Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (New York, NY), pp. 147-150, April 1988

S4.2

SYNTHESIS FILTER OPTIMIZATION AND CODING: APPLICATIONS TO CELP

P. Kabal[1,2], J.-L. Moncet[2], and C. C. Chu[2]

[1]Electrical Engineering
McGill University
Montreal, Quebec H3A 2A7

[2]INRS-Télécommunications
Université du Québec
Verdun, Quebec H3E 1H6

Abstract

The success of Code-Excited Linear Prediction (CELP) for coding speech signals depends on the accurate representation of the pitch structure and the formant structure of the input speech. In this type of coder, an excitation waveform chosen from a dictionary of waveforms drives a cascade of a pitch and a formant synthesis filter. This paper develops the methodology to allow for a joint optimization of the waveform selection process, waveform scaling, and the pitch filter determination. Methods to accommodate high-pitch speakers (pitch lag smaller than the analysis frame size) are given. Additionally, the requirements for coding the synthesis parameters into a bit stream at 4.8 kb/s are discussed.

1. Introduction

This paper examines Code-Excited Linear Prediction (CELP) for the coding of speech signals. This class of coder uses block coding of the excitation signal to achieve high quality speech reproduction at medium to low bit rates [1]. The purpose of this paper is to describe methods that allow for the optimization of the synthesis filters for the excitation waveform chosen. In addition, the coding strategy for the synthesis parameters at a bit rate of 4.8 kb/s is described.

2. Code Excited Linear Predictive Coding

In CELP, each trial waveform is synthesized by passing it through a cascade of synthesis filters. The first part of the cascade, termed the pitch synthesis filter, inserts pitch periodicities into the reconstructed speech. The second filter is the formant synthesis filter which introduces a frequency shaping related to the formant resonances produced by the human vocal tract. The synthesis stage of an CELP coder is shown in Fig. 1.

In CELP, the excitation waveform is chosen from a dictionary of waveforms. Conceptually, each waveform in the dictionary is passed through the synthesis filters to determine which waveform "best" matches the input speech. The optimality criterion uses a frequency weighted mean-square error criterion. The index of the "best" waveform is transmitted to the decoder. In addition, both the formant and pitch filters are updated periodically. The parameters of these filters are sent to the decoder as side information to allow it to form the appropriate synthesis filters.

The CELP coder does not directly need an analysis stage. Ideally the synthesis filters would be optimized for each trial waveform. The formulation of an optimal (in a mean-square sense) formant synthesis filter leads to a highly non-linear set of equations which is not amenable to solution. However, the formant filter can be implemented as the inverse of a filter determined by an analysis step.

The pitch synthesis filter can also be determined by analyzing the input speech. Indeed, some implementations of CELP utilize this form of analysis. However, under certain constraints, the pitch synthesis filter can be chosen to optimize the reconstructed speech signal.

3. Formant Filter

Conventionally the formant synthesis filter is an all-pole structure. This structure is consistent with a vocal tract model and has been shown to be able to produce good quality speech. The usual approach is to derive the corresponding synthesis filter by analyzing the input speech. The inverse filter (an all-zero prediction error filter) is determined by finding the filter which minimizes the mean-square prediction error.

A number of different analysis methods can be used to find the formant prediction error filter. The autocorrelation approach leads to a set of equations which is efficiently solved and which gives a minimum phase solution and hence a stable synthesis filter.

The residual minimizing criterion can be shown to result in a synthesis filter which matches the formant envelope of the original speech. The fit is such that formant peaks are better matched than the valleys. This fit is appropriate for human speech perception, as it has been shown that these features are important for good quality.

The residual matching property can be expressed in the frequency domain for an autocorrelation analysis. The expression for the error is given by [2]

$$\overline{e_d^2} = \frac{g^2}{2\pi} \int_{-\pi}^{\pi} \frac{S(e^{j\omega})^2}{H(e^{j\omega})^2} \, d\omega \, , \qquad (1)$$

where $S(e^{j\omega})$ is the Fourier transform of windowed input sequence, $H(e^{j\omega})$ is the Fourier transform of the synthesis filter, and g is an appropriate constant that will be chosen to normalize the output value. Minimizing the prediction error also minimizes the value of the frequency domain integral in the above equation. This integral can be interpreted as measuring the fit between $H(e^{j\omega})$ and $S(e^{j\omega})$. The contribution will be largest and the fit will be enhanced in those regions of the spectrum in which $S(e^{j\omega})$ is large, i.e. at the formant peaks.

