Application of Wavelet LPC Excitation Model for Speech Compression

[1,2]Armein Z.R. Langi

[1]Research Center on Information and Communication Technology
[2]Information Technology RG, School of Electrical Engineering and Informatics
Institut Teknologi Bandung, Jalan Ganeca 10, Bandung, 40116, Indonesia

Abstract. This paper presents an application of linear predictive coding (LPC) excitation wavelet models for low bit- rate, high-quality speech compression. The compression scheme exploits the model properties, especially magnitude dependent sensitivity, scale dependent sensitivity, and limited frame length. We use the wavelet model in an open-loop dither based codebook scheme. With this approach, the compression yields a signal-to-noise ratio of at least 11 dB at rates of 5 kbit/s and.

Keywords: *LPC; LPC excitation; modelling; speech compression; speech processing; wavelets.*

1 Introduction

Speech signal compression is a necessity in speech communication either because of an operational requirement based on a design constraint, or because of the desire to utilize existing resources efficiently [1]. In a pulse code modulation (PCM) form, real-time telephone-quality speech requires a rate of 64 kbit/s, which is too high for high frequency (HF) radio or practical network channels. At this rate, speech as short as one minute would also occupy large storage space (480 kbytes).

Techniques based on a simple speech production model have successfully reduced the bit rates to below 8 kbits/s, which can be accommodated by the narrow-band channels. In this model, speech is the result of applying an *excitation* to a *vocal tract*. This model becomes practical through techniques such as *linear predictive coding* (LPC). Here, the vocal tract becomes an adaptive filter $H(z)$ called *LPC filter*. In this case, the excitation is called *LPC excitation*. Thus, by efficiently representing both LPC filter and excitation, one can have speech compression. For example, U.S. Federal Standard (FS) 1016 code-excited linear predictive (CELP) and FS-1015 LPC-10e coders efficiently compress the filter parameters and excitation down to 4.8 and 2.4 kbit/s, respectively [2].

LPC excitation models play a critical role to obtain high speech quality at low bit- rates. At the present time, advanced techniques such as line-spectrum pair (LSP) can successfully code the filter parameters at as low as 0.75 - 1 kbit/s, with average spectral distortion less than 1 dB [3-4]. However, that is not the case for the LPC excitation. As it is, it would require a 64 kbit/s rate. There are different techniques to code the excitation based on different models, with a trade off between the resulting quality and the bit rate. One very efficient model used in the LPC-10e consists of a pitch impulse generator, a random impulse generator, a gain controller, and a voiced/unvoiced (V/UV) switch, resulting in a machine-quality speech. The CELP uses a stochastic codebook and an adaptive codebook, resulting in good speech-quality [5]. Another model uses scalar quantization or centre-clipping in conjunction with a pitch filter, as in adaptive differential PCM (ADPCM). This technique results in high speech quality at bit rates of 16 to 32 kbit/s.

The linear combination of wavelets is an attractive model of LPC excitation for speech compression [6-7]. Such a wavelet model of LPC excitation has been shown to have asymmetrical and nonuniform properties that are attractive for speech compression, namely magnitude dependent sensitivity, scale dependent sensitivity, and limited frame length. This paper proposes new speech compression schemes using that model. The schemes exploit those coefficients' asymmetrical properties. Our specific contributions are (1) an ideal scheme through the use of close-loop codebook searching and perceptually weighted measure, as well as (2) a practical scheme through whitening the effect of the quantization noise. Our experiment shows that even in a simple straight-forward implementation, the model indicates promising capability by having SNR 11.03 and 15.33 dB at 5 and 5.5 kbit/s, respectively.

2 The Wavelet Model of LPC Excitation

In this section, we review the wavelet model of LPC excitation as well as its properties.

