telecommunications &
signal processing
laboratory

Time Windows for Linear Prediction of Speech

Peter Kabal

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

October 2003

2003/10/27

Abstract

This report examines the time windows used for linear prediction (LP) analysis of speech. The goal of windowing is to create frames of data each of which will be used to calculate an autocorrelation sequence. Several factors enter into the choice of window. The time and spectral properties of Hamming and Hann windows are examined. We also consider windows based on Discrete Prolate Spherical Sequences. It is demonstrated that with a proper choice of the time-bandwidth parameter, a DPSS window is close to a Hamming window. Multiwindow analysis biases the estimation of the correlation more than single window analysis. Windows for transform coding should be split evenly between analysis and synthesis to maximize the signal-to-(coding)-noise ratio. This report also considers asymmetrical windows as used in modern speech coders. The frequency response of these windows is poor relative to conventional windows. Finally, the presence of a "pedestal" in the time window (as in the case of a Hamming window) is shown to be deleterious to the time evolution of the LP parameters.

Time Windows for Linear Prediction of Speech

1 Introduction

This report examines time windows used in linear prediction (LP) analysis of speech. The goal of the windowing is to create frames of data each of which will be used to calculate an autocorrelation sequence. The low-order correlation values are used to generate a LP fit to the speech spectrum. Several factors enter into the choice of window. Both the time and frequency properties are important. The properties of Hamming and Hann windows are examined. A modified version of these windows is suggested.

We also consider windows based on Discrete Prolate Spherical Sequences (DPSS). It is demonstrated that with a proper choice of the time-bandwidth parameter, a DPSS window is close to a Hamming window. Multiwindow analysis can be implemented using an orthonormal family of DPSS windows. The DPSS windows have the property of maximally concentrating the energy in the main lobe of the frequency response. Multiwindow analysis biases the estimation of the correlation more than single window analysis. This report also considers asymmetrical windows as used in modern speech coders. The frequency response of these windows is poor relative to conventional windows. Finally, the presence of a "pedestal" in the time window (as in the case of a Hamming window) is shown to be deleterious in the time evolution of the LP parameters.

2 Linear Predictive Analysis

Linear predictive analysis fits an all-pole model to the local spectrum of a (speech) signal. The model is derived from the autocorrelation sequence of a segment of the speech. The LP spectral fit is determined by solving a set of linear equations based on the correlation values.

Let the input signal be $x[n]$. This signal is windowed,

$$x_w[n] = w[n]x[n]. \tag{1}$$

The linear prediction formulation minimizes the difference between the windowed signal and a linear combination of past values of the windowed signal,

$$e[n] = x_w[n] - \sum_{k=1}^{N_p} p_k x_w[n-k]. \tag{2}$$

The goal is to minimize the total squared error,

$$\varepsilon = \sum_{n=-\infty}^{\infty} |e[n]|^2. \tag{3}$$

For the case that the window is finite in length, the terms in the sum for the squared error will be non-zero only over a finite interval.

The predictor coefficients (p_k) which minimize ε can be found from the following set of equations

$$\begin{bmatrix} r[0] & r[1] & \cdots & r[N_p-1] \\ r[1] & r[0] & \cdots & r[N_p-2] \\ \vdots & \vdots & \ddots & \vdots \\ r[N_p-1] & r[N_p-2] & \cdots & r[0] \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{N_p} \end{bmatrix} = \begin{bmatrix} r[1] \\ r[2] \\ \vdots \\ r[N_p] \end{bmatrix}. \tag{4}$$

The autocorrelation values are given by

$$r[k] = \sum_{n=-\infty}^{\infty} x_w[n] x_w[n-k]. \tag{5}$$

Again for finite length windows, the sum needs be evaluated only over a finite interval — the rest of the correlation coefficients will be zero. In vector-matrix notation,

$$\boldsymbol{Rc} = \boldsymbol{r}. \tag{6}$$

Let the prediction error filter be denoted by $A(z)$,

$$A(z) = 1 - \sum_{k=1}^{N_p} p_k z^{-k}. \tag{7}$$

The autocorrelation formulation for the optimal prediction coefficients gives a matrix $\boldsymbol{R}$ which is Toeplitz. The Levinson-Durbin algorithm can be used to efficiently solve for the predictor coefficients. The prediction error filter ($A(z)$) will be minimum phase and the corresponding synthesis filter $1/A(z)$ will be stable. The frequency response (power spectrum) of the synthesis filter serves as a model of the signal spectrum.

