Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Computational Geometry & Artificial Vision, Malta, September 15-17, 2005 (pp227-232)

Speech Compression With Masked Modulated Lapped Transform And SPIHT Algorithm

MAITREYEE DUTTA[1], DR. RENU VIG[2]
[1] Computer Science Department
National Institute of Technical Teachers' Training and Research
Sector-26
Chandigarh, INDIA

http://www.nitttrchd.ac.in
[2]Department of Information Technology
Panjab University, Chandigarh, INDIA

Abstract:-A new method of speech compression using the Modulated Lapped Transform(MLT) and Set Partitioning In Hierarchical Trees (SPIHT) algorithm is proposed in this paper. The improvement in the signal with the application of masking with Psychoacoustic model has also been discussed. The proposed scheme is based on the combination of the Modulated Lapped Transform(MLT) and SPIHT. This paper also describes an excitation level based psychoacoustic model to estimate the simultaneous masking threshold for speech coding. The system has the following stages. 1) a windowing function; 2) a time-to-frequency transformation; 3) an excitation level calculation block 4) a correction factor for estimating masking threshold; 5) the inclusion of the absolute masking threshold; 6) the output Signal-to-Masking ratio. We evaluated the performance by integrating the psychoacoustic model into speech coding. Comparisons are also made with Plain LPC Coder, Voice Excited LPC Coder with the coding of the residual signal with DCT, Voice Excited LPC Coder with the coding of the residual signal with MLT, Voice Excited LPC Coder with the coding of the residual signal with MLT and SPIHT. The performance of the coders described has been assessed by computer simulation in terms of

a) Signal –to –noise ratio (SNR)
b) Compression ratio
c) Percent Root mean square Difference(PRD).
d) Informal subjective listening test.

Keywords:
Plain LPC coder, Voice-Excited LPC coder, DCT, MLT, SPIHT, SNR

1 Introduction

The compression of Speech Signals refers to the reduction of the bandwidth required to transmit or store a digitized speech signal. Ideally, the digital representation of a speech signal is coded using a minimum number of bits to achieve a satisfactory quality of the synthesised signal whilst maintaining a reasonable computation complexity.

Most speech coding methods have been designed to remove redundancies and irrelevant information The Set Partitioning In Hierarchical Trees (SPIHT) algorithm [3] is a coding algorithm that allows the transmission of coefficients in a pseudo-sorted fashion where the most significant bits of the largest contained in speech, thus aiming toward producing high quality speech with low bit-rates[1]. The optimization of the bit-rate and quality of the synthesised signal is closely related, where an improvement of one aspect compensates to the degradation of the other. Hence, the main development issue usually evolves around the compromise between the need for low rate digital representation of speech and the demand for high quality speech reconstruction[2].

coefficients are sent first. The sorting is carried out according to the relative importance of the coefficients, determined by Coefficient amplitude, and transmits the amplitudes Partially refining the

transmitted Coefficients continuously until the bit limit is reached[3]. The original design of SPIHT was aimed at image compression, and the intent was to use the algorithm in the frequency domain. However the algorithm may be used in the time domain. The work presented in the paper combined SPIHT with the Modulated Lapped Transform (MLT) to code the LPC residual signal and compared the results to those obtained by using DCT based scheme. Combining psychoacoustic models into speech coding significantly improves the coding efficiency. This masking method has improved the Signal-to-Noise ratio.

2. Set partitioning in hierarchical trees

The Set Partitioning In Hierarchical Trees Algorithm (SPIHT) was introduced by Said and Pearlman [3]. It is a refinement of the algorithm presented by Shapiro in [9]. The algorithm is built on the idea that spectral components with more energy content should be transmitted before other components, allowing the most relevant information to be transmitted using the limited bandwidth available[5]. The algorithm sorts the available coefficients and transmits the sorted coefficients as well as the sorting information. The sorting information transmitted modified a predefined order of coefficients. The algorithm tests available coefficients and set of coefficients to determine if those coefficients are above a given threshold. The coefficients are thus deemed significant or insignificant relative to the current threshold. Significant coefficients are transmitted partially in several stages, bit plane by bit plane.

As SPIHT includes the sorting information as part of the partial transmission of the coefficients, an embedded bit stream is produced, where the most important information is transmitted first. This allows the partial reconstruction of the required coefficients from small sections of the bit stream produced.

