Sociedade de Engenharia de Áudio

Artigo de Congresso

Apresentado no 5º Congresso de Engenharia de Áudio
11ª Convenção Nacional da AES Brasil
21 a 23 de Maio de 2007, São Paulo, SP

A Sequential System for Voice Pitch Modification

Rafael C. D. Paiva,[1] Luiz W. P. Biscainho,[1] and Sergio L. Netto[1]

[1] LPS PEE-COPPE, DEL-Poli,UFRJ Caixa Postal 68504 - Rio de Janeiro, RJ, 21941-972, Brasil

rcdpaiva@lps.ufrj.br, wagner@lps.ufrj.br, sergioln@lps.ufrj.br

ABSTRACT

This paper presents a system for pitch modification of voice signals. In the proposed structure, a recursive least-squares (RLS) adaptive algorithm is employed to determine the linear prediction (LP) model for the vocal tract during the analysis stage. In the subsequent synthesis, an excitation signal with the desired pitch contour is processed by the LP model obtained previously. The method avoids the frame delay inherent to block-processing techniques as well as discontinuities in the LP model. Listening tests evaluate positively the quality of the synthesized signal.

0 INTRODUCTION

Pitch modification and voice transformation have been major subjects for speech processing research in the last years. Perhaps the main reason for this continuous interest is the fact that the human perception of voice is very accurate. Therefore, practical processing of voice signals may require quite intricate techniques for a proper pitch shifting procedure.

Applications of pitch modification systems include voice editing for movies, automatic tuning of musical signals, concatenative synthesis of voice [1], voice morphing [2, 3] or even esophageal voice enhancement [4].

Voice production is often modeled as a source-filter system [5]. In this model, the *excitation* or source signal corresponds to the vocal folds (vibratory or not) output, and appears in guise of either a pseudo-periodic or a random-like waveform. Pseudo-periodic excitation, which can be characterized by a fundamental frequency value, leads to voiced phonemes, whereas random-like excitation leads to unvoiced phonemes. Some expressive information can be also conveyed by the excitation [6, 7]. The *filter*, which represents the vocal tract characteristics (mouth and tongue positions, lips opening, articulatory elements, *etc.*) processes the excitation information to ultimately generate the voice signal. The vocal tract is responsible for the discrimination of different phonemes, which present frequency patterns, often called spectral envelope, that do not change much among different speakers [5], even for distinct excitation characteristics.

The first experiences in pitch modification of speech signals were implemented by speed changes in recorded

signals. Of course, this approach could not be performed without changing the duration of the original signal. The main drawback of this technique was to change the speaker's timbre in a very annoying way, since it distorted the spectrum envelope of the original signal. Those experiences showed that the spectral envelope should be preserved for high quality time and pitch scaling of speech signals.

A more flexible approach is based on the direct spectrum envelope modeling using the *discrete Fourier transform* (DFT), which may lead to the so-called *phasiness* distortion when synthesis blocks are simply put together without proper care. The *phase vocoder* [8] belongs to this class of solution. Another idea developed from speech coding techniques: the pitch shifting can be performed over a *linear prediction* (LP) model [9].

While exhibiting great advantages when compared to primitive speed changing approaches, all the solutions listed before operate on a block-by-block basis: the signal is segmented by an adequate window, the blocks are processed and combined by an *overlap-and-add* (OLA) method. Improvements like the use of variable-length blocks and pitch-synchronization led to very successful methods, including the popular *pitch synchronous OLA* (PSOLA) and its variants [10].

This work proposes applying a sample-oriented solution, *viz.* the *recursive least-squares* (RLS) adaptive algorithm, to determine the LP filter in a pitch shifting system. Section 1 reviews and compares one classic block solution and the sequential RLS solution for LP modeling. Section 2 describes the overall system proposed, and Section 3 illustrates pitch modification using the new technique through practical examples. Section 4 concludes the paper by discussing some performance issues of the system.

