An LPC Excitation Model using Wavelets

[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 a new model of linear predictive coding (LPC) excitation using wavelets for speech signals. The LPC excitation becomes a linear combination of a set of self- similar, orthonormal, band-pass signals with time localization and constant bandwidth in a logarithmic scale. Thus, the set of the coefficients in the linear combination represents the LPC excitation. The discrete wavelet transform (DWT) obtains the coefficients, having several asymmetrical and non-uniform distribution properties that are attractive for speech processing and compression. The properties include magnitude dependent sensitivity, scale dependent sensitivity, and limited frame length, which can be used for having low bit-rate speech. We show that eliminating 8.97% highest magnitude coefficients degrades speech quality down to 1.49dB SNR, while eliminating 27.51% lowest magnitude coefficient maintain speech quality at a level of 27.42 dB SNR. Furthermore eliminating 6.25% coefficients located at a scale associated with 175-630 Hz band severely degrades speech quality down to 4.20 dB SNR. Finally, our results show that optimal frame length for telephony applications is among 32, 64, or 128 samples.

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

1 Introduction

Speech signal processing is a necessity in many applications, including speech recognition, speech communication, and speech analysis [1-2]. Techniques for such applications often deploy speech production models. For example, techniques based on a simple speech production model have successfully reduced the bit rates to below 8 kbits/s. This rate can be accommodated by the narrow-band telephony 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 [3].

In this compression approach, modeling LPC excitation plays a critical role to obtain high speech quality speech. There are different techniques to code the

Received October 17th, 2008.

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. Another model uses scalar quantization or center-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.

This paper proposes a new model of LPC excitation using wavelets. We hypothesize that the wavelet model provides attractive features for speech processing, including compression [4-5]. The LPC excitation becomes a linear combination of a set of highly structured signals, called *wavelet*s. In this combination, we can represent the excitation with a set of real numbers called wavelet coefficients, obtained using discrete wavelet transform (DWT). Our experiments show these coefficients have attractive properties for efficient representation of speech. The properties include magnitude dependent sensitivity, scale dependent sensitivity, and limited frame length, which can be used for having low bit-rate excitation. We show that eliminating 8.97% highest magnitude coefficients degrades speech quality down to 1.49dB SNR, while eliminating 27.51% lowest magnitude coefficient maintain speech quality at a level of 27.42 dB SNR. Furthermore eliminating 6.25% coefficients located at a scale associated with 175-630 Hz band severely degrades speech quality down to 4.20 dB SNR. Finally, our results show that optimal frame length for telephony applications is among 32, 64, or 128 samples. Such asymmetrical and non-uniform distribution properties are suitable for speech compression techniques.

It should be noted that different wavelet models of speech signals have been used for different purposes. For example, a wavelet packet model have been used for speech enhancement [2]. In the model, speech signals are represented with a set of wavelet packet coefficients. A threshold is then applied uniformly to the coefficients, resulting in smoother and cleaner speech. In contrast, our model uses magnitude and scale sensitive thresholding and clipping of wavelet coefficients instead of uniform thresholding.

In another example, a rational wavelet model has been used for phonetic classifications [1]. In this model, speech signals are decomposed through a filter bank. The filter bank accommodates critical-band effect in human auditory system. As a result, the model improves performance of phonetic classifications. In contrast, our model extends the filter to include LPC filter, in addition to wavelet filter bank. This allows us to remove spectral information prior to wavelet analysis, resulting in a more efficient wavelet representation.

2 LPC Excitation

We can derive LPC excitation from a segment of speech. Let $s[n]$ be the value representing the amplitude of speech signal at a sampling instance nT, where T is a sampling period of 0.125 ms. Let us also call LPC excitation at that instance as $t[n]$, and the LPC filter $H(z)$ as [6].

