Efficient Quantization of Speech Excitation Parameters Using Temporal Decomposition

Phu Chien Nguyen, Masato Akagi

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Tatsunokuchi, Ishikawa, 923-1292 Japan
{chien,akagi}@jaist.ac.jp

Abstract

In this paper, we investigate the application of temporal decomposition (TD) technique to describe the temporal patterns of speech excitation parameter contours, i.e. gain, pitch, and voicing. We use a common set of event functions to describe the temporal structure of both spectral and excitation parameters, and then quantize them. Experimental results show that each speech excitation parameter contour can be well described by a set of excitation targets using the event functions obtained from TD analysis of line spectral frequency (LSF) parameters, with considerably low reconstruction error. Moreover, we can efficiently quantize the excitation targets by a combination of two uniform quantizers, one working directly on logarithmic excitation targets and the other working on the difference between current and previous logarithmic excitation targets.

1. Introduction

Temporal decomposition (TD) of speech [1], which is an analysis procedure based on a linear model of the effects of coarticulation, yields a linear approximation of a time sequence of spectral parameters in terms of a series of time-overlapping event functions and an associated series of event targets as given in Equation (1).

$$\hat{\mathbf{y}}(n) = \sum_{k=1}^{K} \mathbf{a}_k \phi_k(n), \qquad 1 \le n \le N \tag{1}$$

where a_k and $\phi_k(n)$ are the kth event target and kth event function, respectively. $\hat{\mathbf{y}}(n)$ is the approximation of $y(n)$, the nth spectral parameter vector, produced by the TD model. N and K are the number of frames and the number of events, respectively.

The modified restricted temporal decomposition (MRTD) determines event targets, $\mathbf{a}_k$, and event functions, $\phi_k(n)$, once the line spectral frequency (LSF) parameters, $\mathbf{y}(n)$, of a speech segment are given. Results of MRTD analysis performed on sentence utterances with LSFs calculated at 10 ms frame intervals using 40 ms Hanning window, show that the event rate should be about 20 events/sec in order to keep the log spectral distortion around 1.5 dB level. The summarization of MRTD is given in the following section. For computational details of MRTD, the readers are referred to [6] .

2. MRTD Algorithm

MRTD employs the restricted second order TD model [2, 4, 6], where only two adjacent event functions can overlap and all event functions at any time sum up to one. The argument for imposing this constraint on the event functions can be found in [2, 6]. Equation (1) is rewritten as

$$\hat{\mathbf{y}}(n) = \mathbf{a}_k \phi_k(n) + \mathbf{a}_{k+1}(1 - \phi_k(n)), \quad n_k \le n < n_{k+1} \tag{2}$$

where n_k and n_{k+1} are the central positions of event k and event $k+1$, respectively.

In order to apply TD to decomposing line spectral frequency (LSF) parameters, the stability of the corresponding linear predictive coding (LPC) synthesis filter after spectral transformation performed by TD must be ensured. The restricted temporal decomposition (RTD) method [4] intends to make LSF parameters possible for TD by enforcing the LSF ordering property on the event targets. However, RTD has not completely solved this problem as indicated in [6]. Moreover, some event functions derived from RTD are ill-shaped, i.e. they have more than one peak, which is undesirable from speech coding point of view. Thus, the modified RTD (MRTD) method [6] has been proposed to overcome the drawbacks imposed on the RTD method.

The initial approximation of event targets is based on a maximum spectral stability criterion. It is assumed that each acoustic event that exists in speech gives rise to a spectrally stable point in its neighborhood. Therefore, the locations of the spectrally stable points and the corresponding spectral parameter vectors can be used as a good approximation to the event locations and event targets, respectively. Here, event localization is done via the local minimal points of a spectral transition measure called spectral feature transition rate (SFTR) [5].

In the result, when once the locations of events n_k, where $k = 1, ..., K$, are known and the corresponding event targets are initialized with the samples of the LSF vector trajectory $\mathbf{y}(n_k)$, we can calculate proper event functions and event targets iteratively in the least mean square sense. However, since the event targets are calculated using the formula $\mathbf{A} = \mathbf{Y}\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{\Phi}^T)^{-1}$ which does not consider the LSF ordering property for them, the estimated event targets may not be interpreted as LSF vectors. Invalid LSF event targets estimated from a LSF vector trajectory cause two serious problem. Firstly, the event targets do not have their own spectra as valid LSF vectors do. It follows that those event targets are regarded as the numerical results, but not as the idealized targets. They also prohibit us from matching the determined events with meaningful phonetic units. Secondly, it is impossible to utilize the advantages of LSF parameters for quantization. The invalid LSF event targets lower the intra/inter-correlations and do not guarantee the stability of

the reconstructed LSF vectors. Therefore, a refinement procedure is applied to the estimated event targets to ensure the LSF ordering property for them with a negligible increase in reconstruction error.

