A Novel Excitation Approach for HMM-Based Speech Synthesis

Ranniery Maia

Report IV

2nd May 2007

Abstract

One of the drawbacks to the speech synthesis technique wherein speech parameters are directly generated from hidden Markov models (HMM-based speech synthesis) is the unnaturalness of the synthesized speech. This problem occurs owing to the rough excitation model employed during the waveform generation stage. This report introduces a new excitation approach that attempt to solve this problem. The proposed scheme consists in feeding the mel log spectrum approximation (MLSA) filter with mixed excitation, obtained through a set of state-dependent filters. The filters are derived from the speech database through a closed-loop procedure where the likelihood of the residual is maximized.

1 Introduction

In the last years the technique in which speech is generated from parameters directly obtained from hidden Markov models [18, 24, 19, 11] has emerged as a good choice for synthesizing speech with different voice styles and characteristics [13, 23, 16, 21, 22]. However, one disadvantage of this technique when compared with unit concatenation-based systems [5] corresponds to the quality of the synthesized speech, which presents a low degree of naturalness. This fact basically occurs because of the simple excitation scheme where the switch between pulse train and random noise are employed to model voiced and unvoiced segments, respectively, generating the excitation which is fed into the MLSA filter [4, 17]. Consequently, the synthesized speech presents the same artifacts which are usually observed in linear predictive (LP) vocoders [3].

In order to solve the problem in question, some approaches have been reported, e.g. [25, 26, 1, 10]. Yoshimura et al [25] proposed the application of the mixed excitation generated by the Mixed Excitation Linear Prediction (MELP) speech coding algorithm [15, 14] to HMM-based

speech synthesis, achieving a significant improvement on the quality when compared to the simple excitation method. The idea consisted in the modeling by HMMs of all the parameters employed by MELP speech coding, namely: bandpass voicing strength parameters, jitter, and Fourier magnitudes, jointly with mel-cepstral coefficients and $F0$. In that case each observation vector was composed of seven streams: (1) mel-cepstral coefficients; (2) $\log(F0)$; (3) $\Delta \log(F0)$; (4) $\Delta\Delta \log(F0)$; (5) bandpass voicing strength coefficients; (6) pulse position jitter; and finally (7) Fourier magnitudes. Later, Zen et al [26] proposed an improved excitation scheme. It consisted in the utilization of the high-quality vocoding method employed by STRAIGHT [8]. In order to synthesize speech without artifacts the bandpass voicing components were modeled by HMMs, in the same way as the method proposed by Yoshimura. The resulting quality was considerably better.

Here a novel excitation approach is proposed. In the present case mixed excitation is generated by inputting pulse train and white noise into two filters which vary according to a sequence of specific states which may be represented, for instance, by:

- leaves of decision-trees generated for the distributions of mel-cepstral coefficients or F0;
- voicing conditions (voiced or unvoiced segments);
- sequence of acoustic units whose durations might be derived through time-alignment, e.g. monophone and triphone.

The filters are derived in a way to maximize the likelihood of residual sequences over the corresponding states. Pulse trains are also optimized in the sense of residual likelihood maximization. The joint procedure comprising filter determination and pulse optimization is conducted iteratively, behaving thus as a closed-loop system. Although some analysis-by-synthesis methods, similar to Code-Excited Linear Prediction (CELP) speech coding algorithms [3], have already been proposed for speech synthesis, e.g. [2], the present approach targets natural residual instead of speech and assumes the error of the system as the unvoiced component of the excitation.

The remaining of this report is organized as follows: Section 2 introduces the idea of the proposed excitation approach for HMM-based speech synthesis; Section 3 concerns the problem formulation; Section 4 shows how the state-dependent filters can be calculated; Section 5 describes pulse train optimization procedure; Section 6 regards the closed-loop algorithm used for filter computation and pulse optimization; and the conclusions are given in Section 7.

