# A Wavelet-Based Technique Towards a More Natural Sounding Synthesized Speech

*Mehmet Ataş, Süleyman Baykut, Tayfun Akgül*

Department of Electronics and Communications Engineering
Istanbul Technical University, Istanbul, Turkey
atasm@itu.edu.tr, baykut@itu.edu.tr, tayfun.akgul@itu.edu.tr

## Abstract

This paper presents a wavelet-based technique to increase the quality and naturalness of LPC based synthesized speech signals. The proposed method is based on wavelet decomposition. We first obtain the wavelet coefficients, and then the variances of the wavelet coefficient at the last four scales (correspond the higher frequency region) of the synthetic speech are replaced by the original variances of the original speech. We apply the technique to synthetic speech. The results suggest that the wavelet-based technique increases the naturalness of the synthesized speech.

## 1. Introduction

Speech coding based on Linear Predictive Coding (LPC) is a successful and very commonly used method for years [1] and has found many application fields from mobile phones to voice mails [2, 3].

The common problem in the LPC based speech coding is to obtain more realistic speech synthesis at the receiver part. The synthesized speech segments may appear as metallic and buzz like due to the relatively smooth synthetic excitation signals and the use of insufficient number of filter coefficients. When we compare the metallic sounding speech with the natural sounding one, the richness and the naturalness are found to be detailed in the high frequency region of the voiced speech, i.e., the power of the high frequency component of natural speech is mostly higher than the metallic one.

In this study, we use wavelet decomposition to rearrange the energy of the high-frequency components to have more realistic synthesized speech.

First, as commonly used in speech coding, the LP coefficients, voiced/unvoiced decision, the gain and the pitch period (if voiced) is extracted and then they are used to synthesize the speech (at the receiver.) In our proposed method, additionally, variances of the wavelet coefficient at the last four scales (correspond to high frequency region) are extracted. These values are used for adjustment of the wavelet coefficients of the synthesized data yielding rich high frequency components. Experiments with real speech data show that the wavelet-based procedure increases the naturalness of the synthesized speech information.

This paper is organized as follows. In Section 2 we give brief background information on LPC analysis, plus wavelet-based analysis and synthesis. In Section 3, the proposed method is explained. The results are given in Section 4 before the conclusion.

## 2. Background

In this section LPC analysis and the wavelet–based analysis of speech are briefly explained.

### 2.1. LPC Analysis

The speech signals, *s(n)*, are assumed to be generated by excitation of a linear filter by a residual source, *r(n)*, as [1]:

$$s(n) = r(n) * h(n) \tag{1}$$

Here *h(n)* is the impulse response of the linear filter that models the vocal tract, * denotes convolution. The filtering procedure in z-domain can be given as:

$$S(z) = R(z)H(z) \tag{2}$$

where *H(z)* is the vocal tract transfer function which is mostly an all-pole model:

$$H(z) = \frac{G}{1 + A(z)} \tag{3}$$

$$A(z) = a_1 z^{-1} + a_2 z^{-2} + \ldots + a_p z^{-p} \tag{4}$$

where $G$ is the gain and $a_k$ are the LP coefficients. Then these coefficients are used for constructing the filter and the inverse filter that model the vocal tract.

### 2.2. Wavelet-based Analysis of Speech

Wavelet transform is an effective tool used in many signal processing applications. In this study we use the wavelet transform to obtain the variances of the wavelet coefficients which reveal the energy levels at different frequency regions. Orthonormal discrete dyadic wavelet transform (DWT) pair is given below [4]:

$$x(t) = \sum_m \sum_n x_n^m \psi_n^m(t) \tag{5}$$

$$x_n^m = \int x(t) \psi_n^m(t) dt \tag{6}$$

Here $x_n^m$ are the wavelet coefficients, $\psi_n^m(t)$ is the normalized dilations and translations of the mother wavelet function $\psi(t)$, $m$ and $n$ are the dilation (scale) and translation indices, respectively. After obtaining wavelet coefficients along scales we calculate the variances of the wavelet coefficients which will be used in the wavelet-based high frequency adjustment technique.

## 3. Proposed Method

The block diagram of the speech analysis/synthesis method is shown in Fig. 1. The speech segments are coded and a code vector is formed in the analysis part and it is used by the synthesis part. After generating the synthesized speech, the richness of this speech is increased by changing the variances of the coefficients of the last four scales of the wavelet transform.

The analysis and the synthesis procedures are summarized below.

