual labels of the resulting segments were mapped onto the states of the above described small phonetic trees. Finally, the appropriately tagged segments were used to train the excitation model. Filter orders were  $M = 512$  and  $L = 256$ .

### 3.1. Effect of the closed-loop training

The right side of Figure 4 shows a segment of natural speech with three corresponding versions synthesized by using: pulse train/white noise switch (simple excitation) - second row; the mixed excitation approach utilized in the HMM-based speech synthesizer for the Blizzard Challenge 2005 [7] - third row; the proposed excitation model - fourth row. The related sources are depicted in the left side of the figure. One can observe that among the excitation and synthesized speech versions, the proposed method approaches more residual and natural speech waveforms, as an effect of the closed-loop training where phase information is also modeled by the voiced filters.

### 3.2. Subjective evaluation

A subjective evaluation was performed with six subjects. Two of the listeners were speech synthesis specialists whereas one of the other four subjects had English as native language. The test consisted in an AB forced preference, with ten utterances randomly selected out of forty sentences, taken from the BTEC corpus [14]. The order in which the utterances were played within each test pair was also randomized. The experiment was executed in a quiet room with the utilization of headphones.

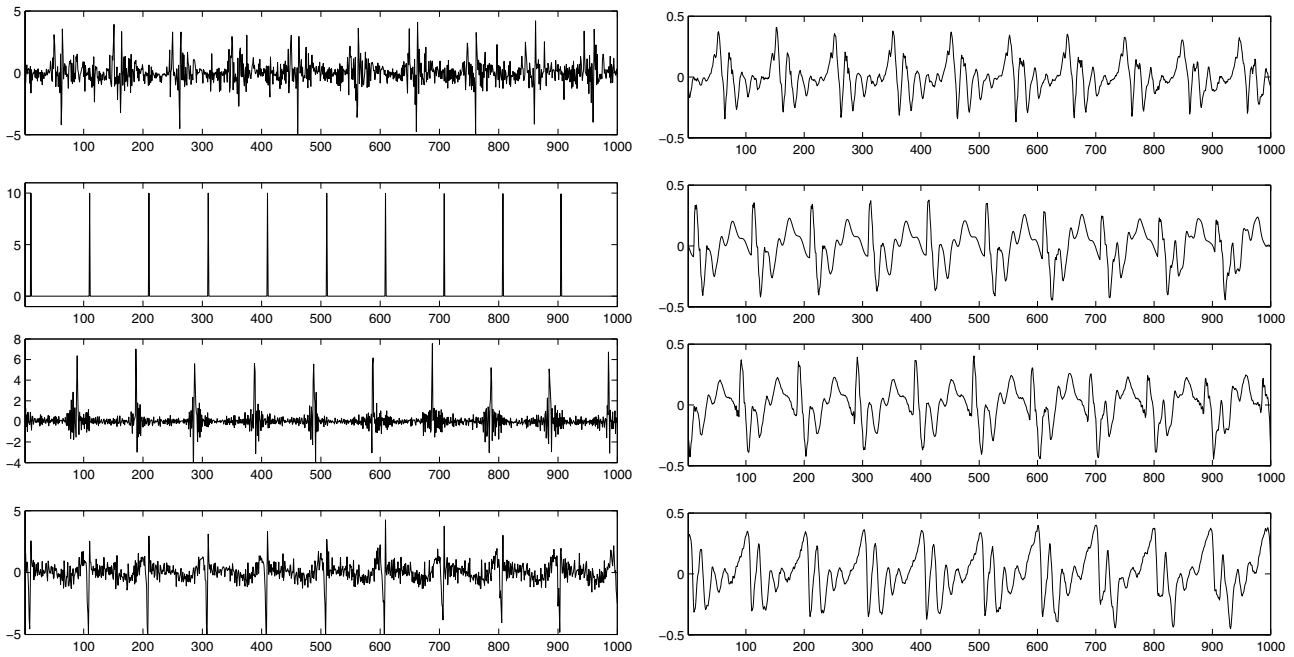


Figure 4: Top: residual (left) and natural speech (right). Second row: simple excitation (left) and corresponding synthesized speech (right). Third row: excitation according to [7] (left) and corresponding synthesized speech (right). Fourth row: excitation given by the proposed method (left) and respective synthesized speech (right).

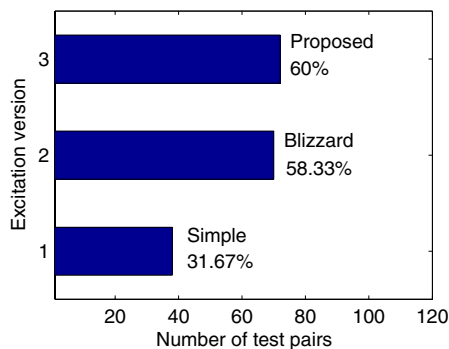


Figure 5: Overall preference for each excitation version.

Figure 5 shows the listener's preference. It can be seen that the proposed approach is equivalent to the Blizzard Challenge 2005 HMM-based synthesizer in terms of quality. In fact, the question asked during the test was: "which utterance from the pair presents better quality?". Though most of times the difference between the proposed and Blizzard versions was noticed, the decision concerning which one presented better quality was usually difficult to be made, according to the listeners.

#### 4. Conclusions

The proposed method synthesizes speech with quality considerably better than the simple excitation baseline. Furthermore, when compared with one of the best approaches reported to eliminate the *buzziness* of HMM-based speech synthesis: the Blizzard Challenge 2005 system as described in [7], the proposed model presents the advantage of producing speech which seems closer to natural. The results have been promising and future steps towards the conclusion of this work include pulse train modeling for the waveform generation part and state clustering using the residual signals themselves.

#### 5. References

- [1] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. of EUROSPEECH*, 1999.
- [2] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, "Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis," in *Proc. of ICASSP*, 2004.
- [3] W. Chu, *Speech Coding Algorithms*. Wiley-Interscience, 2003.
- [4] A. McCree and T. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech and Audio Processing*, vol. 3, no. 4, July 1995.
- [5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed-excitation for HMM-based speech synthesis," in *Proc. of EUROSPEECH*, 2001.
- [6] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "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, Apr. 1999.
- [7] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis for Blizzard Challenge 2005," *IEICE Trans. on Inf. and Systems*, vol. E90-D, no. 1, 2007.
- [8] S. J. Kim and M. Hahn, "Two-band excitation for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, 2007.
- [9] M. Akamine and T. Kagoshima, "Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS drive TTS)," in *Proc. ICSLP*, 1998.
- [10] L. B. Jackson, *Digital filters and signal processing*. Kluwer Academic, 1996.
- [11] S. Kay, *Modern Spectral Estimation*. USA: Prentice-Hall, 1988.
- [12] [http://festvox.org/cmu\\_arctic](http://festvox.org/cmu_arctic).
- [13] H. Duxans and A. Bonafonte, "Residual conversion versus prediction on voice morphing systems," in *Proc. of ICASSP*, 2006.
- [14] [http://www.slt.atr.jp/IWSLT2004\\_whatsnew/archives/000782.html](http://www.slt.atr.jp/IWSLT2004_whatsnew/archives/000782.html).