Multipulse Sequences for Residual Signal Modeling

Ranniery Maia, Heiga Zen, Kate Knill, M. J. F. Gales, Sabine Buchholz

Toshiba Research Europe Ltd., Cambridge Research Lab, Cambridge, UK

{ranniery.maia,heiga.zen,kate.knill,mjfg,sabine.buchholz}@crl.toshiba.co.uk

10.21437/Interspeech.2011-41

Abstract

In source-filter models of speech production, the residual signal - what remains after passing the speech signal through the inverse filter - contains important information for the generation of naturally sounding re-synthesized speech. Typically, the voiced regions of residual signals are regarded as a mixture of glottal pulse and noise. This paper introduces a novel approach to represent the noise component of voiced regions of residual signals through autoregressive filtering of multipulse sequences. The positions and amplitudes of the non-zero samples of these multipulse signals are optimized through a closed-loop procedure. The method in question is applied to excitation modeling in statistical parametric synthesis. Experimental results indicate that the use of multipulse-based noise component construction eliminates the necessity of run-time *ad hoc* procedures such as high-pass filtering and time modulation, common on excitation models for statistical parametric synthesizers, with no loss of synthesized speech quality.

Index Terms: speech synthesis, source modeling, noise modeling, speech analysis, speech coding.

1. Introduction

Source-filter models of speech production have applications in several speech processing fields, e.g. coding, synthesis and voice transformation. Traditionally, it is assumed that the speech signal is generated by the convolution of the impulse response of a minimum-phase time-varying filter with an *excitation* signal [1]. Although estimation of the parameters of the filter is extremely important for the quality of the re-synthesized speech, parameterization of the residual signal - obtained by passing the original speech signal through the inverse filter - is important for synthesizing naturally sounding speech. Usually, the simplest way to model the residual signal is by representing it as either a delta pulse or a white noise sequence for voiced or unvoiced speech segments, respectively. This assumption is the basis of the traditional linear prediction (LP) coefficient analysis theory [1]. In more sophisticated ways to model the excitation signal, voiced excitation is assumed to be composed of a mixture of a glottal pulse and noise. Although more attention has been given to the modeling of the glottal pulse [2], reproduction of the noise component of voiced excitation is also important, since its neglect or over-estimation may result in muffled or harsh re-synthesized speech, respectively. Typically, the construction of the noise component is performed through *ad hoc* adjustments on white noise sequences. Indeed, the mixture of noise with pulse for excitation modeling has been one of the fundamental issues in most of the mixed excitation techniques applied to both speech synthesis and speech coding fields. Justified by observations of the speech spectrum, it is usually assumed that noise mostly constitutes the high frequency part of the excitation signal. Based on this assumption, a number of approaches for speech coding, synthesis and analysis attempt to model the noise component by applying a high-pass filter, among other procedures, to white noise sequences in order to mix it with pulse, e.g. [3, 4]. In terms of analytic approaches to represent the the noise component, [5] shows that is possible to apply the Hilbert and energy envelopes instead of the pitch-synchronous triangular window proposed by [3].

This paper introduces a method to model the noise component of voiced excitation. The noise component of the excitation signal, to be added with pulse in order to mimic the residual, is approximated by the autoregressive (AR) filtering of multipulse sequences. The amplitudes and positions of the non-zero samples of these sparse sequences are optimized through a closed-loop procedure in analogy with code-excited linear predictive (CELP) speech coders [6]. In the proposed method, one multipulse sequence represents pitch-synchronous portions of several noise component segments that share similar phonemic/phonetic characteristics. This cluster dependent optimization process resembles approaches applied to the design of multipulse stochastic codebooks in CELP speech coders, e.g. [7]. When implemented in statistical parametric speech synthesis [8], the proposed model results in a consistent approach for source modeling, suitable for joint source-spectrum estimation frameworks applicable to in speech synthesis [9].

