# Improved Training of Excitation for HMM-based Parametric Speech Synthesis

*Yoshinori Shiga*[†], *Tomoki Toda*[†,‡], *Shinsuke Sakai*[†], *Hisashi Kawai*[†]

[†]National Institute of Information and Communications Technology, Japan
[‡]Nara Institute of Science and Technology, Japan
yoshi.shiga@nict.go.jp

## Abstract

This paper presents an improved method of training for the unvoiced filter that comprises an excitation model, within the framework of parametric speech synthesis based on hidden Markov models. The conventional approach calculates the unvoiced filter response from the differential signal of the residual and voiced excitation estimate. The differential signal, however, includes the error generated by the voiced excitation estimates. Contaminated by the error, the unvoiced filter tends to be overestimated, which causes the synthetic speech to be noisy. In order for unvoiced filter training to obtain targets that are free from the contamination, the improved approach first separates the non-periodic component of residual signal from the periodic component. The unvoiced filter is then trained from the non-periodic component signals. Experimental results show that unvoiced filter responses trained with the new approach are clearly noiseless, in contrast to the responses trained with the conventional approach.

**Index Terms**: speech synthesis, HMM-based speech synthesis, mixed excitation, residual modelling.

## 1. Introduction

Text-to-speech synthesis (TTS) based on hidden Markov models (HMMs) [1] has a great advantage over the other leading speech synthesis techniques in terms of flexibility in synthesising speech with various voice characteristics and speaking styles through the potential use of voice transformation techniques, small corpora and small footprint demand. However, unnatural speech produced through the parametric source-filter model still represents a challenging issue. Attempts at solving this problem have become a research topic of increasing interest.

Many approaches have been reported aiming to improve HMM-based TTS systems through the design of better excitation modelling. Modelling of the *bandpass aperiodicity parameters* and eventual use of the excitation scheme proposed in [2] at run-time is a component of the system described in [3]. Sinusoidal modelling is applied in [4], while the Liljencrants-Fant (LF) glottal-waveform model is used in [5]. Application of glottal inverse filtering is reported in [6].

In this context, we have proposed a trainable excitation model [7][8][9]. The method is based on the principle of analysis-by-synthesis speech coders and consists of the optimisation of state-dependent filter coefficients through iterative minimisation of the difference between synthesised excitation and the residual directly obtained from the speech corpus through inverse filtering. At its synthesis stage, the trained filters are used to generate mixed excitation by inputting pulse train and white noise into the filters.

Synthetic speech from this model, however, contains an excessive amount of noise. In the training stage, the residual of a statistically-optimised voiced filter is dealt with as a target signal for determining the unvoiced filter coefficients. As a result, errors caused in the voiced filter estimation contaminate the unvoiced excitation targets, and consequently influence the unvoiced filter estimation, which makes the final speech output noisy.

To address this problem, in this paper, we employ contamination-free unvoiced-excitation targets for training. In this new training scheme, the 'clean' targets are extracted directly from the residual signals. This extraction is achieved with a periodic signal estimator, which is used to separate the non-periodic component of the residual from the periodic component.

The remainder of this paper is organised as follows: Section 2 outlines the basics of the excitation model; Section 3 explains how the unvoiced excitation training is improved; Section 4 details the results of experiments conducted to confirm the improvements; and the conclusions are in Section 5.

## 2. Trainable excitation model

### 2.1. Generation of excitation signals

Figure 1 shows the synthesis stage of our excitation model [8], where pulse train $t(n)$ and white noise $w(n)$ are passed through voiced and unvoiced filters, $H_{\mathrm{v}}(z)$ and $H_{\mathrm{u}}(z)$. They are added together to result in the excitation signal $\widetilde{e}(n)$. Associated with each HMM state position $s$, each of the filters has the following transfer function:

$$H_{\mathrm{v}}^{s}(z) = \sum_{l=-M/2}^{M/2} h_s(l) z^{-l}, \tag{1}$$

$$H_{\mathrm{u}}^{s}(z) = K_s \Bigg/ \left[ 1 - \sum_{l=1}^{L} g_s(l) z^{-l} \right], \tag{2}$$

where $M$ and $L$ are the respective filter orders. The excitation signal thereby generated will be input into the vocal-tract filter of the source-filter model at the next stage.

### 2.2. Training filters

The excitation model components, the filters $H_{\mathrm{v}}(z)$ and $H_{\mathrm{u}}(z)$, and impulse train $t(n)$, are iteratively calculated to minimise the error between residual and synthetic excitation. Figure 2 illustrates the procedure diagrammatically.

