EFFECT OF THE NUMBER OF LPC COEFFICIENTS ON THE QUALITY OF SYNTHESIZED SOUNDS

Hung Ngo, Texas A&M University-Corpus Christi; Mehrube Mehrubeoglu, Texas A&M University-Corpus Christi

Abstract

Linear Predictive Coding (LPC) is a basic method used in speech signal processing. The purpose of this work is to implement a real-time LPC vocoder on a TMS320C6455 DSP board to assess the effect of the number of coefficients on sound quality for human voice, and bird and vehicle sounds. The experimental results show that there exists a threshold at which the sound quality is degraded. The threshold value may vary depending on the quality of the original input sound, the optimization level of LPC implementation, and the processing speed of the processor. In this work, signal-to-noise ratio (SNR) is used as the quality measure to determine the threshold for C6455 DSP.

Introduction

Linear Predictive Coding (LPC) is an important technique that is commonly used in audio and speech coding [1]-[7]. To produce high-quality speech at low bit rates for storage or transmission, many well-known speech compression and decompression techniques, including Code-Excited Linear Prediction (CELP), use principles of LPC [4]. Although the compression rate of a low-ordered LPC vocoder is very high, the synthesized speech is very unnatural and synthetic [8]. However, the LPC vocoder still enables understandable speech. LPC10, the US standard for linear predictive coding of speech at 2400 bits per second, is a typical example. It is applied in many military applications, which need very low bit rates and do not require high-quality speech [3], [9].

A LPC vocoder processes input signals in separate signal blocks and computes a set of filter coefficients for each block. Since speech sounds are slowly varying in nature, the coefficients of a block are stable in short periods (about 20 ms) [10]. In the implementation used in this study, the block size and hop size were 30ms and 20ms, respectively. Hop size refers to the number of samples in each window from one frame to the next, and can be represented in terms of time, if the sampling rate is set.

A typical LPC-based vocoder has two main tasks: to analyze the original sound and to synthesize the sound based on parameters from the analysis phase. In the analysis phase, the vocoder computes prediction coefficients and residual errors of the input signals. The vocoder then saves these parameters on storage devices or transmits them over networks. The synthesis phase uses stored (or received) coefficients and residual errors to reconstruct the original sounds. In reality, instead of residual errors, pitch period, voiced/unvoiced decision bit, and gain value are calculated and stored/transferred with prediction coefficients [1], [2].

The purpose of this study was to implement a LPC on a TMS320C6455 (Texas Instruments) to assess the effect of the number of coefficients on sound quality based on the metric signal-to-noise ratio (SNR). The input includes both human speech (male and female) and object sounds (car and bird). The quality of synthesized sounds was evaluated using the average SNR values for five different signal segments from each kind of sound recording, namely, people, cars and birds.

The remainder of this study was organized as follows: Section II introduces basic principles of LPC. Section III presents pitch estimation using the auto-correlation method. Section IV shows experimental results and discusses the effect of the number of coefficients on quality of speech sounds. Finally, Section V summarizes the results of the study.

Linear Predictive Coding (LPC)

In LPC, if it is assumed that the present sample of speech is predicted by using the last P samples, the predicted sample $\hat{x}(n)$ of $x(n)$ can be expressed as

$$\hat{x}(n) = a_1 x(n-1) + a_2 x(n-2) + \ldots + a_p x(n-P) = \sum_{k=1}^{P} a_k x(n-k) \quad (1)$$

The error between the actual sample, $x(n)$, and the predict values, $\hat{x}(n)$, can be expressed as

$$\varepsilon(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{P} a_k x(n-k), \quad (2)$$

where coefficients $\{ a_k \}$ are calculated from the following matrix equation [1], [11]:

$$\begin{bmatrix} r(0) & \ldots & r(P-1) \\ r(1) & \ldots & r(P-2) \\ \vdots & \ddots & \vdots \\ r(P-2) & \ldots & r(1) \\ r(P-1) & \ldots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{P-1} \\ a_P \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(P-1) \\ r(P) \end{bmatrix}, \quad (3)$$

and $r(k)$ is computed using autocorrelation method as follows:

$$r(k) = \sum_{k=1}^{N-1-k} x(n)x(n+k)\,. \tag{4}$$

Here, N is the number of samples in each segment or frame. Equation (3) is solved using the Levinson-Durbin recursive algorithm **Error! Reference source not found.**, **Error! Reference source not found.**.