1 LINEAR PREDICTION MODELING

Given a speech signal $s[n]$, suppose one wishes to obtain a set of P filter coefficients a_p such that

$$\hat{s}[n] = \sum_{p=1}^{P} a_p s[n-p] \tag{1}$$

is a good estimate of $s[n]$. The estimation (prediction) error is defined as

$$e[n] = s[n] - \hat{s}[n], \tag{2}$$

which can be seen as the result of submitting $s[n]$ to a filter with transfer function

$$G(z) = 1 - a_1 z^{-1} - a_2 z^{-2} - \ldots - a_P z^{-P}. \tag{3}$$

Conversely, Equations (1) and (2) can describe the so-called *autoregressive* (AR) model, in which $s[n]$ is interpreted as the result of filtering white noise by a filter with transfer function $H(z) = 1/G(z)$. In that sense, the model order P must be high enough to allow the excitation signal (formerly prediction error) $e[n]$ to be considered white. In speech processing techniques, this assumption is not particularly useful, since during voiced portions of speech the voice production model assigns to the excitation the shape of a pseudo-periodic pulse train. In fact, using a reduced-order predictor for voiced speech yields the expected pulsed $e[n]$, with almost equal-amplitude harmonics. Usually this kind of modeling is referred to as *linear prediction coding* (LPC). It leads to an economic way of speech storage and transmission.

There are several practical approaches to obtain the LPC coefficients of a given signal. The most popular and widely used in speech processing work in blocks and are often called just by LPC methods. Classic block solutions include the autocorrelation and the covariance methods [11]. Estimation of LPC coefficients can also be made through adaptive filtering techniques.

The speech spectral envelope is estimated by the LPC model, but it can alternatively be obtained by frequency or cepstrum transforms, as well as other techniques.

The classic autocorrelation block solution and the adaptive RLS solution are revisited in the next subsections.

1.1 Classic Block Solution

The block solution for LPC modeling is performed on a time frame of N samples, considering the signal to be modeled is ergodic and wide-sense stationary (WSS) within this frame.

The error function for this solution is given by

$$e_m[n] = s_m[n] - \widehat{s}_m[n] = s_m[n] - \mathbf{a}_{m,P}^T \, \mathbf{s}_m[n-1], \tag{4}$$

where

$$\mathbf{a}_{m,P} = \left(a_{m,1}\ a_{m,2}\ \ldots\ a_{m,P}\right)^T, \tag{5}$$

$$\mathbf{s}_m[n-1] = \left(s_m[n-1]\ s_m[n-2]\ \ldots\ s_m[n-P]\right)^T \tag{6}$$

with the superscript T denoting matrix transposition.

Considering $n = (P+1), (P+2), \ldots, N$, Equation (4) can be written in matrix form as

$$\mathbf{e}_m = \mathbf{d}_m - \mathbf{S}_m \mathbf{a}_{m,P}, \tag{7}$$

where

$$\mathbf{e}_m = \left(e_m[N]\ e_m[N-1]\ \ldots\ e_m[P]\right)^T, \tag{8}$$

$$\mathbf{d}_m = \left(s_m[N]\ s_m[N-1]\ \ldots\ s_m[P]\right)^T, \tag{9}$$

and $\mathbf{S}_m$ is an $(N-P) \times P$ matrix defined as

$$\mathbf{S}_m = \begin{pmatrix} s_m[N-1] & s_m[N-2] & \ldots & s_m[N-P] \\ s_m[N-2] & s_m[N-3] & \ldots & s_m[N-P-1] \\ \vdots & \vdots & \ddots & \vdots \\ s_m[P] & s_m[P-1] & \ldots & s_m[0] \end{pmatrix} \tag{10}$$

Minimizing the mean squared-error $\mathbf{e}_m^T\mathbf{e}_m$, the optimal coefficient vector is given by

$$\mathbf{a}_{m,P} = \mathbf{R}_{D,m}^{-1}\,\mathbf{p}_{D,m}, \tag{11}$$

where $\mathbf{R}_{D,m} = \mathbf{S}_m^T\mathbf{S}_m$ is the so-called deterministic correlation matrix of the signal s and $\mathbf{p}_{D,m} = \mathbf{S}_m^T\mathbf{d}_m$ is the deterministic cross-correlation vector. For an ergodic WSS signal, the matrix $\mathbf{R}_{D,m}$ is Toeplitz, and efficient methods for calculating its inverse are available.

