ON THE APPLICATION OF RLS ADAPTIVE FILTERING FOR VOICE PITCH MODIFICATION

Rafael C. D. de Paiva, Luiz W. P. Biscainho and Sergio L. Netto

PEE/COPPE, LPS-DEL/Poli
Federal University of Rio de Janeiro
POBox 68504, Rio de Janeiro, RJ, Brazil, 21945-972
(rcdpaiva, wagner, sergioln)@lps.ufrj.br

ABSTRACT

This paper presents a pitch modification scheme, based on the recursive least-squares (RLS) adaptive algorithm, for speech and singing voice signals. The RLS filter is used to determine the linear prediction (LP) model on a sample-by-sample framework, as opposed to the LP-coding (LPC) method, which operates on a block basis. Therefore, an RLS-based approach is able to preserve the natural subtle variations on the vocal tract model, avoiding discontinuities in the synthesized signal and the inherent frame-delay associated to classic methods. The LP residual is modified in the synthesis stage in order to generate the output signal. Listening tests verify the overall quality of the synthesized signal using the RLS approach, indicating that this technique is suitable for real-time applications.

1. INTRODUCTION

Voice analysis and synthesis have been vastly studied in recent years, and many applications and methods have been developed in these areas. Pitch modification, the subject of the present paper, is closely related to voice synthesis, since both systems must consider particular aspects of the voice production system. Applications of voice pitch modification include, for instance, prosody changing, automatic tuning of singing voice and solo to unison transformation. Since it leads to an efficient parameterization of the speech signal, an analysis-and-synthesis scheme for pitch shifting algorithms may also be useful in other applications as concatenative synthesis of voice [1], voice morphing [2, 3, 4], voice transposition [5], or voice enhancement for speakers with vocal disorders [6].

Human speech production is often modeled as a source-filter system [7]. For voiced sounds the source signal may be modeled as a pseudo-periodic pulse train, resulting from the vibration of vocal folds. In such cases, the excitation determines the pitch f_0 (perceived fundamental frequency) as well as other characteristics such as breathness or falsetto emission [8, 9]. Unvoiced sounds are generated without vocal folds vibration. In these cases, the source models the turbulent behavior of the air flow as a noise signal. The filter is responsible for the distinction between phonemes and for the speaker's timbre, though the glottal excitation may influence the timbre too. In frequency domain, voiced speech segments are represented by a train of impulses spaced by f_o with their amplitudes multiplied by the filter's spectral envelope.

Pitch shifting was initially implemented by speed changes in musical recordings. Although this kind of approach was successful for some instruments, it was not possible to change the pitch without changing the duration of the signal. For speech signals, this technique worked very poorly, since it shifted the entire original spectrum, resulting in a very unnatural human voice. Successful experiences with pitch modification in speech result from parametric coding techniques [10, 11], that use the source-filter model discussed above, and from FFT based methods, like the *Phase Vocoder*, that is subject to *phasiness* distortion [12]. A family of non-parametric techniques include the *pitch synchronous overlap-and-add* (PSOLA) and its variations [13]. The PSOLA method segments the signal at pitch periods, then overlaps-and-adds them back to synthesize the output signal with the desired pitch characteristics. Extensions of this technique include combinations with the linear prediction and FFT approaches (LP- and FD-PSOLA) [13, 14].

This paper deals with the pitch shifting problem by using a sequential approach to LP-PSOLA. Instead of classical block techniques [11], it uses the recursive least-squares (RLS) adaptive filter to estimate the LP model in a sample-by-sample manner. This leads to a more natural sounding synthesized signal with smoother variations of the LP model than classical block techniques. Parameterization of PSOLA techniques adds flexibility towards further improvements and different applications.

The paper is organized as follows: the overall pitch-modifying system is presented in Section 2, which comprises a description of the RLS-based LP modeling and the excitation synthesis using LP-PSOLA; experimental results illustrating the system performance are included in Section 3; conclusions and future developments are pointed out in Section 4.

2. PROPOSED SYSTEM

2.1. LP modeling with RLS algorithm

In the LP model, depicted in Figure 1, the **a** FIR-filter coefficients are used to estimate the voice signal spectral envelope, and the prediction residual $e[n]$ is used to recover the original source signal.

