5-6-2000

Linear prediction incorporating simultaneous masking

J Lukasiak
University of Wollongong, jl01@ouw.edu.au

I. S. Burnett
University of Wollongong, ianb@uow.edu.au

Joe F. Chicharo
University of Wollongong, chicharo@uow.edu.au

M. M. Thomson
Motorola Australian Research Centre

Linear prediction incorporating simultaneous masking

Abstract

Whilst linear prediction is the cornerstone of most modern speech coders, few of these coders incorporate the perceptual characteristics of hearing into the calculation of the linear predictor coefficients (LPCs). This paper proposes a method of incorporating simultaneous masking into the calculation of the LPCs. This modification requires only a modest increase in computational complexity and results in the linear predictor removing more perceptually important information from the input speech signal. This results in a filter that better models the formants of the input speech spectrum. The net effect is that an improvement in quality is achieved for a given bit rate or alternately a bit rate reduction can be achieved while maintaining perceived quality. These results have been confirmed through subjective listening tests.

Disciplines

Physical Sciences and Mathematics

Publication Details

This paper originally appeared as: Lukasiak, J, Burnett, IS, Chicharo, JF & Thomson, MM, Linear prediction incorporating simultaneous masking, ICASSP '00. Proceedings. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing, 5-9 June 2000, vol 3, 1471-1474.

LINEAR PREDICTION INCORPORATING SIMULTANEOUS MASKING

*J. Lukasiak, I.S. Burnett, J.F. Chicharo, M.M. Thomson**
Whisper Laboratories, TITR
University of Wollongong
Wollongong, NSW, Australia, 2522
*Motorola Australian Research Centre, Botany, NSW, Australia, 2019

ABSTRACT

Whilst linear prediction is the cornerstone of most modern speech coders, few of these coders incorporate the perceptual characteristics of hearing into the calculation of the linear predictor coefficients (LPC's). This paper proposes a method of incorporating simultaneous masking into the calculation of the LPC's. This modification requires only a modest increase in computational complexity and results in the linear predictor removing more perceptually important information from the input speech signal. This results in a filter that better models the formants of the input speech spectrum. The net effect is that an improvement in quality is achieved for a given bit rate or alternately a bit rate reduction can be achieved while maintaining perceived quality. These results have been confirmed through subjective listening tests.

1. INTRODUCTION

Linear prediction forms an integral part of almost all modern day speech coding or speech compression algorithms. The primary reason for this popularity is that linear prediction provides a relatively simple and well founded technique for removing the redundancy from a speech signal, thus aiding in compression or bit rate reduction. Linear prediction determines and removes redundancy by removing the short term correlations of the input signal.

Whilst linear prediction is widely used in speech coding it was not originally developed specifically for speech coding but rather for the more general field of signal processing. The result of this is that the linear predictor used for speech coding does not exploit many of the well known perceptual properties of hearing. These perceptual properties include the non-linear frequency response of the ear and simultaneous masking amongst many others and are well defined in many texts such as [1]. Previous authors such as [2][3][4] have attempted to incorporate some perceptual properties into the calculation of the linear predictive filter. These methods have reported good results primarily by incorporating the non linear frequency response of the ear into the linear predictive filter analysis. This is achieved by warping the frequency axis to simulate the response of the ear prior to calculating the filter parameters. Hermansky [4] also included equal loudness perception and the intensity-loudness power law into the calculation of the filter. Whilst these attempts reported good results none of them attempted to incorporate simultaneous masking into the calculation. Simultaneous masking occurs in the frequency domain when a high amplitude sound causes adjacent lower amplitude sounds to become inaudible [1]. This property has been widely used in many audio coding techniques as a tool to determine the optimal quantisation step size required to code the input [5]. This reduces the bit rate required for transmission whilst maintaining the perceptual quality of the sound.

