Paper

Source enhanced linear prediction of speech incorporating simultaneously masked spectral weighting

Jason Lukasiak and Ian S. Burnett

Abstract — Linear prediction is the cornerstone of most modern speech compression algorithms. This paper proposes modifying the calculation of the linear predictor coefficients to incorporate a weighting function based on the simultaneous masking property of the ear. The resultant prediction filter better models the perceptual characteristics of the source and results in the removal of more perceptually important information from the input speech signal than a standard LP filter. When employed in a low rate speech codec the net effect is an improvement in subjective quality, with no increase in transmission rate and only a modest increase in computational complexity.

Keywords — linear prediction, psychoacoustics, masking, LPC.

1. Introduction

Linear prediction (LP) 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 human hearing. These perceptual properties include the nonlinear frequency response of the ear and simultaneous masking, amongst others and are well defined in texts such as [1]. Previous authors [2 – 4] have incorporated some perceptual properties into the calculation of the linear predictive filter. These authors have reported good results, primarily by warping the frequency axis to simulate the nonlinear frequency 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 authors reported good results none of them attempted to incorporate simultaneous masking into the filter calculation.

Simultaneous masking occurs in the frequency domain when a high amplitude sound causes adjacent lower amplitude sounds to become inaudible. This property has been widely used in many audio coding techniques, such as MPEG4 [5], as a tool to determine the optimal quantization step size required to code the input and thus allows perceptually transparent compression of the audio signal. However, the use of simultaneous masking in speech coding algorithms has been very limited due to the increased computational demand required to perform quantization of the signal in the frequency domain.

This paper proposes a method for modifying the calculation of the linear prediction coefficients (LPC) to better model the characteristics of the source. This is achieved by incorporating a weighting function based on the simultaneous masking property of the ear into the calculation of the LPC. This approach fits 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, with the resultant residual signal consisting of less perceptually important information. This characteristic allows the subjective quality of the synthesized speech to be improved for a given residual quantization scheme. This paper presents results confirming this characteristic using objective measures and subjective listening tests.

The paper is organized as follows. In Section 2 an overview of linear prediction and human auditory perception is detailed. The new linear prediction method is given in detail in Section 3. Objective and subjective results are provided in Section 4. Finally, the major points are summarized in Section 5.

2. Background

2.1. Linear prediction analysis

The use of a linear predictor in speech coding relies upon the fact that speech can be modeled as the output of a time varying linear system [6]. The development of this model is linked to the use of lossless acoustic tubes to represent the speech production process and is detailed in [6]. Figure 1 represents a simplified representation of this model.

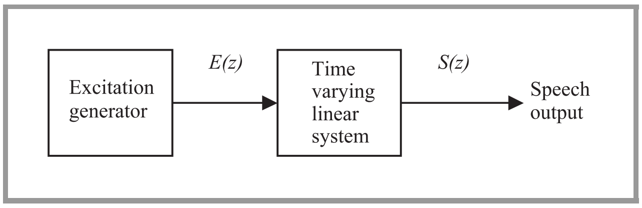

Fig. 1. Source-system model of speech production.

The transfer function representing the linear system in Fig. 1 can be described using an all pole (autoregressive (AR)) system as:

$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \qquad (1)$$

where p is the order of the filter, G is the gain of the system and the a_k are the predictor coefficients.

The popularity of the AR model stems from the fact that if a sufficient order is used the filter can accurately model the speech production system and also the filter parameters can be calculated in a straightforward and efficient manner [6].

To allow linear prediction the input waveform must be a stationary random process. However, as speech is a nonstationary random process some modification is required. It has been determined that a speech signal is stationary over a period of approximately $20-30$ ms [7]. Thus to utilize linear prediction in speech coding it is necessary to divide the input speech into frames of approximately 20 ms length and update the linear prediction coefficients for each of these frames. A number of methods to solve for the predictor coefficients to achieve a minimum mean square error for a given frame have been developed [6]. The most popular of these is the autocorrelation method and this is the method used in this paper.

The mean square error (MSE) solution for the standard LPC's (a_k) can be reduced using the autocorrelation method [8], to:

$$R(l) = \sum_{k=1}^{p} a_k R(l-k), \quad l = 1 \ldots p \qquad (2)$$

where $R()$ is the autocorrelation function of the input speech frame. The recursive Levinson-Durbin [9] algorithm is then used to solve for the filter coefficients a_k.

2.2. Overview of human auditory perception

The human auditory system is a highly complex system. Sounds presented to the ear are not all perceived equally but are governed by a number of nonlinear operations. Humans can hear sounds in the range of approximately 50 Hz $-$ 16 kHz [1], however, these frequencies are not all perceived with equal sensitivity. This phenomenon leads to a set of curves called equal loudness curves [1] which indicate the perceived loudness of a fixed amplitude tone, as the frequency of the tone is varied. Directly related to the equal loudness curves is the threshold of hearing curve. This curve represents the minimum amplitude that is audible for a given frequency.

The frequency scale of the human ear also acts as a set of overlapping bandpass filters. These filters are called critical band filters and the pass band of each individual band pass filters is termed a critical band. The nature of each critical band is that all frequencies within a band are perceived equally by the ear [1]. The critical bands are not of equal width but increase in bandwidth as their center frequency increases. This results in better frequency resolution at lower frequencies. Scharf [10] proposed that the critical bands could be adequately represented by a set of non overlapped rectangular filters. This greatly reduces the complexity of critical band analysis.