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

Then, the speech production model implies (in the z-domain notation)

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

Let us now consider a similar relationship for a segment of speech **s**, which is a vector whose elements are $s[0]$, $s[1]$, ..., $s[N–1]$. With typical values between 7 to 32 ms, NT is time duration short enough for $H(z)$ to be considered stationary. Using *linear prediction* [6], we can obtain a_i in Eq. (1) from **s**. Furthermore, we can approximate $H(z)$ with an $N\times N$ lower-triangular matrix H:

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

Here, h_i are the impulse responses of $H(z)$ [7]. The corresponding LPC excitation is then a vector **t**, whose elements are $t[0]$, $t[1]$, ..., $t[N–1]$. (Clearly, both **s** and **t** are members of a linear space R^N). From the time domain expression of Eq. (2), it can be shown that one $s[n]$ is affected by all $t[m]$, with $m \le n$. Thus, if the additive contribution of all $t[m]$ from the previous segments to the current **s** is a vector **u**, then Eq. (2) becomes

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

This **t** is the LPC excitation that should be modelled and compressed, since a synthesizer using Eq. (4) automatically generates u.

We can compress **t** because it contains less information than **s**, while both use the same numbers of bits. To show this, consider Figure 1, showing the spectra of both segments during a voiced articulation. Clearly, the excitation spectrum does not have the peaks and valleys (called *formants*) contained in **s**. The formants carry the phonemes of speech messages. Since the $H^{-1}(z)$ removes the formants from **t**, the spectrum of **t** is relatively flat. However, this does not mean **t** becomes unimportant. Figure 2 shows both segments in the time domain. Clearly, **t** still carries the periodicity (called *pitch*) of **s**, which is another important feature of speech [6]. Nevertheless, the information measure (*entropy*) of **t** should be lower than **s** because of the formant removal, making compression possible.

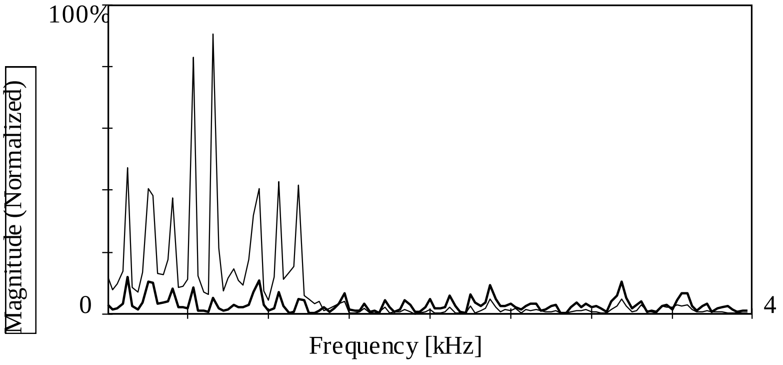

Figure 1 Spectra of both speech and excitation segments.

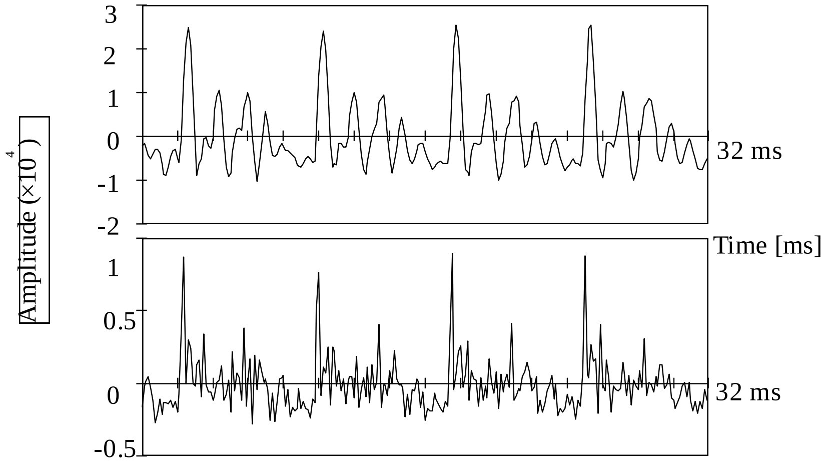

Figure 2 Both the speech and excitation segments in the time domain (N=256).

3 Wavelet Model of LPC Excitation

In this paper, we model the LPC excitation 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 $(\log_2 N)-1$, thus the description of DWT in [8] is directly applicable). 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*, 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.