This paper proposes a method of incorporating simultaneous masking into the calculation of the linear predictive filter. The approach used is to fit the linear predictive spectrum only to the unmasked samples of the input spectrum. The motivation for this technique is to ensure no complexity is wasted modeling the masked regions, thus allowing the unmasked regions to be better represented. This allows the filter to remove more perceptually important information from the signal than the standard technique. The resultant residual signal remaining after exciting the filter with input speech thus consists of less perceptually important information. This characteristic allows the subjective quality of the synthesized speech to be maintained with a more coarsely quantised residual signal. Alternatively the speech quality is improved for a given quantisation level. These results have been confirmed through subjective listening tests.

The paper is organized as follows. In section 2 the method is outlined and a mathematical analysis presented. In section 3 experimental results are presented and discussed. Finally the major points are summarized in section 4.

2. SIMULTANEOUSLY MASKED LINEAR PREDICTIVE COEFFICIENTS (SMLPC)

2.1 Overview of Technique

A Block diagram of the SMLPC method is shown in figure 1. Initially the input speech is transformed to its Power Spectrum via a Fast Fourier Transform (FFT). The power spectrum is then analysed using a psychoacoustic model. This model determines the frequencies that are masked and is based on the model detailed in [6], with the parameters modified to optimise the performance of SMLPC. A modified power spectrum is then produced by taking those frequencies deemed masked and zeroing their value. This results in a power spectrum that contains only unmasked information. Recognising that the

0-7803-6293-4/00/$10.00 ©2000 IEEE.

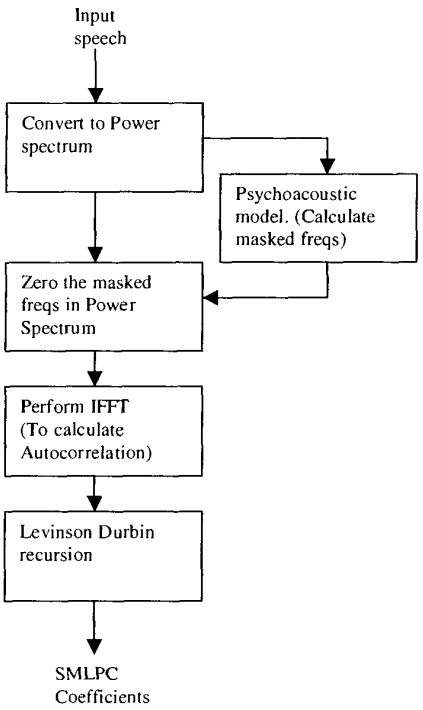

Figure 1. Block diagram of the SMLPC method.

autocorrelation of a discrete stochastic signal is the inverse Discrete Fourier Transform (IDFT) of the power spectrum, the perceptually altered power spectrum is transformed to the autocorrelation function of the unmasked speech. A perceptually altered Linear Predictor can now be easily calculated using the Levinson Durbin recursion [7]. In the forgoing discussion we refer to this modified Linear Predictor scheme as Simultaneous Masked Linear Predictor (SMLPC).

2.2 Mathematical Analysis of SMLPC

In this section we present an analysis of the mathematical operations employed by the SMLPC. The MSE solution for the standard LPC's ($a_p(k)$) can be reduced using the autocorrelation method [8], to:

$$R(l) = \sum_{k=1}^{p} a_p(k)R(l-k) \qquad l = 1.....p \qquad (1)$$

Noting that the autocorrelation values (R(n)) are the inverse discrete Fourier transform of the Power Spectral density P(k) we can state:

$$R(n) = \frac{1}{N}\sum_{k=0}^{N-1} P(k)e^{jwkn/N} \qquad n = 0.....N-1 \qquad (2)$$

If the calculation of R(n) above is modified to only operate on the perceptually important (unmasked) values of k then the autocorrelation becomes:-

$$R(n) = \frac{1}{L}\sum_{unmaskedl} P(l)e^{jwln/N} \qquad n = 0.....N-1 \qquad (3)$$

Where L represents the number of unmasked frequency bands of k from (2).