2.1 LPC excitation

We can use a segment of speech signal *s*[*n*] to obtain LPC excitation *t*[*n*]. Let an LPC filter *H*(*z*) be [8].

$$H(z) = A^{-1}(z) = \left(\sum_{i=0}^{10} a_i z^{-i} \right)^{-1} \tag{1}$$

In the *z*-domain, the speech segment $S(z)$, LPC excitation $T(z)$, and the LPC filter $H(z)$ are related by

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

A similar relationship can be defined in vector notations. Let the segment of speech be **s**, which is a vector whose elements are $s[0]$, $s[1]$, ..., $s[N–1]$. A *linear prediction* procedure [8] can obtain a_i in Eq. (1) from such **s**. Let h_i be the impulse responses of $H(z)$. We can the represent $H(z)$ in an $N×N$ lower-triangular matrix H [9]:

$$H = \begin{bmatrix} h_0 & 0 & \cdots & 0 \\ h_1 & h_0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ h_{N-1} & h_{N-2} & \cdots & h_0 \end{bmatrix} \tag{3}$$

We can now represent the LPC excitation with a vector **t**, whose elements are $t[0]$, $t[1]$, ..., $t[N–1]$, in which $\mathbf{s} = H\,\mathbf{t}$.

However $H(z)$ is an all-pole filter, containing memory. Hence one $s[n]$ is affected by all $t[m]$, with $m \le n$. Thus, there is an additive contribution of all $t[m]$ from the previous segments to the current **s**, denoted as a vector **u**. Taking this into consideration, Eq. (2) becomes

$$\mathbf{t} = H^{-1}\,(\mathbf{s} - \mathbf{u}) \tag{4}$$

In paractive, this **t** should be modelled and compressed without any consideration of **u**, since a speech production filter in Eq. (2) automatically generates **u**.

2.2 The Wavelet Model

The LPC excitation can be seen as a linear-combination of wavelets. Consider a set of signals which are members of R^N, grouped into two subsets $\{\psi_{j,k}[n]\}$ and $\{\phi_{J,k}[n]\}$. Here, J is any integer between 1 and $\log_2 N$. (In this work, we set J to $(\log_2 N)–1$). Index j is called *scale*, ranging from 1, 2, ..., to J, while k is 0, 1, ..., to $(2^{-j}N)–1$. Signals in both subsets are called *wavelet* and *scaling* signals, respectively. Then, there are real numbers $c_{j,k}$ and $d_{J,k}$, called *wavelet coefficients* and *scaling coefficients* [10] defined as

$$c_{j,k} = \sum_{n=0}^{N-1} t[n]\psi_{j,k}[n] \text{ and } d_{J,k} = \sum_{n=0}^{N-1} t[n]\phi_{J,k}[n] \tag{5}$$

With these coefficients, we can express **t** as a linear combination of wavelets as follows.

$$t[n] = \sum_{j=1}^{J} \sum_{k=0}^{2^{-j}N-1} c_{j,k}\psi_{j,k}[n] + \sum_{k=0}^{2^{-j}N-1} d_{J,k}\phi_{J,k}[n] \qquad (6)$$

Eqs. (5) and (6) also represent forward and inverse DWT of **t**, respectively.

2.3 Properties

Other work has shown that the wavelet coefficients have attractive properties:

1. The high-magnitude coefficients are more important than the low-magnitude ones, thus we can coarsely quantize the low-magnitude coefficients. Furthermore, there are more low-magnitude coefficients than the high-magnitude ones, making the bit-rate even lower.

2. The coefficients in a certain scale are more important than the coefficients in the other scales, thus we can coarsely quantize the coefficients in the other scales. Furthermore, the number of important coefficients is less than that of the other coefficients, making it attractive for lossy compression.

3. What is the best length of frame (*N*) for **t** to use? The frame length must be limited to reduce coding delay and system complexity. In discrete Fourier transform (DFT), the answer to this important question determines the uniform sampling resolution in frequency domain. The longer the frame is, the finer the frequency resolution. However, this is not the case in our model. The optimal *N* is among 32, 64, and 128 points.

3 Proposed Compression Schemes

The model can then be used to build compression schemes. The key is to compress the excitation, which is a collection of wavelet coefficients.

