Prosodic Speech Modification Using RELP

Fernando S. Pacheco and Rui Seara
Federal University of Santa Catarina, Florianópolis SC, Brazil

Abstract – **This paper proposes a method to prosodic speech modification based on residual-excited linear predictive coding (RELP) applied to speech synthesis. In this way, pitch and time scale modifications are carried out requiring a very low computational complexity. Formal subjective test comparing the proposed approach with TD-PSOLA, under conditions of increasing and decreasing the pitch, are achieved. For all tested cases, the obtained results show a better performance of the proposed approach as compared with TD-PSOLA.**

I. INTRODUCTION

For text-to-speech (TTS) systems, concatenative synthesis is the approach that has attained better results, so far. In this model, a set of speech segments (diphones, syllables, triphones, for example) is initially recorded and stored. During the synthesis phase, speech is generated by straightforward concatenation of these segments. This set of units, however, is limited in length and does not consider all possible desired intonations and speeds for an appropriate and diversified speech synthesis process. Thus, it is needed to make prosodic modifications (particularly, pitch and duration) at larger or smaller level, depending on the involved context.

A classical technique for temporal series analysis, with application in several areas, is linear prediction (LP) [1]. Particularly in speech processing, this technique can be applied to both coding (linear predictive coding) and speech synthesis. The use of LP in speech synthesis is not a new proposal. Earlier concatenative speech systems have already used such an approach [2]. In this way, a simple excitation signal model has been used (pulse train for voiced segments), which permits both an effective compression of the inventory of units and ease of prosodic alteration. However, the obtained results by this technique have been unsatisfactory in both intelligibility and naturalness of the synthesized speech [3].

With the objective to overcome such drawbacks, different alternatives to the parametric model have been studied and presented in the open literature [2,4-7]. Among these techniques, PSOLA (pitch synchronous overlap and add) [4] is the most popular of them. Such a technique and its variants are widely studied nowadays, mainly the time domain variant (TD-PSOLA) [4].

Fernando S. Pacheco and Rui Seara are with LINSE: Circuits and Signal Processing Laboratory, Department of Electrical Engineering, Federal University of Santa Catarina, Florianópolis, SC, Brazil, 88040-900, Phone: + 55 48 331 9643, Fax: + 55 48 331 9091, E-mails: {fernando, seara}@linse.ufsc.br.

This technique consists of three basic stages: (a) signal decomposition in pitch synchronous segments, with a certain overlap; (b) modification of these segments; and (c) recombination of these segments by means of overlap-adding. The strong point of this technique is its ease of implementation. Although the synthesis quality is satisfactory in general, some artifacts, described as hoarseness and roughness, are sometimes introduced, compromising the synthesized speech quality [8]. Moreover, the spectral mismatch problem is not addressed by TD-PSOLA, leading to significant degradations of quality, mainly, when used short-time segments.

Other approaches have been discussed in the literature: short-time Fourier transform based synthesis [5], harmonic-stochastic model [6], MBROLA [2], sinusoidal synthesis [7], among others. In a general way, these techniques provide better results at the expense of a higher computational load. For applications in which such a feature is a meaningful factor, alternative approaches with a better speech quality and less complexity trade-off should be found. Thus, to retake the LP technique for prosodic modification seems to be an interesting alternative, mainly, due to its relatively lower computational burden along with the possibility of using voice alteration techniques [9] and spectral envelope matching at concatenation points [10]. However, some modifications should be considered to carry out such an aim. Recent results have shown that residual-excited linear predictive coding (RELP) can be used successfully in such an application [11,12]. Particularly in [11], the RELP technique has been used for pitch modification through simple operations of the residual signal. Such an approach was formally evaluated and compared with TD-PSOLA by means of a listening test. As a result, the former one showed a high potential, mainly, for pitch increase.

Taking into account the results presented in [11], in this paper we propose a new prosodic modification procedure based on RELP, which provides perceptually better results than TD-PSOLA for both directions of pitch alteration (increase and decrease).

This paper is organized as follows. Section 2 presents a brief background of the fundamental concepts of the LP approach for speech analysis and synthesis. Section 3 describes the prosodic modification procedure based on

the RELP technique. Results of listening tests are shown and discussed in Section 4. Finally, remarks and conclusions are presented in Section 5.