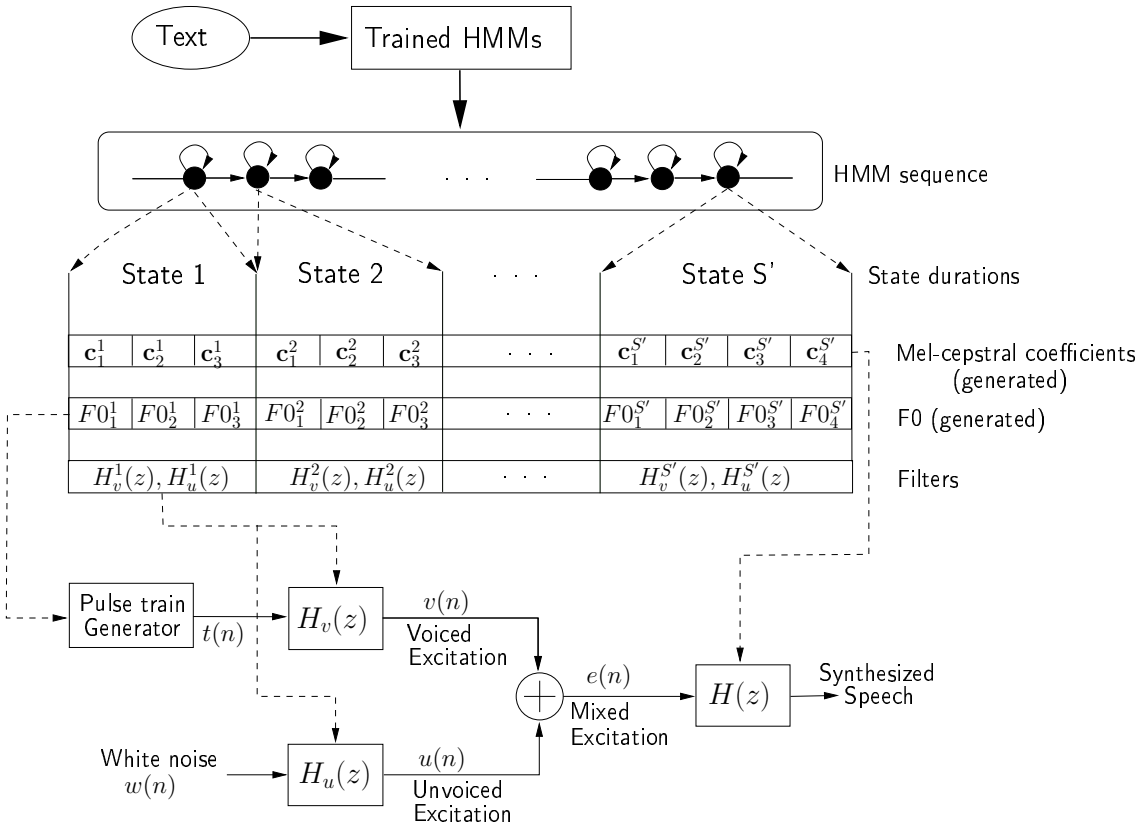

Figure 1: Proposed excitation scheme.

2 The idea

The excitation scheme proposed is depicted in Figure 1. The excitation signal $e(n)$ is constructed by the addition of the pulse train $t(n)$, and white noise, $w(n)$, filtered respectively by the state-dependent voiced and unvoiced filters, $H_v(z)$ and $H_u(z)$. Their transfer functions are given by

$$H_v(z) = \sum_{l=-\frac{M}{2}}^{\frac{M}{2}} h(l) z^{-l}, \tag{1}$$

$$H_u(z) = \frac{K}{1 - \sum_{l=1}^{L} g(l) z^{-l}}. \tag{2}$$

where M and L are the respective the orders. As it can be notice from the illustration, the filters are associated with each state $\{1, \ldots, S'\}$.

2.1 Function of $H_v(z)$

The voiced filter $H_v(z)$ transforms the pulse train $t(n)$ into the voiced excitation $v(n)$, which is as close as possible to the residual sequence $e(n)$. This idea is closely related to the adaptive filtering theory, considering the system identification problem.

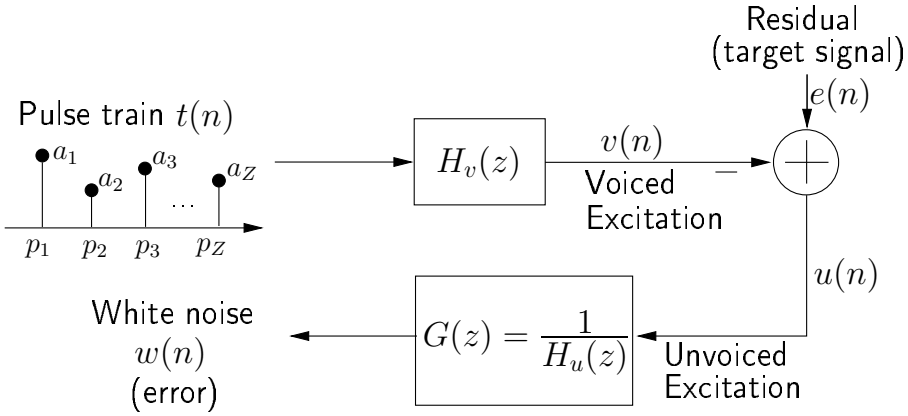

Figure 2: Re-arrangement of the excitation part: input correspond to pulse train and residual whereas white noise is produced at the output.

The property of having a finite impulse response leads to stability. Further, since the final waveform is synthesized offline, an anti-causal structure might achieve better performance through the processing of delayed and advanced samples.

2.2 Function of $H_u(z)$

Since white noise with zero mean and unit variance is the input of the unvoiced filter, the function of $H_u(z)$ is, therefore, to remove the remaining long-term correlation (assuming that the residual $e(n)$ has no short-term correlation) from the difference signal $u(n) = e(n) - v(n)$. For this purpose the all-pole structure based on LP coefficients shown in (2) is a good choice, due to its simplicity in terms of computational complexity. However, in order to perform as expected the order L should be set at least for three pitch periods.

3 Excitation training: 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 represent 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 would

consist in:

1. the determination of the filters $H_v(z)$ and $H_u(z)$;

2. optimization of the positions, $\{p_1, \ldots, p_Z\}$, and amplitudes, $\{a_1, \ldots, a_Z\}$, of $t(n)$.