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

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

This model has many advantages, due to its structure and the signals involved. Firstly, since it is a linear combination, it fits in a highly structured linear system, having many analysis tools. For example, **t** now depends on two sets: the set of wavelet and scaling coefficients $\{c_{j,k}, d_{J,k}\}$, and the set of wavelet and scaling signals $\{\psi_{j,k}[n], \phi_{J,k}[n]\}$, usually called *basis* set. If the basis set is known, the coefficient set sufficiently characterizes **t**. The importance of one coefficient may differ to the other one, which can be exploited for compression. Furthermore, the terms at the right hand side of Eq. (6) by themselves are signals (members of R^N). Thus, we immediately have a signal decomposition, which allows us to study the signal by studying each of the terms.

Secondly, the basis signals are not only orthonormal, but are highly related in a self-similar fashion. The signals in the set $\{\psi_{j,k}[n]\}$ come from a single signal $\psi[n]$ called *wavelet prototype*, which is a bandpass, limited energy signal. The other signals in $\{\phi_{J,k}[n]\}$ come from another signal $\phi[n]$ called *scaling prototype*, which is a lowpass, limited energy signal. Both prototypes are closely related, where one can derive the other [9]. Figure 3 shows examples of Daubechies wavelet prototypes, called *daub4*, *daub12*, and *daub20* [10]. Then, the wavelet and scaling signals are dilated (by scale *j*) and translated (by *j* and *k*) versions of the prototypes, as follows:

$$\psi_{j,k}[n] = \sqrt{2^{-j}}\psi[2^{-j}n - k] = \sqrt{2^{-j}}\psi[2^{-j}(n - 2^{j}k)] \text{ and}$$

$$\phi_{j,k}[n]=\sqrt{2^{-j}}\phi\left[2^{-j}n-k\right]=\sqrt{2^{-j}}\phi\left[2^{-j}\left(n-2^{j}k\right)\right] \tag{7}$$

Clearly, the scale parameter j alters the amplitude, frequency position, and time position, while the parameter k alters the time position, both relative to the prototype. It has been shown that as a basis, the set of signals in Eq. (7) completely defines any signal in R^N [9]. More than that, since each of the terms in RHS of Eq. (6) has a unique combination of j and k, it has a unique position in the time-frequency plane, making time-frequency analysis (decomposition) inherent. This property is well suited for analyzing nonstationary signals such as speech and LPC excitation.

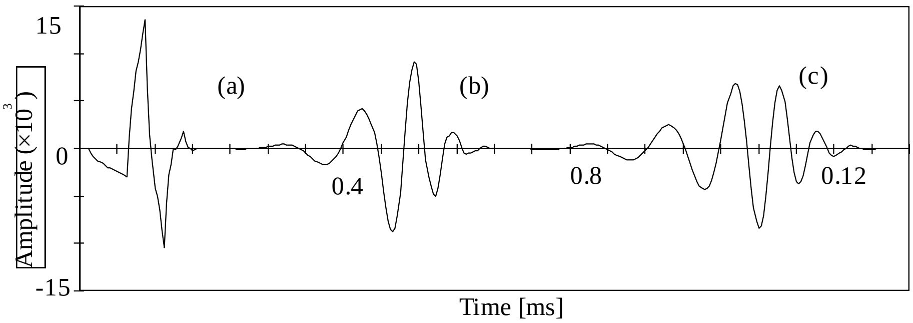

Figure 3 Prototypes of Daubechies wavelets. (a) *daub4*, (b) *daub12*, (c) *daub20*.