3 The compression schemes used

In this paper, first of all speech signal has been compressed with Plain LPC-10 Vocoder and also with voice excited LPC coder with DCT of the residual signal.A block diagram of Plain LPC Vocoder is shown in Figure 1[4].

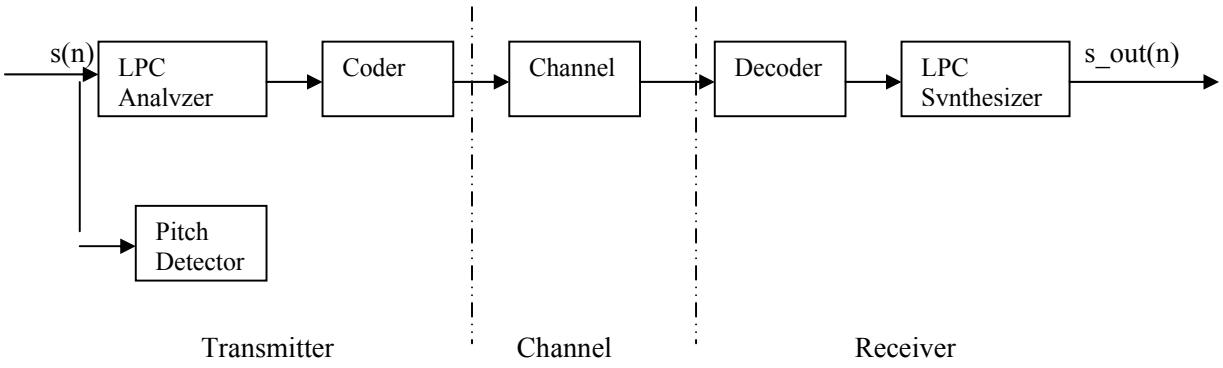

Figure 1: Block Diagram of an LPC vocoder

The principle behind the use of LPC is to minimize the sum of the squared differences between the original speech signal and the estimated speech signal over a finite duration. This could be used to give a unique set of predictor coefficients. These predictor coefficients are estimated every frame, which is normally 20 ms long. The predictor coefficients are represented by a_k. Another important parameter is gain G. The transfer function of the time varying digital filter is given by :

$$H(z)= \frac{G}{1- \sum a_k z^{-k}}$$

The two most commonly used methods to compute the coefficients are, but not limited to, the covariance method and the auto-correlation method. For our implementation, we have used the autocorrelation method. Levinson-Durbin recursion will be utilised to compute the required parameters for the auto-correlation method.

The LPC analysis of each frame also involves the decision regarding if a sound is voiced or unvoiced. If a sound is decided to be voiced, an impulse train is used to represent it, with nonzero taps occuring every pitch period. A pitch-detecting algorithm is employed to determine correct pitch period/frequency. We used the autocorrelation function to estimate the pitch period as proposed in[7]. However, if the frame is unvoiced, then white noise is used to represent it and a pitch period of T=0 is transmitted.From the speech production model it is known that the speech undergoes a spectral tilt of –6dB/oct. To counteract this fact a pre-emphasis filter is used.

3.1 Quantization of LPC coefficients

Usually direct Quantization of the predictor coefficients is not considered. To ensure stability of the coefficients (the poles and zeros must lie within the unit circle in the z plane) a relatively high accuracy (8-10 bits per coefficients) is required. This comes from the effect that small changes in the predictor coefficients lead to relatively large changes in the pole positions. There are two possible alternatives discussed in [7] to avoid the above problem. We had used the partial reflection coefficients(PARCOR). These coefficients are intermediate values during the calculation of the well known Levinson-Durbin recursion. Quantizing the intermediate values is less problematic than quantifying the predictor coefficients directly[10]. Thus, a necessary and sufficient condition for the PARCOR values is $|K_i|<1$.

4 Voice Excited LPC Vocoder

To improve the quality of the sound, the voice – excited LPC coders are used. Systems of this type have been studied by Atal et al. [8] and Weinstein [9]. Figure 2 shows a block diagram of a voice excited LPC vocoder.