Masking occurs when a loud sound causes other softer sounds to become inaudible. Two types of masking can occur, namely simultaneous and temporal masking. Simultaneous masking occurs when a low intensity but audible sound is made inaudible by a higher intensity adjacent sound occurring at a simultaneous moment in time. Temporal masking occurs when a loud tone causes softer tones occurring before and after the tone to become inaudible [1].

The masking and equal loudness phenomena have been researched extensively and have led to the development of psychoacoustic models that allow the auditory system to be accurately modeled. Detailed descriptions of these models can be found in many well recognized texts such as [1].

Incorporating the perceptual characteristics of the auditory system into audio and speech coding allows coders to operate more efficiently at reduced bit rates by distributing the available bits according to the perceptual nature of the signal. This allows the coder to reproduce a perceptually unaltered signal at a greatly reduced bit rate when compared with coders that ignore the perceptual characteristics. An example of this principle is the MPEG4 [5] audio coding standard. Perceptual modeling provides the crux of this coder and allows the coder to produce high quality audio signals at a relatively low bit rate.

3. Linear prediction incorporating simultaneously masked spectral weighting (SMWLPC)

3.1. Motivation

The motivation for this technique was to allow a simple and computationally efficient means of better exploiting the perceptual characteristics of human hearing in low rate LP based speech codecs. Traditionally, perceptual distortion is only exploited in low rate LP speech coding by using a noise shaping weighting filter [11] when coding the residual signal. This filter is used to weight the error signal when searching for the optimal excitation signal to represent the LP residual. This weighting filter de-emphasises the frequency regions corresponding to the formants of the input speech. This de-emphasis exploits the masking characteristic of the ear in that larger errors are imperceptible in louder sections of the input speech than in quieter sections. Whilst the use of this weighting filter has produced good results [12] and been widely accepted, it uses only a basic model of the perceptual characteristics of the ear. Authors such as Sen [13] and Burnett [14] have reported improved performance by employing more sophisticated perceptual models when searching for the excitation signal that minimizes the perceptually weighted MSE to the LP residual. The improved results reported by [13] and [14] have been at the expense of a large increase in computational complexity. This increase in complexity is due to the fact that each perspective excitation signal tested must be transformed to the frequency domain and the error signal then multiplied with the respective perceptual model.

SMWLPC attempts to exploit more sophisticated perceptual models than the filter proposed in [11]. Incorporating these models into the LP filter allows SMWLPC to remove more perceptually important information from the input signal than standard LPC thus resulting in a residual signal that contains less perceptually important information. By exploiting the perceptual models upfront in the LP filter, computational complexity is dramatically reduced when compared to methods where the complex perceptual models are used when quantizing the LP residual [13]. This reduction is due to the fact that only a single transform and weighting multiplication per frame of input speech is required. Also by incorporating the complex perceptual models into the LPC, SMWLPC can be easily adapted to any existing LP based speech coding algorithm by simply replacing the standard LP filter.

The method selected limits the increase in computational complexity over standard LP filtering by maintaining the use of traditional recursive solutions in calculation of the LPC's. This also maintains a stable autoregressive structure that can be directly employed in any LP based speech codec. Standard linear prediction minimizes the error equally across the entire frequency spectrum of the input speech. This approach fails to exploit many of the well known perceptual properties of hearing. The SMWLPC technique employs a method of incorporating simultaneous masking into the calculation of the linear prediction coefficients. This allows the error to be minimized only in the sections of the input spectrum that are unmasked. This is achieved by first determining which frequencies in the input signal are simultaneously masked and then ignoring them in the calculation of the LPC.

3.2. Method

A block diagram of the SMWLPC method is shown in Fig. 2. Initially the power spectrum (frequency domain) of the input speech is calculated via a fast Fourier transform (FFT). A masking threshold function is then calculated for each discrete frequency. The calculation of this function is detailed in Section 3.3. The masked input frequencies are then determined. This is achieved by comparing the power spectrum of each discrete frequency to the masking threshold for that frequency. If the power spectrum is less than the masking threshold or the threshold of hearing, the frequency is deemed masked. A modified power spectrum is then produced by taking those frequencies deemed masked and zeroing their value. This method is equivalent to generating a spectral weighting function whose values are unity for unmasked frequencies and zero for masked frequencies or frequencies whose power is below the threshold of hearing and then multiplying the input spectrum by this weighting function. The result is a power spectrum that contains only unmasked information. Recognizing that the 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 then be easily calculated using the well known Levinson-Durbin recursion [9].

3.3. The masking threshold function

The psychoacoustic model used to calculate the masking threshold function is based on that proposed in [15] with the parameters modified to optimize the performance of the SMWLPC. A block diagram of the method is shown in Fig. 3.

The input power spectrum is segmented into N non overlapped critical bands. Where N represents the number of critical bands that exist within the bandwidth of the input signal. For narrowband speech the input bandwidth is approximately 4 kHz and the number of critical bands is 18. The critical bands are described in Section 2.2 and are given in detail in [10]. The power spectrum lines within each critical band are summed together, this gives an energy estimate for each critical band. The combination of the N energy estimates is called the band energy waveform as referred to in Fig. 3.