3.1 Compressing the Excitation

Before we derive the descriptions for compression, we simplify the notation by defining vectors $\mathbf{v}_i$ as

$$v_i[n] = \begin{cases} \phi_{J,i}[n]; & i = 0,1,\cdots,2^{-J}N-1 \\ \psi_{j(i),k(i)}[n]; & i = 2^{-J}N, 1+2^{-J}N,\cdots,N-1 \end{cases} \tag{7}$$

where

$$j(i) = -\mathrm{int}\left(\log_2\left(\frac{i}{N}\right)\right); \quad k(i) = i - 2^{-j(i)}N \tag{8}$$

Notice that the function int(•) returns the maximum integer value that does not exceed the argument. Consequently, we can define **c** which satisfies both Eq. (4) and Eq. (6), by assigning ordered values of $\{c_{j,k}, d_{J,k}\}$ as the elements of c_i. Clearly, the order must follow that of the scaling and wavelet functions in the $\mathbf{v}_i$ above. Thus, we have Eq. (6) to be

$$t[n] = \sum_{i=0}^{N-1} c_i v_i[n] \tag{9}$$

Here, the DWT becomes a mapping $\Im : (t \to c)$, and its inverse $\Im^{-1}$ is Eq. (9).

To compress **t**, we usually must *approximate* the set of coefficients c_i with $\hat{c}_i$ which uses fewer bits, shown in Figure 1. First the encoder converts the LPC excitation into wavelet coefficients c_i. It then quantizes c_i into $\hat{c}_i$ and compress it. With this approach, we can have an efficient representation of the excitation.

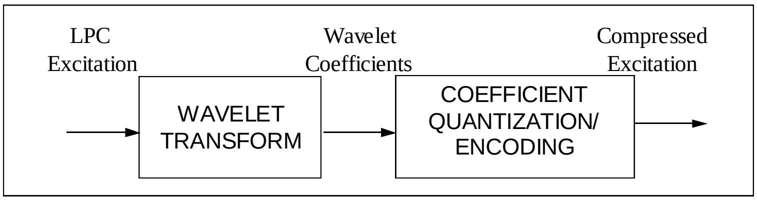

Figure 1 Wavelet encoder.

Figure 2 shows the use of this encoder in a speech compression. The encoder performs LPC analysis on the original speech, resulting in LPC parameters. In inverse LPC filter use the LPC parameters to generate LPC excitation. Wavelet encoder in Figure 1 then produced compressed excitation. Both LPC parameters and compressed excitation are transmitted to a speech decompression.

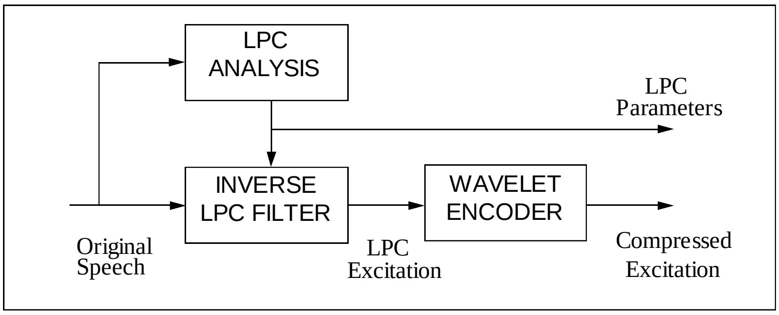

Figure 2 Conceptual speech compressor (analyzer).

In this approach, a speech decompressor would have a scheme as in Figure3. It contains a wavelet decoder. A wavelet decoder first decodes the compressed excitation into the wavelet coefficients. It then inverse transforms the coefficients, resulting in LPC excitation. An LPC filter uses transmitted LPC coefficients to produce reconstructed speech from LPC excitation.

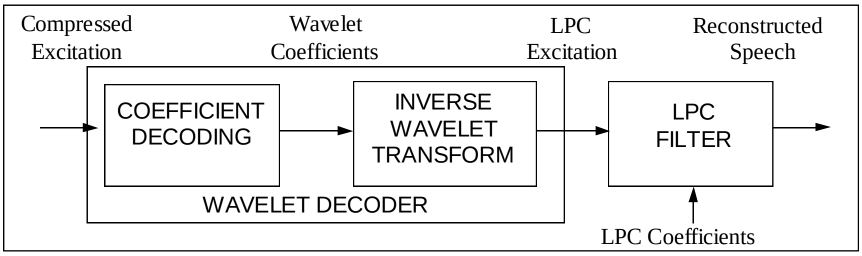

Figure 3 Speech decompressor (synthesizer).