In the next two sections, the procedures in which the state-dependent filters are determined and pulse trains are optimized are described.

4 Filter determination

The filters are determined in a way to maximize the likelihood of the residual[1] given the proposed model of Figure 1. Therefore, the first step is to achieve an expression for the likelihood of $e(n)$ which involves $H_u(z)$, $H_v(z)$ and $t(n)$, and eventually maximize it with respect to the filters.

4.1 Likelihood of the residual $e(n)$ given the excitation model

According to Figure 2 pulse train and speech residual correspond to the input of the system whereas white noise is the output, after being filtered by the inverse unvoiced filter

$$G(z) = \frac{1}{H_u(z)} = \sum_{l=0}^{L} \tilde{g}(l) z^{-l} \tag{3}$$

where

$$\tilde{g}(i) = \begin{cases} \frac{1}{K}, & i = 0, \\ \frac{g(i)}{K}, & 1 \leq i \leq L. \end{cases} \tag{4}$$

If we consider the vector $\mathbf{w} = [w(0) \cdots w(N-1)]^T$ - where N is the database length in number of samples and $[\cdot]^T$ means transposition - as white noise, the probability distribution of the unvoiced excitation vector $\mathbf{u} = [u(0) \cdots u(N-1)]^T$ given the unvoiced filter $H_u(z)$ is

$$P\{\mathbf{u}|H_u(z)\} = \frac{1}{\sqrt{(2\pi)^N |\mathbf{R}|}} e^{-\frac{1}{2}\mathbf{u}^T \mathbf{R}^{-1} \mathbf{u}}, \tag{5}$$

where the covariance matrix $\mathbf{R}$ is

$$\mathbf{R} = \begin{bmatrix} r(0) & \ldots & r(N-1) \\ \vdots & & \vdots \\ r(N-1) & \ldots & r(0) \end{bmatrix}, \tag{6}$$

[1] Obtained from the speech corpus by inverse filtering. The residual signals are extracted so as to present flat spectrum with unit power.

and the autocorrelation sequence, $r(n)$, is given by

$$r(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H_u(e^{jw})|^2 e^{jwn} dw. \tag{7}$$

If we make $\mathbf{u} = \mathbf{e} - \mathbf{v}$ in (5), the likelihood of $\mathbf{u}$ given $H_u(z)$ becomes the likelihood of $\mathbf{e}$ given $H_v(z)$ and $H_u(z)$ since the voiced excitation vector $\mathbf{v} = [v(0) \dots v(N-1)]^T$ is deterministic. Thus,

$$P\left\{\mathbf{e}|H_v(z), H_u(z)\right\} = \frac{1}{\sqrt{(2\pi)^N |\mathbf{R}|}} e^{-\frac{1}{2}[\mathbf{e}-\mathbf{v}]^T \mathbf{R}^{-1}[\mathbf{e}-\mathbf{v}]}, \tag{8}$$

becomes the likelihood of the residual and consequently the function which must be minimized through the determination of the voiced and unvoiced filters. By taking the logarithm of (8), the following expression for the log likelihood is obtained

$$\log P\left\{\mathbf{e}|H_v(z), H_u(z)\right\} = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{R}| - \frac{1}{2}[\mathbf{e}-\mathbf{v}]^T \mathbf{R}^{-1}[\mathbf{e}-\mathbf{v}]. \tag{9}$$

For the derivation of the vector of coefficients for the voiced filter $H_v(z)$,

$$\mathbf{h}_i = \begin{bmatrix} h_i\left(\frac{-M}{2}\right) & \cdots & h_i\left(\frac{M}{2}\right) \end{bmatrix}^T, \tag{10}$$

and the coefficients of the unvoiced filter $H_u(z)$, $\{g_i(1), \dots, g_i(L)\}$, with the related gain K_i, for a specific state i, the log likelihood of the residual written in a form which depend on $\mathbf{h}_i$ and $\{K_i, g_i(1), \dots, g_i(L)\}$ should be taken into account.

4.1.1 Relationship between $\mathbf{R}$ and $G(z)$

Since $G(z) = H_u^{-1}(z)$, then

$$\mathbf{R}^{-1} = \mathbf{G}^T\mathbf{G}, \tag{11}$$

where $\mathbf{G}$ corresponds to the $N \times N$ overall impulse response matrix of the inverse unvoiced filter $G(z)$ including all the states,

$$\mathbf{G} = \begin{bmatrix} \tilde{\mathbf{G}}_{k,j} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{G}}_{i,j} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \tilde{\mathbf{G}}_{n,j} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \tilde{\mathbf{G}}_{i,j+1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \tilde{\mathbf{G}}_{m,j} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \tilde{\mathbf{G}}_{n,j+1} \end{bmatrix}, \tag{12}$$

with $\mathbf{0}$ being a null matrix of any dimension. In this case, the matrices $\tilde{\mathbf{G}}_{i,j}$, $\tilde{\mathbf{G}}_{k,j}$, $\tilde{\mathbf{G}}_{n,j}$, $\tilde{\mathbf{G}}_{m,j}$ contain the impulse responses of the filter $G(z)$ for the j-th segments of the states i, k, n, and m,

respectively. For instance, $\tilde{\mathbf{G}}_{i,j}$, with dimension $(N_{i,j}+L)\times N_{i,j}$, where $N_{i,j}$ is the length of the j-th segment of the state i, has the following shape:

$$\tilde{\mathbf{G}}_{i,j}=\begin{bmatrix}\tilde{g}_i(0) & \cdots & 0\\ \vdots & \ddots & \vdots\\ \tilde{g}_i(L) & & \tilde{g}_i(0)\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \tilde{g}_i(L)\end{bmatrix}, \tag{13}$$

where $\{\tilde{g}_i(0),\ldots,\tilde{g}_i(L)\}$ is the $(L+1)$-size impulse response sequence of the filter $G(z)$ for state i.