To simulate the masking effect between critical bands, the band energy waveform is provided as the input to an inter-band masking calculator. The inter-band masking calculator convolves the band energy waveform with a spreading function (shown in Fig. 4) to produce a spread band energy waveform. The spreading function shown in Fig. 4 is identical to that given in [1] and has been derived from exhaustive psychoacoustic testing.

The spread band energy waveform is then used to determine an initial masking threshold function (IMTF) according to the following formula:

$$IMTF(i) = Energy(i) - O(i)\,, \qquad (3)$$

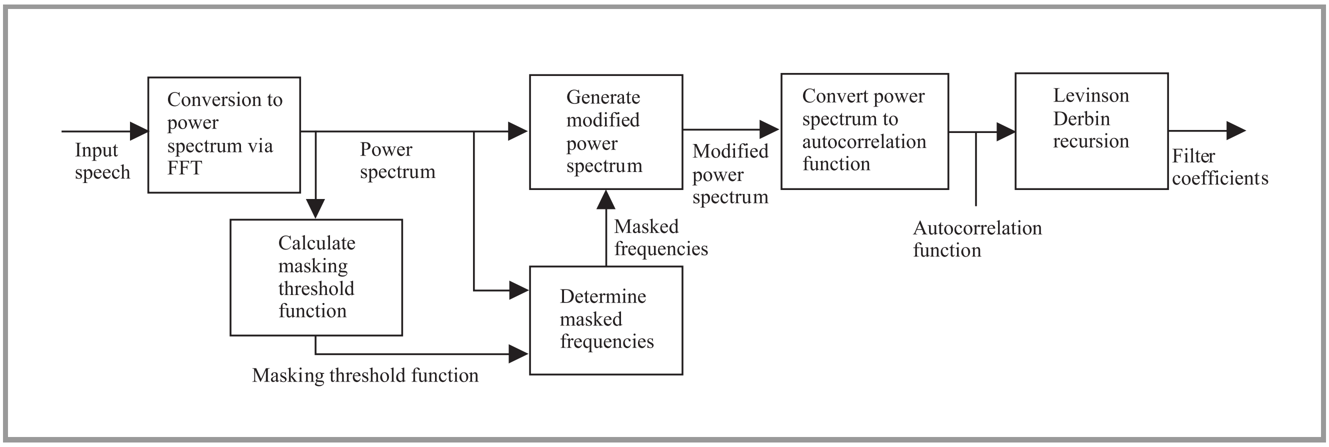
Input speech
Conversion to power spectrum via FFT
Power spectrum
Calculate masking threshold function
Masking threshold function
Generate modified power spectrum
Masked frequencies
Determine masked frequencies
Modified power spectrum
Convert power spectrum to autocorrelation function
Autocorrelation function
Levinson Derbin recursion
Filter coefficients

Fig. 2. Functional block diagram of SMWLPC.

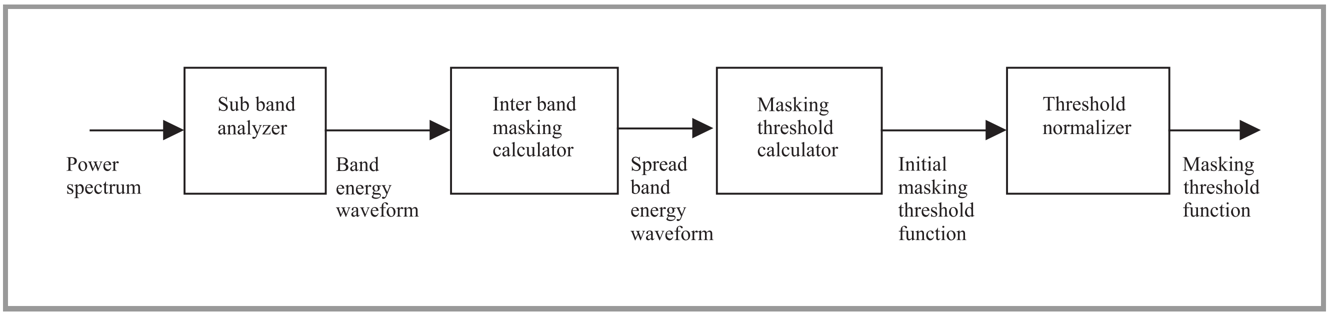
Power spectrum
Sub band analyzer
Band energy waveform
Inter band masking calculator
Spread band energy waveform
Masking threshold calculator
Initial masking threshold function
Threshold normalizer
Masking threshold function

Fig. 3. Block diagram of the masking threshold calculation.

