Pitch Modification of Speech Signal Using Source Filter Model by Linear Prediction for Prosodic Transformations

Ykhlef Fayçal[1], Mhania Guerti[2], Mesaoud Bensebti[3]
[1]*Electronic Department, University of Saad Dahleb*
[2]*Electronic Department, National Polytechnic School of Algiers*
[3]*Electronic Department, University of Saad Dahleb*
[1]*ykhlef_faycal@yahoo.fr*

Abstract

This paper proposes a pitch modification technique based on the use of Source Filter Model (SFM) by Linear Prediction. A complete solution for pitch modification is described, included estimation of model parameters, pitch detection, Unvoiced/Voiced (U/V) classification and speech synthesis at several pitch values. An automatic U/V classification based on the use of the Zero Crossing Rate (ZCR) and the Energy computation is proposed. The satisfactory performance of this technique is evaluated by listening tests using sentences in Arabic Language. The limit range of pitch modification for natural synthesizing is found.

1. Introduction

Make a synthesis of high quality requires an accurate check of vocal quality parameters which depend mainly on the voicing source. The acoustic parameters which carry prosodic information (pitch, duration, energy) must be adjusted and evaluated comparing the perception capacities [1]. An effective method of speech synthesis consists of selecting a class of basic acoustic units, recording them in natural voice and generating utterances by concatenating appropriately modified segments from the inventory of stored units [2]. This method is called concatenative synthesis. Systems based on concatenative speech synthesis are being used in many speech technology applications [3]. For example, personification of Text-To-Speech synthesis systems and preservation of speaker characteristics. One problem of segments concatenation is that it doesn't generalize well to contexts not included in the training process, partly because prosodic variability is very large. There are techniques that allow us to modify the prosody of a unit to match the target prosody. These prosody-modification techniques degrade the quality of the synthetic speech, though the benefits are often greater than the distortion introduced by using them because of the added flexibility [4]. The objective of prosodic modification is to change the amplitude, duration and pitch of a speech segment. Amplitude modification can be easily accomplished by direct multiplication, but duration and pitch changes are not so straightforward.

One category of prosodic modification techniques process directly on the time domain of the speech signal, like the TDPSOLA algorithm [5]. The largest problem in concatenative synthesis using these techniques occurs because of spectral discontinuities due to different pitch values at the unit boundaries.

Other categories process by analysing the vocal signal in order to separate the glottal information from the vocal tract one [6]. This representation is called the Source Filter Model (SFM). The use of SFM allows us to modify the source and filter separately and thus maintain more control over the resulting synthesized signal. The prosodic parameters can be controlled using only the glottal information. This kind of techniques will significantly reduce the problem of spectral discontinuities at unit boundaries [4]. This paper describes a procedure of pitch modification that leads to clear sounding synthesized speech using in our case the Linear Prediction model. At first, we describe the SFM principal and it use in pitch modification. Then, we explain the procedure of speech modelling by Linear Prediction. Afterwards, we present the techniques used in pitch detection and voice classifications. The experiments and results for the analysis and synthesis at several pitch values are presented in details. We complete this paper by a conclusion to our work.

2. Source Filter model

The SFM describes the vocal signal as a time domain glottal signal (with a flat spectral envelope) plus a vocal tract filter containing the information about formants. The source is assumed to carry the

glottal information and the filter represents the vocal tract.

2.1. Pitch modification using the model

We assume that the glottal excitation signal is approximately a pulse train in the case of Voiced speech (quasi periodicity) and a white noise in the case of Unvoiced one. So the spectral envelope of the glottal signal is flat. The shape of the voice spectrum depends only on the vocal tract that acts as a filter which frequency response is the spectral envelope of the output signal [6]. Consequently, if we can design an equivalent filter, we will have a model of the vocal tract. If we keep the vocal tract model as it is and alter only the glottal signal, we will not modify the spectral shape of the voice as long as we preserve the flatness of the glottal spectrum. The pitch modification operation can be done in this case by changing the periodicity of the source signal in the case of Voiced speech (without altering formants witch represent the vocal tract information).

2.2. The Linear Prediction

Since the term Linear Prediction was first coined by N. Wiener in 1966 [7], the technique has become popularly employed in a wide rang of applications.