4. Synthesis Optimization

A new waveform is selected for each block of samples. In addition, the pitch and formant synthesis filters are updated at intervals. For convenience, we will refer to a frame of samples. Each frame of samples will be subdivided into subframes. The formant filter is updated once per frame, while the gain, pitch filter parameters and waveform selection are updated at the subframe level.

The excitation waveform for the current block (subframe) is $x^{(i)}(n)$. This is scaled by the gain factor G and used to drive the pitch synthesis filter. The pitch synthesis filter is specified by a pitch lag M and a set of N_p filter coefficients. The criterion used to measure the error in the synthesized signal is based on a frequency weighting. The transfer function of the weighting filter is given by $W(z) = H(\gamma z)/H(z)$, where γ is a bandwidth expansion factor. The role of the weighting filter is to concentrate the coding noise in the formant regions where it is effectively masked by the speech signal.

With the given form of the weighting filter, the calculation of the frequency weighted error can be rearranged as shown in the block diagram in Fig. 2. In this arrangement, the formant synthesis filter and the weighting filter have been combined to form a bandwidth-expanded synthesis filter. The notation for the signals uses primes (e.g. $s'(n)$) to indicate signals which use the bandwidth-expanded synthesis filter and carets (e.g. $\hat{s}(n)$) to indicate coded signals.

The waveform index i, the gain factor G and the pitch filter parameters will be chosen to minimize the mean-square frequency weighted reconstruction error in the interval $0 \le n < N$.

$$\varepsilon = \sum_{n=0}^{N-1} e_w^2(n) \, . \qquad (2)$$

where the weighted error is given by

$$e_w(n) = s'(n) - \sum_{k=-\infty}^{\infty} \hat{d}(k)h'(n-k) . \tag{3}$$

and $\{h'(k)\}$ denotes the impulse response of the bandwidth-expanded synthesis filter. The output of the pitch synthesis filter can be written as

$$\hat{d}(n) = Gx^{(i)}(n) - \sum_{j=1}^{N_p} \beta_j \hat{d}(n-M-j+1) . \tag{4}$$

It is convenient to rewrite the weighted error $e_w(n)$ in a form with the terms which are not affected by the optimization lumped into a single term.

$$e_w(n) = s''(n) - \sum_{k=0}^{\infty} \hat{d}(k)h'(n-k) . \tag{5}$$

The limits of the convolution sum serve to select a portion of the signal. It is useful to define a window function,

$$w_{[L,U)}(n) = \begin{cases} 1 & L \le n < U \\ 0 & \text{otherwise} . \end{cases} \tag{6}$$

Now the weighted error becomes

$$e_w(n) = s''(n) - G\,\hat{x}^{(i)}_{0,N)}(n) - \sum_{j=1}^{N_p} \beta_j \hat{d}_{0,N)}(n, M-j+1) , \tag{7}$$

where the filtered versions of $x^{(i)}(n)$ and $\hat{d}(n)$ have been defined as

$$\begin{aligned} \hat{x}^{(i)}_{L,U)}(n) &= \sum_{k=-\infty}^{\infty} w_{[L,U)}(k)\, x^{(i)}(k)h'(n-k) \\ \hat{d}_{L,U)}(n,m) &= \sum_{k=-\infty}^{\infty} w_{[L,U)}(k)\, \hat{d}(k-m)h'(n-k) . \end{aligned} \tag{8}$$

Solution for $M \ge N$

The values of the gain factor G and the coefficients $\{\beta_i\}$ which minimize the squared-error are to be found. In matrix form, $\mathbf{\Phi a} = \mathbf{b}$, where

$$\mathbf{\Phi} = \sum_{n=0}^{N-1} \mathbf{v}^{(n)}\mathbf{v}^{(n)T} , \qquad \mathbf{b} = \sum_{n=0}^{N-1} s''(n)\mathbf{v}^{(n)} , \tag{9}$$

and

$$\mathbf{v}^{(n)} = \begin{bmatrix} \hat{x}^{(i)}_{0,N)}(n) \\ \hat{d}_{0,N)}(n, M) \\ \hat{d}_{0,N)}(n, M+1) \\ \vdots \\ \hat{d}_{0,N)}(n, M-N_p+1) \end{bmatrix} , \qquad \mathbf{a} = \begin{bmatrix} G \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{N_p} \end{bmatrix} . \tag{10}$$