II. LP TECHNIQUE

For speech analysis and synthesis, the LP technique is based on the source-filter model of speech production. In this way, the speech signal is decomposed into an excitation signal and an all-pole filter which models the vocal tract. This filter has the following transfer function

$$H(z) = \frac{G}{1 - \sum_{k=1}^{P} a_k z^{-k}} \quad (1)$$

where G denotes a gain parameter, and a_k are the coefficients of the filter $H(z)$ of order P. In the speech synthesis process, the filter parameters are updated at each frame with duration between 5 and 20 ms, period in which the speech signal can be considered approximately stationary. Most common approaches used for calculating these parameters are covariance and autocorrelation methods. The input signal can be modeled by different ways. In the simplest, it is represented by a periodic pulse train for voiced sounds and random noise for unvoiced. An alternative formulation to model the excitation signal is through the residue between the original signal and its estimate obtained by linear prediction. This approach permits an almost perfect reconstruction of the original speech signal.

By considering $\hat{s}(n)$ an estimate of the speech signal, obtained from the original signal $s(n)$ (Eq. (2)), the prediction error (residue) can be obtained as in (3).

$$\hat{s}(n) = \sum_{k=1}^{P} a_k s(n-k) \quad (2)$$

$$e(n) = s(n) - \hat{s}(n) \quad (3)$$

Alternatively, this residue can be obtained by inverse filtering $(1/H(z))$ the original signal. This approach originates a well known technique in the literature, named residual-excited linear predictive coding (RELP), which has been widely used in speech compression applications. In our work, this same framework is used for prosodic modifications.

III. PROSODIC MODIFICATIONS USING RELP

To accomplish prosodic modifications using the residual signal, we must separate the process into three steps. First, the speech signal is partitioned into frames, which are analyzed using the LP technique. This stage computes the filter coefficients and the corresponding residual signal. Once the residual signal is obtained, the required prosodic modifications (duration and pitch) are carried out. This corresponds to the second step of the procedure. In this step, it is still possible to achieve spectral envelope smoothing (matching) operations. At the end, the synthesized signal is obtained by filtering the modified residual signal through the prediction filter $H(z)$, constituting the third step of the process. A block diagram of this procedure is shown in Fig.. In the following, we present a detailed view of each step of the process in question.

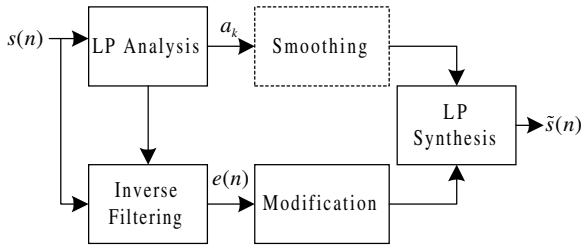

Fig. 1. Block diagram of the prosodic modification procedure.

A. Analysis

Let $s(n)$ be a speech signal segment, corresponding to a unit picked up from the inventory. It is also assumed that this signal is labeled with pitch marks $p(k)$, where $k = 1, 2, ..., M$ is the index of each mark, being M the total number of pitch marks of the unit. To obtain a better matching between frames in the concatenation process, an overlap of *S* samples is used. Thus, an analysis frame $s_k(\ell)$, for $\ell = 1, 2, ..., L$, is obtained from the samples $(p(k) - S)$ to $(p(k+1) - 1)$ of the signal $s(n)$. The signal $s_k(\ell)$ is then windowed by $h_s(\ell)$ (Eq. (4)), generating the signal $s'_k(\ell) = s_k(\ell)\, h_s(\ell)$.

$$h_s(\ell) = \begin{cases} \frac{1}{2}\left(1 - \cos\left(\pi \frac{\ell}{S+1}\right)\right), & 1 \le \ell < S \\ 1, & \ell \ge S \end{cases} \quad (4)$$

An LP analysis is applied to the signal $s'_k(\ell)$, giving rise to the prediction filter coefficients a_k, which permit to obtain the signal $\hat{s}_k(\ell)$. Thus, it is possible to compute the residual signal $e_k(\ell) = s'_k(\ell) - \hat{s}_k(\ell)$. Figure 2 depicts an analysis frame of the vowel [a], the residual signal and the associated spectral envelope. The used sampling rate is 8 kHz. Note that, by the nature of the application in question, the whole analysis procedure can be carried out off-line.