This technique first used for speech analysis and synthesis by Itakura and Saito and Atal and Schroeder in 1968 [7], has produced a very large impact on every aspect of speech research. The importance of Linear Prediction steams from the fact that the speech wave and spectrum characteristics can be efficiently and precisely represented using a very small number of parameters. Additionally, these parameters are obtained by relatively simple calculations. The Linear Prediction is described as a system identification problem, where the parameters of an AutoRegressive (AR) model are estimated from the signal itself. A stationary signal $x(n)$ will be known as AR if it obeys to the following model, also known as AR [8]:

$$x(n) + a_1 x(n-1) + \ldots + a_M x(n-M) = e(n) \qquad (1)$$

where $e(n)$ is a centred white Gaussian noise, of variance J. The coefficients a_i will be known as predictors and the constant M as the prediction order. The relation (1) can be interpreted by the Z transform as follow:

$$X(z) = \frac{1}{A(z)} E(z) \qquad (2)$$

with

$$A(z) = 1 + a_1 z^{-1} + \ldots + a_M z^{-M} \qquad (3)$$

We describe the signal $x(n)$ as an output of the filter with Transfer Function (TF) $1/A(z)$, i.e. of a recursive filter or all pole system. The input of this system is the white noise $e(n)$ that is also interpreted as a prediction error. What justifies seeking the optimal coefficients a_i by minimizing this error or more exactly by minimizing it variance J [9]:

$$J = E\left[e(n)^2\right] = E\left[\sum_{i=0}^{M} a_i x(n-i) \sum_{j=0}^{M} a_j x(n-j)\right] \qquad (4)$$

with $E[.]$ denote the expectation operator.

The minimization of this variance (using autocorrelation method) leads us to the normal equations:

$$\begin{bmatrix} R(0) & R(1) & \cdots & R(M) \\ R(1) & R(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & R(1) \\ R(M) & \cdots & R(1) & R(0) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_M \end{bmatrix} = \begin{bmatrix} J_{\min} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (5)$$

where $R(i)$ represents the autocorrelation function of the signal x_n and $J_{\min}$ the minimized value of the variance error prediction. These equations can be solved using several algorithms. In our case, we have used the Levinson–Durbin algorithm which is suitable for practical implementation of Linear Prediction analysis. The Levinson–Durbin approach finds the solution to the m^{th} order predictor from that of the $(m-1)^{\text{th}}$ order predictor [10].

Table 1: Levinson–Durbin algorithm [10]

- ***Initialisation*** : $l = 0$, ***set*** $J_0 = R[0]$

- *Recursion : for* $l = 1,2,\ldots,M$

Step 1. *Compute* l^{th} *RC*

$$k_l = \frac{1}{J_{l-1}} \left(R[l] + \sum_{i=1}^{l-1} a_i^{(l-1)} R[l-i] \right).$$

Step 2. *Calculate LPCs for the* l^{th} *order predictor:*

$$a_l^{(l)} = -k_l,$$

$$a_i^{(l)} = a_i^{(l-1)} - k_l a_{l-i}^{(l-1)}; \quad i = 1,2,\ldots,l-1.$$

Stop if $l = M$.

Step 3. *Compute the minimum mean-squared prediction error associated with the* l^{th} *-order solution*

$$J_l = J_{l-1}(1 - k_l^2).$$

Set $l \leftarrow l+1$ *; return to Step 1.*

The final LPCs are :

$$a_i = a_i^{(M)}; \quad i = 1,2,\ldots,M.$$

It is an iterative recursive process where the solution of the zero order predictor is first found, which is then used to find the solution of the first order predictor. This process is repeated one step at a time until the desired order is reached (Table 1). Inputs to the algorithm are the autocorrelation coefficients $R[l]$, with the LPCs and RCs (Reflexion Coefficients) k_l the outputs.

