ROBUST CLOSED-LOOP PITCH ESTIMATION FOR HARMONIC CODERS BY TIME SCALE MODIFICATION*

Chunyan Li, Vladimir Cuperman, and Allen Gersho

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106

ABSTRACT

Harmonic coders that synthesize speech without transmitting phase information abandon the benefits of closed-loop parameter estimation via waveform matching. In this paper, we show that effective closed loop parameter estimation can be achieved when a suitable time-scale modification is applied to the speech LP residual in harmonic coders. The concept is demonstrated here specifically for pitch estimation, but is more broadly applicable. For each of a set of pitch candidates generated by a time-domain pitch estimator, the residual is modified to match the pitch contour derived from that candidate. The best candidate is selected by evaluating for each candidate the match between the modified residual and the synthesized residual. The new pitch estimation algorithm significantly reduces gross pitch errors compared to a conventional time-domain pitch estimator and enhances the perceptual performance of a 4 kbps harmonic coder.

I. INTRODUCTION

Low rate sinusoidal coders synthesize speech without transmitting phase information, resulting in a loss of time alignment between the synthetic and the original speech. Time-domain closed-loop parameter estimation is therefore hampered by the inability to do waveform matching. Yet, harmonic coders that do sinusoidal modeling of the speech LP residual, can benefit from waveform matching if a suitable time-scale modification is applied to the original residual. We demonstrate this concept here with a specific method for efficient closed loop pitch estimation.

Synthesized speech quality in harmonic coders depends significantly on the accuracy of the fundamental frequency (pitch) estimation. The lack of reliability of open loop pitch estimators over a wide range of pitch values and input conditions is one of the key obstacles to achieving toll quality at 4 kb/s with harmonic coding. Most pitch estimators do not use analysis by synthesis, i.e., pitch is estimated open-loop based on some reasonable but heuristic criterion, without comparing the resulting synthesized speech with the original speech. This may result in a lack of robustness and mismatches between the original and the synthesized waveforms.

Some closed-loop frequency domain pitch estimators have been propose, based on spectral matching, see for example, [1]. However, for large pitch frequencies, there are only a small number of harmonics, so that time-domain estimation or a combination of time-domain and frequency-domain estimation may be more effective for high performance pitch estimators [2].

In this paper, we propose the use of a nonlinear time scale modification technique, called "signal modification", for solving the waveform matching obstacle in harmonic coders and thereby achieving time-domain closed-loop parameter estimation with application to pitch estimation.

Signal modification has been previously used in analysis-by-synthesis speech coding to directly improve waveform coding efficiency [3]. In this paper, we introduce the use of signal modification for parameter estimation in harmonic coding. Specifically, the original speech signal is modified to match the pitch contour derived for each of a set of pitch candidates generated by a time-domain pitch estimator. The best candidate is selected by evaluating the "degree of matching" between the modified original LP residual signal and the synthetic residual generated with that candidate's pitch contour. Signal modification is performed under constraints that ensure the quality of the input speech will be preserved. An incorrect pitch candidate will either violate the constraints, or result in a mismatch between the modified original and the synthesized signals.

For objective evaluation of the technique, the "true" pitch, obtained manually by spectrum and waveform examination with graphical tools, served as a reference to determine (a) the error rate for gross pitch errors (outliers), and (b) the average pitch error for both clean and noisy speech environments. For subjective evaluation, the algorithm was embedded into a hybrid

* This work was supported by SignalCom, Inc. through the UCSB Internship in Industry program.

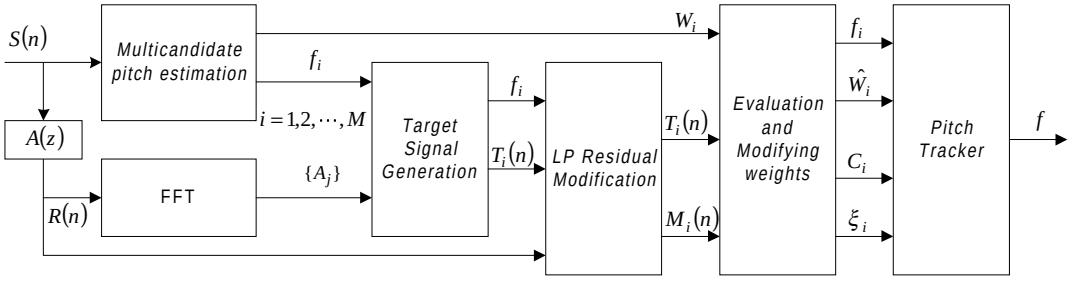

Fig.1. Block diagram for the new pitch estimation algorithm

coder [6] and compared to the pitch estimation technique based on the time-domain normalized autocorrelation.