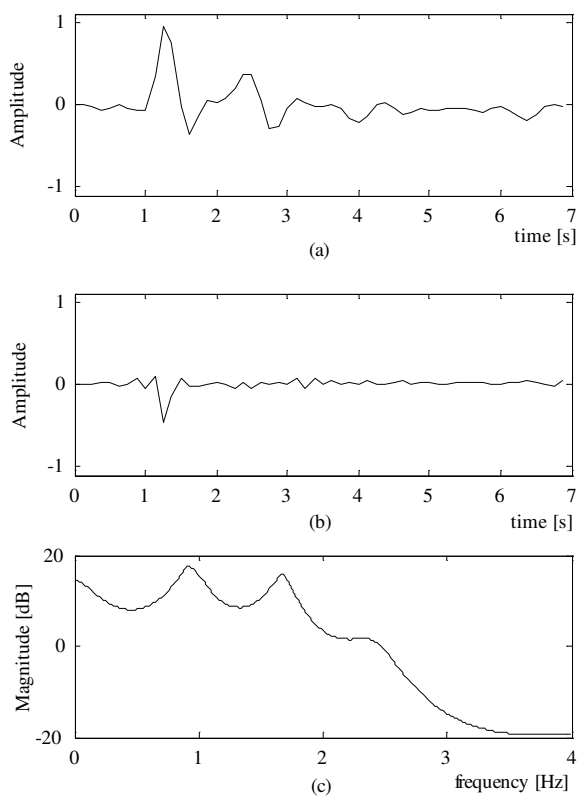

Fig. 2. Analysis frame of the vowel [a]. (a) Waveform; (b) residual signal; (c) spectral envelope (sampling rate of 8 kHz).

B. Modification

On the signal $e_k(\ell)$ a procedure is applied which depends on the pitch alteration direction. Therefore, considering T_k the length (number of samples) of the residual signal $e_k(\ell)$, we desire to modify such a length to obtain a new $e_k(\ell)$ ($e'_k(\ell)$), now with length T'_k. Thus, we use the following procedure:

i) for frequency increasing, the signal $e'_k(\ell)$ is obtained by weighting $e_k(\ell)$ by a window function $h_i(\ell)$, defined as:

$$h_i(\ell) = \begin{cases} 1, & 1 \le \ell < 0.75T'_k \\ \dfrac{1}{2}\left(1+\cos\left(\dfrac{\pi\left(\ell - 0.75T'_k\right)}{0.25T'_k}\right)\right), & 0.75T'_k \le \ell < T'_k \end{cases} \tag{5}$$

then, $e'_k(\ell) = e_k(\ell)\,h_i(\ell)$.

ii) for frequency decreasing, the signal $e'_k(\ell)$ is obtained by extending the signal $e_k(\ell)$ with the last $(T'_k - T_k)$ samples of this same signal.

C. Synthesis

The modified residual signal $e'_k(\ell)$ is processed by the LP filter, resulting an estimate of each speech signal frame $\tilde{s}_k(\ell)$ with modified pitch. By concatenating successive frames with modified pitch, following the same principle presented in Section 3.1, one obtains the modified pitch speech signal.

Figure 3 illustrates the waveform, pitch contour, and spectrogram of the word "passou" with modified pitch, synthesizing an affirmative form. Figure 4 shows the same word for an interrogative form.

IV. SUBJECTIVE ASSESSMENT

A. Procedure

To assess the proposed approach, a subjective listening test has been set up, following the absolute category rating (ACR) method described in [13]. Speech samples, consisting of word pairs, have been presented to listeners, who have been asked to rate each sample on a 5-point scale. This listening-quality scale has grades corresponding to: 5-excellent, 4-good, 3-fair, 2-poor, and 1-bad. This rate of scores is known as mean opinion score (MOS).

The speech material consists of Brazilian Portuguese words. The chosen words (64 words) give rise to a phonetically balanced set. Each word is synthesized by a text-to-speech system[1]. Every word is concatenated by using triphones. These triphones had their pitch periods marked by hand. Each word is submitted to pitch alterations using both the PSOLA algorithm and the proposed approach based on RELP. Alterations follow the directions:

- proportional pitch increase with factors 1.1; 1.2; 1.4; 1.5;
- proportional pitch decrease with factors 0.7; 0.8; 0.9.