The implementation of Equation (2) is called the analysis filter, which can be simply described using the transfer function $A(z)$ as shown in Figure 1 [11]:

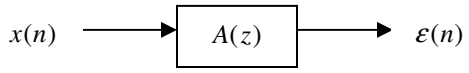


Figure 1. LPC analysis filter

The transfer function $A(z)$ is given by [11]:

$$A(z) = 1 - \sum_{k=1}^{P} a_k z^{-k} \tag{5}$$

If both error sequence and prediction coefficients are available, then the output signal, $x(n)$, can be reconstructed as follows:

$$x(n) = \hat{x}(n) + \varepsilon(n) = \sum_{k=1}^{P} a_k x(n-k) + \varepsilon(n) \tag{6}$$

Equation (5) can be seen as another form of Equation (2). The implementation of Equation (5) is called the synthesis filter which is shown below in Figure 2 [11]:

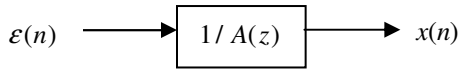


Figure 2. LPC analysis and synthesis filter

$x(n)$ and $\varepsilon(n)$ are the original signal and residual error, respectively.

Pitch Estimation

Pitch determination is very important for many speech coding algorithms [2]. In the implementation for this study, pitch was estimated using an autocorrelation method that detects the highest value of the autocorrelation function in the region of interest (except at $\tau = 0$). The autocorrelation function is given as

$$R(\tau) = \sum_{n=0}^{N-1-\tau} x(n)x(n+\tau)\,, \tag{7}$$

where τ is the lag and N is the number of speech samples in a frame. Pitch period and its harmonics are estimated by determining the value of the lag, τ, at which the value of the autocorrelation value, $R(\tau)$, is highest.

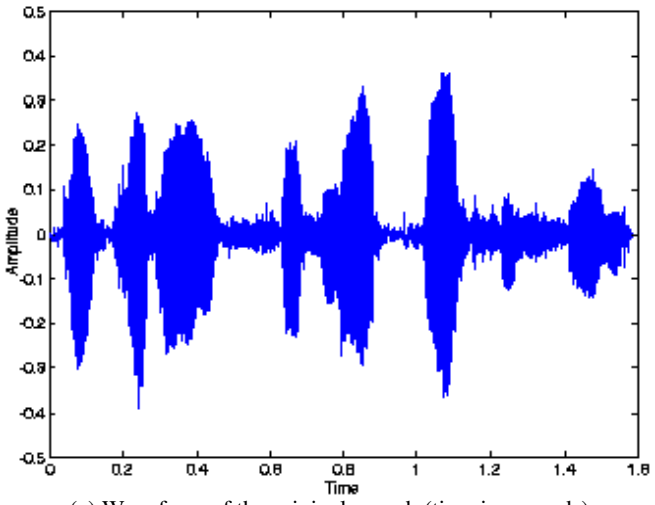
Experimental Methods

The experiment is conducted on a TMS320C6455 DSP board, a high-performance fixed-point DSP developed by Texas Instruments that uses very-long-word-instruction (VLIW) architecture. In the presented implementation, 5-to-8-second recordings of human voices, cars, and birds were used as inputs. The audio sampling rate was set to 16kHz. Each recording was divided into blocks of 30ms. The inter-window hop size was then set to 20ms (overlap segment was 10ms). Two special features, ping-pong buffering and Linked EDMA transfers, were used to ensure uninterrupted audio signals and real-time schedules [12]. The evaluation process was conducted in a quiet room with all settings and configurations kept the same throughout the process.