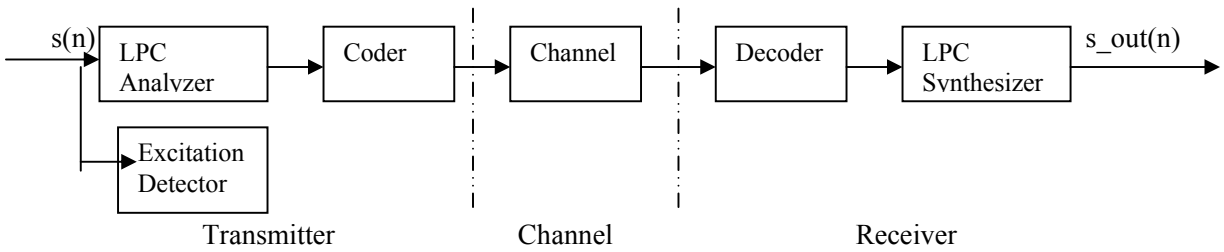

Figure 2: Block Diagram of a voice-excited LPC vocoder

The main idea behind the voice-excitation is to avoid the imprecise detection of the pitch and the use of an impulse train while synthesizing the speech. The input speech signal in each frame is filtered with the estimated transfer function of LPC analyzer. This filtered signal is called residual. If this signal is transmitted to the receiver one can achieve a very good quality.

4.1 DCT of residual signal

For a good reconstruction of the excitation only the low frequencies of the residual signal are needed. To achieve a high compression rate we employed the discrete cosine transform(DCT) of the residual signal[8][10]. It is known, that the DCT concentrates most of the energy of the signal in the first few coefficients. Thus one way to compress the signal is to transfer only the coefficients, which contain most of the energy. Our tests showed that these coefficients could even be quantized using only 4 bits. The receiver simply performs an inverse DCT and uses the resulting signal to excite the voice.

4.2 MLT of the residual signal

The MLT is a uniform M-channel filter bank. In traditional block transform theory, a signal x(n) is

divided into blocks of length M and is transformed by the use of an orthogonal matrix of order M[7]. More general filter banks take a block of length L and transform that block into M coefficients, with the condition that L>M. In order to perform this operation there must be an overlap between consecutive blocks of transformed coefficients. In the case of the modulated lapped transform L is equal to 2M and the overlap is thus M. The basis functions of the modulated lapped transform are given by:

$$a_{n,k}=h(n)\sqrt{\frac{2}{M}}\cos\left[(n+\frac{M+1}{2})(k+\tfrac{1}{2})\pi/M\right] \quad \ldots\ldots(1)$$

Where k=0,……M-1 and n=0,……., 2M-1.
The window chosen is $h(n)=\sin((n+1/2)\pi/2M)$.

4.3 MLT combined with SPIHT

Speech compression is done with the transform coding of LPC residual signal by the combination of the MLT with SPIHT. The residual signal is divided into overlapping frames and the MLT is applied to each frame. The coefficients obtained are transmitted by the SPIHT algorithm which has generated bit stream. At the decoder, SPIHT is used to decode the bit stream received and the inverse transform is used to obtain the reconstructed speech[6].

4.4 Psychoacoustic model

For each spectral component an individual masking threshold is generated. The overall masking threshold follows from superposition of the individual thresholds, which is carried out by simply adding up the threshold at the corresponding frequencies. This masking threshold determines the maximum quantization noise energy that can be added to the original signal to keep the noise inaudible.

5 Results of implementation

Table 1 , Table 2 and Table 3 show the results obtained for complete reconstruction. The results show the comparison of the plain LPC Coder, voice excited LPC coder with DCT of the residual signal, voice excited LPC coder with MLT of residual signal , voice excited LPC coder with MLT-SPIHT of the residual signal and the improved performance using masking method.

Plain LPC Coder				Voice Excited LPC coder with DCT of Residual signal		
Signal	Compression Ratio	SNR	PRD	Compression Ratio	SNR	PRD
Target.wav	14.4:1	-18.08	8.02	16:1	.2023	4.84
John.wav	12.2:1	-21.45	10.12	16.67:1	.2308	4.70
Mace.wav	13.61:1	-20.20	10.02	16.34:1	.2956	5.69

Table 1

Voice Excited Coder with MLT of Residual signal				Voice Excited LPC coder with MLT-SPIHT of Residual signal		
Signal	Compression Ratio	SNR	PRD	Compression Ratio	SNR	PRD
Target.wav	20.54:1	.3232	4.70	21.58:1	.3151	1.01
John.wav	21.63:1	.2319	4.18	21.63:1	.2309	1.03
Mace.wav	20.03:1	.3426	3.89	20.59:1	.3033	.99