Consider the case that $M \ge N$. The filtered signal $\hat{d}_{0,N)}(n,m)$ which appears in $\mathbf{v}^{(n)}$ depends only on the signal $\hat{d}(n)$ for $n < 0$. This part of the signal is known from the previous subframe. For a pitch lag larger than the subframe size, the matrix $\mathbf{\Phi}$ and the righthand side vector $\mathbf{b}$ are known quantities. The determination of the optimum coefficients involves solving the set of linear simultaneous equations.

4.1 Performance with the Optimized Parameters

The solution method considered finds the jointly optimal values of the waveform index i, the pitch lag M, the gain G, and the pitch filter coefficients $\{\beta_i\}$. This is accomplished by finding the optimal coefficient vector for each pair (i, M). The pitch lag M is constrained to be at least as large as N.

The joint optimization can be compared with a strategy which chooses the pitch filter parameters (M and $\{\beta_i\}$) by analyzing the input speech. For this comparison the sampling frequency is 8 kHz. The frame size is 80 samples (10 ms) for the formant filter update and the subframe size is 40 samples (5 ms) for the waveform selection, and gain and pitch filter update. The pitch lag takes on values from 40 to 103 (5–12.9 ms). Only a single pitch coefficient will be used. Both the gain and the pitch coefficient are unquantized. The excitation waveform $x^{(i)}(n)$ is chosen from a repertoire of 32 waveforms. These parameters are appropriate for a CELP coder operating near 5 kb/s.

Figure 3 shows the frequency weighted SNR (signal-to-noise ratio) in dB for a CELP coder using pitch synthesis parameters determined by analyzing the input speech and using synthesis parameters optimized for the synthesis stage. The average frequency weighted SNR increases from 4 dB to 7 when the optimized parameters are used. Comparison of the short-time spectra shows that the harmonic structure is better reproduced with the optimized coefficients. Also the gross spectral information tends to be preserved even towards the high frequencies. The scheme which uses the filter developed by analyzing the input speech offers a poor match especially around the zeroes in the spectral envelope.

As more structure is added to the excitation waveform through the selection of an optimal pitch filter, the size of the codebook becomes less critical. With the proposed scheme operating on blocks of 5 ms duration, dictionary sizes as small as 16 or 32 waveforms produce reasonable speech quality. The quality is directly comparable to a CELP coder with 1024 waveforms but which uses a pitch filter derived at the analysis stage.

4.2 Sequential Optimization

In order to reduce the computational load of the parameter optimization procedure, sequential approaches for determining the synthesis parameters are considered. The sequential algorithm determines the pitch lag M using a zero input from the dictionary ($G = 0$).

$$\varepsilon_{\text{opt}} = \sum_{n=0}^{N-1} s^2(n) - \mathbf{b}^T\mathbf{\Phi}^{-1}\mathbf{b} . \tag{11}$$

The second term in this expression is a function of the pitch lag M through the dependence of $\mathbf{b}$ and $\mathbf{\Phi}$ on $\mathbf{v}^{(n)}$ which in turn depends on M. The optimal value of M can be found by a maximization of the quantity $\mathbf{b}^T\mathbf{\Phi}^{-1}\mathbf{b}$.

With the optimum lag determined for a zero excitation, the lag is kept fixed at this value. Two variants of the basic scheme will be considered. In the first variant, the other parameters of the synthesizer are determined by searching over waveform indices. For each waveform index, the optimum gain and pitch coefficients (assuming the pitch lag already determined) are found. In the second variant, the pitch coefficients are also determined for zero excitation. Keeping the pitch filter fixed (both lag and coefficients), the search is conducted over waveform indices. For each index, the optimum gain is found, assuming the other synthesis parameters are fixed.