1.2 Sequential RLS Solution

The LP modeling using RLS adaptive filtering is made on a sample-by-sample basis. The objective function for the *weighted least-squares* (WLS) formulation is given by

$$\xi_{RLS}[n] = \sum_{i=0}^{n} \lambda^{n-i} e^2[i], \tag{12}$$

where $0 << \lambda < 1$ is the so-called forgetting factor. Equation (12) can be rewritten as

$$\xi_{RLS}[n] = \mathbf{e}^T[n]\Lambda^2[n]\mathbf{e}[n], \tag{13}$$

where

$$\mathbf{e} = (e[n]\; e[n-1]\; \ldots\; e[0])^T, \tag{14}$$

$$\Lambda^2[n] = \text{diag}\left(1, \lambda, \lambda^2, \ldots, \lambda^{n-1}\right). \tag{15}$$

Minimizing $\xi_{RLS}[n]$, the optimal coefficient vector is given by

$$\mathbf{a}_P[n] = \mathbf{R}_D^{-1}[n-1]\,\mathbf{p}_D[n], \tag{16}$$

with

$$\mathbf{R}_D[n-1] = \mathbf{S}^T[n-1]\Lambda^2[n]\mathbf{S}[n-1], \tag{17}$$

$$\mathbf{p}_D[n-1] = \mathbf{S}^T[n-1]\Lambda^2[n]\mathbf{d}[n], \tag{18}$$

where

$$\mathbf{d} = (s[n]\; s[n-1]\; \ldots\; s[0])^T. \tag{19}$$

For the RLS algorithm, the inverse matrix in Equation (16) is determined recursively as

$$\mathbf{R}_D^{-1}[n-1] = \frac{1}{\lambda}\left[\mathbf{R}_D^{-1}[n-2] - \frac{\Psi[n]\Psi^T[n]}{\lambda + \Psi^T[n]\mathbf{s}[n-1]}\right], \tag{20}$$

with

$$\Psi[n] = \mathbf{R}_D^{-1}[n-2]\mathbf{s}[n-1], \tag{21}$$

$$\mathbf{p}_D[n] = \lambda\mathbf{p}_D[n-1] + s[n]\mathbf{s}[n-1]. \tag{22}$$

In practice, one often uses $\mathbf{R}_D^{-1}[-1] = \delta\mathbf{I}$, where δ is a small positive constant, and $\mathbf{p}_D[0]$ is a null vector. For further details on the implementation of the RLS algorithm the reader may refer to [12].

1.3 Comparison of Block and Sequential Solutions

The classic autocorrelation and the sequential RLS solutions are closely related. For instance, one can rewrite the autocorrelation solution as

$$\mathbf{a}_{m,P} = (\mathbf{S}^T[n-1]\Lambda_m'[n]\mathbf{S}[n-1])^{-1}\mathbf{S}^T[n-1]\Lambda_m'[n]\mathbf{d}[n], \tag{23}$$

which closely resembles the RLS solution given in Equation (16) with

$$\Lambda_m'[n] = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N\times N} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}. \tag{24}$$

This indicates that both solutions are built around deterministic correlation matrices and vectors including weighting terms: $\Lambda_m'[n]$, in the LS solution of the block method; $\Lambda[n]$, in the WLS solution of the sequential method. Furthermore, one can see that both methods consider only part of the signal at a given time instant: the classic autocorrelation method uses a square window, whereas the RLS algorithm applies an exponential window over past samples, resulting in a smaller contribution of older observations of $\mathbf{s}[n]$.