4.1.2 The voiced excitation vector v

Because there is a different voiced filter for each state, the overall voiced excitation $\mathbf{v}$ is given by

$$\mathbf{v}=\mathbf{A}_1\mathbf{h}_1+\ldots+\mathbf{A}_S\mathbf{h}_S=\sum_{i=1}^{S}\mathbf{A}_i\mathbf{h}_i, \tag{14}$$

where S is the total number of states. The term $\mathbf{A}_i$ is the overall pulse train matrix where only the pulse train samples belonging to the state i are non-zero. For example, considering the case of the j-th segment of the i-th state, shown in (12), the corresponding matrix $\mathbf{A}_i$ would be

$$\mathbf{A}_i=\begin{bmatrix}\mathbf{0}\\ \mathbf{A}_{i,j}\\ \mathbf{0}\\ \mathbf{A}_{i,j+1}\\ \mathbf{0}\\ \vdots\\ \mathbf{0}\end{bmatrix}, \tag{15}$$

where the $(M+1)\times(N_{i,j}+M)$ matrix $\mathbf{A}_{i,j}$ is

$$\mathbf{A}_{i,j}=\begin{bmatrix}t_{i,j}(0) & \cdots & 0\\ \vdots & \ddots & \vdots\\ t_{i,j}(\frac{M}{2}) & \cdots & t_{i,j}(0)\\ \vdots & \ddots & \vdots\\ t_{i,j}(N_{i,j}-1) & & t_{i,j}(\frac{M}{2})\\ \vdots & \ddots & \vdots\\ 0 & \cdots & t_{i,j}(N_{i,j}-1)\end{bmatrix}. \tag{16}$$

The vector $\{t_{i,j}(0) \cdots t_{s,j}(N_{i,j}-1)\}$ corresponds to the pulse train for the j-th segment of state i, and the overall pulse train matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \mathbf{A}_1 + \ldots + \mathbf{A}_S = \sum_{i=1}^{S} \mathbf{A}_i. \tag{17}$$

4.1.3 Likelihood of $e(n)$ in terms of $\mathbf{h}_i$, $\mathbf{A}_i$ and $\mathbf{G}$

By substituting (11) and (14) into (9), the following expression for the log likelihood is finally obtained

$$\log P\left\{\mathbf{e}|H_v(z), H_u(z)\right\} = -\frac{N}{2}\log 2\pi + \frac{1}{2}\log|\mathbf{G}^T\mathbf{G}| - \\ - \frac{1}{2}\left[\mathbf{e} - \sum_{i=1}^{S}\mathbf{A}_i\mathbf{h}_i\right]^T \mathbf{G}^T\mathbf{G}\left[\mathbf{e} - \sum_{i=1}^{S}\mathbf{A}_i\mathbf{h}_i\right], \tag{18}$$

which corresponds to the term to be maximized along the process for the computation of the filters.

4.2 Determination of the voiced filter $H_v(z)$

To determine the impulse response of $H_v(z)$ for a particular state i, the vector of coefficients $\mathbf{h}_i$ which maximizes the log likelihood in (18) can be obtained from

$$\frac{\partial \log P\left[\mathbf{e}|H_v(z), H_u(z)\right]}{\partial \mathbf{h}_i} = 0, \tag{19}$$

which results into the following linear system

$$\mathbf{X}_i\mathbf{h}_i = \mathbf{y}_i, \tag{20}$$

where

$$\mathbf{X}_i = \mathbf{A}_i^T\mathbf{G}^T\mathbf{G}\mathbf{A}_i, \tag{21}$$

$$\mathbf{y}_i = \mathbf{A}_i^T\mathbf{G}^T\mathbf{G}\left[\mathbf{e} - \sum_{\substack{k=1 \\ k\neq i}}^{S}\mathbf{A}_k\mathbf{h}_k\right]. \tag{22}$$

Therefore, if the $(M+1)\times(M+1)$ matrix $\mathbf{X}_i$ is singular the solution for $\mathbf{h}_i$ is unique. In fact, (20) corresponds to the least-squares formulation for the design of a filter through the solution of an over-determined linear system [7].