Using matrices and vectors, with $N$ being the total number of samples of the entire database, the filters are determined in a

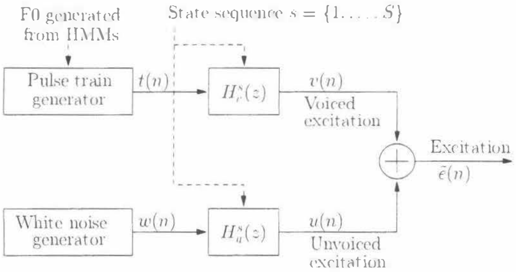

Figure 1: *Excitation signal generation: filters $H_\text{v}(z)$ and $H_\text{u}(z)$ are associated with each HMM state s.*

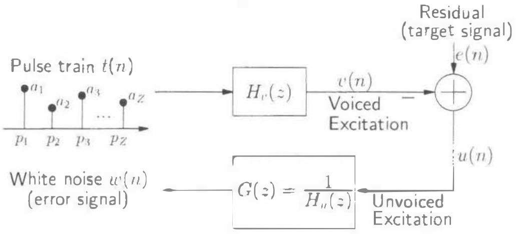

Figure 2: *Excitation model training: both filters are computed based on an analysis-by-synthesis optimisation.*

way that minimises the mean squared error $\varepsilon$, given by

$$\varepsilon = \frac{1}{N}\left(e - \sum_{s=1}^{S}\mathbf{A}_s\boldsymbol{h}_s\right)^{\mathsf{T}}\mathbf{G}^{\mathsf{T}}\mathbf{G}\left(e - \sum_{s=1}^{S}\mathbf{A}_s\boldsymbol{h}_s\right), \quad (3)$$

where $\mathbf{G}$ is an $N \times N$ matrix containing the impulse response of the inverse unvoiced filter $G(z)$, $\boldsymbol{h}_s = [h_s(-M/2)\cdots h_s(M/2)]^{\mathsf{T}}$ 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. In this case, each state $s = \{1, \dots, S\}$ corresponds to a different HMM state-position covering the entire database, after the Viterbi alignment.

Voiced filter coefficients for a given state $s$ are obtained by making $\partial\varepsilon/\partial\boldsymbol{h}_s = 0$, which results in a linear system for the solution of $\boldsymbol{h}_s$ [8]. On the other hand, the unvoiced filter coefficients for state $s$, $\{g_s(1), \dots, g_s(L)\}$, and related gain $K_s$, are determined by performing linear prediction analysis on the unvoiced excitation signal $\tilde{u}(n) = e(n) - v(n)$ over segments tagged as state $s$.

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 (3). The procedure to determine the positions and amplitudes resembles multipulse excitation linear prediction coding algorithms [10].

The overall procedure for the design of the filters and optimisation of $t(n)$ is performed in an interchanging way, with the convergence criterion being either the filter coefficient variation or the mean squared error reduction.

### 2.3. Tree-based state definition

The filters vary according to each HMM state and their coefficients are optimised using a residual signal ML criterion [7]. The excitation training process can be enumerated through the following steps: (1) state definition; (2) residual segment classification according to the defined states; (3) iterative filter calculation for each cluster of residual segments using the procedure described in the previous section.

Assuming that the noise sequence $w(n)$ output by filter $G(z)$ in Figure 2 is a Gaussian process, the log likelihood of the signal $u(n)$, also a Gaussian process, is given by

$$\log P[\boldsymbol{u}|\mathbf{H}_\text{u}] = -\frac{N}{2}\log 2\pi + \frac{1}{2}\log|\mathbf{G}^{\mathsf{T}}\mathbf{G}| - \frac{1}{2}\boldsymbol{u}^{\mathsf{T}}\mathbf{G}^{\mathsf{T}}\mathbf{G}\boldsymbol{u}, \quad (4)$$

where $N$ is the number of samples of the entire database, $\boldsymbol{u} = [u(0)\cdots u(N-1)]^{\mathsf{T}}$, $\mathbf{G} = [\tilde{g}^{(0)} \vdots \cdots \vdots \tilde{g}^{(N-1)}]$, and

$$\tilde{\boldsymbol{g}}^{(m)} = \Big[\underbrace{0 \cdots 0}_{m \text{ terms}} \ \frac{1}{K} \ \frac{g(1)}{K} \ \cdots \ \frac{g(L-1)}{K} \ \underbrace{0 \cdots 0}_{N-m-1 \text{ terms}}\Big]^{\mathsf{T}}.$$

By arranging (4) with some approximations [9], the likelihood of $e(n)$ given the excitation model is simply a function of the unvoiced filter gain component $K$ as

$$\log P[\boldsymbol{e}|\mathbf{H}_\text{v}, \mathbf{H}_\text{u}, \boldsymbol{t}] \approx -\frac{N}{2}\log 2\pi - N\left(\log K + \frac{K^2}{2}\right). \quad (5)$$

Note that $P[u(n)|H_u(z)] \Leftrightarrow P[e(n)|H_v(z), H_u(z), t(n)]$.