However, this approximation introduces error (distortion) that should be minimized. Notice that in the wavelet decoder, the coefficient set results in excitation $\hat{t}[n]$, which would produce $\hat{s}[n]$ instead of $s[n]$ in Eq. (2). Since the distortion occurs at the excitation, the LPC filter will enhanced the error according to speech magnitudes. In other words, the distortion correlates with the speech. This results in disturbing and unpleasant speech distortion.

To measure the distortion, we can use a Euclidean error measure with a perceptual weighting filter $W(z)$ as used in [5]. The $W(z)$ enhances the Euclidean measure to exploit the masking property in human perception. With the weighting filter is represented by it impulse response matrix W, the error measure becomes:

$$d(s,\hat{s}) = \|W(s-\hat{s})\| = \sqrt{\sum_{i=0}^{N-1}\left(\sum_{j=0}^{N-1} w_{i,j}\left(s[j]-\hat{s}[j]\right)\right)^2} \tag{10}$$

We can reformulate the error measure in Eq. (13) in terms of **c**, as also derived in [11].

$$d(s,\hat{s}) = \|W(s-\hat{s})\| = \left\|WH\left(t-\hat{t}\right)\right\| = \left\|WH\Im^{-1}(c-\hat{c})\right\| \tag{11}$$

If Q be an $N\times N$ matrix whose i-th column is $\mathbf{v}_i$, we immediately have Eq. (9) to be

$$\mathbf{t} = Q\mathbf{c} \tag{12}$$

We can now simplify Eq. (10) by first defining $T[\mathbf{c}]$ as a mapping of **c** as

$$T[\mathbf{c}] = W\,H\,Q\,\mathbf{c} \tag{13}$$

and then rewrite Eq. (10) as

$$d(s,\hat{s}) = \|T[c-\hat{c}]\| \tag{14}$$

Since $T[]$ is linear, the upper-bound of the error is

$$d(s,\hat{s}) \le \|T\|\|c-\hat{c}\| \tag{15}$$

(The norm definition of Eq (10) must be one that is compatible with Euclidean norm of vectors). Clearly we must minimize $\|c-\hat{c}\|$ so that we minimize the upper-bound. However, this rather simple minimization is not sufficient, because T changes with **s**. There are cases where minimizing $\|c-\hat{c}\|$ does not minimize Eq. (14), because $c-\hat{c}$ is not generally an eigenvector of T. Thus, we must focus on minimizing Eq. (14) instead of minimizing $\|c-\hat{c}\|$ alone.

Although we can use the scheme in Figure 6, the distortion correlates to speech. In practice we can improve the quality using two options of quantizing c. First option is a close-loop searching through a set of codebooks. Second option is an open-loop scheme through noise whitening.

3.2 An Ideal Closed-Loop Scheme

Although the quantization is performed for c, this scheme ensures the minimization of $d(s,\hat{s})$ instead of $d(c,\hat{c})$ through a closed loop approach. The scheme uses a set of codebooks to store a limited set of $\hat{c}$ in the place of coefficient decoding (see Figure 7). However, the analyzer becomes a scheme in Figure 4. Here, the compressed excitation is search from the codebooks through the close-loop trials. This closed-loop search uses minimum $d(s,\hat{s})$ as its criterion, ensuring its closeness.

However such quality comes with the expense of very high computing requirements. Searching the codebook through close-loop trials involves inverse transform, LPC filtering as well distance measures for each trial. As a result, such an ideal solution is not practical.