II. PITCH ESTIMATION ALGORITHM

In signal modification (also known as *time warping*), the time scale of a signal is altered so that the signal will match a reference signal, called the *target* signal. With properly selected constraints to the allowed warping, the perceptual quality of the modified speech signal can be preserved. Our proposed pitch estimation method is based on the fact that for a correct pitch estimate, the modified speech should match the synthetic speech with the constraint that the modified speech is perceptually close or identical to the original speech. The degree of matching between the modified original speech and the synthetic speech is evaluated for each pitch candidate and affects the final pitch decision.

A simplified block diagram for the proposed pitch estimation algorithm is shown in Fig. 1. The pitch candidates are generated by time-domain pitch estimation based on the normalized autocorrelation function. The pitch candidates, f_i, correspond to the local maxima of the autocorrelation function. In the time-domain pitch estimator, the input speech *S(n)* is first low-pass filtered to a bandwidth of 800 Hz. The low-pass filtered signal is inverse filtered by a 2nd order LPC inverse filter to give a spectrally flattened signal for which the autocorrelation function is computed. A number of M pitch candidates are selected from the local maxima of the normalized autocorrelation function. The value of the normalized autocorrelation function corresponding to each pitch candidate will be used as its weight and denoted by W_i.

For each pitch candidate f_i, a target signal, $T_i(n)$, is generated by synthesizing the reconstructed LP residual with the sinusoidal model

$$T_i(n) = \sum_k A_k \cos[\theta_k(n)]$$

where the spectral amplitudes $\{A_k\}$ are obtained by sampling the LP residual signal spectrum at the harmonics of that pitch candidate. The spectral phases $\{\theta_k\}$ are derived from the previous frame pitch $f^{(-1)}$ and the current pitch candidate f_i, assuming a linear pitch contour

$$\theta_k(n) = \frac{2\pi k}{F_s}\left[f^{(-1)} n + \frac{(f_i - f^{(-1)})}{2N} n^2 + \varphi_0 \right]$$

where N is the frame size in sample, F_s is the sampling frequency, and φ_0 is the initial linear phase. This target signal will be exactly aligned with the synthetic excitation generated by the decoder if the current pitch candidate would be used as the pitch for the current frame.

The signal modification is performed on the LP residual signal using the target signal as a reference signal. The modification is performed by shifting each pulse in the LP residual such that the corresponding pulse in the modified residual will match a pulse in the target signal. This shifting procedure is constrained to ensure the modified residual signal will give speech quality as good as the original one. The original residual signal is divided into several small segments where each segment contains at most one significant pulse. An accumulated shift parameter τ_{acc} is adjusted for each segment to match the corresponding segment of the target signal. At the boundaries between these segments, part of signal is either omitted or repeated whenever the shift parameter changes. It is important that the significant features of the original signal, such as pitch pulses, will not be placed on the boundary when the shift segment boundaries are determined. The shift adjustment for the current segment is limited by a procedure similar to that used in the EVRC coder [4]. Since τ_{acc} may be a fractional number, the shifting procedure is performed at a resolution higher than that provided by the 8kHz sampling rate. The procedure is described below.

First, a temporary modified residual $\hat{R}_{tmp}(n)$ is obtained by using the integer shift obtained by rounding the shift τ_{acc} to the nearest integer. Then, an integer energy correlation vector $CORI(k)$ is computed between

the temporary modified residual $R_{tmp}(n)$ and the target signal $T_i(n)$, where k varies in the range of acceptable right and left shifts. Finally, a fractional energy correlation vector $CORF(t)$ is obtained by interpolating $CORI(k)$. The optimal shift, τ_{opt}, that will match the temporary modified residual to the target signal is then defined as the index t that maximize $CORF(t)$. The accumulated shift τ_{acc} is adjusted based on τ_{opt}, if the normalized $CORF(t)$ is larger than an empirical threshold which is 0.5 in our experiments. Once τ_{acc} has been determined, the original LP residual signal is shifted by τ_{acc} to create the modified residual signal. During the above procedure, the shift range is bounded to ensure that the quality of the signal will not be affected. If the matching requires a shift outside the range bounds, a no shift decision will be made and τ_{acc} will be unchanged for the current segment. This may lead to a misaligment between the target signal and the modified residual signal.