2.3. Source filter by Linear Prediction

In what precedes, the signal $e(n)$ was considered as an error prediction which we minimized the variance to calculate the Linear Prediction coefficients a_i. The first representation shows that we can rebuild the residue $e(n)$ of the estimation starting from the signal $x(n)$ using a non- recursive filter represented by the TF $A(z)$. Conversely and is the second representation, it will be noted that the residue $e(n)$ can be considered as an excitation signal that is being used to create the signal $x(n)$ using a recursive all poles filter $1/A(z)$. In the case of a speech, the excitation signal can be periodic (Voiced sounds) or random (Unvoiced sounds) [1]. The model generally adopted to create sounds artificially is crude compared to the complexity of the phonate system but it is completely satisfactory for our need (Figure 1). This model includes:

- a periodic generator of impulses units;
- a random numbers generator with null average value and a unit variance;
- a switch being used to choose the Voiced or Unvoiced sounds;
- a factor σ called gain model, it is selected equal to the minimized value of the variance error prediction;
- an all poles filter.

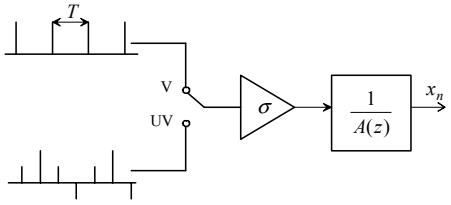


Figure 1: Vocal tract model

2.4. Pitch modification using Linear Prediction

The pitch modification of the segments stored in LPCs form is a straightforward operation. The period T_0 of the vocal signal being a parameter of the model, it is enough to impose directly the synthesis value T_0, what causes to modify the period of the impulses periodic train used in excitation of the Voiced zones [1]. If we want that this operation doesn't affect the power of the signal, it is necessary to modify simultaneously the value of the gain factor σ starting from its initial value σ_0 (corresponding to T_0).

$$\sigma = \sigma_0 \sqrt{T_0 / T} \tag{6}$$

3. Pitch detection

The pitch estimation plays a very important role in speech compression, speech coding, speech recognition and synthesis, as well as in voice modification. A good estimation of the pitch period is crucial to improving the performance of speech analysis and synthesis systems [11]. Many algorithms had been reported in the literature for pitch detection which can be classified in tree main categories:

- Time domain: such as traditional auto-correlation, Average Magnitude Difference Function (AMDF) [12], and DAta ReDuction method (DARD) [13],
- Frequency domain methods: such as CEpstral Pitch detector CEP [14].
- Hybrid method: which incorporates features of both Time-domain and Frequency-domain approaches such as the Simplified Inverse Filtering Technique (SIFT) [15] and Spectral equalization LPC method using Newton's transformation [16]. Each technique present advantages and inconveniences which are due to several problems quoted at reference [16]. In our work, since we use the autocorrelation function to compute LPC parameters, we have preferred to estimate the pitch from the residual $e(n)$ of the estimation using the traditional autocorrelation technique. This approach enables us to avoid the vocal tract interaction where the measurement (the residual excitation) contains only information on the excitation source (pitch). In order to separate the components of the speech signal, we need to classify each speech segment into two categories: Voiced or Unvoiced for both pitch detection and the chose of the LPC model generators used in excitation. This decision is based on the measurement of the Zero Crossing Rate (ZCR) and the Energy of the speech segments.

The decision rule for U/V classification is based on the following criteria:

- If the Energy (equ. 7) is smaller than an Energy threshold then the segment is Unvoiced otherwise we compute the ZCR (equ. 8).

$$E = \frac{1}{N} \sum_{n=0}^{N-1} x_n{}^2 \tag{7}$$

$$ZCR_i = ZCR_{i-1} + \frac{1}{2}\left|(sign(x(i)) - sign(x(i-1))\right| \quad (8)$$

where x denotes N samples of a speech signal.

The function sign(x) returns 1 if the element is greater than zero, 0 if it equals zero and -1 if it is less than zero. The initial value of ZCR is null.

- If the ZCR is larger then a ZCR threshold then the segment is Unvoiced, otherwise the segment is decided to be Voiced.

4. Experiments and results

In this section, we will establish a speech signal model of analysis and synthesis for pitch modification based on the following steps:

- acquisition of the signal;
- emphasis of the speech signal using a transmittance filter $(1-0.95z^{-1})$;
- blocks constitution of "N" samples with shifts of "L" samples;
- U/V classification and pitch detection of Voiced segments;
- estimation of the model parameters for each segment;
- synthesis by Add and Overlap using Hamming window.