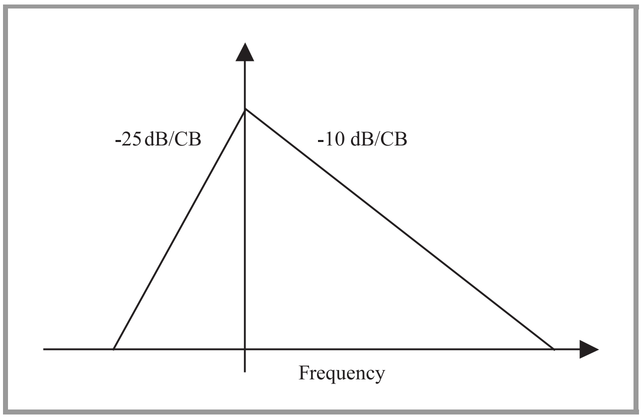

Fig. 4. Inter-band spreading function.

where $Energy(i)$ represents the total energy of the ith band of the spread band energy waveform measured in decibels; $O(i)$ is given by:

$$\begin{array}{ll} O(i) = 50\,, & \alpha < 0.2 \\ O(i) = \alpha(\beta + i) + (1-\alpha)\gamma, & \alpha \geq 0.2 \end{array} \tag{3.1}$$

where:

$$\alpha = \min\left(\frac{SFM}{SFM_{\max}}, 1\right), \tag{3.2}$$

$$SFM = 10\,\log\frac{G_m}{A_m}, \tag{3.3}$$

$SFM_{\max}$ is an empirically determined value; G_m is the geometric mean of the power spectrum; A_m is the arithmetic mean of the power spectrum; β and γ are empirically determined constant values that represent the tone masking noise and noise masking tone thresholds respectively.

The value of $SFM_{\max}$ suggested in [15] was -60 dB however, upon testing with a pure sine wave at 1 kHz the required $SFM_{\max}$ to give an alpha value of 1 was determined to be -40 dB. In Eq. (3.1) β and γ are set to 14.5 and 7 respectively; α is a measure of the flatness of the power spectrum, a value of 1 indicates a purely tonal signal and 0 represents pure noise. Equation (3.1) utilizes α to ensure that the correct mix of noise and tone thresholds is selected. The value for $O(i)$ in Eq. (3.1) differs greatly from that suggest in [15] where the definition is given as:

$$O(i) = \alpha(\beta + i) + (1-\alpha)\gamma \quad \text{for all } \alpha\,. \tag{4}$$

Setting Eq. (3.1) to a very large constant value for a very noise like signal $(\alpha < 0.2)$ ensures that for this type of input the IMTF is made small and designates virtually the entire spectrum unmasked. This overcomes the situation where Eq. (4) designates virtually the entire spectrum masked for such a signal thus leaving too few samples to successfully generate the filter coefficients. This characteristic was reported to cause distortions in the reconstructed speech in [16] and the modification overcomes this problem. Also in Eq. (4) $\gamma = 5.5$, contrasting with the increased $\gamma = 7$ in Eq. (3.1). This modification enhances the performance of

SMWLPC and was determined empirically through informal listening tests.

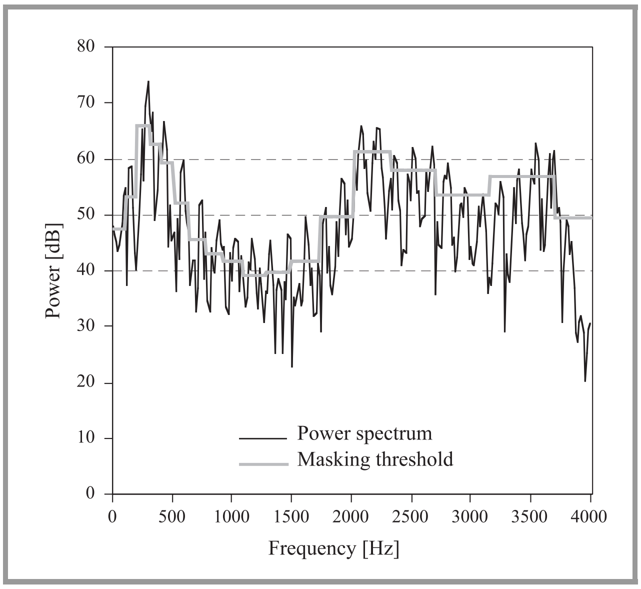

Fig. 5. Example of the masking threshold function.

The IMTF is then adjusted by the threshold normalizer to account for misestimation of the $Energy(i)$ values resulting from the shape of the spreading function. This results in a masking threshold function an example of which (with the corresponding power spectrum) is shown in Fig. 5.

3.4. Mathematical analysis of SMWLPC

SMWLPC is now analyzed mathematically to explore and contrast the differences between this approach and standard LPC.

The MSE solution for the standard linear predictive coefficients using the autocorrelation method was given in (2). As the input frame of speech is assumed to be a stationary random process the autocorrelation values $(R(n))$ can be computed via an inverse discrete Fourier transform of the power spectral density $P(k)$ [17]:

$$R(n) = \frac{1}{N}\sum_{k=0}^{N-1} P(k)e^{jwkn/N} \quad n = 0\ldots N-1. \qquad (5)$$

If the calculation of $R(n)$ in (5) is modified to only operate on the perceptually important (unmasked) values of k then the autocorrelation becomes:

$$R(n) = \frac{1}{L}\sum_{unmasked\ \ l} P(l)e^{jw\ln/N} \quad n = 0\ldots N-1, \qquad (6)$$

where L represents the number of unmasked frequency bands of N.