Thirdly, there is a possibility to group the RHS terms according to their scale. As implied by Eq. (7), all signals $\psi_{j,k}[n]$ in the same scale j have the same frequency position and duration. Let us now define

$$\tilde{t}_j[n]=\sum_{k=0}^{2^{-j}N-1}c_{j,k}\psi_{j,k}[n] \text{ and } \tilde{t}_{LP}[n]=\sum_{k=0}^{2^{-j}N-1}c_{j,k}\psi_{j,k}[n] \tag{8}$$

We can then decompose **t** to signals with different frequency bands, according to

$$t[n]=\tilde{t}_{LP}[n]+\sum_{j=1}^{J}\tilde{t}_j[n] \tag{9}$$

In this special decomposition, **t** consists of J bandpass signals $\tilde{t}_j[n]$ (each has a different frequency position but the same frequency duration in the logarithmic scale), and one lowpass signal $\tilde{t}_{LP}[n]$. For example, if N is 64 and J is 5, scale j is 1, 2, 3, 4, and 5. Figure 4 shows the corresponding decomposition of **t**. The plots are ordered from the lowest frequency band ($\tilde{t}_{LP}[n]$) to the highest one ($\tilde{t}_1[n]$). Note that the superposition of all decomposing signals is exactly the same with the original signal.

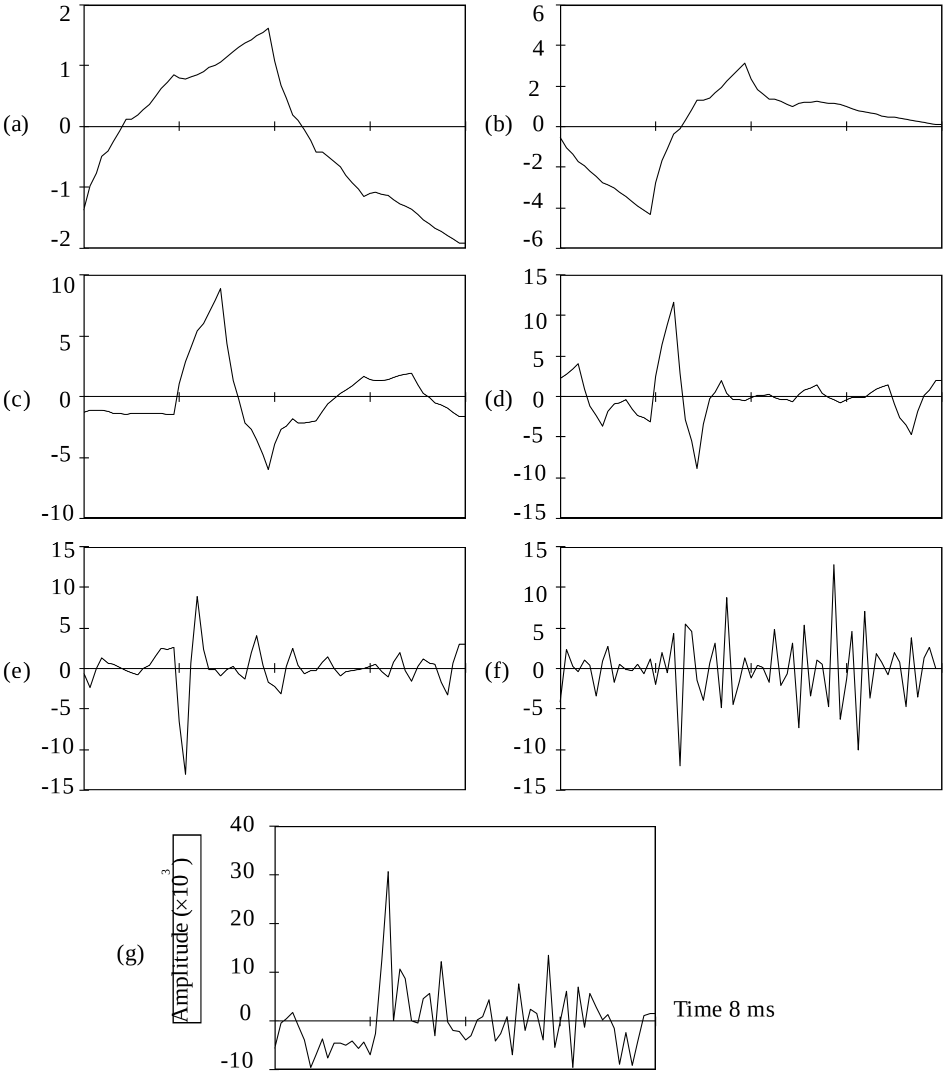

Figure 4 Wavelet decomposition of an 8-ms excitation: (a) the low-pass section, (b) scale 5, (c) scale 4, (d) scale 3, (e) scale 2, (f) scale 1, and (g) superposition of all the above signals. (All amplitudes $\cdot 10^3$.)