Generally, we estimate that the TF of the model must comprise a pair of poles by KHz of band-width; the glottal excitation and the radiation of the lips require together 3 or 4 poles. Then if the sampling frequency F_e is equal to 10 KHz, we choose M equal to 13 or 14 [9]. In our case, F_e is chosen equal to 16 kHz and the order M to 19 (16 for the TF and 3 for the glottal excitation and radiations with the lips). The duration of the signal segments depend on the selected method and conditions under which it is applied. The practice shows that the window must encroach over several fundamental periods for the Voiced sounds.

We usually use windows of 30 ms shifted by 20 ms hence at this duration the speech signal is considered as a stationary signal. For F_e equal to 16 KHz, the analysis segments are setup to 480 samples shifted by 320. The same chose is remained for Unvoiced segments. The following phonetically representative sentences, stored on a disk, were used as samples to produce synthesized speech with different pith values:

1. "The seventh course" الدرس السابع
2. "The paradise Family" أهل الجنة
3. "The Arabic press" الصحافة العربية

These sentences were spoken by adult male speaker, in Arabic language and sampled at 16 KHz.

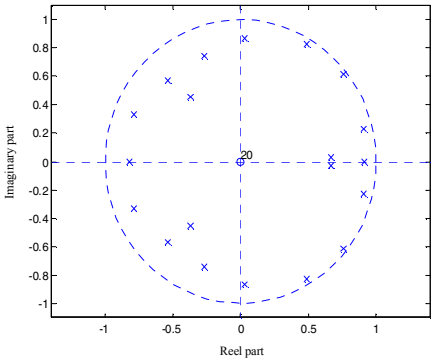


Figure 2: TF Poles of the phoneme [a] over 20 ms with AR order M=20 from the context ''الدرس'' of the first sentence

The LPCs Coefficients represent a z^{-1} polynomial which is not other than the TF denominator of the vocal tract. As this last is essentially stable, the poles must be inside the unit circle (Figure 2).

As we have explained in section 3, our pitch detection method is based on the use of the autocorrelation function from the residual $e(n)$ of the estimation. The noise effects problems can be attenuated by using a low pass filter.

We have used an 8^{th} order Butterworth low pass filter with a cut-off frequency of approximately 800 Hz. The Energy and the ZCR thresholds used in the U/V classification are given for each input signal starting from practical tests which give better results.

In order to automate the system, we have proposed a technique to determine these thresholds. We start first by establishing a row consisted of various calculated Energies of all input signal segments by an ascending order. Then, among these various successions of Energy, we determine the one which represents the 25^{th} percent of the row indices (positions of the Energy row), which indicates finally the desired threshold Energy.

The ZCR threshold is calculated in the same way by determining in this time the rate which represents the 75^{th} percent of the row indices (positions of the ZCR row) (Figure 3).

The decision rule for U/V thresholds is based on the principle that about 75% of Arabic speech sounds are Voiced and remain 25% Unvoiced sounds [1]. The pitch evaluation of the input signal (the first sentence ''السابع الدرس'') according to the analysis blocks is represented in the Figure 4

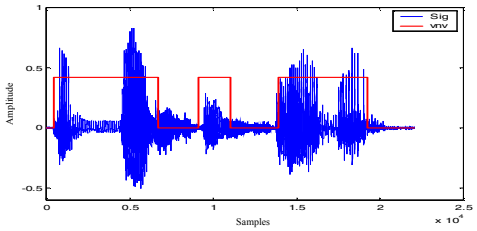


Figure 3: U/V classification of the sentence "الدرس السابع"

The synthesized phoneme [a] and the original one taken from the context ''الدرس'' at a duration of 10 ms are represented in the Figure 5, where sig - synth denote the synthesized signal and sig - org the original one. The delay observed is due to the filtering operations.