One interpretation of the operation of the sequential approaches is as follows. The excitation signal which is used to drive the formant synthesis filter is composed of two components. The first is a scaled and delayed version of the previous excitation signal. In voiced speech, this approach supplies the pitch component. The gain-scaled waveform from the dictionary fills in details that are missing in the excitation signal. It also supplies the startup component for the pitch excitation in transition regions (unvoiced or silence to voiced).

One can expect that the performance of the sequential approaches to be inferior to the optimal joint solution. The methods were compared using unquantized coefficient values. The system parameters are as before.

The first variant produces speech which can be described as smoother than the optimal method, but which lacks a certain fullness. In addition, the energy variations are not rendered quite as accurately. The second variant produces a pitch coefficient which is within 10% of the value given by the first variant in steady voiced speech. Larger differences are observed in silence and transition regions as well as in voiced segments with rapid formant changes. The overall differences between the optimal scheme and the first variant are small enough that the computational savings associated with the sequential approach are attractive.[†]

Solution for $M < N$

The limitation that the pitch lag be greater than the subframe size causes some problems for high pitched female speech. The pitch period in our female samples can become as low as 28 samples (3.5 ms, corresponding to a 285 Hz pitch frequency). One can argue that pitch doubling can capture this short pitch period. However, some wavering in the speech can be observed whenever the pitch period hovers around the 40 sample value. This is caused by the pitch lag changing suddenly between its fundamental

[†] Quantization effects mask the differences further.

value and its pitch doubled value. In addition, one can note an improvement in the harmonic structure when pitch doubled values are used.

The basic problem in solving for the gain and pitch coefficient for lags less than the subframe size is that the equations become nonlinear in the coefficients for $M > N$. This is due to the fact that both the matrix Φ and the vector $\mathbf{b}$ contain terms in $\hat{d}(n)$ for $n \geq 0$. These terms in turn depend on the coefficients. The general solution of the nonlinear set of equations is impractical.

Consider the case that a single pitch coefficient is being sought ($N_t = 1$) for a zero input from the dictionary ($G = 0$). Also let the pitch lag lie in the interval $N/2 \leq M < N$, where N is the subframe size. The excitation signal takes on one of two forms

$$\hat{d}(n) = \begin{cases} \beta \hat{d}(n-M) & 0 \leq n < M \\ \beta^2 \hat{d}(n-2M) & M \leq n < N \end{cases} \tag{12}$$

The squared-error sum can be expressed as

$$\begin{aligned} \varepsilon = & \sum_{n=0}^{N-1} s''(n)^2 - 2\beta \sum_{n=0}^{N-1} s''(n)\hat{d}_{0,M}(n,M) - \beta^2 \sum_{n=0}^{N-1} \hat{d}_{0,M}(n,M)^2 \\ & - 2\beta^2 \sum_{n=M}^{N-1} s''(n)\hat{d}_{M,N}(n,2M) - 2\beta^3 \sum_{n=M}^{N-1} \hat{d}_{0,M}(n,M)\hat{d}_{M,N}(n,2M) \\ & - \beta^4 \sum_{n=M}^{N-1} \hat{d}_{M,N}(n,2M)^2 . \end{aligned} \tag{13}$$

Setting the derivative to zero gives a cubic in β which can be solved in closed form. However, the solution of the cubic involves transcendental functions.

The proposed method for finding the optimum value for β takes a short cut based on using quantized values for β. In this scheme, the sum terms are precomputed. Each of the possible quantized values for β is substituted into the equation. The value of β which gives the smallest value for ε is chosen. For a relatively small number of quantized values, this approach is computationally attractive.

A second method for calculating the pitch coefficient for $M < N$ is more empirical. In this scheme, the past pitch filter output is periodically continued.

$$\hat{d}(n) = \begin{cases} \beta \hat{d}(n-M) & \text{for } 0 \leq n < M \\ \beta \hat{d}(n-2M) & \text{for } M \leq n < N \end{cases} \tag{14}$$

This scheme embodies an automatic pitch doubling for part of the subframe. With this formulation, the solution for β results in a linear equation.

4.3 Results for an expanded pitch lag range

The original spectra for cases in which the pitch lag is doubled show spurious energy between harmonics (Fig. 4). These effects disappear with the expanded pitch lag range (pitch coefficient quantized to 16 values). The resulting speech is considerably improved in quality for those speakers for which the pitch lag falls below 40 samples.