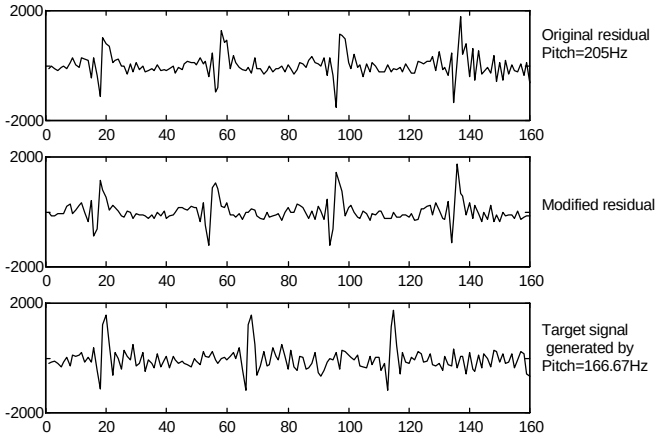

Fig. 2. Failure to match target signal during the signal modification procedure

Figure 2 gives an example of an incorrect pitch estimate leading to a misalignment which can not be corrected by the modification process. In the figure, the original LP residual signal has a pitch of 205Hz. However, the target signal is generated based on the first pitch candidate produced by the pitch estimator which has a value of 166.67 Hz. For the first pitch pulse in this frame (situated around sample 20), in order to match the corresponding pulse in the target signal, τ_{acc} is adjusted to –1.875 samples. When the modification procedure is applied to the second segment, the matching requires a right shift of about 6 samples. That results in a 7.875 samples adjustment of τ_{acc}, which is out of the range of permissible shift adjustments. In this case, τ_{acc} is kept unchanged for this segment. Same thing happens to the following segments in this frame. Therefore, the modification procedure keeps τ_{acc} as –1.875 samples for the whole frame and leaves a large error between the target signal and the modified signal.

The study of a large quantity of experimental data shows that the modified residual signal will not match the target signal if the pitch candidate is not correct. On the other hand, very good alignment between the modified residual and the target is obtained when a pitch candidate results in a well-fitted pitch contour.

To quantitatively evaluate the quality of matching in signal modification, we developed an empirical algorithm based on two criteria: for each pitch candidate f_i, we compute the normalized correlation (NCOR), C_i, and the normalized mean square error (NMSE), ξ_i, between the target signal and the modified residual signal. These two criteria are related, however, we found experimentally that the use of both criteria may improve results in some difficult cases such as weak periodic sounds and frames with fast pitch changes. As a result of the signal matching evaluation, we change the weights of the pitch estimates by increasing the weights of those candidates that lead to good matching.

A brief description of the empirical procedure follows. First, we determine the pitch candidate $f_{\hat{\xi}}$ with minimum NMSE $\hat{\xi}$, the pitch candidate f_C with maximum NCOR $\hat{C}$, and the pitch candidate f_W with maximum weight $\hat{W}$. We define the signal modification procedure as "reliable" if $\hat{\xi}$ is less than an empirical threshold and $\hat{C}$ is larger than an empirical threshold. If the measurement is not reliable, the weights for all pitch candidates will be kept unchanged.

Next, we remove the "bad" pitch candidates defined as pitch candidates f_i satisfying the following conditions:

$$\begin{cases} \xi_i > MAX(\bar{\xi}, \xi_T) & and \\ C_i < MIN(\bar{C}, C_T) \end{cases}$$

where $\bar{\xi}$ and $\bar{C}$ are the average value of the NMSE and NCOR respectively computed over the set of all candidates, and ξ_T and C_T are predefined thresholds.

The decision on increasing the weight of a pitch candidate f_i is based on a criterion $\lambda(f_i)$ which combines candidate's weight with its NMSE and NCOR as follows:

$$\lambda(f_i) = \frac{\alpha}{\xi_i} + \beta C_i + \gamma W_i$$

where α, β and γ are all fixed numbers determined experimentally on a very large database. We increase the weights of candidates which have $\lambda(f_i)$ larger than one of $\lambda(f_{\hat{\xi}})$, $\lambda(f_C)$ and $\lambda(f_W)$.