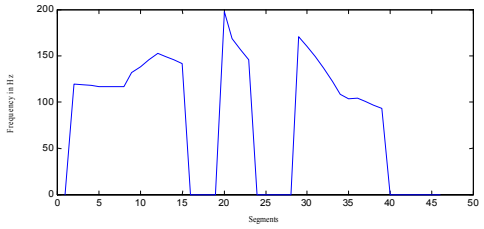


Figure 4: Pitch evaluation of the sentence "الدرس السابع"

For obtaining a synthetic signal of different pitch values, we have to change simultaneously the periodic generator of impulses units used as a source excitation of Voiced sounds and the gain factor given by equ. 6.

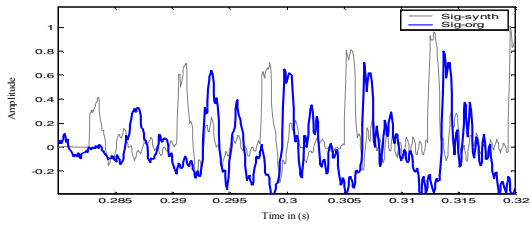
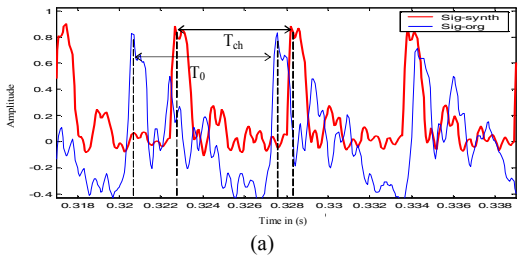


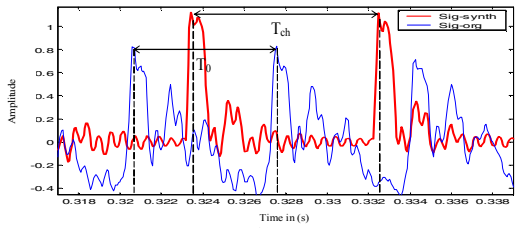
Figure 5: Temporal representation of the original and synthesized phoneme [a]

The Figure 6 shows the original phoneme [a] taken from the context "الدرس" and the shifted version synthesized using modifications factors of 0.8 and 1.3.

The legends sig - org and sig - synth denote respectively the original signal and shifted one. Also, the legends T_0 and T_{ch} denote respectively the original and the shifted pitch periods using previously modifications factors. The practical tests show that using the SFM by Linear Prediction, the synthesized signal suffer from amplitude distortion even though we have used the gain corrector. Furthermore, the natural of the synthesized signal is not well reproduced essentially for nasals and transistorises sounds. These problems are basically due to the AR model and sources errors and also the accuracy of the pitch detection and classification algorithms. In fact, there are glottal excitations models more sophisticated such as those of Rosenberg or Liljencrants-Fant who approximate more the shape of the excitation signal [17]. The U/V classification used in this work does not consider mixed sounds (like Voiced Fricatives). The use of a mixed excitation returns to the synthesized signal a certain roundness which has been missed in the U/V binary model classification. It is very difficult to improve the quality or the natural of synthesized signal by increasing the sampling frequency or the model order. The advantage of this technique is to change the pitch of the speech signal without altering the spectral envelope which preserves the voice timbre as well as possible (Figure 7). The legends DSP0p8 and DSP1p3 denote respectively the spectral envelope power (in dB) of the synthesized signal using 0.8 and 1.3 factors modifications and DSP1 the spectral envelope power of the synthesized signal without changing the pitch value.



(a)



(b)

Figure 6: Pitch modification of the phon. [a]
(a) Down modification (factor=0.8)
(b) Up modification (factor=1.3)

Furthermore, the spectral discontinuities due to different pitch values at the unit boundaries are well diminished by the SFM. The thresholds of the modifications factors reached by any technique of pitch modification are theoretically included between two ends, one maximal and other minimal. To find these limits, we have used the sentences given at the top, spoken by adult male speaker, in Arabic language to evaluate the capability of this technique.