Therefore, 896 words with modified pitch are generated (64 words modified by seven pitch factors using both methods). In addition, following ITU-T P.800 Recommendation [13], samples with different SNR conditions are also provided. These reference conditions are used to uphold comparisons of the subjective test results obtained at different conditions and/or different time. In this way, we include into the test word set some words without pitch alterations in which different levels of noise are added. Such conditions are obtained by use of the modulated noise reference unit (MNRU) [14], which is a unit that intends to introduce controlled degradations in speech signals. The used SNR levels are the following: 10, 18, 24 and 30 dB. Words without prosodic modifications (raw concatenative synthesis from the text-to-speech system) are also included.

[1] Such a system has been developed at the Electronic Instrumentation Laboratory: Circuits and Signal Processing (LINSE) of the Department of Electrical Engineering at the Federal University of Santa Catarina.

Consequently, the total number of reference stimuli (words) is 320.

A set of 24 listeners participated in the experiment. Each listens to 76 speech samples over headphones, and they are asked to rate the samples according to their general impression following the previously presented scale. The order of pair presentation is randomized. Each pair (a sample) is composed of two words modified by the same pitch factor and the same method. The listening test has lasted approximately 7 minutes for each listener.

B. Results

Following the described procedure, 1824 listening opinions have been obtained. For each pitch alteration factor and method used, the final scores are obtained by averaging the 24 listener scores.

In Table I, MOS values for different pitch alteration factors resulting from both methods are presented. Note that for both pitch modification directions (all tested cases), the proposed approach yields a better score than the one obtained by TD-PSOLA.

TABLE I
MOS FOR DIFFERENT PITCH ALTERATION FACTORS FROM PSOLA AND RELP METHODS

Pitch Alteration	MOS	
	PSOLA	RELP
0.7	2.78	2.85
0.8	2.85	3.18
0.9	3.46	3.83
1.1	3.14	3.47
1.2	2.49	3.10
1.4	2.26	2.50
1.5	2.00	2.32

As part of our experiment, Table II shows the obtained scores for the reference stimuli. Through these results it is possible to establish a relative scale for the subjective degradation level of each method, besides permitting a more effective comparison between the obtained results for different conditions.

TABLE II
MOS FOR REFERENCE CONDITIONS (MNRU) WITH DIFFERENT SNR

SNR(dB)	MOS
10	1.81
18	2.58
24	3.04
30	3.77
Original	3.96

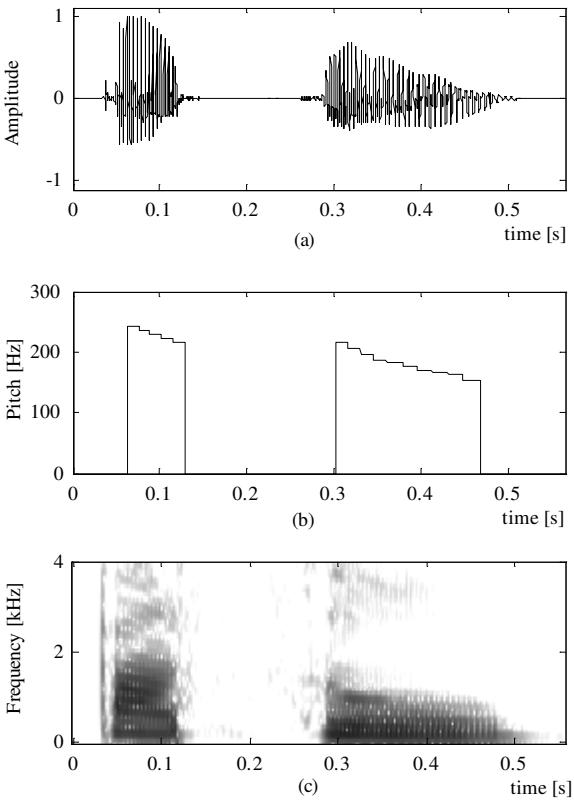

Fig. 3. Modified pitch of the word "passou" (affirmative form). (a) Waveform; (b) pitch contour; (c) spectrogram.