4.2.1 Segment basis

Since the matrix $\mathbf{A}_i$ is as shown in (15), $\mathbf{X}_i$ and $\mathbf{y}_i$ can thus be written as

$$\mathbf{X}_i = \sum_{j=1}^{N_i} \mathbf{A}_{i,j}^T \mathbf{G}_{i,j}^T \mathbf{G}_{i,j} \mathbf{A}_{i,j}, \tag{23}$$

$$\mathbf{y}_i = \sum_{j=1}^{N_i} \mathbf{A}_{i,j}^T \mathbf{G}_{i,j}^T \mathbf{G}_{i,j} \left[\mathbf{e}_{i,j} - \sum_{\substack{k=1 \\ k \neq i}}^{S} \sum_{l=1}^{N_k} \mathbf{A}_{k,l} \mathbf{h}_k \right], \tag{24}$$

where N_i and N_k are the number of segments which belong to the states i and k, respectively. The $(N_{i,j} + M + L) \times (N_{i,j} + M)$ matrix $\mathbf{G}_{i,j}$ is

$$\mathbf{G}_{i,j} = \begin{bmatrix} \mathbf{G}_{i,j}^p & \tilde{\mathbf{G}}_{i,j} & \mathbf{G}_{i,j}^s \end{bmatrix} \tag{25}$$

where $\mathbf{G}_{i,j}^p$ and $\mathbf{G}_{i,j}^s$ contain, respectively, impulse responses of the inverse unvoiced filters whose states are covered by $\frac{M}{2}$ samples before and after segment j of state i. Finally, the residual vector for the j-th segment of the state i, $\mathbf{e}_{i,j}$, is given by

$$\mathbf{e}_{i,j} = \begin{bmatrix} e_{i,j}\left(-\frac{M}{2}\right) & \cdots & e_{i,j}\left(N_{i,j} + \frac{M}{2} - 1\right) \end{bmatrix}. \tag{26}$$

4.3 Determination of $H_u(z)$

To visualize the problem for the computation of the coefficients of $H_u(z)$, another expression which may represent the log likelihood function should be considered.

It can be noticed that

$$[\mathbf{e} - \mathbf{v}]^T \mathbf{R}^{-1} [\mathbf{e} - \mathbf{v}] = \mathbf{u}^T \mathbf{G}^T \mathbf{G} \mathbf{u} = \mathbf{w}^T \mathbf{w} = \frac{1}{K^2} \sum_{k=0}^{N-1} \left\{ u(k) - \sum_{l=1}^{L} g(l) u(k-l) \right\}^2, \tag{27}$$

and it can be verified [9] that

$$|\mathbf{R}| = \prod_{k=0}^{N-1} |H_u(e^{j\omega_k})|^2 = \prod_{k=0}^{N-1} \frac{K^2}{\left| 1 - \sum_{l=1}^{L} g(l) e^{-j\omega_k l} \right|^2}. \tag{28}$$

If we substitute (27) and (28) into (8), and take the logarithm of the resulting expression, the log likelihood function becomes

$$\log P\{\mathbf{e}|H_u(z)\} = \sum_{k=0}^{N-1} \log \left| 1 - \sum_{l=1}^{L} g(l) e^{-j\omega_k l} \right| - \frac{1}{2} \sum_{k=1}^{N-1} \left(\log 2\pi K^2 + \frac{1}{K^2} \left\{ u(k) - \sum_{l=1}^{L} g(l) u(k-l) \right\}^2 \right). \tag{29}$$

Since $G(z)$ is minimum-phase, the first term in the right side of (29) becomes zero. By making

$$\frac{\partial \log P\{\mathbf{e}|H_u(z)\}}{\partial K} = 0, \tag{30}$$

the gain K_m which maximizes (29) can be derived,

$$K_m = \sqrt{\varepsilon_m}, \tag{31}$$

where the minimum energy ε_m is given by

$$\varepsilon_m = \min_{g(1),\ldots,g(l)} \left\{ u(k) - \sum_{l=1}^{L} g(l)u(k-l) \right\}^2. \tag{32}$$

Therefore, the problem can be interpreted as the autoregressive (AR) spectral estimation of $u(n)$ [12, 9, 6].

4.3.1 Segment basis

To perform state-dependent AR spectral analysis (LP analysis) on $u(n)$ a mean autocorrelation (or covariance) sequence representing each corresponding state should be taken into account. Among several methods, the following ones are considered:

- **method 1:** considering all segments of a particular state i as ensembles of a wide-sense stationary process, the mean autocorrelation function for i can be computed as the average of all short-time autocorrelation functions from all the segments belonging to i (making an analogy to the method presented in [20] for the periodogram), i.e.,

$$\bar{\phi}_i(k) = \frac{1}{\sum_{j=1}^{N_i} F_j} \sum_{j=1}^{N_i} \sum_{l=1}^{F_j} \phi_{i,j,l}(k), \tag{33}$$

  where $\phi_{s,j,k}(k)$ is the short-term autocorrelation sequence obtained from the l-th analysis frame of the j-th segment of the state i; F_j is the number of analysis frames, and N_i is the number of segments of state i.

- **method 2:** the mean autocorrelation sequence is derived as the average from all the "segmental" sequences,

$$\bar{\phi}_i(k) = \frac{1}{N_i} \sum_{j=1}^{N_i} \left\{ \frac{1}{F_j} \sum_{l=1}^{F_j} \phi_{i,j,l}(k) \right\}. \tag{34}$$

Finally, the coefficients $\{g_i(1), \ldots, g_i(L)\}$ can be derived from $\bar{\phi}_i(k)$ by using, e.g., the Levinson-Durbin algorithm, with K_i being the corresponding LP gain [12].