The method employing the periodic continuation of the pitch filter output was also tried. From the formulation, one can see that method does not allow for pitch pulses in a subframe which change in amplitude from one pulse to another. It can underestimate the impact of coefficient values greater than one. This represents a potential cause for local degradations of the synthetic speech. More obvious was the presence of occasional artifacts in the reconstructed speech due to sudden bursts of the high frequencies.

5. Parameter Coding

The target bit rate is 4800 bits/sec for 8 kHz sampled speech. The bit allocations used are summarized in Table 1. The allocation allows for a dictionary of 32 waveforms of 40 samples each (1000 b/s).

5.1 Formant filter coding

Our quantization scheme is based on a line spectral frequency (LSF) parameterization of the formant synthesis filter. This LSF coding uses 21 bits to code 10 formant coefficients. The coefficients are computed for frames of 120 samples (15 ms). The scheme uses a combination of intra- and inter-frame coding. This part of the scheme is conceptually similar to one described by Crosmer and Barnwell [3]. The quantizers are designed based on histograms of occurrences, with care taken to avoid LSF crossovers. To help lock-in after transitions from silence/unvoiced segments to voiced segments an adaptive prediction scheme is used for the inter-frame coder. The step size adaptation uses extra cues taken from the speech energy. In addition, prediction leakage and automatic reset in silence mitigate error propagation in the presence of channel errors.

Parameter	Transmission Rate		
	bits	update	b/s
pitch filter	10	40	2000
formant filter	24	240	800
waveform index	5	40	1000
gain factor	5	40	1000
		total	4800

Table 1 Bit allocations for a 4800 b/s CELP coder

To further reduce the bit-rate, only every second frame of coefficients is coded. The missing frames are filled by interpolating adjacent frames. A 3-bit code specifies which interpolation value is to be used. Which of the 8 possible interpolations to use is determined based on a frequency weighted criterion. The criterion combines weights based on the spectral frequency and the distance from neighbours.

The formant filter coding scheme uses a total of 24 bits every 240 samples, corresponding to a bit rate of 800 b/s.

5.2 Pitch parameter coding

The remaining 3000 b/s is allocated to the gain, the pitch lag, and the pitch coefficient(s). The target bit rate only allows for the use of a single pitch coefficient. Our experience is that fairly rapid update of the pitch predictor is necessary for good quality. The pitch predictor is updated every 5 ms. The gain and pitch filter allocation then is 15 bits every 5 ms.

The scheme adopted uses 5 bits to code the gain parameter. The gain parameter is coded as sign and differential magnitude. The pitch lag and the pitch coefficient are coded together with 10 bits. The pitch lag takes on one of 73 values and the single pitch coefficient takes on one of 14 values. The configuration chosen allows the pitch to cover the range from 31 samples to 103 samples, corresponding to pitch frequencies of 78–258 Hz.

6. Summary and Conclusions

The CELP coder runs as a simulation in floating point. However, most of the components have been previously implemented in fixed-point arithmetic. In recent work, the conversion from direct form coefficients to LSF's and vice-versa has been simulated on a fixed-point signal processor. With the proposed method for pitch filter optimization, the computational rate is compatible with a fixed-point signal processor with a 100 ns cycle time.

The coder generates high quality speech at 4.8 kb/s and is relatively robust to channel errors. It has been tested at channel error rates of 0.001, with only minor degradations in the resulting speech. In addition, an adaptive postfilter has been added to achieve a small increase in perceived speech quality [4]. Further work is in progress to evaluate the performance for degraded speech input.

References

1. M. R. Schroeder and B. S. Atal, "Code-Excited Linear Prediction (CELP): high quality speech at very low bit rates", *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tampa, Florida, pp. 937–940, April 1985.
2. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
3. J. R. Crosmer and T. P. Barnwell, III, "A low bit rate segment vocoder based on line spectrum pairs", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Tampa, Florida, pp. 7.2.1–7.2.4, April 1985.
4. J.-H. Chen and A. Gersho, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 2185–2188, Dallas, Texas, April 1987.