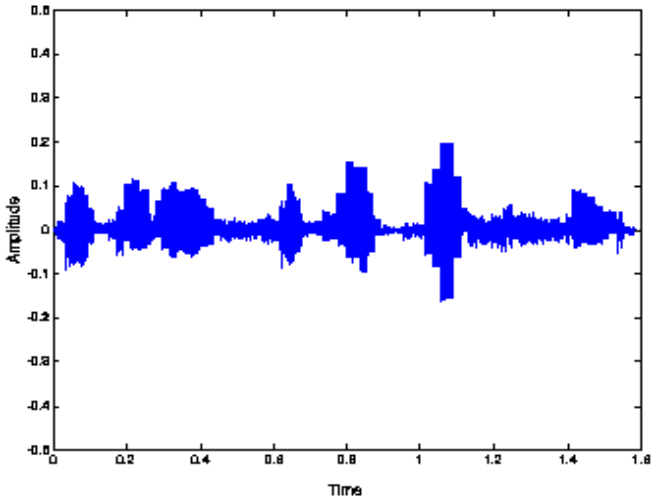
Figure 3 demonstrates the original (a) and synthesized (b) female voice signal acquired when the subject was saying "digital signal processing." Figure 4 shows the spectra for the two (original and synthesized) speech signals. The synthesized speech sounds were evaluated using SNR. SNR is defined as the ratio between the signal's energy and the noise's energy in dB. The larger the SNR ratio, the better the sound quality. The SNR is given in Equation (8) as [1]

$$SNR = 10\log_{10}\left(\frac{\sum_{n}(x[n])^2}{\sum_{n}(x[n]-y[n])^2}\right), \tag{8}$$

where $x[n]$ and $y[n]$ are the amplitude of the original speech signal, $x(n)$, and the synthetic version, $y(n)$, at discrete time, *n,* respectively.



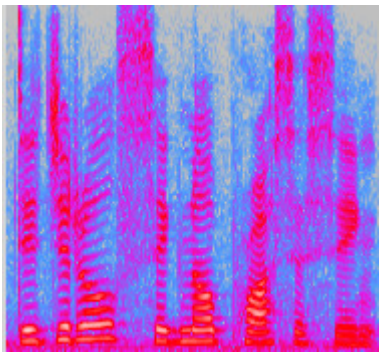
(a) Waveform of the original speech (time in seconds)



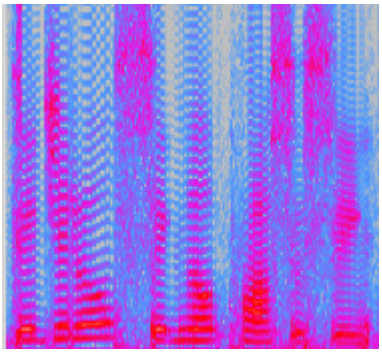
(b) Waveform of the synthesized speech (time in seconds)

Figure 3. Time waveform of a female voice saying "digital signal processing."

Each signal was segmented into 480-sample frames (30ms) with 320-sample overlap (20ms hop size). To assess the quality of sound solely based on the number of coefficients, the recording environment and devices were kept the same throughout the experimental process. All the original sounds were sampled at 16kHz. Also, block size and hop size were fixed at 480 samples and 320 samples, respectively. To measure the quality of the outputs, the SNR values for speech samples was calculated from 5 different individuals for each sex for each choice of prediction order. Each person was asked to read the same sentence that is about 5 to 7 seconds long depending on personal reading speeds. The human voice signals were acquired in a quiet room using a microphone. SNR values were computed from 5 to 30 LPC coefficients. The SNR results are presented in the next section.



(a) Spectrogram of the original speech



(b) Spectrogram of the synthesized speech

Figure 4. Spectrogram of a female voice saying "digital signal processing."