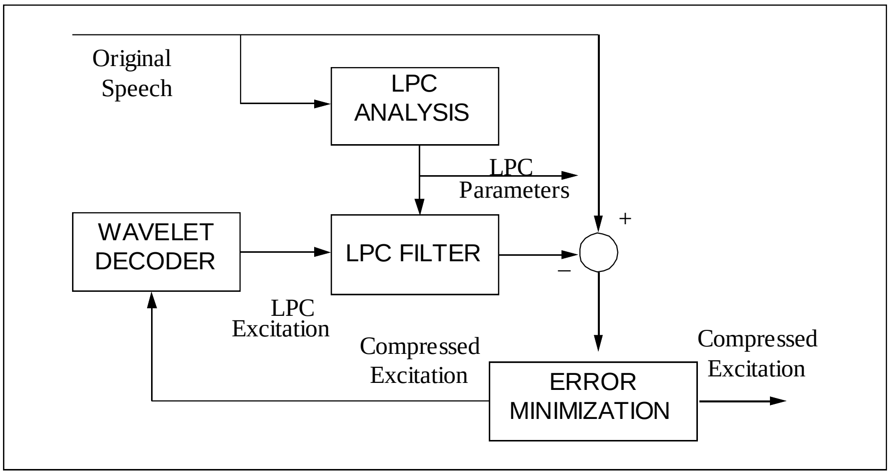

Figure 4 A closed-loop analyzer.

3.3 A Practical Open-Loop Scheme

Alternatively we should look at an open-loop approach but still maintains uncorrelated distortion. The second scheme rearranges the analyzer in Figure 2 to obtain white noise effect (instead of correlated one) of the quantization noise on the resulting speech. If the quantization noise has correlation with the speech, the noise is more perceivable [5]. Although the quantization itself can result in coefficient error that is uncorrelated with the coefficients, the speech error still correlates with speech signal, because the filter H(z) shapes the error spectrum.

To avoid it, we can rearrange the compressor as shown in Figure 5. Assume that the wavelet encoder introduces e, an additive, uncorrelated error of coefficients ($\hat{c} = c + e$). One can easily show (see [8]) that with the redesigned analyzer, e is uncorrelated with s, and

$$\hat{s} = s + e \tag{16}$$

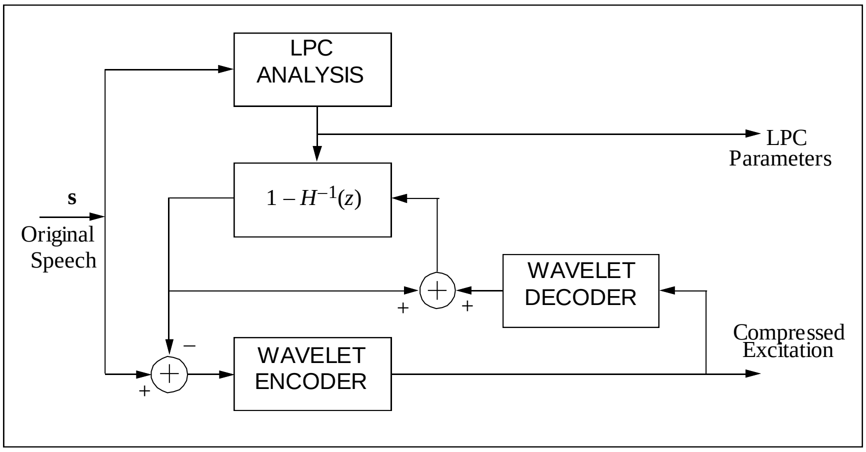

Figure 5 Speech compressor, with white-noise effect on the reconstructed speech.

4 Experimental Results

There are many schemes that can be used to exploit the properties described earlier. In principle, every scheme that uses LPC excitation can adopt the model. Here, we simply use the scheme as in Figure 2 and Figure 3 for our experiment, with slight modifications in the wavelet encoder/decoder. We incorporated LSP coding for the LPC coefficients at a rate of 1 kbit/s.

The wavelet encoder consists of a normalizer, a wavelet transformer, and a limited size codebook, as shown in Figure 6. The normalizer computes the gain factor of the LPC excitation and extracts that from the LPC excitation, so that the variance of the input of the wavelet transformer is one.