In the proposed scheme, the RLS adaptive algorithm is employed to determine the **a** coefficients that minimize the objective function

$$\xi_{RLS}[n] = \sum_{i=0}^{n} \lambda^{n-i} e^2[i] = \sum_{i=0}^{n} \lambda^{n-i} \left(s[i] - \widehat{s}[i]\right)^2, \quad (1)$$

where $0 \ll \lambda < 1$ is the so-called forgetting factor and

$$\widehat{s}[n] = \sum_{p=1}^{P} a_p s[n-p] = \mathbf{a}^T \mathbf{s}[n-1], \quad (2)$$

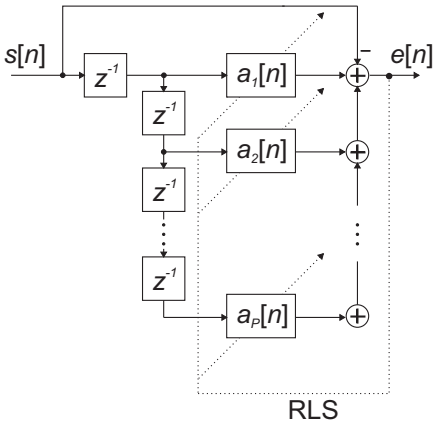


Figure 1: Block diagram of the LP scheme.

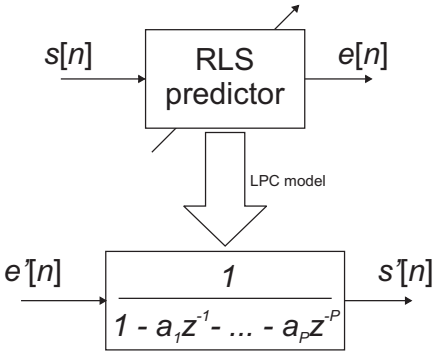


Figure 2: Analysis and synthesis modes using the RLS-LP model.

where

$$\mathbf{s}[n-1] = \left[\; s[n-1] \quad s[n-2] \quad \ldots \quad s[n-P] \;\right]^T, \quad (3)$$

$$\mathbf{a} = \left[\; a_1 \quad a_2 \quad \ldots \quad a_P \;\right]^T. \quad (4)$$

The result is a coefficient vector given by

$$\mathbf{a}[n] = \mathbf{R}^{-1}[n-1]\,\mathbf{p}[n] \quad (5)$$

where

$$\mathbf{R}[n-1] = \sum_{i=0}^{n} \lambda^{n-i}\mathbf{s}[i-1]\mathbf{s}[i-1], \quad (6)$$

$$\mathbf{p}[n] = \sum_{i=0}^{n} \lambda^{n-i}\mathbf{s}[i-1]s[i]. \quad (7)$$

The values of $\mathbf{R}^{-1}[n-1]$ and $\mathbf{p}[n]$ in Equation (5) can be calculated in a recursive way, to avoid extra computational burden, as given by

$$\mathbf{p}[n] = \mathbf{s}[n-1]s[n] + \lambda\mathbf{p}[n-1], \quad (8)$$

$$\mathbf{R}^{-1}[n-1] = \frac{1}{\lambda}\left[\mathbf{R}^{-1}[n-2] - \frac{\Psi[n]\Psi^T[n]}{\lambda + \Psi^T[n]\mathbf{s}[n-1]}\right], \quad (9)$$

where

$$\Psi[n] = \mathbf{R}^{-1}[n-2]\mathbf{s}[n-1]. \quad (10)$$

For further details on the RLS implementation the reader may refer to [15].

Using the RLS coefficients obtained in the analysis cycle, the LP model can be employed in the synthesis cycle with a new excitation signal $e'[n]$ to get the modified signal $s'[n]$ with the desired pitch characteristics, such that

$$s'[n] = e'[n] - \mathbf{a}^T[n]\mathbf{s}'[n-1], \quad (11)$$

as represented in Figure 2. The signal $e'[n]$ is obtained from $e[n]$ using PSOLA algorithm as detailed in Section 2.2.