This paper is organized as follows. Section 2 describes the proposed noise component model; Section 3 shows how the approach in question can be applied to statistical parametric speech synthesis; Section 4 presents some experiments; and the conclusions are given in Section 5.

2. Noise component modeling

2.1. The noise component model

The idea is to approximate the noise component signal $u(n)$ by the following convolution

$$\hat{u}(n) = h_u(n) * w(n), \tag{1}$$

where $\left\{w\left(-\frac{T}{2}\right), \ldots, w\left(\frac{T}{2}\right)\right\}$, is a $T+1$-length non-causal multipulse sequence, and $\{h_u(0), \ldots, h_u(P)\}$ is a P-th order approximation of the impulse response of an AR filter,

$$H_u(z) = \frac{1}{\sum_{l=0}^{L} g(l) z^{-i}}, \tag{2}$$

whose coefficients are $\{g(0), \ldots, g(L)\}$. The sequence $w(n)$ is defined by the positions, $\{l_0, \ldots, l_{Z-1}\}$, and amplitudes, $\{m_0, \ldots, m_{Z-1}\}$, of its Z non-zero samples. The convolution $w(n) * h_u(n)$ is assumed to represent pitch-synchronous portions of segments of the noise component $u(n)$.

In order to track a variety of different sounds, segments that share the same phonetic/phonemic properties are clustered together, and one multipulse sequence is optimized so as to represent the noise component of the cluster.

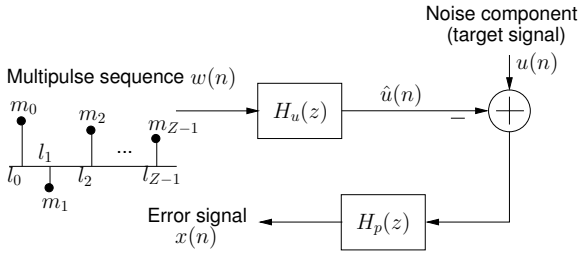

Figure 1: *Illustration of the multipulse optimization process using analysis-by-synthesis.*

2.2. Multipulse sequence optimization

Figure 1 illustrates the optimization procedure in which the positions and amplitudes of $w(n)$ are determined, where a strong resemblance to multipulse linear predictive speech coders [6] can be noticed. The error weighting filter $H_p(z)$ emphasizes the spectrum of the error signal $x(n) = u(n) - \hat{u}(n)$ over specific frequencies. The coefficients $\{g(0), \ldots, g(L)\}$ of the AR filter $H_u(z)$ are obtained by LP analysis [1] of $u(n)$.

2.2.1. MSE criterion

Figure 2 illustrates the convolution $\hat{u}(n) = h_u(n) * \hat{w}(n)$ in matrix notation, for all the sentences of the database concatenated in one vector with J pitch onsets. All the samples belong to the same cluster. The reconstructed noise component vector $\hat{\boldsymbol{u}}$ is given by

$$\hat{\boldsymbol{u}} = \sum_{j=0}^{J-1} \hat{\boldsymbol{u}}_j = \sum_{j=0}^{J-1} \boldsymbol{H}_{u,j}\boldsymbol{w}, \tag{3}$$

where $\{\hat{\boldsymbol{u}}_0, \ldots, \hat{\boldsymbol{u}}_{J-1}\}$ are pitch-synchronous segments produced by the product $\boldsymbol{H}_{u,j}\boldsymbol{w}$, with $\boldsymbol{H}_{u,j}$ being a matrix as illustrated in Figure 2 and $\boldsymbol{w} = \left[w\left(-\frac{T}{2}\right) \quad \cdots \quad w\left(\frac{T}{2}\right)\right]^\top$ being the multipulse sparse sequence to be optimized. Using the analysis-by-synthesis optimization procedure illustrated in Figure 1, the mean square error (MSE) is

$$\varepsilon = \frac{1}{N}\left[\boldsymbol{u} - \sum_{j=0}^{J-1} \boldsymbol{H}_{u,j}\boldsymbol{w}\right]^\top \left[\boldsymbol{u} - \sum_{j=0}^{J-1} \boldsymbol{H}_{u,j}\boldsymbol{w}\right], \tag{4}$$