3 Time Windows in Linear Prediction Analysis of Speech

The focus is on discrete-time windows, but lessons from continuous-time still apply. A continuous-time window which is discontinuous (for example, a rectangular window) has a frequency response that falls off as $1/f$ asymptotically. A continuous-time window which is continuous but with a discontinuous first derivative (for example, a triangular window) has a frequency response that falls off as $1/f^2$. A continuous-time window which is discontinuous in the second derivative (for example, a Hann window) has a frequency response that falls off asymptotically as $1/f^3$. Smoothness implies a more rapid fall off. However, smoothness also implies that the effective length of the window can be substantially less than the full length.

The effect of a time window can be described in the frequency domain as a convolution of the frequency response of the window with the frequency response of the signal. The convolution smears frequency features, with the amount of smearing depending on the width of the main lobe of the window frequency response. In addition, spectral leakage from distant frequency components will occur if the sidelobe level of the window response is too large.

Additional bandwidth expansion prior to LP analysis can implemented by lag windowing the autocorrelation sequence, often with a Gaussian window. Lag windowing will not be explored in this report. See [1] for more on bandwidth expansion.

One important property of windows is the window length. In speech coding, a window length of 30 ms (240 samples at a sampling rate of 8 kHz) has been found to be a reasonable compromise in terms of the dynamics of speech production. The window has to be long enough that correlation values can be estimated by averaging lagged values, but not too long such that the local statistical properties of the signal change significantly within the window span. In fact, the update interval is often smaller than the window lengths (20 ms or 160 samples, for instance). This means that the windows used for LP analysis of adjacent frames overlap. This window length will be a constant for our window comparisons. It is to be noted, that speech coding systems offer some robustness to poor LP analysis. For most speech coders, the LP residual signal is coded for transmission and can compensate somewhat for inadequate spectral (LP) modelling.

3.1 Hann and Hamming Windows

Hann and Hamming windows fall in the class of raised-cosine windows. They are both commonly used in speech and audio processing. In Appendix A, both standard and modified versions of the raised-cosine windows are developed by sampling a continuous-time raised-cosine window. The standard definition for an N-sample window is

$$w[n] = \begin{cases} \frac{1+\alpha}{2} - \frac{1-\alpha}{2} \cos(\frac{2\pi n}{N-1}), & 0 \le n \le N-1, \\ 0, & \text{elsewhere.} \end{cases} \tag{8}$$

Altering α allows the characteristics to change from a rectangular window ($\alpha = 1$), to a Hamming window ($\alpha = 0.08$), to the Hann window ($\alpha = 0$). Note that for the standard Hann window, the end points are zero.

The (modified) Hann and Hamming windows are given by[1]

$$w[n] = \begin{cases} \frac{1+\alpha}{2} - \frac{1-\alpha}{2} \cos\bigl(\frac{\pi(2n+1)}{N}\bigr), & 0 \le n \le N-1, \\ 0, & \text{elsewhere.} \end{cases} \tag{9}$$

Appendix A analyzes the modified Hann windows both from the point of view of a product of time sequences (modulation) and as a convolution of time sequences.

The overall normalized frequency response for the Hann window is plotted on a dB scale in Fig. 1. As expected, the smoothness of the window (continuous-time version smooth up to the second derivative), leads to a steep fall-off of the sidelobes, although the first sidelobe is uncomfortably large at only 31.5 dB down.

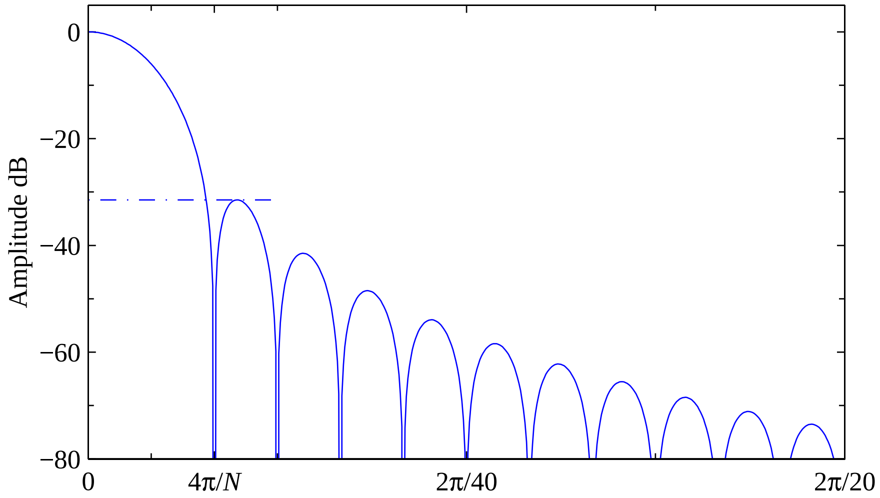

Fig. 1 Normalized frequency response of a modified Hann window ($N = 240$). The broken horizontal line is at −31.5 dB.