Substituting the above autocorrelation sequence (3) into (1) gives:-

$$R(n) = \sum_{k=1}^{p} a_p(k)\left(\frac{1}{L}\sum_{unmaskedl} P(l)e^{jwl(n-k)/N}\right) \qquad n = 1.....p \qquad (4)$$

It is clear that (4) solves the mean square solution for $a_p(k)$ using only the unmasked values of K. Also by interchanging the order of operation it is evident that 1/L is common to both the right and left hand sides of (4) and thus can be removed. This results in each summation term being equal to only the sum of the unmasked values of P(k) multiplied by the respective harmonic component. The sum of only the unmasked values of P(k) is identical in value to the sum over all K with the masked values of P(k) set to zero.

The above analysis confirms that the zero masked LPC fits only to unmasked regions and simply ignores the masked regions in its calculation of the LP coefficients. The fact that only the unmasked regions are modeled allows the SMLPC to achieve a better fit to these regions as complexity is not wasted in attempting to model masked regions.

An alternate approach to analysing the effect of the SMLPC is to view the predictor error in the frequency domain. The mean squared prediction error can be expressed as [9] :

$$E = \frac{G^2}{2\pi}\int_{-\pi}^{\pi} \frac{\left|S(e^{jw})\right|^2}{\left|H(e^{jw})\right|^2} dw \qquad (5)$$

Where G is filter gain, $S(e^{jw})$ is the input speech in the frequency domain and $H(e^{jw})$ is the frequency response of the filter. From (5) it can be deduced that minimizing E is equivalent to minimizing the ratio of the input energy spectrum to the squared magnitude of the frequency response of the filter. It can be seen that zeroing the power spectrum (numerator of equation) at any particular frequency, causes the difference between the model and the spectrum at that frequency to have no contribution to the integral of the ratio over the entire spectrum. The result is that the zeroed (masked) regions have no effect in calculating the linear predictive coefficients.

The preceding analysis was confirmed experimentally by modifying the IDFT to only operate on the unmasked coefficients and comparing the result to that obtained by zeroing the masked

coefficients in a standard IDFT. The results obtained were identical.

2.3 Computational Complexity

The computational complexity of the SMLPC is increased when compared to the standard LPC. However, this includes calculation of the psychoacoustic model parameters which remain available for other coding tasks such as quantisation. In standard LPC, calculation of the autocorrelation requires $(p+1)N_w$ operations [9]. Where p is filter order and N_w is the window size. The SMLPC uses an FFT and requires $N_f \log_2 N_f$ multiplications plus $N_f/2$ comparisons to calculate the autocorrelation function. Where N_f is the FFT length used. The SMLPC also requires approximately $2N_f + 700$ operations in calculation of the psychoacoustic parameters. Both methods require approximately p^2 operations to solve the matrix equations. The configuration in this paper used $N_w = 240$, $p = 10$ and $N_f = 512$. The complexities in this case are SMLPC = 5892 operations and standard LPC = 2740 operations. The computational demand of SMLPC can be made approximately equal to that of the standard LPC by using FFT of length 256. This size transform has little effect on the performance of SMLPC for 4Khz band limited speech.

3. EXPERIMENTAL RESULTS

3.1 LP Spectral Estimate

It is well known that the spectrum of a LP filter provides a good estimate to the spectrum of the input speech. To examine the effect of SMLPC on the accuracy of the spectral estimate, 10^{th} order LPC and SMLPC were calculated for a number of voiced and unvoiced speech segments. The spectra produced by both methods were then compared to the actual speech spectrum. A typical example of the spectrum produced is shown in Figure 2. The masked frequencies are indicated by shading. It is clearly evident that the SMLPC spectra is a more accurate representation of the input speech spectra in unmasked regions. The increased accuracy often results in the SMLPC modeling 2 distinct peaks of the input spectrum whilst the standard LPC produces only a single peak between the two peaks in the input spectrum. The poles of the SMLPC are also generally shifted away from largely masked sections of the spectrum.