Table 2

Voice Excited LPC coder with MLT-SPIHT of Residual signal with masking of the signal		
Compression Ratio	SNR	PRD
21.58:1	.8151	1.01
21.63:1	.6212	1.03
20.59:1	.8123	.99

Table 3

5.1 Comparison

A comparison of the original speech sentences against the LPC reconstructed speech and the voice –excited LPC methods were studied. In both cases, the reconstructed speech has a lower quality than the input speech sentences. Both the reconstructed signals sound mechanized and noisy with the output of plain LPC vocoder being nearly unintelligible. The LPC reconstructed speech sounds guttural with a lower pitch than the original sound. The sound seems to be whispered. The noisy feeling is very strong. The voice –excited LPC reconstructed file sounds more spoken and less whispered. The guttural feeling is also less and the words are much easier to understand. The speech that was reconstructed using Voice-excited LPC with the MLT-SPIHT of the residual signal sounded better, but still sounded muffled.

Looking at the SNR computed in Table 1 and Table 2 , it is obvious that the first sound is very noisy, having a negative SNR. The noise in this file is even stronger than the actual signal . The voice excited LPC encoded sound is far better, and its SNR, although barely , is on the positive side. However, even the speech coded with the improved voice-excited LPC e.g. Voice-excited LPC with the MLT-SPIHT of the residual signal does not sound exactly like the original signal.

The LPC method to transmit speech sounds has some very good aspects, as well as some drawbacks. The advantage of vocoders is a very low bit rate compared to what is achieved for sound transmission. On the other hand, the speech quality achieved is quite poor. The Voice-excited LPC with the MLT-SPIHT of the residual signal gives better compression ratio and SNR but even then the quality achieved is still not very good.

The application of psychoacoustic model of masking has improved the quality of the signal as seen from the result shown in Table 3.

6 Conclusion:

This paper has presented a comparison of different speech compression techniques based on Modulated Lapped Transform and SPIHT. The results show clearly that significant savings may be obtained if the MLT is used in place of DCT. The results presented have also highlighted the advantage of the SPIHT algorithm, combined with relevant transform coefficient relationships, to scalable coding of speech as the algorithm is designed with the aim of producing an embedded bit stream.The reduction in bandwidth can be more significant if the masking model is included in the coding.

References:

[1] M.H.Johnson and A Alwan, " Speech Coding: Fundamentalsand Applications," in the Encyclopedia of Telecommunication, Wiley, December, 2002

[2] Ozgu Ozun, Philipp Steurer and Deniel Thell," Wideband Speech Coding with Linear Predictive Coding (LPC) " in EE214A: Digital Speech Processsing , Winter 2002

[3]Amir Said and Willium A Pearlman, "A new, fast and efficient image codec based on Set Partitioning in hierarchical trees," IEEE Transactions on Circuits and Systems For Video Technology, vol. 6, no. 3, pp. 243-250, June 1996

[4] H.Farsi, A.M.Kondoz, "A preprocessing method for pitch Smoothing and more speech regularity, IEE Electronics Letters, Vol.37, no.21,Oct 2001

[5] Lu Z, Kim DY, Pearlman WA, " Wavelet Compression of ECG signals by the Set Partitioning in Hierarchical Trees Algorithm", in IEEE Transactions on Biomedical Engineering Vol. 47, no. 7 PP. 849-856, June 2000.

[6] M.Hans and R.W Schafer, " Lossless Compression of digital audio," IEEE Signal Proc. Mag. Vol. 18, PP 21-32, July 2001

[7] Changyou Jing and Heng-MingTai," A fast algorithm for computing modulated lapped transform , in Electronics Letters, 7th June 2001 vol.37 no.12

[8] De4bargha Mukherjee, Sanjit .K.Mitra, Image resizing in the Compressed domain using Subband DCT, in IEEE Transactions on Circuits and Systems for Video Technology, Vol. 12 No. 7, July 2002

[9] Shapiro J.M. (1993), "Embedded image coding using zerotrees of wavelet coefficients " in IEEE Transactions on signal Processing, Vol. 41, No.12,pp.3445-3462.

[10] C.Sturt, S.Villette, A Kondoz, " LSF quantisation for pitch synchronous speech coders", IEEE International Conference on Acoustic Speech and Signal processing, Hong Kong, April 2003.