By taking into account the state-dependency of the filter coefficients, (5) can be re-written as

$$\log P[\boldsymbol{e}|\mathbf{H}_\text{v}, \mathbf{H}_\text{u}, \boldsymbol{t}] = -\frac{N}{2}\log 2\pi + \sum_{j=1}^{S}\mathfrak{L}_j, \quad (6)$$

where $\mathfrak{L}_j$ is the likelihood of $e(n)$ under state $s_j$ and given as

$$\mathfrak{L}_j = -N_j\left(\log K_j + \frac{K_j^2}{2}\right). \quad (7)$$

In (7), $N_j$ is its corresponding number of samples, $K_j$ is the corresponding unvoiced filter gain, and $S$ is the number of states (or clusters for tied states).

From Figure 2, initially voiced filter coefficients are computed, followed by the determination of $u(n)$, finally leading to gain component $K_{s_j}$. The process of splitting one cluster into two can thus be sketched as follows: (1) split $s_j$ into $s_{j1}$ and $s_{j2}$ given a candidate question; (2) calculate voiced filter coefficients, $\boldsymbol{h}_{s_{j1}}$ and $\boldsymbol{h}_{s_{j2}}$, for the new clusters $s_{j1}$ and $s_{j1}$, respectively; (3) compute unvoiced filter coefficients with corresponding gain components, $\boldsymbol{g}_{j1}$, $K_{j1}$, $\boldsymbol{g}_{j2}$, and $K_{j2}$, respectively for $s_{j1}$ and $s_{j2}$. After calculating $\mathfrak{L}_{j1}$ and $\mathfrak{L}_{j2}$ from $K_{j1}$ and $K_{j2}$, respectively, according to (7), the likelihood increment due to the split can be measured by

$$\mathfrak{L}_\text{inc} = \mathfrak{L}_\text{after} - \mathfrak{L}_\text{before} = \mathfrak{L}_{j1} + \mathfrak{L}_{j2} - \mathfrak{L}_{s_j}. \quad (8)$$

The minimum description length (MDL) [11] is adopted as a stop criterion. Refer to [9] for details.

The determination of voiced filters and unvoiced filter gain components for $s_{jx}$ implies computationally-expensive optimisation of filter coefficients and pulse trains for the new clusters. To reduce the complexity, this iterative optimisation is replaced by a single calculation of voiced filters followed by linear prediction analysis of the unvoiced excitation signal under segments belonging to $s_{jx}$ to derive the gain component $K_{jx}$ [9].

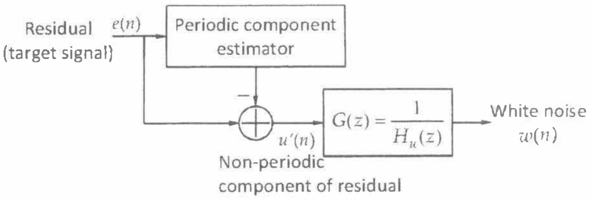
Residual
(target signal)
$e(n)$
Periodic component estimator
−
$u'(n)$
Non-periodic component of residual
$G(z) = \frac{1}{H_u(z)}$
White noise
$w(n)$

Figure 3: *Proposed training scheme for unvoiced filter coefficients*

# 3. Improved training of unvoiced excitation

## 3.1. Modification to unvoiced-filter training

As is clear from Fig. 2, $w(n)$ is not actually *whitened* during the training stage. The error of voiced component estimates contaminates the unvoiced excitation target $u(n)$, which causes the final speech output to be 'rough' (i.e., noisy). Conventionally, we avoided the noisiness by attenuating the noise component during the synthesis stage. This was being done mainly by passing the synthesised unvoiced excitation through a high-pass filter (HPF) with cut-off frequency 2 kHz before it is mixed with the voiced counterpart [7].

This remedy, however, does not remove the root cause of the rough speech problem. The total volume of perceptible noise can certainly be reduced, but while unvoiced information is almost eliminated for the range below the HPF cut-off frequency, the influence of the voiced component estimation error remains in the range above the frequency.