[1] The modified versions of the Hann and Hamming windows appear in Mitra [2], although there they are symmetrical about zero and have an odd number of coefficients.

For the Hamming window $\alpha = 0.08$. In effect the window sits on a rectangular pedestal. The pedestal increases the attenuation of the near-in sidelobes in the frequency response (see Appendix A). However, the addition of the pedestal makes the overall function discontinuous. Thus the Hamming window has better near-in sidelobe suppression at the expense of poorer far-out suppression. This is illustrated in Fig. 2 which can be compared to the corresponding figure for the Hann window (Fig. 1). The minimum sidelobe attenuation is 42.7 dB.

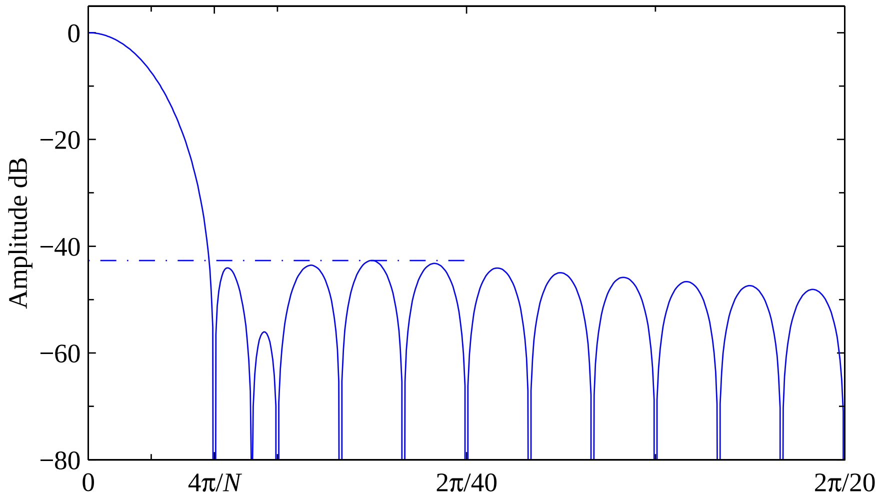

Fig. 2 Normalized frequency response of a modified Hamming window ($N = 240$). The broken horizontal line is at −42.7 dB.

For a Hann or Hamming window with 8 kHz sampling, the main lobe width of the frequency response is 133 Hz between zero crossings.

4 Discrete Prolate Spheroidal Sequences

The discrete prolate spheroidal sequences (DPSS's) are those finite length sequences which concentrate the maximum amount of energy in a given bandwidth [3]. The DPSS's are parameterized by the time bandwidth product NW, where W is the normalized one-sided bandwidth in Hz. There is no closed form for these windows.[2] However, windows are most often pre-computed and stored, so the complexity of the functional description of the window shape need not be a concern.[3]

Like the Hamming window, the DPSS's also sit on pedestals. The DPSS window which has the most energy within $|\omega| \leq 3.5\pi/N$ ($NW = 1.75$) closely matches the central part of a Hamming window. These two windows are plotted in Fig. 3. The solid line is the modified Hamming window

[2] The discrete prolate spheroidal sequences can be calculated using the `dpss` function in Matlab.

[3] The Kaiser window [4] is an approximation to the zeroth order DPSS.

($N = 240$). The dashed line is the discrete prolate spheroidal sequence of the same length. The DPSS sequence has been normalized such that the interpolated value mid-way between the largest samples is unity.