The non-speech sounds are acquired from a free website (freespeech.org) and are included as a separate evaluation in this implementation. The non-speech sounds utilized were comprised of bird and car sounds which lasted up to 50 seconds. Similar to human speech sounds, recordings of birds and cars have been assessed for a number of LPC coefficients ranging from 5 to 30.

The Mean Opinion Score (MOS) is used as an independent measure to evaluate the synthesized sounds from all categories based on the human perception of sound quality. A group of five people was asked to rate the quality of synthesized sounds based on a scale of 1 to 5 (1 = Bad, 2 Poor, 3 = Fair, 4 = Good, 5 = Excellent). MOS is particularly valuable in putting quantitative SNR values into perspective with perceived quality of the signals. The MOS results are presented in the next section.

Results and Discussion

To quantify the quality of the synthesized signals, SNR was used as the metric for the tested number of LPC coefficients for each of the four categories of sound signals. For

the human voice signals, the final SNR shown in Table 1 below represents the average value of the 5 SNRs computed from each person's speech.

It can be seen in Table 1 and Figure 5 that the SNR values of both male and female voice signals greatly improve when prediction orders increase from 5 to 10. After that, SNR values increase slowly for the orders from 10 to 25. At the order of 25, SNR values start decreasing due to the limit of processing speed. At this point, the large amount of signal input cannot be computed in real-time and the synthesized output suffers from increased noise. If the prediction order is increased, the quality of the synthesized sounds continues to decrease dramatically.

Table 1. SNR values of human speech

Order	Male	Female
5	5.5	4.9
10	6.7	5.9
15	7.1	6.2
20	7.2	6.4
25	7.2	6.5
30	6.2	5.8

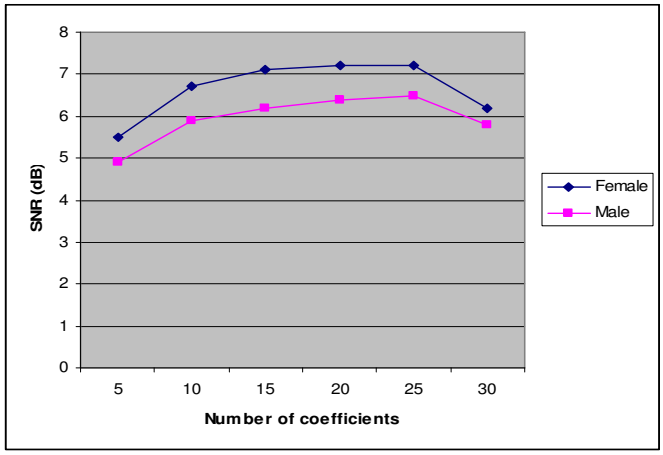


Figure 5. Effect of the number of prediction orders on quality of synthesized speech sounds

The non-speech sounds are also evaluated in this implementation. Similar to speech sounds, recordings of birds and cars were assessed for prediction orders ranging from 5 to 30. The results are shown in Table 2 and Figure 6.

Table 2. SNR values of non-speech sounds

Order	Bird	Car
5	5.3	5.7
10	6.1	6.3
15	6.3	6.4
20	6.4	6.5
25	6.5	6.5
30	5.7	5.6

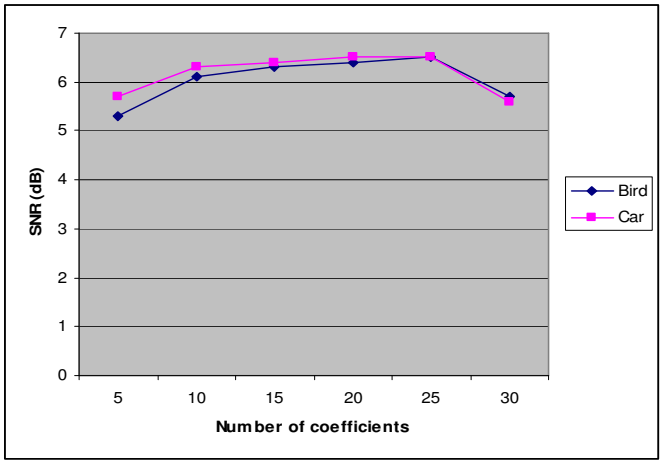


Figure 6. Effect of the number of prediction orders on quality of synthesized non-speech sounds