4 Experimental Results

We are interested in finding asymmetrical and nonuniform speech representations. Compression schemes are usually able to exploit such unbalanced representations. Our experiments show that wavelet representation of **t**, i.e., $\{c_{j,k}, d_{J,k}\}$, has several properties that are attractive for compression. The compression is based on exploiting nonuniform distribution of various aspects of the coefficients $\{c_{j,k}, d_{J,k}\}$. Our investigation reveals such asymmetrical properties. We describe our experimental results as follow.

4.1 Magnitude Dependent Sensitivity

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. To show this property, we (i) alter the coefficients depending on its magnitude, (ii) use the altered coefficients to reconstruct the speech using Eq. (2), (4), and (6), and (iii) measure the resulting distortion in terms of SNR. To alter the coefficient, we set a range of magnitude, and if the magnitude of a coefficient lies in the range, we scale the coefficient by 50%. The process is repeated for different ranges. Table 1 shows the effect of altering the magnitude to the SNR. The table also shows the results of similar experiment with scaling factor of 0% (complete truncation).

Table 1 Magnitude sensitivity of coefficients.

Magnitude Ranges	SNR for 50% Scaled (dB)	SNR for 0% scaled (dB)	Number of Coefficients (%)
0-100	31.79	27.42	27/51
100-299	25.78	20.32	18.63
200-300	23.32	17.59	11.88
300-450	19.77	13.87	11.62
450-700	16.57	10.62	11.41
700-1200	13.32	7.34	9.95
>1200	7.45	1.49	8.97

Notice the asymmetrical coefficient-significance according to its magnitude. Throwing away 27.51% of the coefficients degrades the SNR down to 27.42 dB only, as long as those coefficients have low magnitudes. However, truncating only 8.97% of the coefficients with high magnitudes severely degrades the SNR down to 1.49 dB. The reason is the LPC excitation itself is magnitude sensitive, as concluded in [11]. Since high-magnitude coefficients are usually responsible for high-magnitude excitation, they becomes more significant than the other coefficients. This means we should use more bits for higher magnitude coefficients. Fortunately, Table 1 also shows that there are few high magnitude coefficients.

4.2 Scale Dependent Sensitivity

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. To show this property, we (i) alter the coefficients depending on its scale as in Eq. (9), (ii) use the altered coefficients to reconstruct the speech, and (iii) measure the resulting distortion, as before. To alter a coefficient, we set the coefficient to zero. The process is repeated for different scale *j*. Table 2 shows the effect of altering the magnitude to the SNR.

Table 2 Scale sensitivity of coefficients.

Scale	3dB Bandwidth (Hz)	SNR when clipped (dB)	Number of coefficients (%)
1	1500-4000	15.94	50
2	700-2550	9.63	25
3	350-1260	5.78	12.5
4	175-630	4.20	6.25
5	90-310	7.73	3.125
Lowpass	0-150	10.70	3.125

Again, notice the asymmetrical coefficient-significance according to its scale. Altering 50% of the coefficients results in 15.94 dB SNR, when they are in scale 1. However, altering 5.78% of coefficients severely degrades SNR to 4.20 dB, when they are in the scale 4. The reason is the LPC excitation is responsible for pitch information, while bands in scales 4 and 5 contain almost all human pitch frequency. Consequently, we must use more bits for coefficients in scales such as 3, 4, and 5. Fortunately, Table 2 also shows that there are few such coefficients.

4.3 Effect of the Frame Length

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.