Substituting the autocorrelation sequence (6) into (2) gives:

$$\frac{1}{L}\sum_{unmasked\ \ l} P(l)e^{jw\ln/N} =$$
$$= \sum_{k=1}^{p} a_k\left(\frac{1}{L}\sum_{unmasked\ \ l} P(l)e^{jwl(n-k)/N}\right),\ n = 1\ldots p. \quad (7)$$

It is clear that the above equations solve the mean square solution for a_p using only the unmasked values of k. Also as $\frac{1}{L}$ is a common factor it can be removed from the equations. 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, which is identical in value to the sum over all k with the masked values of $P(K)$ set to zero.

The above analysis demonstrates that SMWLPC 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 SMWLPC to achieve a better fit to these regions as complexity is not wasted attempting to model masked regions.

An alternate approach to examining the effect of the SMWLPC is to view the predictor error in the frequency domain. This can be expressed as [6]:

$$E = \frac{G^2}{2\pi}\int_{-\pi}^{\pi}\frac{|S(e^{jw})|^2}{|H(e^{jw})|^2}\,dw. \qquad (8)$$

Equation (8) shows that minimizing E is equivalent to minimizing the ratio of the input energy spectrum $(S(e^{jw}))$ to the squared magnitude of the frequency response $(H(e^{jw}))$ of the model. 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 contribute nothing 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.

3.5. Computational complexity

The computational complexity of SMWLPC 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 quantization. In standard LPC, calculation of the autocorrelation requires $(p+1)N_w$ operations [6], where p is filter order and N_w is the window size. The SMWLPC 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 SMWLPC 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 SMWLPC = 5892 operations and standard LPC = 2740 operations. The computational demand of SMWLPC can be made approximately equal to that of the standard LPC by using an FFT of length 256. This size transform has little effect on the performance of SMWLPC for 4 kHz band limited speech.

3.6. Data windowing requirements

The psychoacoustic model given in [15] was based on audio signals sampled at 32 kHz or greater. Due to this sampling rate and also the ability to accept large delays in audio coding as they seldom operate in real time, the transform length is set to 2048. Using this length window produces frequency bins that are separated by less than 16 Hz and as the lowest frequency in the audio range is > 50 Hz, spectral leakage between frequency bins [17] has little effect even if a rectangular window were used. To adapt this model for use in narrowband speech coding with a sampling rate of only 8 kHz and a constraint on the maximum delay due to the need to operate in real time, the number of samples in the window is greatly reduced. This short transform length causes the frequency separation between adjacent frequency bins to become almost equal to the lowest pitch value for voiced speech of approximately 50 Hz. This characteristic causes spectral leakage across the frequency bins to have a large effect for low pitch speech. The leakage can act as an initial spreading function for low pitched speech and thus causes the masking threshold generated for this speech to become distorted. The authors have found that if a Hamming window of length 240 samples is used the effects of leakage are minimized and results have shown the masking threshold to be consistent across a range of pitch values. This window length is towards the upper limits used in speech coding but is common in low rate speech coders such as the FS1016 [18] 4.8 kbps CELP coder. If a shorter window length is used the spectral leakage can cause the masking thresholds to become inconsistent across the range of possible pitch values.

4. Experimental results

4.1. Objective results

4.1.1. LPC spectral estimate

The spectra of the linear predictive filter provides a good estimate of the spectra of the input speech. This relationship is clearly evident when examining Eq. (8) which shows that the predictor coefficients are calculated by minimising the ratio of squared error between the speech and filter spectra. This property of linear predictive filtering is widely exploited in harmonic coders to provide a bit rate effective means of transmitting the spectral envelope.

To examine the effect of SMWLPC on the accuracy of the spectral estimate, 10th order LPC and SMWLPC analyses were performed 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 Fig. 6. The masked frequencies are indicated by shading. It is clearly evident in Fig. 6 that the SMWLPC spectra is a more accurate representation of the input speech spectra in unmasked formant regions. As can be seen at around 800 Hz in Fig. 6 the increased accuracy often results in the SMWLPC modeling 2 distinct formant peaks of the input spectrum whilst the standard LPC produces only a single peak between the two peaks of the input spectrum. This better modeling of the perceptually important formant regions allows SMWLPC to remove more of this perceptually important information from the input speech that a standard LP filter.

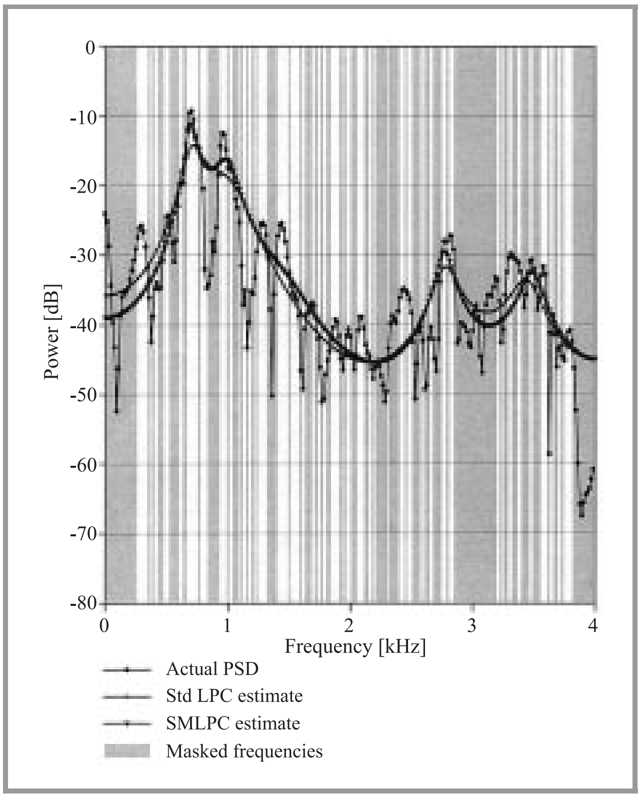

Fig. 6. Comparison of SMWLPC and standard LPC spectral estimates.