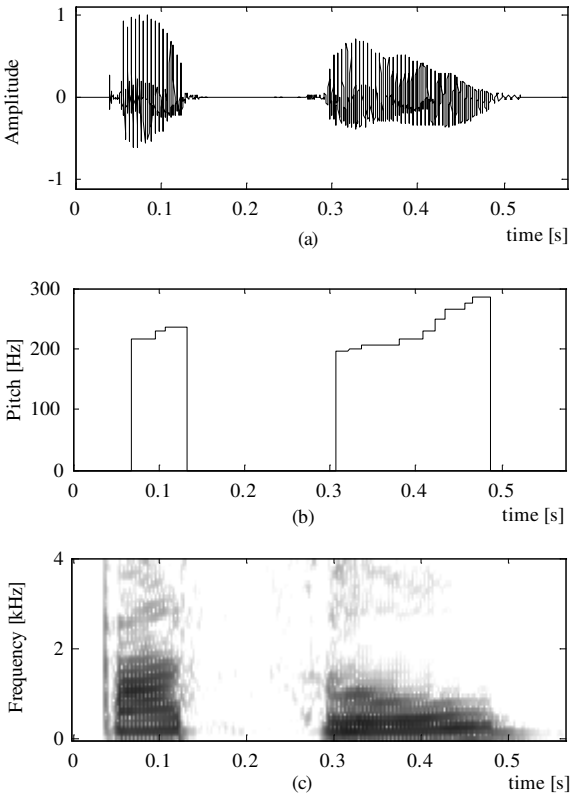

Fig. 4. Modified pitch of the word "passou" (interrogative form). (a) Waveform; (b) pitch contour; (c) spectrogram.

V. Remarks and Conclusions

In this paper we presented a new prosodic speech modification algorithm based on residual-excited linear prediction. In this approach, pitch and duration modifications are obtained by requiring a very low computational burden. Moreover, as such an algorithm is based on LP technique, it permits: efficient segment set compression; alterations in the spectral envelope to reduce discontinuities at concatenation points; use of alteration voice techniques.

In a formal subjective test, the proposed algorithm has been compared with TD-PSOLA for pitch modifications in increasing and decreasing directions. For both cases, the proposed approach has outperformed the TD-PSOLA technique.

References

[1] J. Makhoul. "Linear Prediction: A Tutorial Review", *Proc. IEEE*, vol. 63, no. 4, pp. 561-580, Apr. 1975.

[2] B. S. Atal and S. L. Hanauer. "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *J. Acoustical Soc. of America*, vol. 50, no. 2(2), pp. 637-655, 1971.

[3] T. Dutoit. "High Quality Text-to-Speech Synthesis: A Comparison of Four Candidate Algorithms", *Proc. ICASSP 94*, vol. I, pp. 565-568, Adelaide, Australia.

[4] F. Charpentier and E. Moulines. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", *Proc. Eurospeech 89*, vol. II, pp. 13-19, Paris, France.

[5] R. Veldhuis and H. Haiyan. "Time-scale and pitch modifications of speech signals and resynthesis from the discrete short-time Fourier transform", *Speech Communication*, vol. 18, no. 3, pp. 257-279, May 1996.

[6] F. Violaro and O. Böeffard. "A Hybrid Model for Text-to-Speech Synthesis," *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 5, pp. 426-434, Sept. 1998.

[7] E. B. George and M. J. T. Smith. "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. on Speech and Audio Processing*, vol. 5, no. 5, pp. 389–406, Sept. 1997.

[8] R. W. L. Kortekaas and A. Kohlrausch. "Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli", *J. Acoustical Soc. of America*, vol. 101, no. 4, pp. 2202-2213, Apr. 1997.

[9] P-F. Yang and Y. Stylianou. "Real Time Voice Alteration Based on Linear Prediction", *Proc. ICSLP 98*, Sydney, Australia.

[10] D. T. Chappell and J. H. L. Hansen. "A comparison of spectral smoothing methods for segment concatenation based speech synthesis", *Speech Communication*, vol. 36, no. 3-4, pp. 343-373, Mar. 2002.

[11] H. T. Bunnell, D. Yarrington and K. E. Barner. "Pitch Control in Diphone Synthesis", *Proc. II ESCA/IEEE Workshop on Speech Synthesis*, pp. 127-130, Sept. 1994, New Paltz, NY, USA.

[12] E. Rank and H. Pirker. "VIECTOS-Speech Synthesizer, Technical Overview," Technical Report OEFAI-TR-98-13, Austrian Research Institute for Artificial Intelligence, Vienna, Austria, Apr. 1998.

[13] Recommendation ITU-T P.800, Methods for subjective determination of transmission quality, Int'l Telecommunication Union, Geneva, Switzerland, Aug. 1996.

[14] Recommendation ITU-T P.810, Modulated Noise Reference Unit (MNRU), Int'l Telecommunication Union, Geneva, Switzerland, Feb. 1996.