Advantages of the RLS algorithm include avoiding the inversion of the correlation matrix, by employing the recursive estimate given by Equation (20), and allowing an estimate of the optimal coefficient vector for each time instant n, instead of working in blocks as in the classic method.

In order to provide a quick comparison between the classic block and RLS methods, we consider a low-order LP modeling of a segment of nonstationary speech. Figure 1 shows the pole tracking in the z-plane for each method: (a) Classic block model using 20-ms frames with no time overlap between consecutive frames; (b) The same as (a), but with a 5-ms linear interpolation on the LP models; (c) RLS model. In these figures, the filled circles indicate initial positions of the LP model and the rhombus marks correspond to the final positions. From these plots, one can observe a similar path being followed by all three schemes, with the RLS yielding the trajectory with better continuity.

2 Proposed System

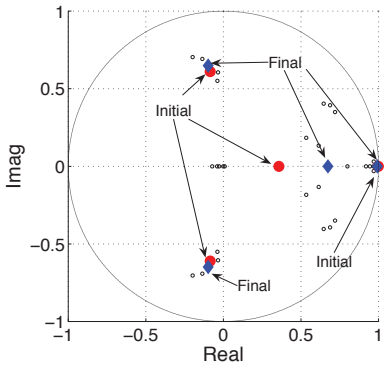
The proposed pitch modification algorithm uses the LP model $\mathbf{a}_P$ obtained with by RLS algorithm to synthesize the voice signal $s'[n]$ with the desired pitch information, as indicated in Figure 2, by

$$s'[n] = e'[n] - \mathbf{a}_P^T[n]\mathbf{s}'[n-1], \tag{25}$$

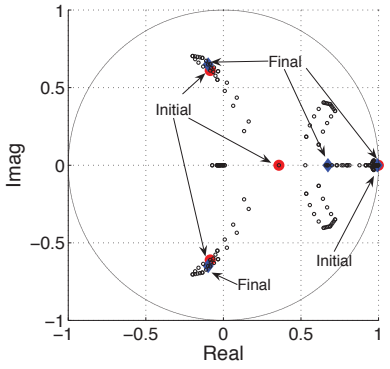
where

$$\mathbf{s}'[n-1] = (s'[n-1]\; s'[n-2]\; \ldots\; s'[n-P])^T. \tag{26}$$

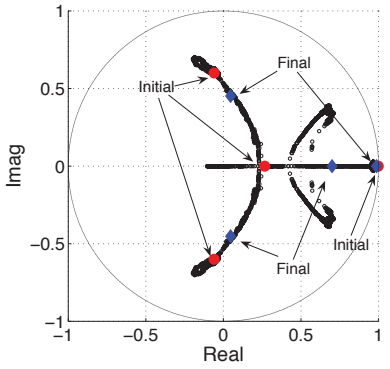
In practice, for low sampling frequency (around 8 kHz) one must use a model order $10 \leq P \leq 15$, and



(a)



(b)



(c)

Figure 1: Pole evolutions: (a) Block algorithm with 20-ms frame; (b) Block algorithm with 20-ms frames, with 1:4 coefficient interpolation; (c) RLS algorithm.

for high quality audio (sampling frequencies between 30 and 44.1 kHz) a larger P is required.

A desired pitch-period contour $p'[n]$ can be determined from a scaling transformation $\beta[n]$ on the original pitch period $p[n]$. To do that, $p[n]$ and the corresponding pitch marks $p_m[n]$, associated to the vocal folds closures, as illustrated in Figure 3, must be known.