### 3.1. Analysis Part

The analysis part has the following steps:

#### *3.1.1. Windowing and Pre-Emphasizing*

The sampling frequency of the speech signals used in this study is 16 kHz. The speech signal is segmented into 32ms windows so that the windowed data is now assumed to be stationary.

#### *3.1.2. LP Coefficients*

LP analysis is applied to the speech segments and the LP coefficients are stored. In this study the order of the LP analysis is chosen as 16.

#### *3.1.3. Wavelet-Based Analysis*

The wavelet coefficients of the speech segments are obtained by equation (6) and the variances are calculated. Variances of the coefficients at the last four scales are stored for the use of the synthesis part.

In our labeling scheme, the higher scale coefficients correspond to the higher frequency regions. Therefore, we additionally store the variances of the last four scales which correspond the $\pi/8 - \pi$ [radian] frequency band for the use of synthesis part.

#### *3.1.4. Inverse Filtering*

Inverse of the vocal tract filter is modeled by using the LP coefficients and the inverse filtering is performed to extract the residual signal. The variance of the residual also stored for the synthesis part.

#### *3.1.5. Voiced/Unvoiced Detection*

Speech signals can be classified in two basic characteristics: Voiced and unvoiced. Voiced speech signals show a pseudo-periodic structure whereas unvoiced speech signals are white noise-like. The other main difference between the voiced and unvoiced speech is the short-time energy. Voiced speech signals have higher energy than unvoiced speech signals. On the other hand, zero-crossing number of unvoiced speech signals is approximately 10 times higher than voiced speech signals [3]. In this study the voiced/unvoiced decision is made according to these criteria.

#### *3.1.6. Pitch Period Estimation*

If the corresponding speech segment is labeled as voiced, the pitch period of the speech is determined by cepstral pitch detection method [2].

#### *3.1.7. Forming the Code Vector*

The code vector contains the information to synthesize the speech. LP coefficients are the most important components of the code vector. Other components of the code vector are voiced/unvoiced decision, pitch period, the variance of the speech segment, and the variances of the wavelet coefficients at the last four scales. These values are then use to adjust the high frequency components in the synthesized speech.

### 3.2. Synthesis Section

The synthesis of speech signals are presented in this section.

#### *3.2.1. Unvoiced Speech Synthesis*

The vocal tract filter is modeled by using the LP coefficients. Then the filter is excited by white Gaussian noise with the same variance carried in the code vector. For voiced speech segment, the filter is excited by a periodic residual signal. This signal represents the glottal flow.

#### ***3.2.1.1 Glottal Pulse Model***

Glottal pulse shape selection is one of the most important aspects of speech synthesis. There are many types of models used to model glottal flow. In this study, commonly used Liljencrants-Fant (LF) model is used to represent the voice source [5].

### 3.3. High Frequency Adjustment in the Wavelet Domain

The synthesized speech sounds as metallic and buzz like due to the relatively smooth synthetic excitation signals [6, 7]. The high frequency components of the synthesized speech have relatively low energy compared to the original speech. In this study we adjust the high frequency components by replacing the variances of the wavelet coefficients at the last four scales with the ones in the code vector. The procedure is shown in Fig. 2. The resulting speech now sounds more natural and rich.

## 4. Simulation on Real Speech Data

We apply the wavelet-based technique to 32 ms long synthesized voiced speech segment (/EE/) sampled at 16 kHz. The original, the synthesized and the adjusted high frequency components speech segments are given in Fig. 3- (a), (b) and (c), respectively. Even the visual inspection of Fig. 3-(b) and (c) show that Fig. 3-(c) has more detail (high frequency component) compared to Fig. 3-(b).

We apply the high frequency adjustment technique to longer speech signal. The signal is segmented into 32ms windows. After the voiced/unvoiced decision, and the pre-emphasizing filtering, the LP coefficients are calculated. If the segment is voiced, pitch period is estimated as well. The variances of the residual signals and the variances of the last four scales of the wavelet coefficients are also determined for every segment and the code vector is formed as explained in Section 3.1. In Fig. 4 - (a), the waveform of the Turkish sentence “Sayısal İşaret İşleme” which means “Digital Signal Processing” in English, is given. The sentence is articulated by an adult female speaker. The synthesized speech is given in Fig. 4 - (b) and the speech after our proposed method is applied is given in Fig. 4 - (c). These examples will be made available to listen for the audience during the presentation of this paper in the conference.

## 5. Conclusions

LPC based speech synthesis makes speech communication possible at low bit rates. However, a problem of metallic and buzz like speech is faced due to the relatively smooth synthetic excitation signals. In this study, the high frequency components of the synthesized speech is adjusted and enhanced by a wavelet-based technique to improve the naturalness of the synthesized speech. The results show that the technique gives promising results.