where N is the number of elements of $\boldsymbol{u}$. Assuming that $\boldsymbol{u}$ can be split into J pitch synchronous overlapping segments, then

$$\boldsymbol{u} = \sum_{j=0}^{J-1} \boldsymbol{u}_j. \tag{5}$$

Also, it can be noticed from Figure 2 that

$$\boldsymbol{H}_{u,j}\boldsymbol{w} = \sum_{z=0}^{Z-1} m_z \boldsymbol{h}_u^{(z)}, \tag{6}$$

where $\{m_0, \ldots, m_{Z-1}\}$ are the amplitudes of the Z non-zero samples of $w(n)$, and

$$\boldsymbol{h}_u^{(z)} = \left[\underbrace{0 \cdots 0}_{z} \quad h_u(0) \quad \cdots \quad h_u(P) \quad \underbrace{0 \cdots 0}_{T-z}\right]^\top. \tag{7}$$

After substituting (5) and (6) into (4), the MSE ε becomes

$$\varepsilon = \frac{1}{N}\sum_{j=0}^{J-1}\left[\tilde{\boldsymbol{u}}_j - \sum_{z=0}^{Z-1} m_z \boldsymbol{h}_u^{(z)}\right]^\top \left[\tilde{\boldsymbol{u}}_j - \sum_{z=0}^{Z-1} m_z \boldsymbol{h}_u^{(z)}\right]. \tag{8}$$

In (8), $\tilde{\boldsymbol{u}}_j$ is the portion of $\boldsymbol{u}_j$ whose samples are non-zero, i.e.,

$$\tilde{\boldsymbol{u}}_j = \left[u_j\left(p_j - \tfrac{T}{2}\right) \quad \cdots \quad u_j\left(p_j + \tfrac{T}{2} + P\right)\right]^\top. \tag{9}$$

2.2.2. Position and amplitude determination

The single pulse amplitude $\hat{m}_z$ that minimizes (8) can be derived by making $\frac{\partial \varepsilon}{\partial m_z} = 0$, i.e.,

$$\hat{m}_z = \frac{\boldsymbol{h}_u^{(z)\top}\left[\bar{\boldsymbol{u}} - \sum_{\substack{r=0 \\ r\neq z}}^{Z-1} m_r \boldsymbol{h}_u^{(r)}\right]}{\boldsymbol{h}_u^{(z)\top}\boldsymbol{h}_u^{(z)}}, \tag{10}$$

where it can be demonstrated that

$$\bar{\boldsymbol{u}} = \frac{1}{J}\sum_{j=0}^{J-1} \tilde{\boldsymbol{u}}_j. \tag{11}$$

The best position of the z-th pulse, $\hat{l}_z$, can be found by minimizing the MSE, after making $m_z = \hat{m}_z$ in (8), resulting in

$$\hat{l}_z = \underset{l_z = -\frac{T}{2}, \ldots, \frac{T}{2}}{\arg\max} \left\{ \frac{\left(\boldsymbol{h}_u^{(z)\top}\left[\bar{\boldsymbol{u}} - \sum_{\substack{r=0 \\ r\neq z}}^{Z-1} m_r \boldsymbol{h}_u^{(r)}\right]\right)^2}{\boldsymbol{h}_u^{(z)\top}\boldsymbol{h}_u^{(z)}} \right\}. \tag{12}$$