One may notice that the RLS-LP model is determined for each time sample n, leading to smooth transitions between consecutive models. The resulting model quality depends on the choices of the number of LP coefficients P and the forgetting factor λ. It is possible to show that the RLS sequential solution is a special case of the classical LP block solutions, with λ controlling the equivalent length of an exponential analysis window [16]. Proper values of P and λ may vary for distinct sampling rates.

2.2. Source implementation

There are several approaches to generate the excitation signal $e'[n]$. The most straightforward is to use an impulse train plus noise for voiced segments, but it often leads to artificial speech results. Several works employ a glottal pulse model [8, 9, 17, 18] to emulate the vocal effort, the parameterization of which constitutes a cumbersome task. One sample of glottal pulse from $e[n]$ can also be used as a pulse model to generate $e'[n]$ [16].

In this work, the LP error $e[n]$ is used to generate the modified excitation signal $e'[n]$, as illustrated in Figure 2. The desired pitch is introduced in $e'[n]$ by means of a PSOLA technique applied to the residual signal $e[n]$. This procedure constitutes the so-called LP-PSOLA technique [13]. To do that, one applies a peak tracking algorithm to determine the instants of glottal closure, which correspond to changes in the statistical properties of the signal. Figure 3 illustrates the result of a pitch marking procedure on a speech signal $s[n]$. In this figure, $p_m[n]$ indicates the pitch marks and the intervals $p[n]$ are the corresponding pitch periods. The procedure can be implemented using wavelets [19], by observing variations

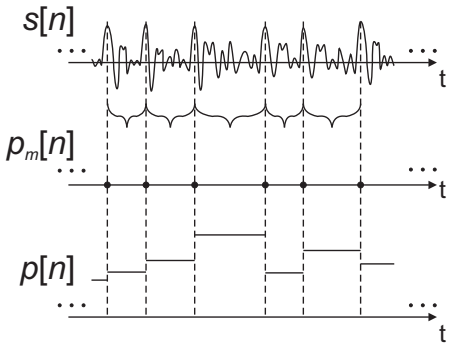


Figure 3: Pitch marks $p_m[n]$ and associated pitch detection $p[n]$ of a speech signal $s[n]$.

of statistical properties [20], or simply by direct observation of the amplitude envelope.

Once reliable pitch marks $p_m[n]$ and pitch periods $p[n]$ of the original signal are determined, the desired pitch contour can be modified as desired. For that purpose, new pitch marks $p'_m[n]$ are determined corresponding to a new pitch period $p'[n]$, such that

$$p'[n] = \beta[n]p[n], \tag{12}$$

where $\beta[n]$ is the pitch-period modification factor, which can be made variable for natural prosody modification, automatic pitch correction, vibrato synthesis, and so on. The new pitch marks $p'_m[n]$ are determined by forcing an interval of $p'[n]$ samples between two consecutive marks, such that a pitch mark will be placed at position $n + p'[n]$ if n has a pitch mark (*i.e.* $p'_m\,[n + p'[n]] = 1$ if $p'_m[n] = 1$, where pitch mark positions are indicated by 1).

The next step is to link each new pitch mark $p'_m[n]$ with its corresponding closest peak on the original signal $p_m[n]$. This is done straightforwardly by comparing the time index of $p_m[n]$ and $p'_m[n]$, as illustrated in Figure 4.

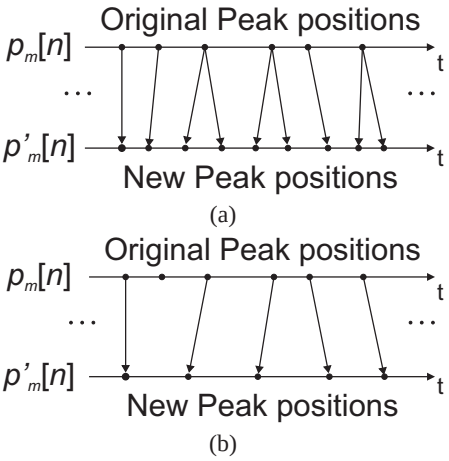


Figure 4: Pitch mark association for the synthesized source signal with: (a) Increased pitch; (b) Decreased pitch.

In the final step of the new source generation, each peak in the original signal is then segmented, by two half-hanning windows, starting at the preceding pitch mark and ending at the next one. The resulting source segments are put together by an overlap-and-add procedure according to the new pitch period $p'[n]$ obtained previously, as given in Figure 5.