Although non-speech sounds cover a wider range of frequency compared to speech sounds, the results were similar. The SNR values rapidly improve for orders from 5 to 10, slowly increase with orders from 10 to 25, and start decreasing with orders greater than 25.

With both speech and non-speech sounds, the synthesized outputs were degraded at a prediction order of 25, due to the processing speed of the processor. Since the implementation optimization also affects the performance, the order threshold can be increased with another implementation that is more optimized; however, the order of 10 is usually chosen since there is not significant improvement in sound quality for orders greater than 10 [1], [7], [10].

If the processing speed is not a constraint, it is obvious that an increase of the LPC order will improve the quality of the synthesized speech, but at the cost of increased bandwidth. In this case, bandwidth and processing speed limitations of the TMS320C66455 sets a very important quality threshold. This threshold, however, does not have much of an effect on the results, since the improvement in the synthe-

sized speech quality is insignificant for the number of coefficients greater than 15 (analogous to the law of diminishing returns). This is shown in the experimental results for both speech and non-speech sounds in Figures 5 and 6.

The Mean Opinion Score is also used to evaluate the synthesized sounds. Table 3 and Figure 7 show the average MOS values for female, male, bird, and car sounds.

Table 3. Mean Opinion Score for synthesized sounds

Order	Female	Male	Bird	Car
5	1	1	1	1
10	3	2	2	2
15	4	3	3	3
20	4	3	4	3
25	4	3	4	2
30	3	2	2	1

The MOS values display some differences from the SNR values above. For example, the perceived quality of synthesized human sounds does not improve when the number of coefficients increases from 15 to 25 for either female or male voice signals, although such improvement, though small, is apparent in SNR values. The perceived quality of synthesized car and bird sounds also follow slightly different trends when compared with the SNR results. However, the human-perceived quality of all synthesized sounds starts decreasing from the order of 25 to 30, in line with SNR results. The MOS values above support the use of 15 coefficients for the real-time implementation presented here.

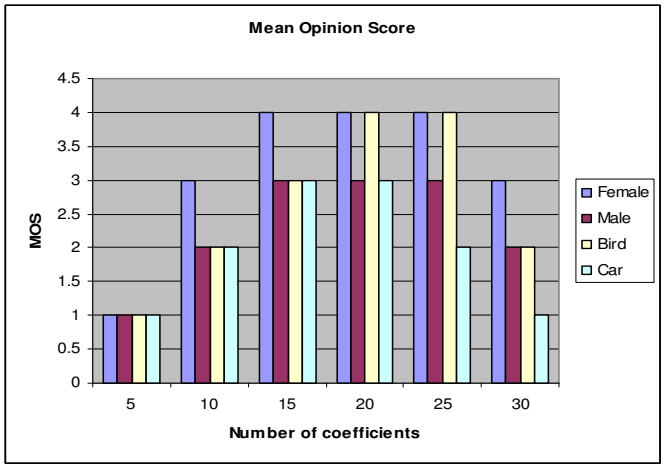


Figure 7. Graph of Mean Opinion Score

It is important to note that the relatively low SNR values are comparable with some of the earlier findings for similar metric measures used to evaluate sound quality using LPC analysis with different hardware [13]. This simple implementation is suitable for differentiating among sounds from different categories such as male voice, female voice, bird, and vehicle, but may not be sufficient to perform more stringent identification such as voice recognition, and bird or vehicle type. LPC is suitable for application areas that require limited bandwidths.