Once $\hat{l}_z$ is determined, the $z+1$-amplitude vector

$$\boldsymbol{m} = \left[m_0 \quad \cdots \quad m_z\right]^\top, \tag{13}$$

is calculated by solving

$$\boldsymbol{m} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}, \tag{14}$$

$$\boldsymbol{\Gamma} = \begin{bmatrix} \boldsymbol{h}_u^{(0)\top}\boldsymbol{h}_u^{(0)} & \cdots & \boldsymbol{h}_u^{(0)\top}\boldsymbol{h}_u^{(z)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{h}_u^{(z)\top}\boldsymbol{h}_u^{(0)} & \cdots & \boldsymbol{h}_u^{(z)\top}\boldsymbol{h}_u^{(z)} \end{bmatrix}, \tag{15}$$

$$\boldsymbol{\gamma} = \left[\boldsymbol{h}_u^{(0)\top}\bar{\boldsymbol{u}} \quad \cdots \quad \boldsymbol{h}_u^{(z)\top}\bar{\boldsymbol{u}}\right]^\top. \tag{16}$$

2.2.3. Estimation of $\boldsymbol{u}_j$ and training procedure

Decomposition of $\boldsymbol{u}$ into its pitch synchronous components $\{\boldsymbol{u}_0, \ldots, \boldsymbol{u}_{J-1}\}$ is conducted as follows. After pitch onset detection (pitch marking), pitch periods are modified to a normalized pitch period, with $\frac{T}{2}$ samples, by using the Pitch-Synchronous Overlap and Add (PSOLA) [10]. After that, hanning windows are applied to $T+1$ samples, with center at each pitch onset p_j to obtain $\left\{\tilde{u}_j\left(-\frac{T}{2}\right), \ldots, \tilde{u}_j\left(\frac{T}{2}\right)\right\}$. Samples $\left\{\tilde{u}_j\left(\frac{T}{2}+1\right), \ldots, \tilde{u}_j\left(\frac{T}{2}+P\right)\right\}$ are derived by exciting the AR filter $H_u(z)$ to the zero input with non-zero memory.

Optimization of the multipulse sequence is done through the following steps:

1. estimate $\tilde{\boldsymbol{u}}_j$ for each pitch onset;
2. for pulse $z = \{0, \ldots, Z-1\}$:
   (a) determine the best position l_z according to (12);
   (b) calculate the $z+1$ amplitudes using (14).

The normalized pitch period, $\frac{T}{2}$, and number of pulses, Z, are set beforehand.

2.3. Synthesis

Synthesis of the noise component $\hat{u}(n)$ is done as follows. First, equally spaced pitch marks are constructed according to the normalized pitch period $\frac{T}{2}$. After that, copies of $w(n)$ are placed at the center of each auxiliary pitch mark and the resulting signal filtered through $H_u(z)$. Finally, as the last step, the PSOLA algorithm is utilized to modify the equally spaced pitch onsets into the original ones, $\{p_0, \ldots, p_{J-1}\}$.