To resolve this more effectively, we employ contamination-free unvoiced-excitation targets for training. The 'clean' targets are extracted directly from the residual signals $e(n)$. The overall training scheme for the unvoiced filter is schematically shown in Fig. 3. The extraction is achieved with a periodic component estimator, which is used to separate the non-periodic component of the residual from the periodic component.

It is true that periodic component estimation also introduces a certain level of error, which smears the resulting non-periodic component directly. However, the contamination should be minor because the periodic component is estimated *locally for each speech segment* whereas in the conventional approach the voiced filter is optimised *over the entire database*.

## 3.2. Periodic/non-periodic decomposition of residual signals

Since the decomposition is a part in the offline training process, one may employ a relatively computationally-expensive approach. We adopt the following model to represent the periodic (i.e., harmonic) component of the residual:

$$\widetilde{s}(t) = \sum_{k=-J}^{J} A_k(t) e^{j[\Theta_k(t) + \phi_k]}, \tag{9}$$

where $A_k(t) = \alpha_k t + \beta_k$ and $\Theta_k(t) = \omega_k(\gamma t^2 + t)$ with $\omega_k = \omega_0 k = 2\pi f_0 k$ and the fundamental frequency $f_0$. Represented by $J$ is the number of harmonics. Obviously, in this model both the frequency and amplitude of each harmonic are approximated in a piecewise linear sense.

The problem is to find $\alpha_k$, $\beta_k$, $\gamma$ and $\phi_k$ that minimise

$$\delta = \sum_{t=t_0-N_w}^{t_0+N_w} w^2(t)\left[s(t) - \widetilde{s}(t)\right]^2, \tag{10}$$

where $s(t)$ is the original signal and $w(t)$ is a window function whose length is $2N_w + 1$. The solution described in [12] is applied to the problem above.

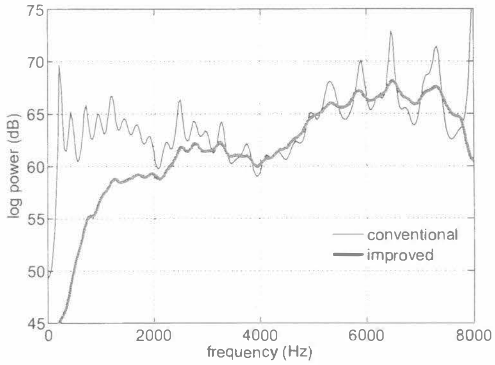

Figure 4: *Comparing unvoiced filter responses from the conventional (thin black line) and improved (thick red line) training schemes*

## 3.3. Modified unvoiced filter training

The unvoiced filter coefficients for state $s$, $\{g_s(1), \ldots, g_s(L)\}$, and related gain $K_s$, can be determined using linear prediction analysis on the unvoiced excitation signal $\widetilde{u}'(n)$ over segments tagged as state $s$. The states are defined using the same decision-tree-based technique as in Section 2.3. Thus, a different tree is constructed for the unvoiced filter additionally to the one for the voiced tree.

Figure 4 shows typical unvoiced filter responses obtained from the improved training and from the conventional training. These responses correspond to the second state of the 5-state HMM of English sound /i:/ in a certain context. It can be observed from this figure that spectral energy in the low frequency range is sufficiently low for the improved training, but not for the conventional training.

# 4. Experiments

We conducted a subjective evaluation to confirm the effectiveness of the modified training for unvoiced excitation.

## 4.1. Conditions and procedure

A listening test was performed with five subjects consisting of four speech synthesis experts and one with no experience in speech research. The test took the form of an AB forced preference, with the utterances of 20 sentences taken from the Blizzard Challenge 2009 test set (the first ten sentences from each of the 'news' and 'novel' categories), with the aim of comparing the quality of speech from the conventional and improved training both with and without the application of the HPF mentioned in Section 3.1. The test was carried out in a quiet room, and the listeners used headphones.

The speech data used for training were 4014 English utterances by a British male speaker. They were recorded by Phonetic Arts Ltd., U.K. and released for Blizzard Challenge 2010 [13]. $F_0$s and spectral envelopes were estimated from the recordings (16-kHz sampling) using the Snack Sound Toolkit and the STRAIGHT analysis [14], respectively, with 5-ms frame shifts. Each of the spectral envelopes was then converted