3. MRTD of Excitation Parameters

3.1. Determination of Excitation Targets

The MRTD technique is employed to describe the temporal characteristics of speech excitation parameters, i.e gain, pitch and voicing. The same event functions evaluated for LSF parameters are also used to describe the temporal pattern of the gain, pitch and voicing parameters. We are motivated by the fact that the speech production mechanism is assumed to be a synchronously controlled process with respect to the movement of different articulators, i.e. jaws, tongue, larynx, glottis etc. Therefore, we expect that the temporal evolutionary patterns of different properties of speech, i.e. spectrum, pitch, gain and voicing, can be described by a common set of event functions.

Let $b(n)$ be a excitation parameter, i.e. gain, pitch or voicing. Then $b(n)$ is approximated by $\hat{b}(n)$, the reconstructed excitation parameter for the nth frame, as follows in terms of excitation targets, b_ks, and event functions, $\phi_k(n)$s.

$$\hat{b}(n) = \sum_{k=1}^{K} b_k \phi_k(n), \qquad 1 \le n \le N \tag{3}$$

In Equation (3), the event functions, $\phi_k(n)$s, are known and therefore the excitation targets, b_ks, are determined by minimizing the sum squared error between the original excitation parameters and the reconstructed excitation parameters as follows.

$$E_b = \sum_{n=1}^{N} \left(b(n) - \sum_{k=1}^{K} b_k \phi_k(n) \right)^2$$

By setting the partial derivative of E_b with respect to b_r to zero;

$$\frac{\partial E_b}{\partial b_r} = \sum_{n=1}^{N} \left(b(n) - \sum_{k=1}^{K} b_k \phi_k(n) \right) (-2\phi_r(n)) = 0$$

$$\sum_{k=1}^{K} b_k \sum_{n=1}^{N} \phi_k(n)\phi_r(n) = \sum_{n=1}^{N} b(n)\phi_r(n), \qquad 1 \le r \le K \tag{4}$$

Equation (4) gives a set of K variable simultaneous equations, using which b_k, where $1 \le k \le K$, could be evaluated.

In the case of pitch parameters, linear interpolation was used within the unvoiced segments to form a continuous pitch contour. In the case of voicing parameters, a hard limiter with a threshold value of 0.5 was used to determine the reconstructed binary voicing parameters and binary voicing targets, from the non-binary results of Equations (3) and (4), respectively.

3.2. Simulation Results

The gain, pitch and voicing parameters, hereafter indicated by $g(n)$, $p(n)$, and $v(n)$, respectively, were calculated at 10 ms frame intervals with a 40 ms analysis window, for the sentence utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai*," of the ATR Japanese speech database. Each parameter contour was MRTD analyzed according to the procedure described above using the event functions obtained from MRTD analysis of LSF parameters.

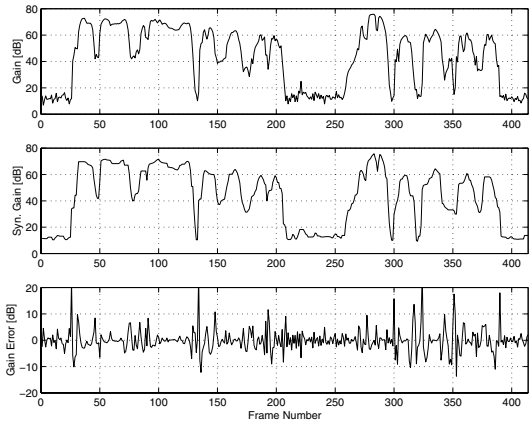

Figure 1: Original gain parameters, $g(n)$, reconstructed gain parameters, $\hat{g}(n)$, for the sentence utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai*." The RMS gain error is 4.37 dB.