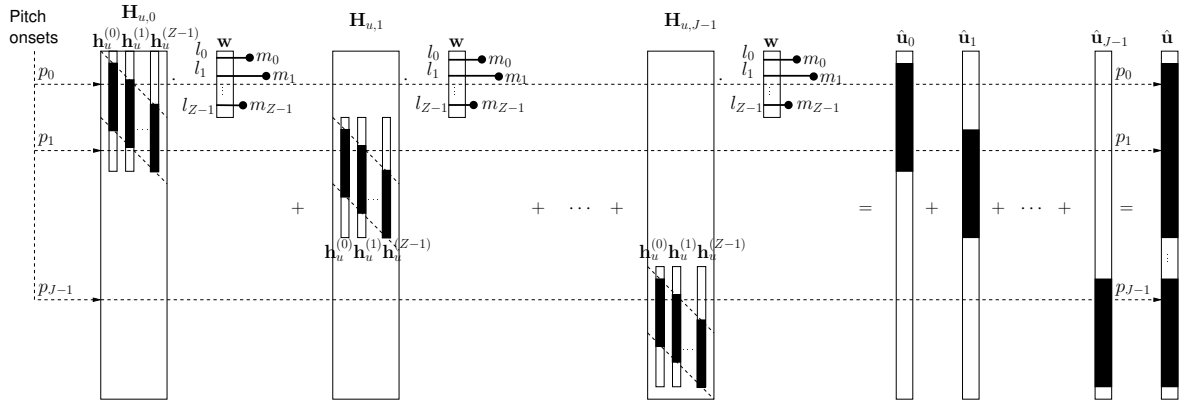

Figure 2: *Illustration of how the noise component is modeled through the convolution $\hat{u}(n) = h_u(n) * w(n)$, using matrix notation. Parts in dark represent the possible non-zero samples belonging to a given phonetic cluster. It can be seen that the convolution is performed pitch synchronously, with the resulting partial vectors $\{\hat{\mathbf{u}}_0, \ldots, \hat{\mathbf{u}}_{J-1}\}$ being overlapped and added to form the vector $\hat{\mathbf{u}}$.*

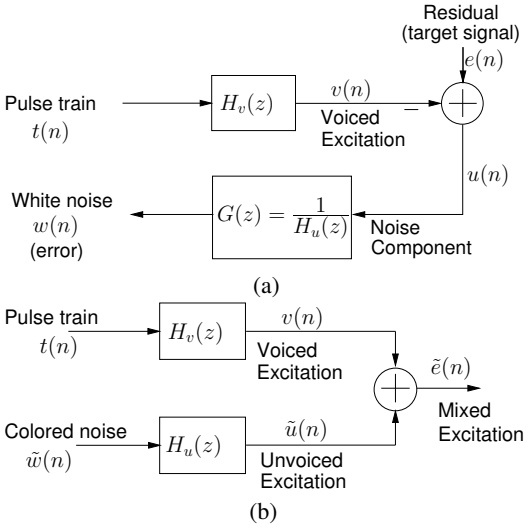

Figure 3: *Mixed excitation based on state dependent filtering of noise and pulse train: (a) training part; (b) synthesis part.*

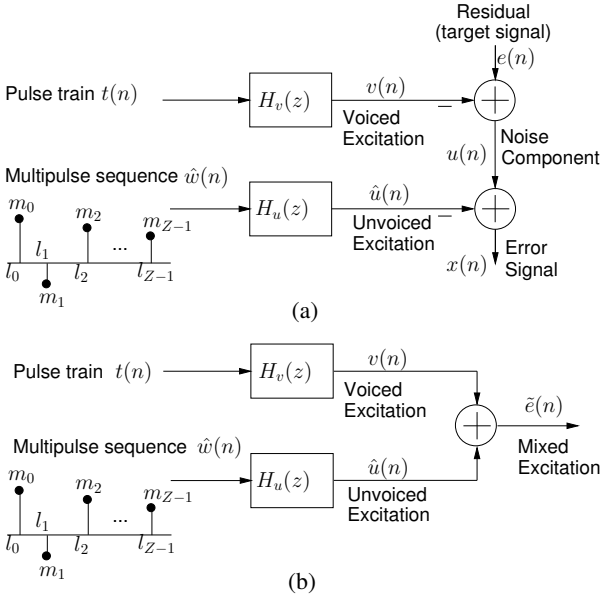

Figure 4: *The state dependent mixed excitation framework augmented with multipulse sequences for noise modeling, where the error of the system is $x(n)$ and $w(n)$ is approximated by a state-dependent sparse sequence $\hat{w}(n)$: (a) training part; (b) synthesis part.*

3. Application to speech synthesis

The proposed noise component model can be applied to excitation modeling in statistical parametric speech synthesis [8].