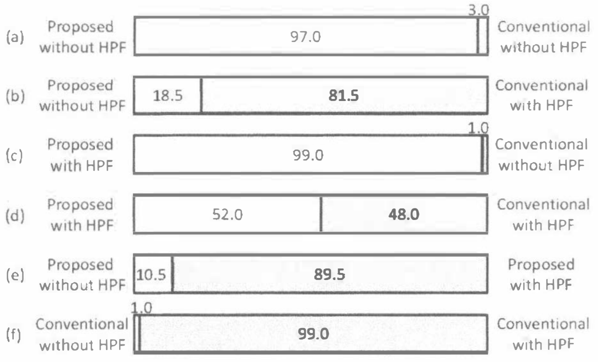

Figure 5: *Subjective evaluation results: the figures indicate preference scores (%).*

into the $39^{th}$-order mel-cepstrum using the Speech Signal Processing Toolkit (SPTK). Residual signals as excitation-training targets are extracted by passing speech through the mel-log-spectrum approximation (MLSA) filter [15]. The periodic component of the residual is estimated on the technique described in Section 3 using 20-ms-width window and 5-ms frame shift.

Five-state left-to-right no-skip HSMMs for duration, $F_0$ and mel-cepstral coefficients were trained on the basis of the trajectory training scheme under global-variance constraint [16]. The orders of excitation filters were $M = 512$ and $L = 64$.

### 4.2. Results and Discussion

The number of terminal nodes of the resulting trees (i.e., the number of clusters) were 83 and 254 for the voiced and unvoiced filters, respectively. Figure 5 shows the listeners' preference for each type of test pair. Since the unvoiced filter is estimated separately from the voiced filter estimation, clear speech with little noise is synthesised from the excitation model trained with the improved method. For this reason, as shown in Figure 5(a), speech from the improved training is preferred by listeners in 97% of all cases, if no HPF is applied to the synthesised unvoiced excitation signals during synthesis.

On the other hand, Figs. 5(b)–(f) show that the HPF is still necessary even for models from the improved approach, and that when the HPF is applied for both approaches, the model can produce speech of slightly better quality, although this is not significant. Careful listening by the authors revealed that the types of noise perceptible in the background differ depending on the approach. With respect to the improved training, it is an intermittent type of noise arising segmentally, whereas the noise from the conventional method is rather stationary. The former noise is considered to be generated at state boundaries, where the excitation filter response can change dramatically, because no dynamic features are used in the current model of excitation.

## 5. Conclusions

We have investigated an improved training framework for our mixed excitation model, where unvoiced excitation signals are generated through a filter whose coefficients are trained directly from a non-periodic component of the residual signals.

Some problems still remain with our mixed excitation training scheme. First, during synthesis, the unvoiced filter response changes state by state, which causes perceptible noise in synthetic speech. To attenuate that noise, the dynamic features of the filter response may be effective. Second, in the current framework, the power of the voiced excitation cannot be determined. The amplitude of the optimised impulses should be taken into account for the determination. Resolving these will be part of our future work.

## 6. Acknowledgements

The authors would like to thank Keiichi Tokuda and Minoru Tsuzaki for their stimulating and helpful discussions on issues dealt with in this paper. Further, we thank Satoshi Nakamura for his support in conducting this study.

## 7. 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. EUROSPEECH'99*, (Budapest, Hungary), pp. 2347–2350, Sept. 1999.

[2] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *MAVEBA*, 2001.

[3] 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, Jan. 2007.

[4] O. Abdel-Hamid, S. Abdou, and M. Rashwan, "Improving the Arabic HMM based speech synthesis quality," in *ICSLP*, 2006.

[5] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an improved modeling of the glottal source in statistical parametric speech synthesis," in *SSW6*, 2007.

[6] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "HMM-based Finnish text-to-speech system using glottal inverse filtering," in *Interspeech*, 2008.

[7] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," in *SSW6*, 2007.

[8] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "A trainable excitation model for HMM-based speech synthesis," in *INTERSPEECH*, 2007.

[9] R. Maia, T. Toda, K. Tokuda, S. Sakai, and S. Nakamura, "A decision tree-based clustering approach to state definition in an excitation modeling framework for HMM-based speech synthesis," in *INTERSPEECH*, 2009.

[10] W. Chu, *Speech Coding Algorithms*. Wiley-Interscience, 2003.

[11] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. on Information Theory*, vol. IT-30, July 1984.

[12] Y. Stylianou, *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*. PhD thesis, Ecole Nat. Supèrieure Télécommun., France, Jan. 1996.

[13] http://www.synsig.org/index.php/Blizzard_Challenge_2010.

[14] 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, Apr. 1999.

[15] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. of ICASSP*, 1992.

[16] T. Toda and S. Young, "Trajectory training considering global variance for HMM-based speech synthesis," in *ICASSP*, 2009.