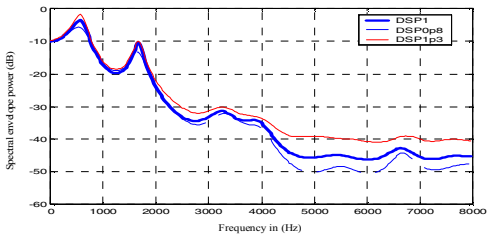


Figure 7: Spectral envelopes of the interpolated and the synthetic phoneme [a] on a 30 ms interval

After listening tests, we have noticed that ones we are out of a minimum value of 0.4 and a maximum one of 2 (modifications factors) the vocal signal obtained by this technique is becoming increasingly robot-like from where a loss of the naturalness is appearing.

5. Conclusion

The performance of the Source Filter Model by Linear Prediction designed for pitch modification of speech signal was investigated by synthesizing sentences in Arabic Language. We have proposed in this work a complete solution included the algorithms used for pitch detection, U/V classification, AR parameters estimations and synthesizing using Add and Overlap method. The pitch modification operation using this technique is well done by preserving the spectral envelope and the timbre of the speech signal. The synthesized sentences suffer from synthesizing errors due essentially to AR modelling, pitch detection, U/V classifications and the excitations sources. The optimization of all criterions is a very difficult operation. We can conclude that the pitch modification using SFM by Linear Prediction technique is an elementary operation. The complexity of this technique is located in the analysis approach. The use of combined pitch modification method between Source Filter Model and time domain technique will improve the natural speech synthesizing.

10. References

[1] F. Ykhlef, Pitch modification for Text To Speech synthesis of Arabic Language (French Lang.). Master These, Saad Dahleb University, Blida, Algeria, September 2005.

[2] R. Ansari, D. Kahn, M. Macchi, "Pitch modification of speech using a low-sensitivity inverse filter approach," *IEEE Signal Processing Letters*, vol. 5, no. 3, pp. 60-62, March 1998.

[3] A. Syrdal, R. Bennett, and S. Greenspan, *Applied Speech Technology.* Boca Raton, FL: CRC, 1995.

[4] Xuedong Huang, Alex Acero and Hsiao-Wuen Hon, Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice-Hall Inc., 2001.

[5] X. Sun, "Voice quality conversion in TD-PSOLA speech synthesis," *IEEE Proc. of the Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 953-956, 2000.

[6] Patrick Bastien, Pitch Shifting & Voice Transformation. VoicePro Whitepapers, tc- helicon vocal technologies, reachable from www.tc-helicon.com/.

[7] Sadaoki Furui, *Digital Speech Processing, Synthesis, and Recognition* .Second Edition (Revised and Expanded), Marcel Dekker, New York, 2001.

[8] Calliope, Speech and it Automatic Processing (French Lang.). Ed. J.P. Tubach, Masson, Paris, 1989.

[9] R. Boite, H. Bourlard, T. Dutoit, J. Hancq et H. Leich, *Speech Processing (French Lang.)*, Polytechnic Press and Romandes University, Electrical Collection, 2000.

[10] W. C. Chu, Speech Coding Algorithms: Foundation and Evolution of Standardized Coders, Hoboken, NJ: Wiley Interscience, 2003.

[11] Li Hui, Bei-Qian Dai, Lu Wie, "Pitch detection algorithm based on AMDF and ACF," *IEEE Proc. of the Acoustics, Speech, and Signal Processing*, Vol. 1, pp. I-I, 2006.

[12] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-22, pp. 353-362, Oct. 1974.

[13] N. J. Miller, "Pitch detection by data reduction," IEEE Trans. Acoust., Speech, Signal Processing (Special Issue on IEEE Symposium on Speech Recognition), vol. ASSP-23, pp. 72-79, Feb. 1975.

[14] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. SOC. Amer.*, vol. 41, pp. 293-309, Feb. 1967.

[15] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367-377, Dec. 1972.

[16] Rabiner, L.Cheng, M, Rosenberg, A, McGonegal, C, "A comparative performance study of several pitch detection algorithms," *Acoustics, Speech, and Signal Processing (see also IEEE Transactions on Signal Processing), IEEE Transactions on*, Vol. 24, Issue 5, pp. 399- 418, Oct. 1976.

[17] C. d'Alessandro, *Lectures in Automatic Speech Processing (French Lang.),* Wavelets Summer School, LIMSI-CNRS, Toulouse, France, 1993.