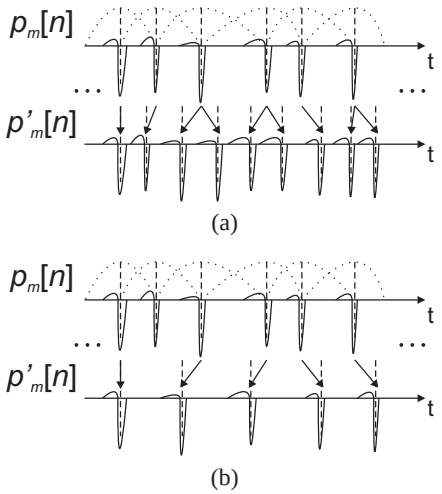


Figure 5: Composition of the new source signal with: (a) Increased pitch; (b) Decreased pitch. In each case, the dashed lines correspond to the segmentation windows centered at each original pitch mark.

3. EXPERIMENTAL RESULTS

This section describes some practical experiments performed with the proposed pitch-modification system.

Example 1: A portion of a song recorded by a female Brazilian singer was modified using $\beta[n] = 2$ and $\beta[n] = 0.5$. Figure 6 shows a small portion of the original and modified signals, whereas Figures 7 shows their corresponding spectrograms.

The proper pitch modification can be inferred from Figure 6 by noticing how the modified peaks become more separated or closer when $\beta[n] = 2$ and $\beta[n] = 0.5$, respectively. A similar conclusion can be drawn observing the fringes in the modified spectrograms in Figure 7.

Example 2: Once again the pitch characteristic of a song was modified with $\beta[n] = 2$ and $\beta[n] = 0.5$. This time, however, the recorded voice of a male singer was employed. The time- and frequency-domain representations of the resulting signals are depicted in Figures 8 and 9, respectively.

Once again, from these figures, it is easy to identify the desired pitch modification, while the original spectral envelope is kept essentially unchanged in all cases.

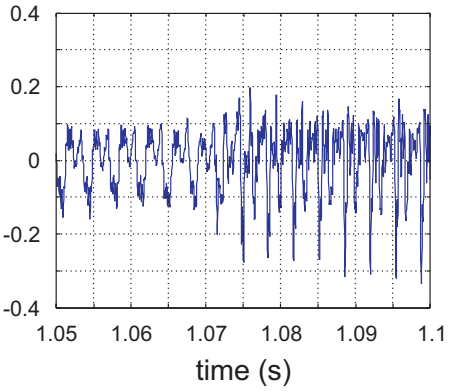
Example 3: Figure 10 compares the results of the proposed method and PSOLA for a one-octave decreasing of pitch, *i.e.* $\beta[n] = 2$, on the same signal employed in Example 1. This figure illustrates a major drawback of the PSOLA algorithm, which is the significant energy decrease in between consecutive peak marks when $\beta[n] > 1$. Although larger analysis windows could be employed in such cases, they could lead to spurious peaks on the synthesized signal, since adjacent peaks would not be sufficiently reduced by the analysis windows. These spurious peaks might lead to roughness on the modified signal.

It is worth noting that, instead of directly overlapping portions of the output signal as in PSOLA, the LP-PSOLA intrinsically keeps the responses to each excitation pulse individualized, as in the voice production model itself. This feature, coupled with the smoothly tracked RLS-LP model, allows one to expect that the proposed system can produce a more natural synthesized voice.

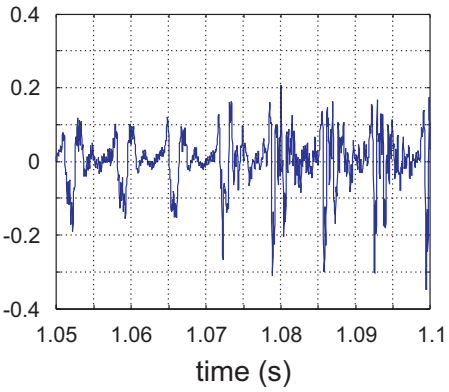
4. CONCLUSIONS

A complete system for pitch modification of voice signals was presented in this paper. Spectral envelope modeling is performed by an adaptive RLS filter, leading to a sample-by-sample estimation of the LP model. This results in smooth transitions in the estimated model, thus yielding a more natural synthesized signal.