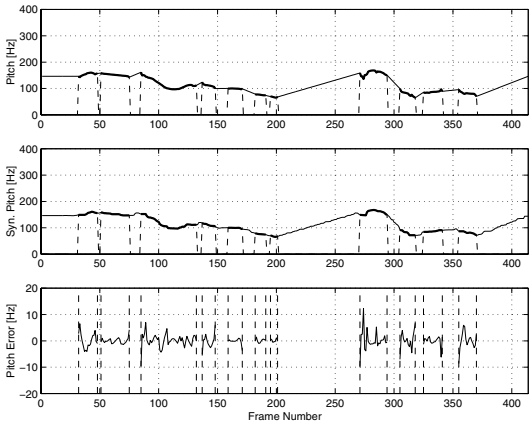

Figure 2: Original pitch parameters, $p(n)$, reconstructed pitch parameters, $\hat{p}(n)$, for the sentence utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai*." The RMS pitch error is 2.09 Hz.

Fig. 1 shows the plots of original and reconstructed gain parameters and the plot of frame-wise gain error, $e_g(n)$, where $e_g(n) = \hat{g}(n) - g(n)$. The RMS gain error, $\sqrt{E_g}$, where $E_g = \frac{1}{N}\sum_{n=1}^{N} {e_g}^2(n)$, was found to be about 4.37 dB.

Fig. 2 shows the plots of original and reconstructed pitch frequency parameters and the plot of frame-wise pitch frequency error, $e_p(n)$, where $e_p(n) = \hat{p}(n) - p(n)$. The RMS pitch error, $\sqrt{E_p}$, where $E_p = \frac{1}{N}\sum_{n=1}^{N} {e_p}^2(n)$, was found to be about 2.09 Hz.

In the case of binary voicing parameters, the voicing error, $e_v(n)$, where $e_v(n) = \hat{v}(n) - v(n)$, appeared only at, but not all, voiced/unvoiced boundaries as error spikes of mostly 1 frame. The percentage number of frames with voicing errors was found to be about 4.59%. Fig. 3 shows the plots of original and reconstructed voicing parameters and the plot of frame-wise voicing error, $e_v(n)$.

Moreover, we have also evaluated the performance of MRTD in terms of excitation parameters over a set of 250 sen-

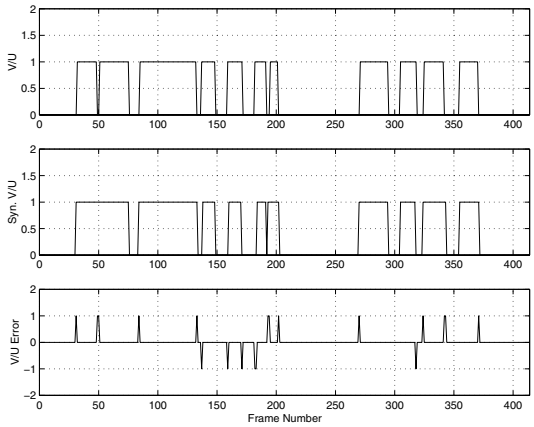

Figure 3: Original binary voicing parameters, $v(n)$, reconstructed binary voicing parameters, $\hat{p}(n)$, for the sentence utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai*." The percentage number of frames with voicing errors is 4.59%.

tence utterances of the ATR Japanese speech database. This speech data set consists of about 20 minutes of speech spoken by 10 speakers (5 males & 5 females) resampled at 8 kHz sampling frequency. The RMS gain error, RMS pitch error and percentage number of frames with voicing errors were found to be about 4.01 dB, 5.75 Hz and 4.74%, respectively. It was observed that the RMS gain error and RMS pitch error can be mainly attributed to some discrete time points, where the corresponding frame-wise gain error and pitch error obtained very high values. Meanwhile, no voicing errors were observed during continuous voiced and unvoiced segments, except for the points of voicing transitions.

The significant match between the original and reconstructed excitation parameters results in the fact that a common set of event functions can be used to describe the temporal patterns of both spectral and excitation parameters.

Also, we have evaluated the performance of the original restricted temporal decomposition (RTD) method [4] over the speech data set mentioned above. Experimental results obtained were about 4.03 dB, 5.8 Hz, and 4.75% for RMS gain error, RMS pitch error, and percentage of frames with voicing error, respectively, slightly higher than those obtained from the MRTD method.

4. Quantization of Excitation Targets

Since voicing targets can be quantized at 1 bit/target, in this section only the quantization of gain and pitch targets is presented.

4.1. Quantization scheme for excitation targets

Fig. 4 shows gain and pitch target contours for the sentence utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai*." Note that the slow evolution of gain target and pitch target in voiced segments, interrupted by sudden jumps in unvoiced segments.