Using 10^{th} order filters and hamming windowed speech segments, the log spectral distortion between the input speech and the respective LPC estimates in the unmasked regions of the spectrum were calculated. The results for a number of sentences from the Timit database spoken by both male and female speakers are shown in Table 1. The spectral distortion is computed as:

$$sd(m) = \sqrt{\frac{1}{L/2} \sum_{l\ unmasked} \left[10 \log \frac{|S(k)|^2}{|H(k)|^2} \right]^2} \qquad (6)$$

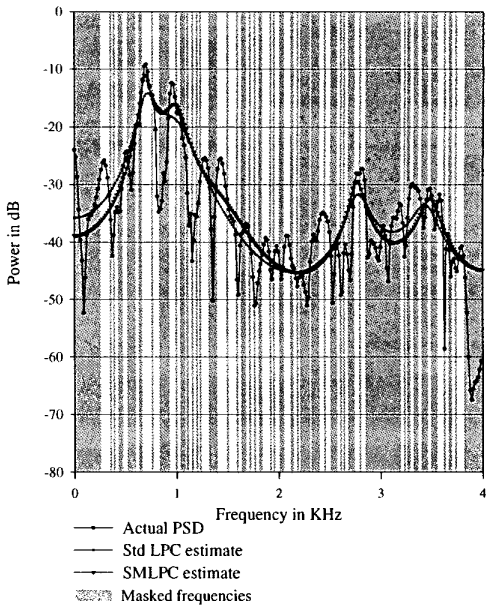

Figure 2. Spectral estimates. The solid line is the input PSD, dashed line is SMLPC estimate and dotted is standard LPC estimate. The masked frequencies are indicated at the bottom of the graph.

Gender of Speaker	SMLPC unmasked SD	Std LPC unmasked SD	Percentage Improvement
Male	2.94	3.02	2.7211
Male	3.25	3.47	6.7692
Female	4.08	4.28	4.902
Female	3.47	3.54	2.0173

Table 1. Spectral Distortion of LP estimates

Where S(k) represents the input speech spectrum, H(k) represents the LP filter spectrum and L is the number of unmasked coefficients.

$$SD = \frac{1}{M} \sum_{m=1}^{M} sd(m) \qquad (7)$$

Where M represents the number of frames. The results indicate that the SMLPC reduces the spectral distortion in the unmasked regions of the spectrum. This supports the claim that SMLPC provides a more accurate spectral estimate thus allowing the filter to remove more of the perceptually important information from the input speech than a standard Linear predictor.

3.2 Analysis of the LPC Residual

Figure 3 shows the difference between the residual signal power spectrums for a standard LP filter and the SMLPC filter over a typical speech segment. A positive value indicates that the standard LPC residual has greater power and a negative signal indicates that the SMLPC residual is of higher power. The figure shows that in ranges of frequency that are largely free of masking or exhibit regular spaced masking (strongly voiced) such as between 200Hz and 1300Hz, the SMLPC residual has lower power than the standard LPC residual. Also in regions that are heavily masked such as between 2700Hz and 3500Hz the SMLPC residual has greater magnitude than the standard LPC residual. These results reinforce the claims that the SMLPC removes more of the perceptually important unmasked information from the signal than a standard LPC.

3.3 Subjective Listening Tests

To test the performance of the SMLPC in an existing speech coder, a version of the Federal standard 1016 CELP coder [10] was modified to use the SMLPC in place of the standard LPC. All other parameters including codebooks were left unaltered.