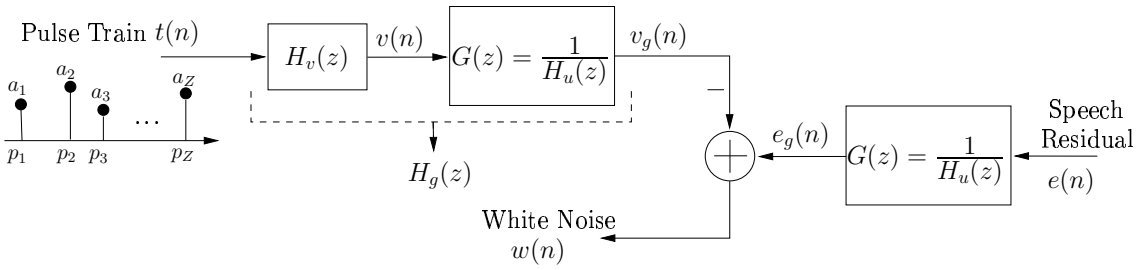

Figure 3: Scheme for the amplitude and position optimization of the non-zero samples of $t(n)$.

5 Pulse optimization

As the filters are determined through a closed-loop algorithm, the pulse positions and amplitudes of $t(n)$ are optimized. The procedure is conducted by keeping $H_v(z)$ and $H_u(z)$ for each corresponding state i constant, and minimizing the mean squared error of the system of Figure 2,

$$\varepsilon = \mathbf{w}^T \mathbf{w}. \tag{35}$$

It can be noticed that considering the pulse optimization this error minimization is the same as maximizing (8).

The goal of the pulse optimization is to approach $v(n)$ to $e(n)$ as much as possible, in a way to remove the short and long-term correlation of $u(n)$ during the filter computation process. The procedure is carried out in a similar way to the approach employed by *Multipulse Excited Linear Prediction* speech coders [3]. These algorithms attempt to construct glottal excitations which can synthesize speech by using a few position and amplitude optimized pulses. In our particular case, the optimization is performed in the neighborhood of the pulse positions, that are firstly obtained from pitch marks.

5.1 The procedure

To visualize the way the pulses are optimized, Figure 3 should be considered. The error of the system $\mathbf{w}$ is given by

$$\mathbf{w} = \mathbf{e}_g - \mathbf{v}_g = \mathbf{H}_g \mathbf{t}, \tag{36}$$

where

$$\mathbf{e}_g = \begin{bmatrix} e_g(0) & \cdots & e_g(N-1) & \cdots & e_g(N+L) \end{bmatrix}^T, \tag{37}$$

is the $N+L$ length vector containing the overall residual signal $e(n)$ filtered by $G(z)$. The impulse response matrix $\mathbf{H}_g$ is

$$\mathbf{H}_g = \begin{bmatrix} \mathbf{h}_{g1} & \mathbf{h}_{g2} & \cdots & \mathbf{h}_{gN+L-1} \end{bmatrix}, \tag{38}$$

with each respective column given by

$$\mathbf{h}_{gj} = \begin{bmatrix} 0 & \cdots & 0 & h_g\left(-\frac{M}{2}\right) & \cdots & \mathbf{h}_g\left(\frac{M}{2}+L\right) & 0 & \cdots & 0 \end{bmatrix}^T. \tag{39}$$

The vector $\mathbf{t}$ contains non-zero samples only at certain positions, i.e,

$$\mathbf{t} = \begin{bmatrix} 0 & \cdots & 0 & a_i & 0 & \cdots & 0 & a_{i+1} & \cdots & 0 \end{bmatrix}^T. \tag{40}$$

Therefore, the voiced excitation vector $\mathbf{v}$ can be written as

$$\mathbf{v} = \mathbf{H}_g\mathbf{t} = \sum_{i=1}^{Z} a_i \mathbf{h}_{gi}, \tag{41}$$

where

$$\{a_1, \ldots, a_Z\}, \tag{42}$$

$$\{p_1, \ldots, p_Z\}, \tag{43}$$

are respectively the Z amplitudes and positions of $t(n)$ which are aimed to be optimized.

5.1.1 Amplitude determination

The error to be minimized is

$$\varepsilon = \mathbf{w}^T\mathbf{w} = \left[\mathbf{e}_g - \mathbf{H}_g\mathbf{t}\right]^T\left[\mathbf{e}_g - \mathbf{H}_g\mathbf{t}\right]. \tag{44}$$

Substituting (41) into (44), the following expression results

$$\varepsilon = \mathbf{e}_g^T\mathbf{e}_g - 2\mathbf{e}_g\left[\sum_{i=1}^{Z} a_i\mathbf{h}_{gi}\right] + \sum_{i=1}^{Z} a_i^2\mathbf{h}_{gi}^T\mathbf{h}_{gi} + \sum_{i=1}^{Z} a_i\mathbf{h}_{gi}\left[\sum_{\substack{j=1 \\ j\neq i}}^{Z} a_j\mathbf{h}_{gj}\right]. \tag{45}$$