## 6. References

[1] Atal B. S., Hanauer S., “Speech Analysis and Synthesis by Linear Prediction of the Speech Wave,” *The Journal of the Acoustical Society of America*, vol.50, pp. 637-655, April 1971.

[2] Deller, J. R., Hansen, J. H. I., Proakis, J. G., “Discrete-Time Processing of Speech Signals,” *IEEE Press*, New York, NY, 2000

[3] Quateri, T., F., **“**Discrete–Time Speech Signal Processing Principles and Practice,” *Prentice Hall Inc.*, 2002.

[4] G. W. Wornell, “Wavelet-Based Representations for the *1/f* Family of Fractal Processes,” *Proc. of IEEE*, vol. 81, no. 10, pp. 1428-1450, Oct. 1993.

[5] Fant, G. and Lin, Q., “A Four Parameter Model of Glottal Flow,” *STL-QPSR*, 85(2), 1-13., 1988.

[6] Lee, Y., **“**Speech Quality Enhancement by Exploring 1/*f* Nature of Speech Residual,” *MS Thesis*, Drexel University, PA, 1998.

[7] Aoki, N., Ifukube, T., “Enhancing the Naturalness of Synthesized Speech by Using the Random Fractalness of Vowel Source Signals,” *Electronics and Communications in Japan,* vol. 84, no. 1*, pp.* 11-20., 2001.