3.1. Implementation under state dependent filtering

Figure 3(a) shows the block diagram of the training part of the state-dependent mixed excitation method of [11], where pitch marks, residual and state configuration are utilized to train the state-dependent filters $H_v(z)$ and $H_u(z)$. At synthesis time, shown in Figure 3(b), the trained filter coefficients, a F_0-generated pulse train, and the noise signal $\tilde{w}(n)$ are used to construct the excitation signal $\tilde{e}(n)$. *Excitation states* can be regarded as terminal nodes of the decision trees for the spectrum. At training time, shown in Figure 3(a), the filter coefficients are calculated by assuming that the error of the system $w(n)$ is a Gaussian white noise signal. However, due to limitations of the model $w(n)$ is usually colored, i.e., the noise component $u(n)$ still retains significant residual information. Although some work has been done on this problem, e.g. [12], this issue can also be significantly reduced at synthesis time by replacing the theoretical white noise sequence $w(n)$ by a colored noise $\tilde{w}(n)$. A typical way to derive $\tilde{w}(n)$ is to apply a high-pass filter to a white noise sequence followed by pitch-synchronous time modulation using a triangular window in the same way as in harmonic plus noise modeling of speech [3]. This coloring procedure is considered in Section 4.

3.2. Augmented model

The application of the noise component model to the mixed excitation framework in question is done via the augmented excitation model shown in Figure 4. Note that the multipulse sequence, denoted by $\hat{w}(n)$ in this case, represents an approximation to the ideal noise signal $w(n)$ of Figure 3(a). The unvoiced excitation $u(n) = e(n) - h_v(n) * t(n)$ is regarded as the target signal for the noise model, and the pulse positions $\{p_0, \ldots, p_{J-1}\}$ of $t(n)$ are the pitch onsets. The training procedure of this new model is composed of three steps as follows: (1) $t(n)$ optimization; (2) $H_v(z)$ and $H_u(z)$ coefficient calculation; (3) $\hat{w}(n)$ optimization. Details of $t(n)$ optimization and determination of the filter coefficients can be found in [11].

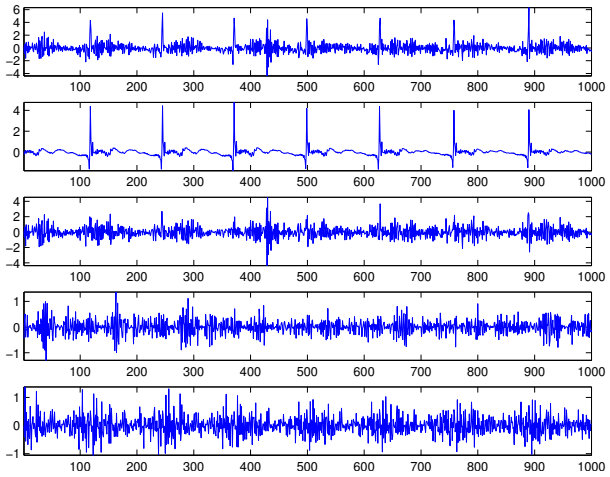

Figure 5: *Waveform examples. (see Figures 3 and 4). From top to bottom: residual $e(n)$; voiced excitation $v(n)$; ideal noise component $u(n) = e(n) - v(n)$; noise component produced by the baseline system (with coloring); noise component produced by the augmented system.*

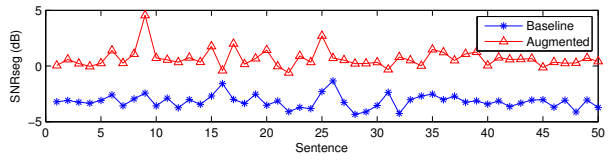

Figure 6: *Segmented SNR ratio for 50 re-synthesized sentences.*

4. Experimental results

A database of 4898 sentences sampled at 16 kHz, uttered by an American English female speaker was utilized to verify the effectiveness of the multipulse-based noise component model when applied to speech synthesis. For that, a statistical parametric synthesizer was trained. For spectral parameterization, 40 mel-cepstral coefficients were calculated from smooth spectra. Mel-cepstral coefficients, F_0, and their respective dynamic features were used as observation vectors of hidden semi-Markov models. The training procedure followed the configurations of the demos released as part of the hidden Markov model-based speech synthesis toolkit [13]. After that, an excitation model as outlined in Section 3.1 was trained, where leaves of the trees for the spectrum were regarded as filter states. Voiced and unvoiced filter orders were 256 and 64, respectively. Finally, state dependent multipulse sequences for the noise component, regarded as the signal $u(n) = e(n) - v(n)$ from the excitation model, were trained. The order of the multipulse sequences was set to twice the average pitch period of the database: $T = 198$ samples, and the number of pulses was $Z = 41$. The error weighting filter $H_p(z)$ comprised a 32-th order linear phase high-pass filter with cutoff frequency at 2 kHz, chosen so as to weight the error $x(n)$ mostly on the high frequencies.