To obtain an objective measure for the amount of extra perceptually important information that is removed by SMWLPC the average weighted unmasked residual energy ($WURE$) was calculated using:

$$WURE = \frac{1}{L} \sum_{unmasked\ l} w(l) P_r(l)\,, \qquad (9)$$

where $P_r()$ is the PSD of the residual signal, $w()$ is a weighting function equal to a 40th order LP spectrum of the input speech normalized to have unity maximum and L is the total number of unmasked spectral lines. The use of the weighting function $w()$ in Eq. (9) places greater emphasis on the perceptually important formant regions of the input spectrum.

10th order LP analysis using both SMWLPC and standard LPC was performed on 10 input sentences (5 male/5 female). Frames of length 200 samples were used in the analysis with a linear predictive Hamming window of 240 samples having an overlap of 20 samples between frames. The $WURE$ of each frame with an α (from Eq. (3.2)) greater than 0.2 was calculated via Eq. (9) and the values averaged across the entire sentence. An α greater than 0.2 was used as these are the frames for which SMWLPC differs from standard LPC as explained in Section 3.3. The results of the analysis are shown in Table 1.

Table 1
Percentage greater WURE removed by SMWLPC

Sentence number	Speaker gender	SMWLPC % improvement
1	Male	4.7
2	Male	17.52
3	Male	3.95
4	Male	3.55
5	Male	10.5
6	Female	2.87
7	Female	3.84
8	Female	5.82
9	Female	2.51
10	Female	3.13

The results in Table 1 show the average percentage reduction in *WURE* for SMWLPC compared to standard LP for each input sentence. The results demonstrate that more perceptually important information was removed by SMWLPC for each of the input files. The average improvement for all sentences was 5.84%, this represents a significant improvement and indicates that SMWLPC removes significantly more perceptually important information from the input signal than standard LPC.

A typical example of the difference between the weighted residual power spectrums for a standard LP filter and the SMWLPC filter over a typical speech segment is shown in Fig. 7. A positive value indicates that the SMWLPC residual has greater power and a negative signal indicates that the standard LPC residual is of higher power. The masked frequencies are shaded. Figure 7 shows that in ranges of frequency that are largely free of masking or exhibit regular spaced masking (strongly voiced) such as between 0 Hz and 1500 Hz, the SMWLPC residual has lower power than the standard LPC residual. Also in regions that are heavily masked such as between 2700 Hz and 3500 Hz the SMWLPC residual has greater magnitude than the standard LPC residual. These results reinforce the claims that the SMWLPC removes more of the perceptually important unmasked information from the signal than a standard LPC.

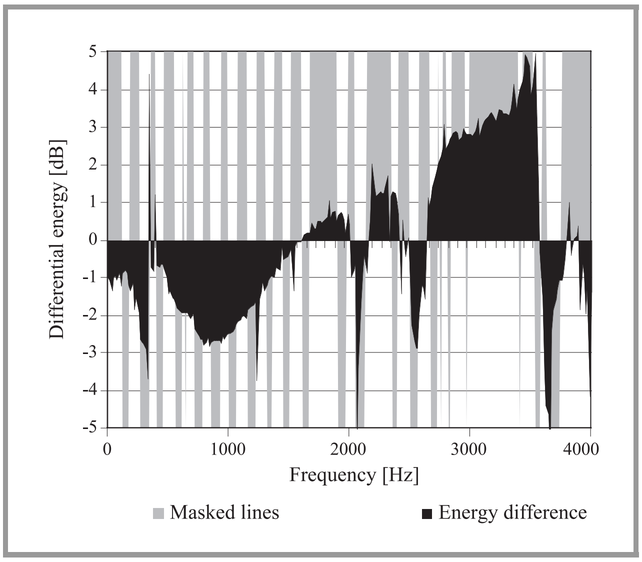

Fig. 7. Difference in weighted residual energy.

4.1.2. Quantization properties

The direct form LP coefficients shown in the calculations of Section 2.1 are susceptible to quantization noise [6]. Due to this characteristic they are rarely used in speech coding [19]. The most popular representation of the LP coefficients are line spectral frequencies (LSF). The LSF's are calculated from the direct form coefficients and their characteristics make them suitable for quantization. These characteristics include monotonically increasing order, strong intra and inter frame correlation and clustering together at formant frequencies [7].