## 7. Figures

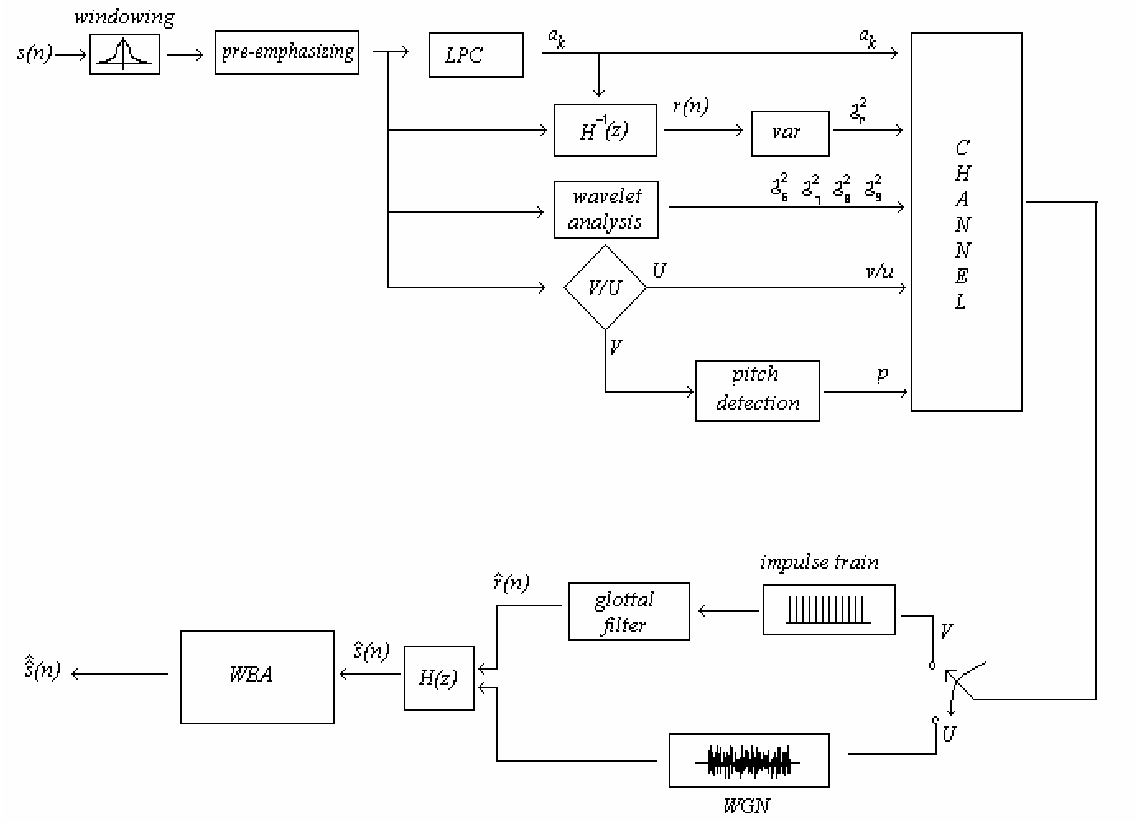

**Figure 1:** Block Diagram of the Speech Analysis/Synthesis Procedure.

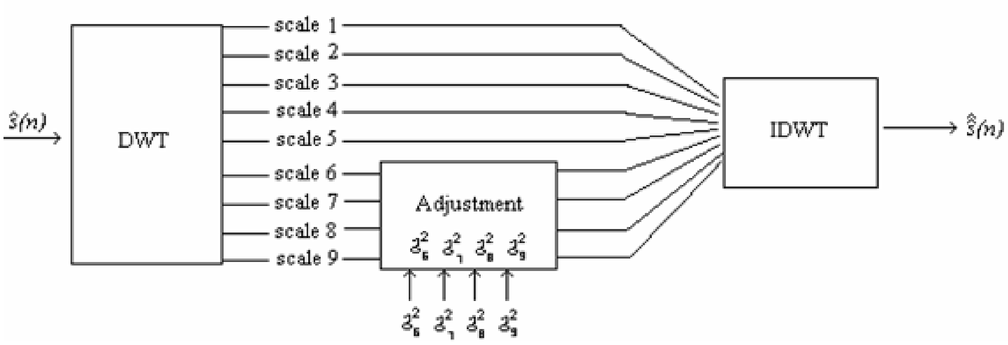
$\hat{s}(n)$
DWT
scale 1
scale 2
scale 3
scale 4
scale 5
scale 6
scale 7
scale 8
scale 9
Adjustment
$\hat{\sigma}_6^2$ $\hat{\sigma}_7^2$ $\hat{\sigma}_8^2$ $\hat{\sigma}_9^2$
$\sigma_6^2$ $\sigma_7^2$ $\sigma_8^2$ $\sigma_9^2$
IDWT
$\hat{\hat{s}}(n)$

**Figure 2:** Wavelet-based high frequency component adjustment.

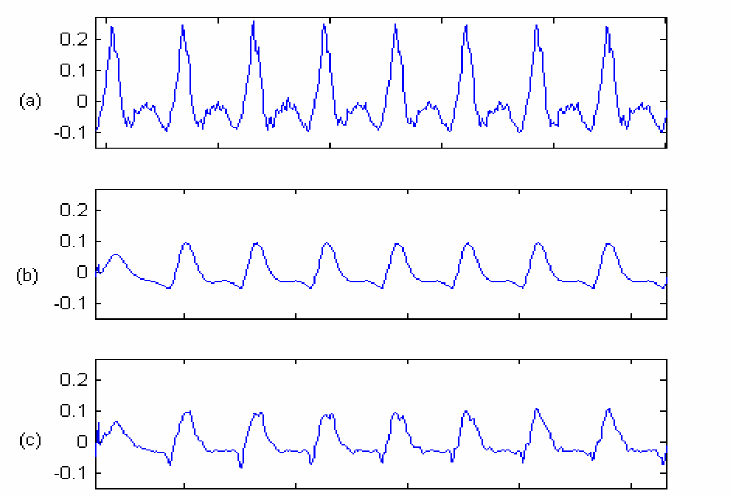

**Figure 3:** (a) A voiced speech segment /EE/ from the original speech data, (b) synthesized speech segment, (c) synthesized speech segment with the adjusted high frequency components.

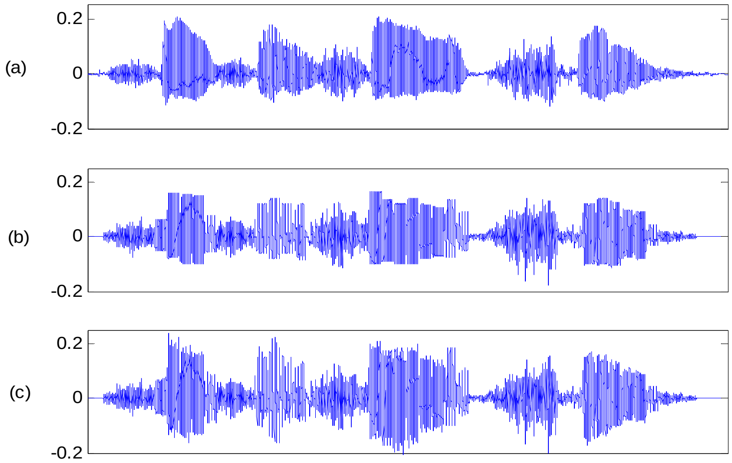

**Figure 4:** (a) The original speech waveform of “Sayısal İşaret İşleme” (in Turkish), (b) the synthesized speech, (c) resulting speech after the high frequency adjustment process.