Conclusion

In this study, an LPC vocoder for TMS320C6455 DSP was implemented. The experimental results for both human speech (male and female) and object sounds (car and bird) show that the number of coefficients greatly impacts the quality of both speech and non-speech sounds. Using SNR value as the quality metric, in this implementation, the quality of synthesized sounds starts degrading at the prediction order of 25; however, this threshold value may vary depending on the quality of the original input sound, the optimization level of the LPC implementation, and processing speed of the DSP processor. Although an order of 10 is used in typical real-time applications, in this implementation, based on the metrics SNR and MOS for these experimental results, an order of 15 coefficients were suitable for male, female, bird and car sound synthesis. This implementation satisfies low-bandwidth applications such as basic environmental monitoring to detect the presence or absence of sound and its main category.

Acknowledgments

This project has been partially funded by Texas Research Development Fund, Texas A&M University-Corpus Christi.

References

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

[2] Huang, X., Acero, A., and Hon, H. W., Spoken Language Processing: A Guide to Theory, Algorithm, and System De-velopment, Prentice Hall, 2001.

[3] Itakura F., "Line spectrum representation of linear predic-tive coefficients of speech signals," J. Acoust. Soc. Am., 57, 537(A), 1975.

[4] Liu, Y. J., "On reducing the bit rate of a CELP-based speech coder," IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 49-52, 1992.

[5] Makhoul J., "Linear prediction: A tutorial review," Pro-ceedings of the IEEE, 1975, pp. 561-580.

[6] Makhoul J., "Correction to Linear prediction: A tutorial review," Proceedings of the IEEE, 1976, p. 285.
[7] Vallabha, G. and Tuller, B., "Choice of Filter Order in LPC Analysis of Vowels," From Sound to Sense, MIT, 2004, pp. C-203 – C-208.
[8] McCree, A.V. and Barnwell, T.P. III, "Improving the per-formance of a mixed excitation LPC vocoder in acoustic noise," Acoustics, Speech, and Signal Processing, ICASSP-92, 1992, pp. 137-140.
[9] Tremain, T., "The government standard linear predictive coding algorithm: LPC - 10," Speech Technology, 1982, pp. 40 - 49.
[10] Cook, P. R., Real Sound Synthesis for Interactive Applica-tions, A K Peters, 2002.
[11] Park, S. W., Gomez, M., and Khastri, R., "Speech Compression Using Line Spectrum Pair frequencies and wavelet transform," Proc. Intelligent Multimedia, Video and Speech processing, 2001, pp. 437-440.
[12] DSK6455 Technical Reference Manual, http://c6000.spectrumdigital.com/dsk6455/v2/files/6455_dsk_techref.pdf. Accessed on August 15, 2009.
[13] Serizawa, M. and Gersho, A., "Joint Optimization of LPC and Closed-Loop Pitch Parameters in CELP Coders," IEEE Signal Processing Letters, 6(3), pp.52-54, March 1999.

Biographies

MEHRUBE MEHRUBEOGLU is a faculty member in the Mechanical Engineering and Engineering Technology program at Texas A&M University-Corpus Christi. She received her B.S. degree in Electrical Engineering from the University of Texas at Austin. She received her M.S. degree in Bioengineering and Ph.D. degree in Electrical Engineering from Texas A&M University. She was awarded ONR/ASEE summer faculty fellowships in 2009 and 2010. Her research interests include imaging, signal and image processing, and applications of spectroscopy in engineering and science. She is also interested in effective teaching and learning pedagogies. She can be reached at Ruby.Mehrubeoglu@tamucc.edu

HUNG NGO received his B.S. in Computer Science from University of Natural Sciences, Vietnam in August 2006. He worked in industry for two years before going back to school to pursue his M.S. in Computer Science. He is now a graduate student in the Department of Computing Sciences at Texas A&M University – Corpus Christi. His current research interests include Wireless Sensor Networks, Wireless Security, and Digital Signal Processing. Hung Ngo can be reached at ngolehung84@yahoo.com