To show this, we recall that N determines the number of j (bands or scales) that the DWT produces for a segment **t**, which is $\log_2(N)$ (including the lowpass section). Thus, doubling *N* increases the number of sections by one. However, what is happening is the lowpass section breaks into a new bandpass section and a new lowpass section. In other words, the resolution increases at the low frequency only, and frequency bands of scale *j* in both *N* and 2*N* frame lengths are the same. Table 3 shows our coarse measurement of the 3 dB cut-off frequencies of the low-pass section and its closest band-pass section, when using *daub4* wavelets. If we deal with telephone bandwidth signals (300-3300 Hz), there is no need to increase resolution at bands below 300 Hz. Hence, there is no need to use *N* more than 32. Furthermore, LPC excitation mostly contains pitch information, and pitch frequencies are more than 80 Hz. Thus even for wide band speech, a number of 128 is sufficient for *N*.

Table 3 Cut-off (3 dB) frequencies of the low-pass (LP) section and its closest band-pass (BP) section, approximated for *daub4* wavelet.

N	Total Bands	LP cut-off (Hz)	BP Left-side cut-off (Hz)	BP Right-side cut-off (Hz)
4	2	2400	1547	4000
8	3	1300	700	2550
16	4	600	350	1260
32	5	300	175	630
64	6	150	90	310
128	7	80	45	155
256	8	40	22	78
512	9	20	11	39
1024	10	10	5	20
2048	11	5	2	10
4096	12	2.5	1	5

5 Conclusions

The linear combination of wavelets is an attractive model of LPC excitation for speech compression. We have described a model of LPC excitation as a linear combination of a set of self-similar bandpass signals known as wavelets. Here, the coefficients of the linear combination (obtained through DWT) represent the LPC excitation. The coefficients posses properties that are useful for speech compression: magnitude dependent sensitivity, scale dependent sensitivity, and limited frame length.

We show the coefficients resulting from this model have different significance for speech quality. High magnitude coefficients are more important than low magnitude ones. Eliminating 8.97% highest magnitude coefficients degrades speech quality down to 1.49dB SNR, while eliminating 27.51% lowest magnitude coefficient maintain speech quality at a level of 27.42 dB SNR. Furthermore, coefficients at middle scale are more important than those in other scales. Eliminating 6.25% coefficients located at a scale associated with 175-630 Hz band severely degrades speech quality down to 4.20 dB SNR. Finally, our results show that optimal frame length for telephony applications is among 32, 64, or 128 samples. Because of such asymmetrical and non-uniform distribution properties, the model is suitable for speech compression techniques.

Acknowledgements

This work was supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada, and Riset Unggulan Terpadu (RUT).

References

[1] Choueiter, G.F. & Glass, J.R., *An Implementation of Rational Wavelets and Filter Design for Phonetic Classification*, IEEE Transactions on Audio, Speech, and Language Processing, **15**(3), pp.939-948, March 2007.

[2] 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, pp.492-492, 5-7 Sept. 2007.

[3] 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**, pp. 235-244, 19-21 Sept. 2001.

[4] Najih, A.M.M.A., Ramli, A.R., Ibrahim, A. & Syed, A.R., *Comparing Speech Compression Using Wavelets with Other Speech Compression Schemes*, Proc. Student Conference on Research and Development, 2003, SCORED 2003, pp. 55-58, 25-26 Aug. 2003.

[5] Najih, A.M.M.A., Ramli, A.R., Prakash, V. & Syed, A.R., *Speech Compression Using Discreet Wavelet Transform*, NCTT 2003 Proc. 4th National Conference on Telecommunication Technology, pp. 1- 4, 14-15 Jan. 2003.

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

[7] 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.

[8] Press, W.H., *Wavelet transforms, Harvard-Smithsonian Centre for Astrophysics Preprint No. 3184*, 1991. (Available through anonymous ftp at 128.103.40.79, directory /pub).

[9] Mallat, S., *A Theory for Multiresolution Signal Decomposition: The Wavelet Representation*, IEEE Trans. Patt. Anal. Machine Intell., **11**(7), pp. 84-95, July 1989.

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

[11] Atal, B.S & Remde, J.R., *A new model of LPC excitation for producing natural-sounding speech at low bit rates*, Proc.IEEE Int. Conf. Acoust. Speech and Signal Proc., Paris IEEE CH1746-7/82, pp. 614-617, 1982.