The optimal pulse amplitude a_i which minimizes (44) is thus given by

$$\frac{\partial\varepsilon}{\partial a_i} = 0, \tag{46}$$

which leads to

$$a_i = \frac{\mathbf{h}_{gi}^T\left[\mathbf{e}_g - \sum_{\substack{j=1 \\ j\neq i}}^{Z} a_j\mathbf{h}_{gj}\right]}{\mathbf{h}_{gi}^T\mathbf{h}_{gi}}. \tag{47}$$

5.1.2 Position determination

By substituting (47) into (48), an expression for the error considering the optimal amplitude is achieved,

$$\varepsilon_a = \mathbf{e}_g^T\mathbf{e}_g - 2\mathbf{e}_g^T\sum_{\substack{j=1\\ j\neq i}}^{Z} a_j\mathbf{h}_{gj} + \sum_{\substack{j=1\\ j\neq i}}^{Z} a_j^2\mathbf{h}_{gj}^T\mathbf{h}_{gj} + \\ + \sum_{\substack{j=1\\ j\neq i}}^{Z} a_j\mathbf{h}_{gj}^T\left[\sum_{\substack{r=1\\ r\neq j}}^{Z} a_r\mathbf{h}_{gr}\right] - \frac{\left\{\mathbf{h}_{gi}^T\left[\mathbf{e}_g - \sum_{\substack{j=1\\ j\neq i}}^{Z} a_j\mathbf{h}_{gj}\right]\right\}^2}{\mathbf{h}_{gi}^T\mathbf{h}_{gi}}, \quad (48)$$

where it can be seen that the only term which depends on p_i is the last one. Therefore, the best position p_i is that one which minimizes ε_a, that is,

$$p_i = \underset{p_i=1,\ldots,N}{\arg\max} \frac{\left[\mathbf{h}_{gi}^T\left(\mathbf{e}_g - \sum_{\substack{j=1\\ j\neq i}}^{Z} a_j\mathbf{h}_{gj}\right)\right]^2}{\mathbf{h}_{gi}^T\mathbf{h}_{gi}}. \quad (49)$$

6 Closed-loop 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 described in Table 1. Pitch marks may represent the best choice to construct the initial pulse trains $t(n)$. The convergence criterion is the variation of the voiced filter coefficients.

7 Conclusion

This report introduces a novel *trainable* excitation scheme for HMM-based speech synthesis. The proposed technique consists in determining voiced and unvoiced filters for each predefined state which may be represented, for instance, by leaves of phonetic decision-trees for mel-cepstral coefficients. Some experiments have shown that this new approach can synthesize speech without the artifacts imposed by the conventional simple excitation model. Furthermore, since the scheme in question is derived through an iterative procedure in which the distortion between constructed excitation and residual sequences is minimized, synthesized speech sounds closer to its natural version.

Table 1: Algorithm for joint filter computation and pulse optimization. $\mathbf{I}_X$ means identity matrix of $X \times X$ order.

$t(n)$ initialization

1) For each utterance $l = 1$ to $l = U$ do

1.1) Initialize $\{p_{l_1}, \ldots, p_{l_Z}\}$ based on the pitch marks

1.2) Optimize $\{p_{l_1}, \ldots, p_{l_Z}\}$ according to (49), considering $\mathbf{H}_g = \mathbf{I}_{N+M+L}$

1.3) Calculate $\{a_{l_1}, \ldots, a_{l_Z}\}$ according to (47), considering $\mathbf{H}_g = \mathbf{I}_{N+M+L}$

$H_v(z)$ initialization

1) For each state from $i = 1$ to $i = S$ do

1.1) Compute $\mathbf{X}_i$ and $\mathbf{y}_i$ according to (23) and (24), considering $\mathbf{G}_{i,j} = \mathbf{I}_{N_{i,j}+M+L}$

1.2) Obtain initial $\mathbf{h}_i$ by solving $\mathbf{X}_i \mathbf{h}_i = \mathbf{y}_i$

2) Set voiced filter variation tolerance: ϵ_v

3) Set the number of iterations: N_{iter} and $N_{\text{iter}_{\text{max}}}$

Recursion

1) For each state from $i = 1$ to $i = S$ do

1.1) Make $\mathbf{h}_i{}^a = \mathbf{h}_i$

1.2) Make $\varepsilon_v = 0$

1.3) Compute $\mathbf{X}_i$ and $\mathbf{y}_i$ according to (23) and (24)

1.4) Obtain $\mathbf{h}_i$ by solving $\mathbf{X}_i \mathbf{h}_i = \mathbf{y}_i$

1.5) Compute the voiced filter variation $\varepsilon_v = \varepsilon_v + (\mathbf{h}_i{}^a - \mathbf{h}_i)^T (\mathbf{h}_i{}^a - \mathbf{h}_i)$

2) For each state from $i = 1$ to $i = S$ do

2.1) Obtain the mean autocorrelation, $\bar{\phi}_i(k)$ according to (33) or (34)

2.2) Compute $\mathbf{g}_i$ and K_i using the Levinson-Durbin algorithm

3) If $\varepsilon_v < \epsilon_v$ or $N_{\text{iter}} = N_{\text{iter}_{\text{max}}}$, go to (6)

4) For each utterance $l = 1$ to U do

4.1) Optimize $\{p_{1_l}, \ldots, p_{Z_l}\}$ according to (49)

4.2) Calculate $\{a_{1_l}, \ldots, a_{Z_l}\}$ according to (47)

5) Return to (1)

6) End

References

[1] ABDEL-HAMID, O., ABDOU, S., AND RASHWAN, M. Improving the Arabic HMM based

speech synthesis quality. In *Proc. of ICSLP* (2006).