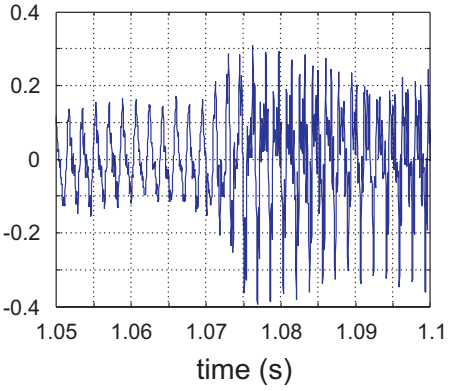
The source signal with the desired pitch characteristics is ob-



(a)

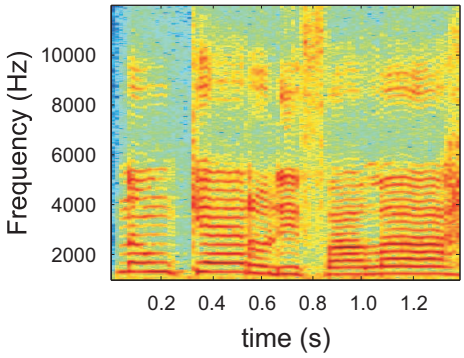


(b)

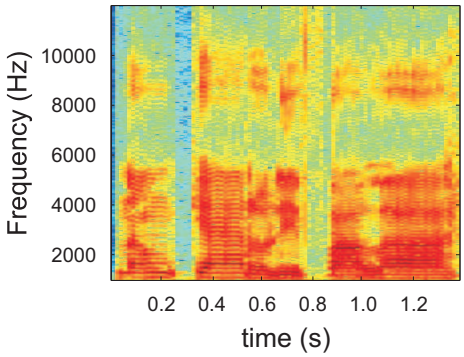


(c)

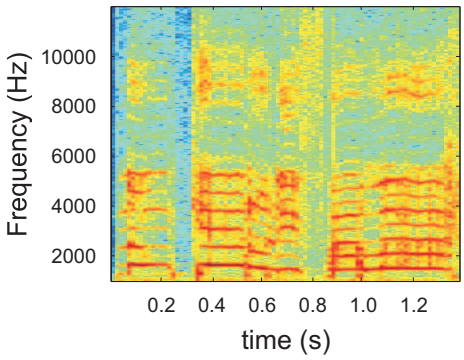
Figure 6: Extracts of signals in Example 1: (a) Original signal; (b) Modified signal with $\beta[n] = 2$; (c) Modified signal with $\beta[n] = 0.5$.



(a)



(b)



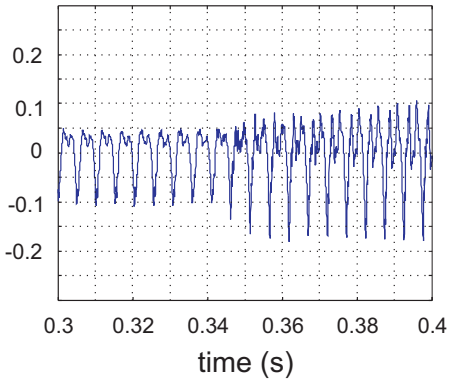
(c)

Figure 7: Spectrograms of signals in Example 1 showing that the original spectral envelope is preserved in all cases: (a) Original spectrogram; (b) Modified spectrogram with $\beta[n] = 2$; (c) Modified spectrogram with $\beta[n] = 0.5$.

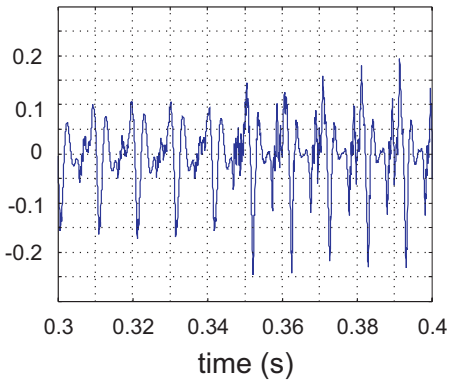
tained by applying a PSOLA algorithm on the RLS residual error. The advantage of this method is to preserve the individuality of each glottal pulse. Furthermore, in case of insufficient LP-model order, part of the spectral envelope information is carried by the RLS residual error, thus reducing the envelope modeling error in the synthesized signal. Additionally, the proposed system is able to sustain signal information in between pitch marks even for large pitch-modification factors $\beta[n]$.