The transformer produces coefficients, and using a neural network, we can obtain codewords from the codebooks that is the closest in Euclidean sense to the set of coefficients. For each band, we use one codebook. Thus the compressed speech contains the LSP parameters, gain factor, and indices to the codebooks.

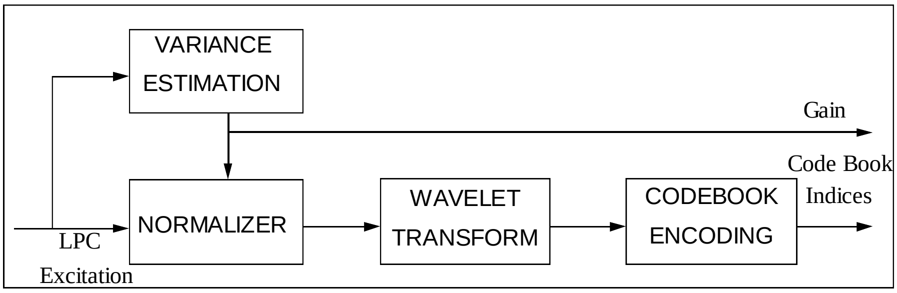

Figure 6 Wavelet encoder.

To reconstruct the speech, we use the inverse process depicted in Figure 7. The process passes the parameters to the codebooks, inverses transforms the resulting codeword, scales the resulting excitation signal according to the gain factor, and applies the resulting signal to an LPC filter.

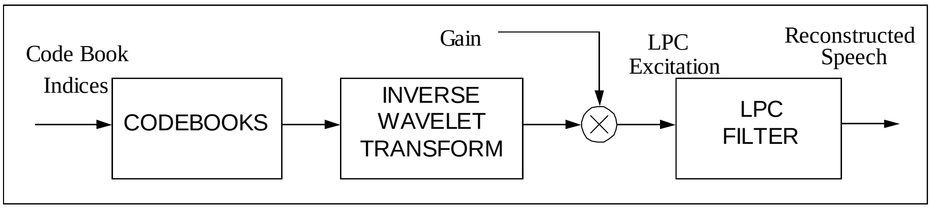

Figure 7 Speech decompressor.

We design a codebook for every scale using frequency-sensitive competitive learning neural network [12]. Thus, for the frame length of 64, there should be 6 codebooks. However, based on the properties discussed above, we decided to include scale 1, 2, 3, and 4 only, and omit scale 5 and the lowpass section. Thus, we have designed four codebooks with two different sizes, 128 and 256, and trained them using the coefficients obtained from training sentences.

By combining the codebooks, we can have different sets of codebooks with different numbers of bits required between 28 to 32 bits per 64 samples. For two sets with 28 and 32 bits per 64 samples, we need 3.5 and 4.0 kbit/s, respectively. Assuming that the gain factor requires 4 bits per 64 samples, i.e., 0.5 kbit/s, and LPC coefficients require 1 kbit/s, the two sets result in 5 and 5.5 kbit/s, respectively.

The performance test showed the promising capability of using wavelet to model the LPC excitation. To measure the performance, we set *daub4* as the wavelet prototype [13] and male spoken speech signal containing 17 Harvard sentences [5] for training as well as test. We then perform speech compression and decompression in two different sets of codebooks. The codebooks have a size of 128 and 256, respectively. The neural network was able to distribute the codewords among the training set. For the given codebook sizes, the SNR of the coefficients were low, as depicted in Table 1. Those are SNR for wavelet coefficients related to excitations.

Table 1 SNR of wavelet coefficients for each scale codebook.

Scale code book	SNR for size 128 (dB)	SNR for size 256 (dB)
1	4.37	8.92
2	6.25	10.77
3	8.97	14.92
4	14.4	21.8

However, when the excitations are applied to LPC filter, the speech SNR improves significantly. Thank to the power of the model, the speech SNR measurement with 128 and 256 sizes of codebooks results in 11.03 and 15.33 dB, respectively, which are quite high for their bit rates. Although these results are preliminary due to the limited number of test sentences, they show the promising potential of the wavelet model.