To examine the effect that SMWLPC has on the correlation properties of the LSF's, the inter and intra frame correlation for both standard and SMWLPC LSF parameters were compared. This comparison showed no significant differences in the correlation values between the two methods. To verify this finding in a practical situation a vector linear predictor as proposed in [20] was calculated for both the standard LPC and SMWLPC LSFs respectively. The predictor produced is a square matrix that uses the LSF vector from the previous frame to estimate the LSF vector for the current frame by exploiting both intra and inter frame correlation. The spectral distortion between the predicted vector and the actual vector was calculated for each frame of a test sequence of 1000 frames and was then averaged across all frames. The spectral distortion was calculated via Eq. (10) and the results are shown in Table 2.

$$sd = \sqrt{\frac{1}{\frac{N}{2}} \sum_{k=0}^{\frac{N}{2}-1} \left[10 \log \frac{|S(k)|^2}{|H(k)|^2} \right]^2}, \qquad (10)$$

where N is the FFT length, $S(k)$ is the actual LPC spectrum and $H(k)$ is the predicted LPC spectrum.

Table 2
Average spectral distortion for the predicted LSF vector

	Average spectral distortion [dB]
SMWLPC	2.38
STD LPC	2.31

The results shown in Table 2 indicate that the spectral distortion was virtually identical for both LP methods. The difference of 0.07 dB is statistically insignificant as the resultant error would then be vector quantized and a final spectral distortion of less than 1 dB is known to produce transparent results for speech coding [21]. Achieving a virtually identical spectral distortion indicates that in practical situations SMWLPC maintains the high inter and intra frame correlation values of standard LPC LSF's and thus are suitable for high compression quantization schemes such as vector linear prediction.

4.2. Subjective listening tests

To test the performance of the SMWLPC in existing speech codecs, a version of the 4.8 kbps FS 1016 CELP coder [18] and a WI [22] coder operating at 2 kbps [16] were modified to use the SMWLPC in place of the standard LPC. The motivation for selecting the CELP and WI coders was to test the performance of SMWLPC in structures that code the LP residual signal in a closed loop and open loop method respectively. As the WI coder uses vector quantization of the LSF parameters the coder was set to operate with non quantized LSF's. This removed the need to retrain the LSF codebook for the SMWLPC and also ensured an unbiased evaluation of SMWLPC's effect on the perceptual content of the residual signal, with no effect from quantization errors of the LPC parameters. This modification was not necessary for the CELP coder as it uses scalar quantization of the LPC parameters which were found to match both the standard LPC and SMWLPC. All other parameters including codebooks were left unaltered.

Each of the coders was used to generate synthesized speech for 10 input speech sentences (5 male, 5 female) from the TIMIT database using both the standard LPC and SMWLPC. Subjective forced A/B comparison testing comprising 20 untrained listeners was conducted. To avoid statistical bias in the results, each sentence pair was played twice in each test with the order of the sentences being reversed. Thus the total test comprized the comparison of some 800 sentence pairs. The results are shown in Tables 3 and 4.

Table 3
A/B comparison results for the FS1016 CELP coder

Speaker gender	SMWLPC [%]	STD LPC [%]
Female	59.5	40.5
Male	53.5	46.5
Total	56.5	43.5

Table 4
A/B comparison results for the WI coder

Speaker gender	SMWLPC [%]	STD LPC [%]
Female	54.5	45.5
Male	57.5	42.5
Total	56	44

An alternative view of the results is to look at the majority listener preference for the particular sentences. These results are shown in Table 5.

Table 5
Majority preferred sentences

	SMWLPC [%]	STD LPC [%]	No preference [%]
CELP	70	30	0
WI	60	20	20
Total	65	25	10

The results clearly indicate a preference for the SMWLPC coded speech in all instances and for both coders. This clear preference is despite the fact that the coding structures for both coders were left unaltered. Modifying the quantization procedures for the residual signal to suit the SMWLPC characteristics by, for example, retraining codebooks and introducing search weighting functions that suit the SMWLPC characteristics such as that proposed in [13] could be expected to show further substantial improvements in the performance of the coders when using SMWLPC.

It is interesting to note that for the CELP coder the preference for female speakers using SMWLPC was higher than for males and for the WI coder this was reversed. It is a well known property that CELP coders sound better for male speakers due to the retention of phase (temporal) information but poor modeling of the harmonic structure in the coding process [23]. Conversely harmonic type speech coders such as WI coders are better suited to female speakers due to the retention of the harmonic structure but loss of the phase information [23]. It appears that by removing more of the perceptually important information from the input speech before the residual is coded, SMWLPC is able to overcome some of the short comings of a particular low rate coding algorithm.

The results presented have extended and support those reported in [16] and [24] where a significant preference for SMWLPC coded sentences was reported. In [16] mean opinion score (MOS) testing was conducted using a 2 kbps WI coder and the results showed an improvement in MOS score from 3.31 to 3.45 when the standard LPC was replaced by SMWLPC.

5. Conclusion

A new technique which modifies the calculation of the LPC to better model the source for low rate speech coding has been developed. The technique involves the use of a psychoacoustic model to determine the simultaneously masked frequencies and also the frequencies whose power falls below the threshold of hearing. This information is then used to weight the power spectrum of the input speech, producing a modified power spectrum that contains only unmasked information. A modified autocorrelation function is then generated via a DFT operation and standard recursive algorithms are used to solve for the LPC. Retaining the use of the standard recursive algorithms limits any increase in computational complexity and also ensures that a stable all pole filter is produced.