Informal subjective tests have shown good results for pitch scale modification in the range $0.5 \leq \beta[n] \leq 2$, which, in musical terms, means from an octave downwards to an octave upwards. This system is currently being tested against standard pitch modification methods.

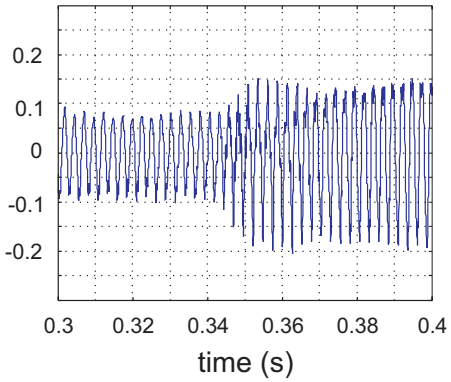
The parameterization inherent to the described method suggests the system can be extended to fit voice *morphing* applications, *e.g.* male-female conversion, voice transposition, and non-human voice synthesis for voice editing in cartoons.



(a)

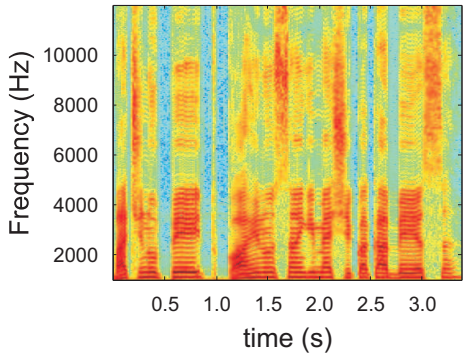


(b)

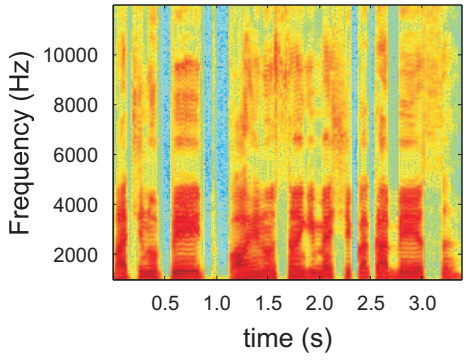


(c)

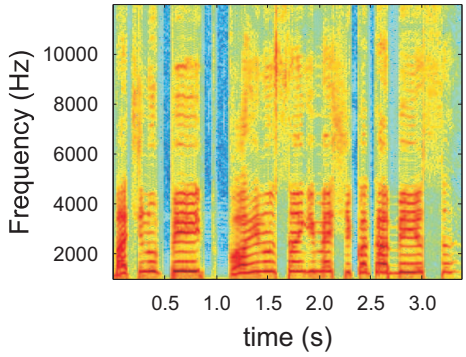
Figure 8: Extracts of signals in Example 2: (a) Original signal; (b) Modified signal with $\beta[n] = 2$; (c) Modified signal with $\beta[n] = 0.5$.



(a)



(b)



(c)

Figure 9: Spectrograms of signals in Example 1 showing that the original spectral envelope is preserved in all cases: (a) Original spectrogram; (b) Modified spectrogram with $\beta[n] = 2$; (c) Modified spectrogram with $\beta[n] = 0.5$.

5. ACKNOWLEDGMENTS

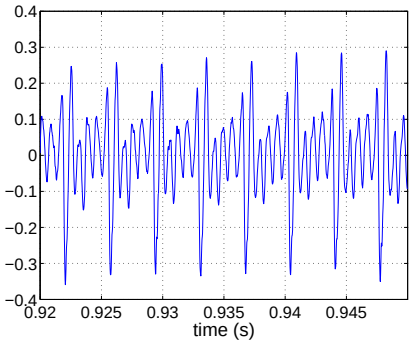
The authors would like to thank the Brazilian sponsors of this work, *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) and *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro* (FAPERJ), for their support.