[2] AKAMINE, M., AND KAGOSHIMA, T. Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS drive TTS). In *Proc. ICSLP* (1998).

[3] CHU, W. *Speech Coding Algorithms*. Wiley-Interscience, USA, 2003.

[4] FUKADA, T., TOKUDA, K., KOBAYASHI, T., AND IMAI, S. An adaptive algorithm for mel-cepstral analysis of speech. In *Proc. of ICASSP* (1992).

[5] HUNT, A., AND BLACK, A. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of ICASSP* (1996).

[6] ITAKURA, F., AND SAITO, S. A statistical method for estimation of speech spectral density and formant frequencies. In *Speech Analysis*, R. Schafer and J. Markel, Eds. IEEE Press, 1979.

[7] JACKSON, L. B. *Digital Filters and Signal Processing*, 3 ed. Kluwer Academic Publishers, 1996.

[8] KAWAHARA, H., MASUDA-KATSUSE, I., AND DE CHEVEIGNÉ, A. 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 27*, 3-4 (Apr. 1999).

[9] KAY, S. *Modern Spectral Estimation*. Prentice-Hall, USA, 1988.

[10] KIM, S. J., AND HAHN, M. Two-band excitation for HMM-based speech synthesis. *IEICE Trans. Inf. & Syst. E90-D* (2007).

[11] MAIA, R., ZEN, H., TOKUDA, K., KITAMURA, T., AND RESENDE, F. G. Towards the development of a Brazilian Portuguese text-to-speech system based on HMM. In *Proc. of EUROSPEECH* (2003).

[12] MARKEL, J., AND GRAY, A. *Linear Prediction of Speech*. Springer-Verlag, New York, 1982.

[13] MASUKO, T., TOKUDA, K., KOBAYSHI, T., AND IMAI, S. Voice characteristics conversion for HMM-based speech synthesis system. In *Proc. of ICASSP* (1997).

[14] MCCREE, A., TRUONG, K., GEORGE, E., BARNWELL, T., AND VISWANATHAN, V. A 2.4 kbits/s MELP candidate for the U.S. Fdereal Standard. In *Proc. of ICASSP* (2006).

[15] MCREE, A., AND BARNWELL III, T. A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Trans. on Speech and Audio Processing 3*, 4 (July 1995).

[16] SHICHIRI, K., SAWABE, A., YOSHIMURA, T., TOKUDA, K., MASUKO, T., KOBAYASHI, T., AND KITAMURA, T. Eigenvoices for HMM-based speech synthesis. In *Proc. of ICSLP* (2002).

[17] TOKUDA, K., KOBAYASHI, T., AND IMAI, S. Adaptive cepstral analysis of speech. *IEEE Trans. Speech and AUdio Processing 3*, 6 (Nov. 1995).

[18] TOKUDA, K., YOSHIMURA, T., MASUKO, T., KOBAYASHI, T., AND KITAMURA, T. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. of ICASSP* (2000).

[19] TOKUDA, K., ZEN, H., AND BLACK, A. W. An HMM-based speech synthesis applied to English. In *Proc. of IEEE Workshop in Speech Synthesis* (2002).

[20] WELCH, P. The use of Fast Fourier Transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Trans. Audio and Electroacoustics 15*, 2 (June 1967).

[21] YAMAGISHI, J., ONISHI, K., MASUKO, T., AND KOBAYASHI, T. Modeling of various speaking styles and emotions for HMM-based speech synthesis. In *Proc. of EUROSPEECH* (2003).

[22] YAMAGISHI, J., TACHIBANA, M., MASUKO, T., AND KOBAYASHI, T. Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis. In *Proc. of ICASSP* (2004).

[23] YOSHIMURA, T., TOKUDA, K., MASUKO, T., KOBAYASHI, T., AND KITAMURA, T. Speaker interpolation in HMM-based speech synthesis. In *Proc. of EUROSPEECH* (1997).

[24] YOSHIMURA, T., TOKUDA, K., MASUKO, T., KOBAYASHI, T., AND KITAMURA, T. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. of EUROSPEECH* (1999).

[25] YOSHIMURA, T., TOKUDA, K., MASUKO, T., KOBAYASHI, T., AND KITAMURA, T. Mixed-excitation for HMM-based speech synthesis. In *Proc. of EUROSPEECH* (2001).

[26] ZEN, H., TODA, T., NAKAMURA, M., AND TOKUDA, K. Details of the Nitech HMM-based speech synthesis for Blizzard Challenge 2005. *IEICE Trans. on Inf. and Systems E90-D*, 1 (2007).