In this subsection, we propose a differential and logarithmic quantization scheme for gain and pitch targets. The logarithm of gain target as well as the logarithm of pitch target are quantized both in differential and a memoryless quantizer, and the best

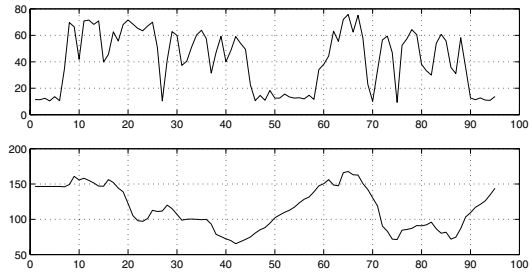

Figure 4: Top: gain target contour and bottom: pitch target contour, for the sentence utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai*."

of two output values is transmitted to the receiver. The major advantage of this scheme is that the high correlation of consecutive gain and pitch targets during voiced segments can be exploited, without losing performance of unvoiced speech, where consecutive gain and pitch targets have low correlation. Another advantage is that the proposed scheme quantizes rapidly and slowly changing gain and pitch targets in separate quantizers, which allows for different resolution for these two cases.

We applied the method of pitch quantization proposed in [3] to quantize gain and pitch targets. The function of the algorithm is as follows: First the logarithm of gain target (respectively the logarithm of pitch target) is computed. The logarithmic gain target (respectively the logarithmic pitch target) is used in two branches of the algorithm. In the first branch, the gain target (respectively pitch target) is directly quantized in a uniform quantizer. In the second branch, the previous value of the quantized gain target (respectively pitch target) is subtracted before quantization, and added back after quantization, to form a differential quantization scheme. The output of the two branches are compared to the unquantized gain target (pitch target), and the best is selected for transmission. The full algorithm for quantization of gain target (respectively pitch target) using 5 bits is given as follows. Notice that t is denoted for gain target (respectively pitch target).

Step 1. Initialization (this is only done for the first call), t_{min} and t_{max} are the minimum and maximum gain target (respectively pitch target), respectively.

$t_{range} = \log t_{max} - \log t_{min}$ (the range of logarithmic gain or pitch target).

$\varepsilon_1 = t_{range}/20 \times \{0, 1, 2, ..., 20\} + \log t_{min}$ (21 entries, index 0-20).

$\varepsilon_2 = \log 1.06 \times \{-5, -4, ..., 4, 5\}$ (11 entries, index 21-31).

Step 2. Get an estimated gain target (respectively pitch target) $T(n)$, and compute the logarithmic gain target (respectively pitch target), $t(n) = \log T(n)$.

Step 3. Find the closest value to $t(n)$ in ε_1, $\tilde{t}_1(n) = \arg\min_{c \in \varepsilon_1} \mid t(n) - c \mid^2$

Step 4. Find the closest value to $d(n) = t(n) - \tilde{t}(n-1)$ in ε_2, $\tilde{t}_2(n) = \arg\min_{c \in \varepsilon_2} \mid d(n) - c \mid^2 + \tilde{t}(n-1)$

Step 5. Compare $\tilde{t}_1(n)$ and $\tilde{t}_2(n)$ to $t(n)$, and select the best.

The index to the selected codebook entry (see definition of ε_1 and ε_2 for the index assignment) is output.

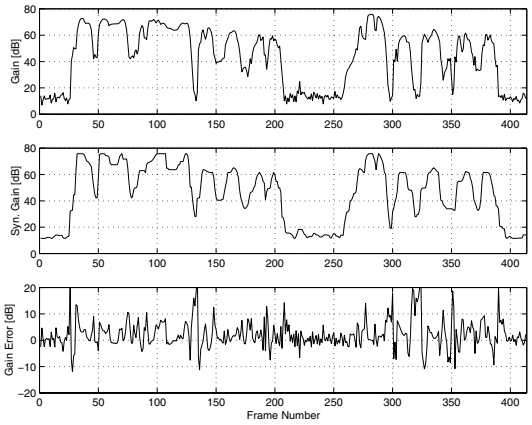

Figure 5: Original gain parameters, $g(n)$, reconstructed gain parameters after quantization, $\tilde{g}(n)$, for the sentence utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai*." The RMS gain error is 5.74 dB.