Table 2 SNR of the synthesized speech.

Code book size	Bit rate (kbit/s)	SNR (dB)
128	5	11.03
256	5.5	15.33

5 Conclusions

The linear combination of wavelets is an attractive model of LPC excitation for speech compression. We have applied a wavelet model of LPC excitation for speech compression. The scheme exploits coefficients' asymmetrical properties: magnitude dependent sensitivity, scale dependent sensitivity, and limited frame length. We have described an ideal scheme through the use of close-loop codebook searching and perceptually weighted measure, as well as a practical scheme through whitening the effect of the quantization noise. Our experiment shows that even in a simple straight-forward scheme, the model indicates promising capability by having SNR 11.03 and 15.33 dB at 5 and 5.5 kbit/s, respectively.

Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and Riset Unggulan Terpadu (RUT). The author acknowledges Dr. Ken Ferens, University of Manitoba, Canada, for his contribution in neural network based codebook design, used in this result.

References

[1] Markovic, M.Z, *Speech Compression - Recent Advances and Standardization*, 5th International Conference on Telecommunications in Modern Satellite, Cable and Broadcasting Service, 2001, TELSIKS 2001, **1**, Page(s): 235 – 244, 19-21 Sept. 2001.

[2] Campbell, Jr. J. P., Tremain, T. E. & Welch, V. C., *The proposed Federal Standard 1016 4800 bps voice coder: CELP*, Speech Technology, pp. 58-64, Apr./May 1990.

[3] Paliwal, K.K. & Atal, B.S., *Efficient Vector Quantization of LPC Parameters at 24 bits/ frame*, IEEE Trans. Speech Audio Proc., IEEE 1063-6676/93, vol. 1, no. 1, pp. 3-14, Jan 1993.

[4] Soong, F. K. & Juang, B-H, *Optimal Quantization of LSP Parameters*, IEEE Trans. Speech Audio Proc., IEEE 1063-6676/93, **1**(1), pp. 15-24, Jan 1993.

[5] Langi, A., *Code-Excited Linear Predictive Coding for High-Quality And Low Bit-Rate Speech*, M.Sc. Thesis, University of Manitoba, Winnipeg, MB, Canada, 138 pp., 1992.

[6] Najih, A.M.M.A., Ramli, A.R., Ibrahim, A., Syed, A.R., *Comparing speech compression using wavelets with other speech compression schemes*, Proceedings, Student Conference on Research and Development, 2003, SCORED 2003, P:55 – 58, 25-26 Aug. 2003.

[7] Najih, A.M.M.A., bin Ramli, A.R., Prakash, V., Syed, A.R., *Speech compression using discreet wavelet transform*, NCTT 2003 Proceedings. 4th National Conference on Telecommunication Technology, pp1 – 4, 14-15 Jan. 2003.

[8] Parsons, T. W., *Voice and Speech Processing*. New York: McGraw-Hill, 402 pp., 1986.

[9] Atal, B.S., *A model of LPC excitation in terms of eigenvectors of the autocorrelation matrix of the impulse response of the LPC filter*, in Proc. IEEE ICASSP, CH2673-2/89, pp. 45-48, 1989.

[10] Xu, Y., Wang, G., Gu, Y. & Liu, H., *A Novel Wavelet Packet Speech Enhancement Algorithm Based On Time-Frequency Threshold*, ICICIC '07, Second International Conference on Innovative Computing, Information and Control, 2007., P:492 – 492, 5-7 Sept. 2007.

[11] Ofer, E. D., Malah, Dembo, A., *A Unified Framework for LPC Excitation Representation in Residual Speech Coders*, in *Proc. IEEE ICASSP*, CH2673-2/89, pp. 41-44, 1990.

[12] Ferens, K. & Kinsner, W., *Energy and Frequency Adaptive Wavelet Subband Coding for Wideband Audio Compression*, Proc. 9th Int. Conf. on Math. and Comp. Modelling, 1993.

[13] Daubechies, I., *Ten Lectures on Wavelets*, Philadelphia, Penn: SIAM, 357 pp., 1992.