Experimental results have shown that the technique better models the spectrum in the unmasked formant regions and thus removes more of the perceptually important information from the input speech signal than a standard LP filter. Subjective listening tests using both CELP and WI coders has confirmed that this property improves the perceptual quality of the synthesized speech for a given residual coding method.

Acknowledgements

Jason 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.

References

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

[2] H. W. Strube, "Linear prediction on a warped frequency scale", *J. Acoust. Soc. Am.*, vol. 68, no. 4, pp. 1071–1076, 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", in *Proc. ICSLP*, 1998.

[4] H. Hermansky, "Perceptual linear predictive analysis of speech", *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1753, 1990.

[5] "MPEG4", ISO/IEC FCD 14496-3.

[6] L. B. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. New Jersey: Prentice Hall, 1978.

[7] A. M. Kondoz, *Digital Speech*. New York: Wiley, 1995.

[8] J. Makhoul and J. Wolf, "Linear prediction and the spectral analysis of speech", BBN report, no. 2304, August 1972.

[9] J. Makhoul, "Linear prediction: a tutorial review", *Proc. IEEE*, vol. 63, pp. 561–580, 1975.

[10] B. Scharf, "Critical bands", in *Foundations of Modern Auditory Theory*, J. Tobias, Ed. New York: Academic Press, 1970, pp. 159–202.

[11] M. Schroeder and B. S. Atal, "Predictive coding of speech signals and subjective error criteria", in *IEEE Trans. ASSP*, 1979, pp. 247–254.

[12] P. Kroon and E. F. Deprettere, "A class of analysis by synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbits/s", *IEEE J. Selec. Areas Commun.*, vol. 62, pp. 353–363, 1988.

[13] D. Sen, D. H. Irving, and W. H. Holmes, "PERCELP – perceptually enhanced random codebook excited linear prediction", in *Proc. IEEE W/shop Speech Cod. Telecommun.*, 1993, pp. 101–102.

[14] I. S. Burnett, "Hybrid techniques for speech coding", Ph.D. thesis, University of Bath, 1992.

[15] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria", *IEEE J. Selec. Areas Commun.*, vol. 6, pp. 314–323, 1988.

[16] J. Lukasiak and I. S. Burnett, "Exploiting simultaneously masked linear prediction in a WI speech coder", in *Proc. IEEE W/shop Speech Cod.*, 2000, pp. 11–13.

[17] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing*. New Jersey: Prentice Hall, 1996.

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

[19] G. S. Kang and L. J. Fransen, "Low-bit rate speech encoders based on line spectrum frequencies (LSFs)", NRL report, no. 8857, Naval Research Lab., Washington D.C., Jan. 1985.

[20] M. Yong, G. Davidson, and A. Gersho, "Encoding of LPC spectral parameters using switched adaptive inter frame vector prediction", in *Proc. ICASSP*, 1988, vol. 1, pp. 402–405.

[21] K. K. Paliwal and B. S. Atal, "Efficient vector quantisation of LPC parameters at 24 bits/frame", *IEEE Trans. Speech Audio Proc.*, vol. 1, no. 1, pp. 3–14, 1993.

[22] W. B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms", in *Proc. ICASSP*, 1995, vol. 1, pp. 508–511.

[23] J. Skoglund and W. B. Kleijn, "On time frequency masking in voiced speed", *IEEE Trans. Speech Audio Proc.*, vol. 8, no. 4, pp. 361–369, 2000.

[24] J. Lukasiak, I. S. Burnett, J. F. Chicharo, and M. M. Thomson, "Linear prediction incorporating simultaneous masking", in *Proc. ICASSP*, 2000, vol. 3, pp. 1471–1474.

Jason Lukasiak is currently a Ph.D. student in the Institute for Telecommunications Research at the University of Wollongong, Australia. He received a B.E. (Hons.) from the University of Wollongong in 1998 and immediately commenced studying for a Ph.D. He worked for BHP Slab and Plate products from 1987 to 1997 where his positions ranged from computer network technician to Electrical Project Engineer. During this time he studied part time receiving an electrical trades certificate, advanced certificate in computer technology and associate diploma of electrical engineering. Jason's Ph.D. research topic is scalable speech compression over a range of bit rates from $1-8$ kbps.
e-mail: j101@ouw.edu.au
Whisper Laboratories, TITR
University of Wollongong
Wollongong NSW 2522, Australia

Ian S. Burnett is Director of the Telecommunications Research Centre and a Senior Lecturer at the University of Wollongong, Australia. He received B.Sc. and M.Eng. degrees from the University of Bath, UK in 1987 and 1988, respectively. From 1987 he was with GEC Marconi Secure Radio working on digital communications and speech compression. In 1989 he returned to the University of Bath under a Vodafone scholarship, and completed a Ph.D. in "Hybrid Techniques for Speech Coding" in 1992. In 1993 he worked for Loughborough Sound Images on various real-time signal processing systems, before taking his present position. His current research interests lie primarily in speech/audio compression, audio scene analysis and multimedia. He is currently active in the ISO MPEG standardization concentrating on MPEG-21 and MPEG-7-Audio.
e-mail: j101@ouw.edu.au
Whisper Laboratories, TITR
University of Wollongong
Wollongong NSW 2522, Australia