Figure 5 shows some waveforms of residual, voiced excitation and noise components produced by the baseline with coloring procedure and augmented excitation models. The natural modulation at pitch onsets in the multipulse based noise component is visible. Figure 6 shows the segmented signal-to-noise ratios (SNRseg) between natural and re-synthesized closed sentences for the baseline and augmented models. Speech synthesized by the augmented system is closer to the natural versions. For the re-synthesis, natural spectrum and F_0, Viterbi-aligned state durations and optimized pitch marks were used.

Table 1: *Subjective test results where the mean preferences (%) and p-values are shown.*

Test 1: baseline with no noise coloring			
Baseline	Augmented	No Pref.	p-value
22.61	52.55	24.85	0

Test 2: baseline with noise coloring			
Baseline	Augmented	No Pref.	p-value
35.4	31.3	33.3	0.183

Two preference tests were conducted with 25 synthesized sentences under a real text-to-speech scenario. The baseline system with and without coloring procedure for the noise component was used in tests one and two, respectively. The augmented system was the same on both tests, with the noise component being synthesized as described in Section 2.3. The tests were conducted through a web recruitment system with 34 and 26 subjects, respectively. Table 1 shows the results, where the baseline only becomes comparable in quality to the augmented system (there is no significant difference between the systems according to the p-value bigger than 0.05), when its noise signal is obtained with the coloring process described in Section 3.1.

5. Conclusions

Optimized multipulse sequences were shown to be effective to model the noise component of residual signals. When applied to source modeling in statistical parametric synthesis, the need to empirically adjust the noise at run-time was removed without degrading the quality of the synthesized speech. This consistent approach is suitable for joint modeling frameworks in statistical speech synthesis [9].

6. References

[1] J. R. Deller, Jr., J. H. L. Hansen, and J. G. Proaks, *Discrete-Time Processing of Speech Signals*. IEEE Press Classic Reissue, 2000.

[2] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of the glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 001–013, 1985.

[3] Y. Stylianou, J. Laroche, and E. Moulines, "High-quality speech modification based on a harmonic + noise model," in *Proc. of Eurospeech*, pp. 451–454, 1995.

[4] P. Lanchantin, G. Derottex, and X. Rodet, "An HMM-based speech synthesis system using a new glottal source and vocal-tract separation method," in *Proc. of ICASSP*, pp. 4630–4633, 2010.

[5] Y. Pantazis and Y. Stylianou, "Improving the modeling of the noise part in the harmonic plus noise model of speech," in *Proc. of ICASSP*, pp. 4609–4612, 2008.

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

[7] L. da Silva and A. Alcaim, "Multipulse stochastic codebook for CELP speech coders (in Portuguese)," *Journal of the Brazilian Telecomm. Society*, vol. 13, pp. 83–91, Dec. 1998.

[8] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Comm.*, vol. 51, pp. 1039–1064, Nov. 2009.

[9] R. Maia, H. Zen, and M. Gales, "Statistical parametric speech synthesis with joint estimation of acoustic and excitation model parameters," in *Proc. of ISCA SSW7*, pp. 88–93, 2010.

[10] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Comm.*, vol. 9, pp. 453–467, Dec. 1990.

[11] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," in *Proc. of ISCA SSW6*, pp. 131–136, 2007.

[12] Y. Shiga, T. Toda, S. Sakai, and H. Kawai, "Improved training of excitation for HMM-based parametric speech synthesis," in *Proc. of Interspeech*, pp. 809–812, 2010.

[13] http://hts.sp.nitech.ac.jp. As of 30 March 2011.