6. REFERENCES

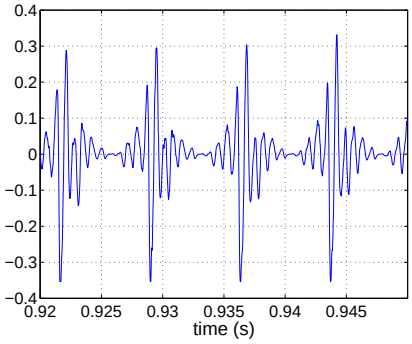
[1] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Processing Magazine*, vol. 24, no. 2, pp. 67–79, March 2007.

[2] L. Fabig and J. Janer, "Transforming singing voice expression - the sweetness effect," in *Proc. of the DAFx04 - 7th International Conference on Digital Audio Effects*, Naples, Italy, October 2004.

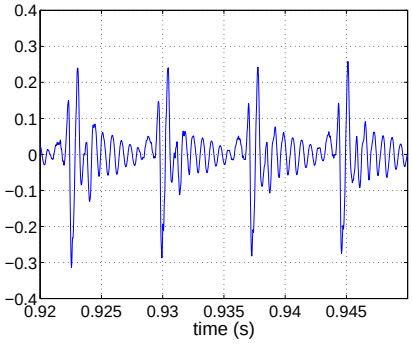
[3] A. Loscos and J. Bonada, "Emulating rough and growl voice



(a)



(b)



(c)

Figure 10: Extracts of signals in Example 3: (a) Original signal; (b) Signal modified by PSOLA; (c) Signal modified by the proposed method.

in spectral domain," in *Proc. of the DAFx04 - 7th International Conference on Digital Audio Effects*, Naples, Italy, October 2006.

[4] P. Depalle, G. Garcia, and X. Rodet, "The recreation of a castrato voice, farinelli's voice," in *Proc. of the WASPAA'95 - Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, October 1995, IEEE, pp. 15–18.

[5] E. Turajlic, D. Rentzos, S. Vaseghi, and C. H. Ho, "Evaluation of methods for parameteric formant transformation in voice conversion," in *Proc. of the ICASSP'03 - International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, Hong Kong, April 2003, IEEE, pp. 724–727.

[6] J. Bonada and A. Loscos, "Esophageal voice enhancement by modeling radiated pulses in frequency domain," *121st Audio Engineering Society Convention*, October 2006, Preprint 6952.

[7] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-Time Processing of Speech Signals*, Wiley-IEEE, 1999.

[8] Q. Fu and P. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 492–501, March 2006.

[9] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, February 1990.

[10] J. A. Moorer, "The use of linear prediction of speech in computer music applications," *Journal of Audio Engineering Society JAES*, vol. 27, no. 3, pp. 134–140, March 1979.

[11] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, April 1975.

[12] J. Laroche and M. Dolson, "Phase-vocoder: about this phasiness business," in *Proc. of the WASPAA'97 - Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, October 1997, IEEE.

[13] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175–205, February 1995.

[14] K. S. Rao and B. Yegnanarayana, "Prosody modification using instants of significant excitation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 972–980, May 2006.

[15] P. S. R. Diniz, *Adaptive Filtering: Algorithms and Practical Implementations*, Kluwer, 2 edition, 2002.

[16] R. C. D. de Paiva, L. W. P. Biscainho, and S. L. Netto, "A sequential system for voice pitch modification," in *Proc. of the AES-Brazil'07, 5th AES-Brazil Conference*, São Paulo, Brazil, May 2007, Audio Engineering Society, pp. 11 – 16.

[17] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," STL-QPSR 26 4, Dept. for Speech, Music and Hearing - Royal Institute of Technology, Stockholm, Sweden, 1985.

[18] H. L. Lu, *Toward a High-Quality Singing Synthesizer with Vocal Texture Control*, Ph.D. thesis, Dept. of Electrical Engineering - Stanford University, Palo Alto, USA, July 2002.

[19] S. Kadambe and G. F. Boudreaux-Bartels, "Application of the wavelet transform for pitch detection of speech signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 917–924, March 1992.

[20] C. Ma, Y. Kamp, and L. F. Willems, "A Frobenius norm approach to glottal closure detection from the speech signal," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 258–265, April 1994.