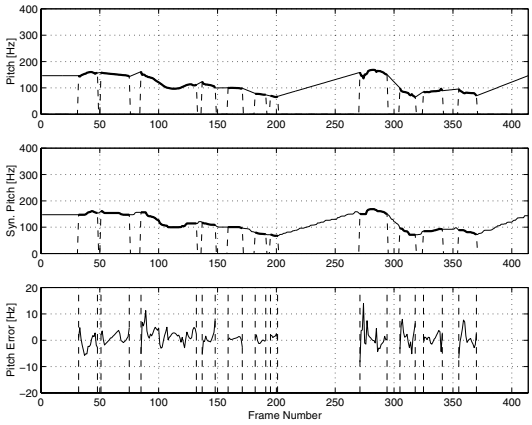

Figure 6: Original pitch parameters, $p(n)$, reconstructed pitch parameters after quantization, $\tilde{p}(n)$, for the sentence utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai*." The RMS pitch error is 2.61 Hz.

4.2. Simulation results

The gain and pitch parameters $g(n)$ and $p(n)$, respectively, were calculated at 10 ms frame intervals with a 40 ms analysis window, for the sentence utterance "*kantan na shiryou wo ookuri shimasu node, shibaraku omachi kudasai*." Each parameter contour was MRTD analyzed using the event functions obtained from MRTD analysis of LSF parameters. After that, gain and pitch targets were quantized and transmitted using the procedure described above.

Fig. 5 shows the plots of original gain parameters and reconstructed gain parameters (after quantization), and the plot of frame-wise gain error, $e_g(n)$, where $e_g(n) = \tilde{g}(n) - g(n)$. The RMS gain error, $\sqrt{E_g}$, where $E_g = \frac{1}{N}\sum_{n=1}^{N} {e_g}^2(n)$, was found to be about 5.74 dB.

Fig. 6 shows the plots of original pitch parameters and reconstructed pitch parameters (after quantization), and the plot of frame-wise pitch error, $e_p(n)$, where $e_p(n) = \tilde{p}(n) - p(n)$. The RMS pitch error, $\sqrt{E_p}$, where $E_p = \frac{1}{N}\sum_{n=1}^{N} {e_p}^2(n)$, was found to be about 2.61 Hz.

In the case of binary voicing parameters, results obtained after quantization are the same in comparison with those obtained from MRTD analysis only since binary voicing target can be transmitted accurately using 1 bit/target.

In addition, we have evaluated the performance of MRTD analysis and quantization in terms of gain and pitch parameters over the set of 250 Japanese sentence utterances mentioned earlier. The RMS gain error, RMS pitch error were found to be about 6.27 dB, 9.59 Hz, respectively. It was also observed that the RMS gain error and RMS pitch error can be mainly attributed to some discrete time points, where the corresponding frame-wise gain error and pitch error obtained very high values.

The significant match between the original and reconstructed excitation parameters after MRTD analysis and quantization results in the fact that MRTD can be regarded as a reasonable technique of analyzing and quantizing excitation information of speech.

5. Conclusion

We have extended the MRTD technique to describe the temporal patterns of speech excitation parameters. Results from this description, namely, excitation targets can be quantized efficiently at 11 bits/target, which results in a bit-rate of 220 bps required for encoding excitation information of speech. The low reconstruction error in excitation parameters after MRTD analysis as well as excitation target quantization justifies the fact that speech excitation parameters can not only be well represented by excitation targets using a common set of event functions derived from MRTD analysis of LSF parameters, but also can be encoded efficiently by quantizing and transmitting the excitation targets.

6. References

[1] B. S. Atal, "Efficient coding of LPC parameters by temporal decomposition," Proc. ICASSP'83, pp. 81-84, 1983.

[2] P.J. Dix and G. Bloothooft, "A breakpoint analysis procedure based on temporal decomposition," IEEE Trans. Speech and Audio Proc., 2(1): 9-17, 1994.

[3] T. Eriksson and H. G. Kang, "Pitch quantization in low bit-rate speech coding," Proc. ICASSP'99, pp. 489-492, 1999.

[4] S.J. Kim and Y.H. Oh, "Efficient quantization method for LSF parameters based on restricted temporal decomposition," Electron. Lett., 35(12): 962-964, 1999.

[5] A.C.R. Nandasena, P.C. Nguyen, and M. Akagi, "Spectral stability based event localizing temporal decomposition," Computer Speech and Language, 15(4): 381-401, 2001.

[6] P. C. Nguyen and M. Akagi, "Improvement of the restricted temporal decomposition method for line spectral frequency parameters," Proc. ICASSP2002, pp. 265-268, 2002.

[7] P.C. Nguyen, T. Ochi, and M. Akagi, "Coding speech at very low rates using STRAIGHT and temporal decomposition," Proc. ICSLP2002, pp. 1849-1852, 2002.