For that purpose, the pitch period $p[n]$ can be calculated using classic autocorrelation method [5], cepstrum-based method [13], the YIN estimator [14], and event-based techniques. The latter methods look

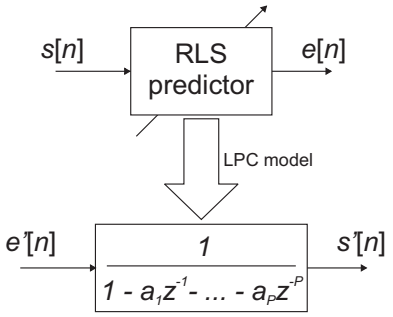


Figure 2: Analysis and synthesis scheme using the RLS algorithm.

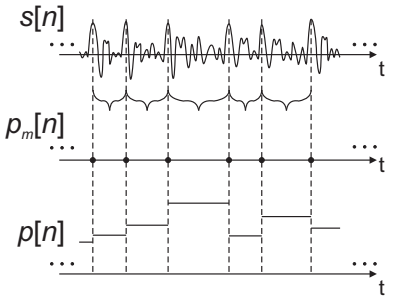


Figure 3: Example of pitch marking and associated pitch detection.

for periodicity in the amplitude envelope, and try to find pitch marks that would correspond to the instants of significant glottal excitation. Such methods identify each pitch period individually, requiring precise heuristic information on the signal at hand. Implementation of event-based techniques for speech includes wavelets [15], characteristics of autocorrelation matrix [16], and simpler methods like amplitude envelope calculation.

New pitch marks $p'_m[n]$ are then obtained incorporating the new pitch-period information $p'[n]$, as represented in Figure 4, where c is an auxiliary sample

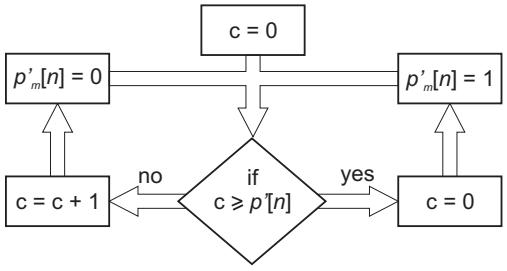


Figure 4: Search algorithm for modified pitch marks $p'_m[n]$ according to modified pitch period $p'[n]$.

counter. If $p'[n]$ is not an integer number, it may be rounded to its closest integer.

The excitation signal $e'[n]$ can then be formed by

concatenating several copies of a standard pitch interval of the RLS residual signal $e[n]$ according to the new pitch marks $p'_m[n]$. Alternative approaches include using a typical model of the glotal pulse [6, 7] or combining a pitch period of the LP residual with the PSOLA [10] algorithm. The overall procedure for composing $e'[n]$ is depicted in Figure 5, where $\beta[n]$ can be made variable for a wide range of pitch-modification scenarios.

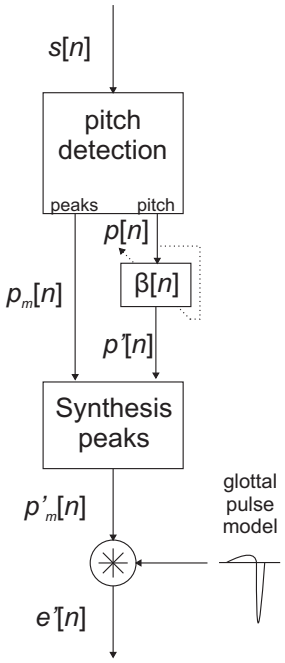
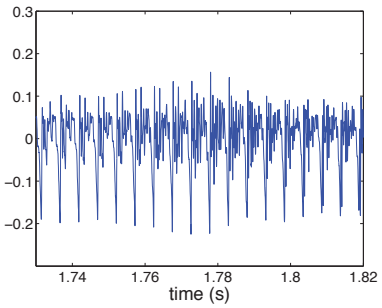
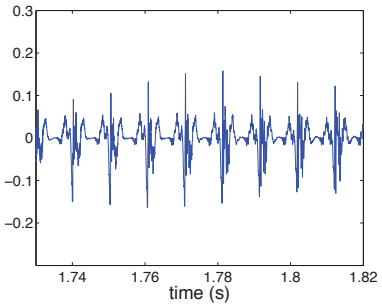