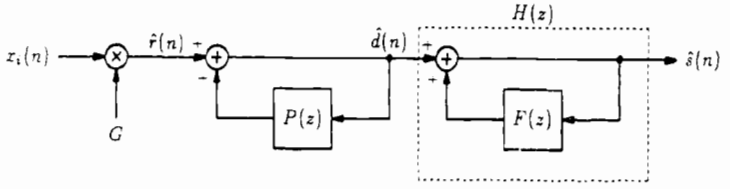

Fig. 1 Synthesis stage for an CELP coder

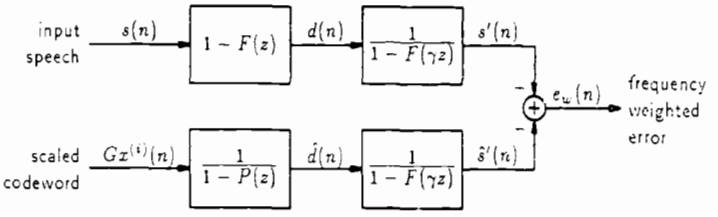

Fig. 2 Model for the calculation of the frequency weighted error

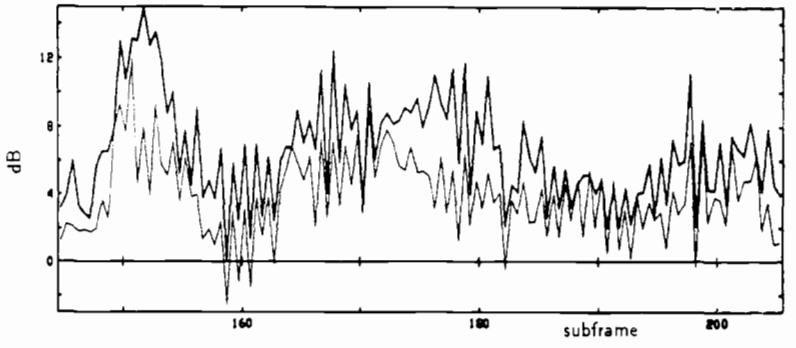

Fig. 3 Segmental frequency weighted SNR (dB). (thin line for parameters developed by analyzing the input speech, thick line for parameters optimized for the synthesis stage).

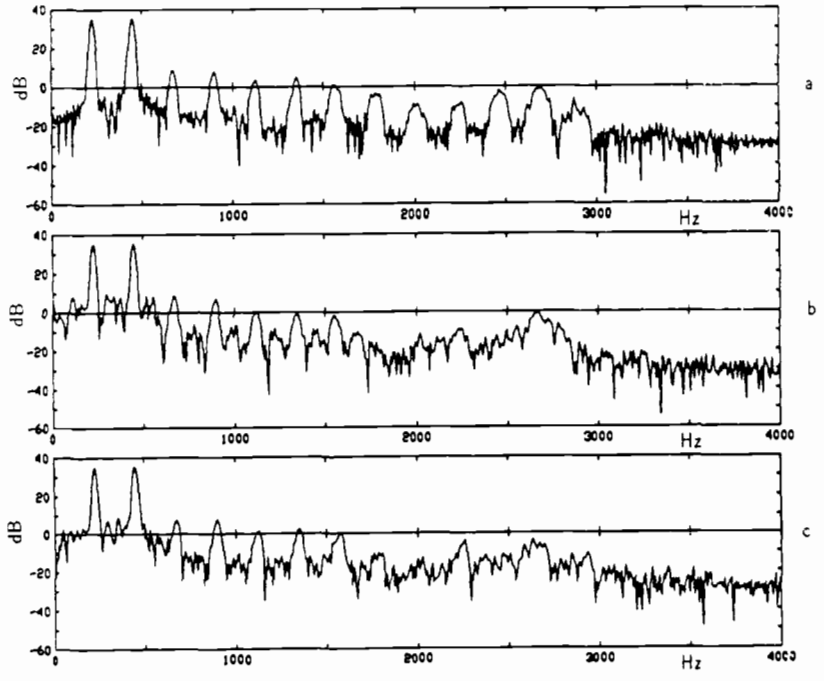

Fig. 4 Short-term spectra for female speech.
(a) original speech,
(b) synthesized speech with the pitch lag constrained to be larger than 40 samples, and
(c) synthesized speech with the pitch lag allowed to fall below 40 samples.