Synthesised speech was produced for a variety of male and female speakers. Double blind comparative A/B tests where A and B were played twice in opposite order and the listener had to indicate their preference for A, B or neither each time, were conducted using a substantial listener base. The results obtained indicated that the SMLPC synthesized speech was preferred for 55% of male speech whilst the standard CELP was preferred on only 17.5% of occasions. For female speech no clear preference was evident. The results clearly indicate that SMLPC offers a significant improvement for male speech whilst not degrading the perceptual quality of female speech. The differential in improvement between male and female speech may be attributed to the fact that at low frequencies the bandwidths of the critical bands are only around 100Hz. This results in some critical bands containing no pitch harmonics for female speech. Thus the masking threshold for these bands is very small and few frequencies are deemed masked even though the information within the band may be perceptually unimportant. One possible solution to improve the performance of SMLPC for female speech would be modifying the masking function according to pitch. This approach is similar to that proposed by Chen [11] where the masking function is modified to follow the pitch harmonics as well as the formant peaks.

4. CONCLUSION

A new technique to incorporate simultaneous masking into the calculation of a Linear Predictive filter has been developed. The technique involves use of a psychoacoustic model to determine the masked frequencies and then modifies the autocorrelation function to utilize these masked frequencies. This is achieved by zeroing the masked coefficients in the power spectrum before transforming this to the autocorrelation function via an IDFT operation.

Experimental results have shown that the technique better models the spectrum in the unmasked regions and thus removes more of the perceptually important information from the input speech signal than a standard LPC.

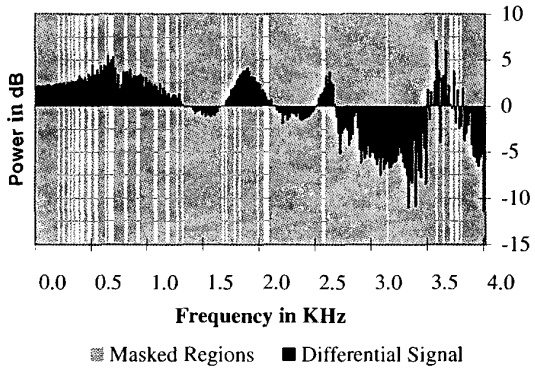

Figure 3. Difference between standard LPC and SMLPC residual power spectrum for a typical speech segment. The shaded areas indicate the masked frequencies.

5. ACKNOWLEDGEMENTS

J.Lukasiak is in receipt of an Australian Postgraduate Award (Industry) and a Motorola (Australia) Partnerships in research Grant. Whisper Laboratories is funded by Motorola and the Australian Research Council.

6. REFERENCES

[1] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, Sydney 1997.

[2] H.W. Strube, "Linear Prediction on a Warped Frequency Scale", Journal of the Acoustical society of America, Vol.68, no.4, pp1071-1076,1980.

[3] Y. Nakatoh, T. Norimatsu, A.Heng Low and H. Matsumoto, "Low Bit Rate Coding for Speech and Audio using Mel Linear Predictive Coding(MLPC) Analysis", Proc. Of ICSLPA, 1998.

[4] H. Hermansky, "Perceptual Linear Predictive Analysis of Speech", Journal of the Acoustical Society of America, Vol.87,no.4, pp1738-1753, April 1990.

[5] N. Jayant, J. Johnston and R. Safranek, "Signal Compression Based on Models of Human Perception", Proc. of IEEE, Vol. 81, No.10, October 1993.

[6] J.D. Johnston, "Transform coding of audio signals using perceptual noise criteria", IEEE J. on selected Areas in Comm., vol.6, pp314-323, February 1988.

[7] J. Makhoul, "Linear Prediction: A Tutorial Review", Proc. IEEE, Vol. 63, pp.561-580, 1975.

[8] J. Makhoul and J. Wolf, "Linear prediction and the Spectral Analysis of Speech", BBN Report No.2304, August 1972.

[9] L.B. Rabiner and R.W. Schafer, Digital Processing of speech Signals, Prentice Hall, New Jersey, 1978.

[10] National Communication System, details to assist in implementation of Federal Standard 1016 CELP, Office of the manager National Communication System, Arlington.

[11] J. Chen and A. Gersho, "Adaptive post filtering for quality enhancement of coded speech", IEEE transactions on speech and Audio processing, Vol.31, pp. 59-71, Jan 1995.