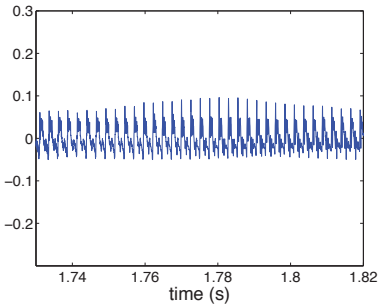
Figure 5: Synthesis of the modified excitation signal.



(a)



(b)



(c)

Figure 6: Time plot of: (a) Original signal; (b) Modified signal with lower pitch, $\beta = 2$; (c) Modified signal with higher pitch, $\beta = 0.5$.

3 Experimental Results

The proposed system was implemented with $P = 30$ RLS coefficients and $\lambda = 0.999$ for a signal sampled at 44.1 kHz. Figure 6 shows the results for a voiced segment of a spoken phrase and Figure 7 shows the correspoding spectrogram. From these figures, one can clearly see that the spectral envelope is preserved in the generated signals, whereas the pitch information is modified as desired. Listening tests have validated the quality of the overall result for pitch scaling factors in the 2-octave range $0.5 \leq \beta \leq 2$.

4 Conclusions

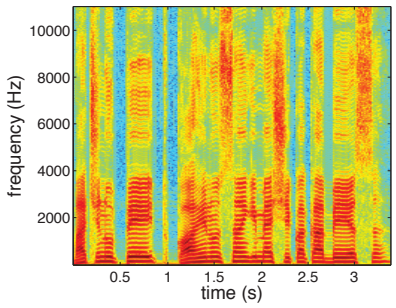
This paper presented a pitch shifting scheme using adaptive filtering techniques for spectral envelope estimation. The adaptive RLS filter has proved to be a good solution when compared to conventional block solutions for estimating of the LPC model of speech signals, since it guarantees model smoothness through time.

The work has illustrated the applicability of this system on a pitch scaling procedure, which is the basis of automatic tuning, solo to unison transformation or prosody transposition. Listening tests have verified the overall quality of the algorithm for a wide range of pitch transformation.

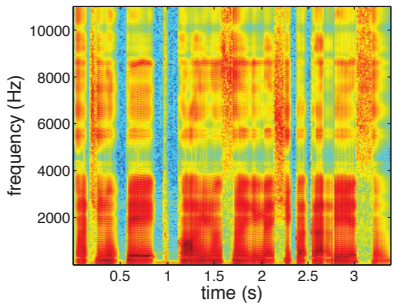
The next step of this work is to perform the systematic comparison of the proposed technique with others available in the literature. For this purpose, a wider voice corpus, including speakers and singers of different gender and age as well as different phonation types, will be employed.

5 Acknowledgments

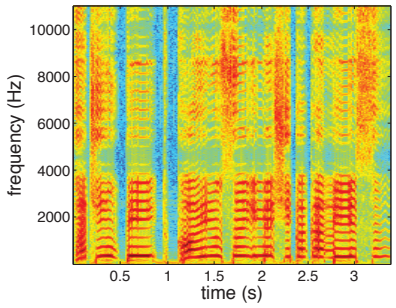
The authors would like to thank the Brazilian sponsors of this work *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) and *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*(FAPERJ) for their support.



(a)



(b)



(c)

Figure 7: Spectrogram of: (a) Original signal; (b) Modified signal with lower pitch, $\beta = 2$; (c) Modified signal with higher pitch, $\beta = 0.5$.

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

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

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293 – 309, February 1967.

[14] A. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, April 2002.

[15] 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.

[16] 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.