To achieve a smooth pitch contour, a pitch tracker is used to select the final pitch from the remaining pitch

candidates with modified weights. Dynamic programming is applied to select the best pitch value by employing a combination of local and contextual evidence, including the pitch value for the previous frame, pitch candidates of the current frame, pitch candidates of future frames (look-ahead), and the corresponding weights for each pitch. Defined in a similar way as in [5], the transition cost function, which is a combination of the pitch weight and the pitch smoothness cost, is computed for each pitch candidate. Based on the transition cost functions, two pitch candidates f_{t1} and f_{t2}, which give lowest cumulative transition costs up to the 2nd future frame, are selected from *M* time-domain pitch candidates. When the difference between the modified weights of the two candidates f_{t1} and f_{t2} is small, we turn again to signal modification for the final decision. Normally, the candidate with the lowest cummulative transition cost, f_{t1}, will be chosen as the final pitch value, however there are two exception conditions under which f_{t2} will be chosen as the final pitch value:

1. If the measurement from signal modification is reliable and the second candidate has much lower NMSE: $\xi_{t1} - \xi_{t2} > 0.6$.
2. If the measurement from signal modifcation is reliable, and the second candidate achieves both the best NMSE and the best NCOR, $t2 = \arg\min(\xi_i) = \arg\max(C_i)$ and $\xi_{t1} > \overline{\xi}$.

III. EXPERIMENTAL RESULTS

Objective performance was evaluated for the proposed pitch estimator by comparing its performance with the open-loop normalized autocorrelation estimator and pitch tracker described in Sec. 2 but without signal modification. To determine errors in each pitch estimation, the "true" pitch was manually estimated from observations of the waveform and the spectrum of the residual using a graphical tool. A pitch for voiced speech was thus estimated every 10 ms.

Table I. Objective results for pitch algorithm

	Method	FMSE (Hz) NPE ≤ 10%	% of Outliers	
			10%< NPE < 30%	NPE ≥ 30%
Clean	TMS	2.13	0.87%	2.06%
	TS	2.42	1.79%	5.51%
Office noise	TMS	2.31	0.77%	2.79%
	TS	2.55	1.15%	6.38%
Harmonic noise	TMS	2.30	0.99%	2.76%
	TS	2.62	1.87%	6.72%
Babble noise	TMS	2.35	1.10%	2.65%
	TS	2.64	2.04%	6.34%

*NPE: Normalized Pitch Error

The results of the objective performance evaluation are presented in Table I. The time-domain estimator without signal modification is denoted in Table I by TS, and our proposed pitch estimation algorithm including signal modification is called TMS. The normalized pitch error (NPE) is computed by dividing the pitch error by the pitch value and is expressed in percentages. Outliers are defined as gross pitch errors with NPE>10% and are divided in two categories, 10%<NPE<30% and NPE≥30%. The fine pitch mean square error (FMSE) is the MSE computed after eliminating all outliers.

The results in Table I indicate that the proposed approach (TMS) achieves a rather minor reduction in the FMSE when compared with the TS method. However, the reduction in outliers is very significant under both clean and noisy conditions.

To corroborate the objective results, both pitch estimation methods were embedded into a harmonic coder [6] for subjective testing. The test speech sentences included both flat and IRS filtered speech. The proposed pitch estimation algorithm was found to give better synthesized quality under both clean and noisy environments. For example, the subjective tests for clean speech indicate a preference of about 51.04% for the proposed pitch estimation versus 30.21% for TS method with 18.75% indicating no preference. These results confirmed the validity and benefit of the new procedure.

REFERENCE

[1] D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder", *IEEE Trans. ASSP*, 1988, Vol. 36, No. 8, pp.664-678

[2]. S. Yeldener, J. C. De Martin, and V. Viswanathan, "A Mixed Sinusoidally Excited Linear Prediction Coder at 4 kb/s and Below", *Proc. ICASSP*, 1998, pp.589-592.

[3]. W. B. Kleijin, P. Kroon, D. Nahumi, "The RCELP Speech-Coding Algorithm", *European Trans. On Telecommunications and Related Technologies*, Vol. 5, Sept.-Oct., pp.573-582, 1994.

[4]. TIA Draft standard, TIA/EIA/IS-127, Enhanced Variable Rate Codec (EVRC) 1996.

[5]. H. Ney, "Dynamic Programming Algorithm for Optimal Estimation of Speech Parameter Contours", *IEEE Trans. On Systems, Man and Cybernetics*, Vol. SMC-13, No.3, pp.208-214, March/April, 1983.

[6]. E. Shlomot, V. Cuperman, and A. Gersho, "Combined Harmonic and Waveform Coding of Speech at Low Bit Rates", in *Proc. Int. Conf. Acoust. Speech Sign. Process.*, Vol. 2, pp. 585-588, May, 1998.