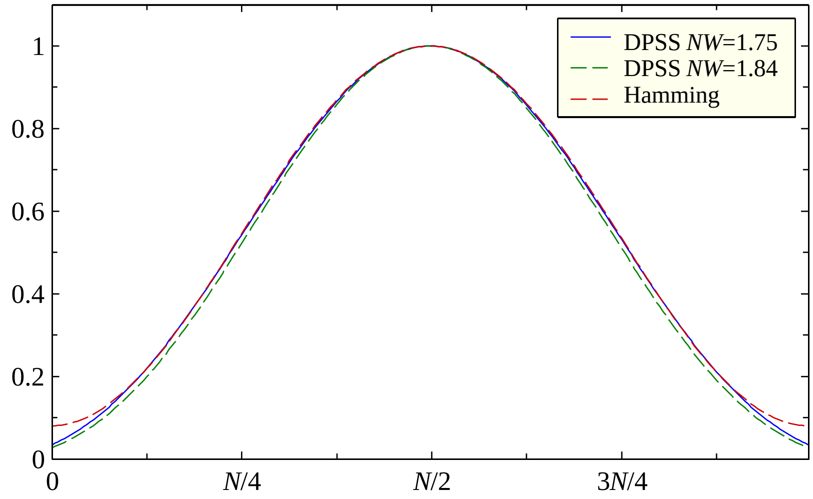

Fig. 3 Discrete prolate spheroidal sequence window, time-bandwidth product $NW = 1.75$ (solid line) and DPSS window, time-bandwidth product $NW = 1.84$ (lower dashed line). Also shown is the Hamming window (upper dashed line). The DPSS windows have been normalized to have a maximum height of unity. All windows have length $N = 240$.

The normalized frequency response of the DPSS window is plotted in Fig. 4. The main lobe width is slightly smaller than that for either the Hann or Hamming window. The first sidelobe attenuation (38.8 dB) is between that for a Hann window (31.5 dB) and that for a Hamming window (42.7 dB). Likewise, the rate of fall-off of the side-bands is also between that of the Hann and Hamming windows.

For the sample DPSS above, the bandwidth was chosen so that the central part of the time window closely matches the Hamming window. Choosing the time-bandwidth product to be $NW = 1.84$ gives a main lobe width in the frequency domain equal to that of the Hamming window. For this choice of bandwidth, the first sidelobe attenuation is now 40.5 dB. The pedestal on which the window sits is reduced (see Fig. 3).

4.1 Multiwindow Analysis with DPSS Windows

The DPSS window above is the first in a family of orthogonal windows for multiwindow (multitaper) spectral analysis [3]. The number of windows (M) with significant concentration of energy in the

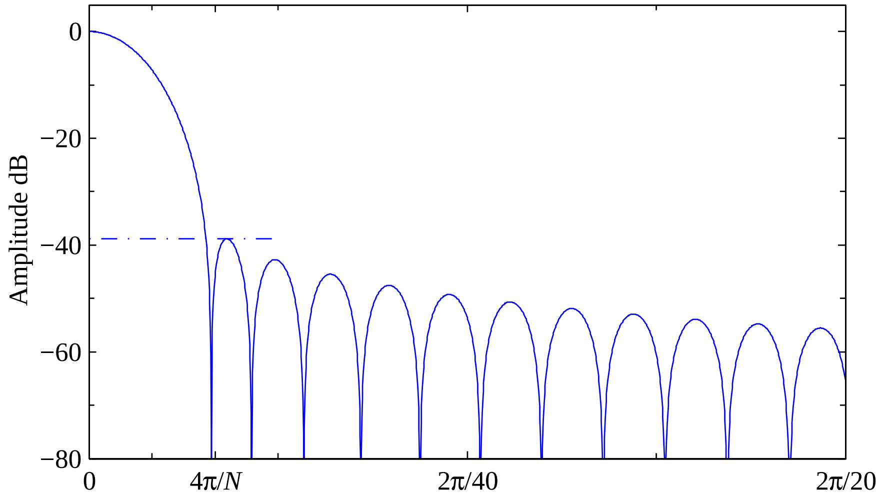

Fig. 4 Normalized frequency response of a discrete prolate spheroidal sequence window (time-bandwidth product $NW = 1.75$, $N = 240$). The broken horizontal line is at -38.8 dB.

given bandwidth is determined by the time-bandwidth product

$$M \leq 2NW - 1, \tag{10}$$

where W is the one-sided bandwidth in Hz. Multiwindow analysis estimates the power spectrum by averaging the spectral estimates using each window. The resulting averaged estimate can have a reduced variance. However, in the speech processing application that is the focus of this work, it is the low-order correlation terms that need to be estimated. A lower variance in the power spectral estimate does not necessarily translate to a better correlation estimate.

For the case of $NW = 1.75$ and $N = 240$, the first three windows are shown in Fig. 5. The first window has 99.98 % of the energy within the design bandwidth. The corresponding figures for the second and third windows are 99.09 % and 88.64 %. If a fourth window had been used, it would have less than 50 % of its energy within the design bandwidth.

The autocorrelation can be calculated for each of the windows and the final correlation obtained as the average of these correlations. For this analysis, it is assumed that each of the windows, $w_m[n]$ is normalized to unit energy. The correlation estimate using window m is (c.f. Eq. (1) and Eq. (5)),

$$r_m[k] = \sum_{n=-\infty}^{\infty} w_m[n]x[n]\, w_m[n-k]x[n-k]. \tag{11}$$

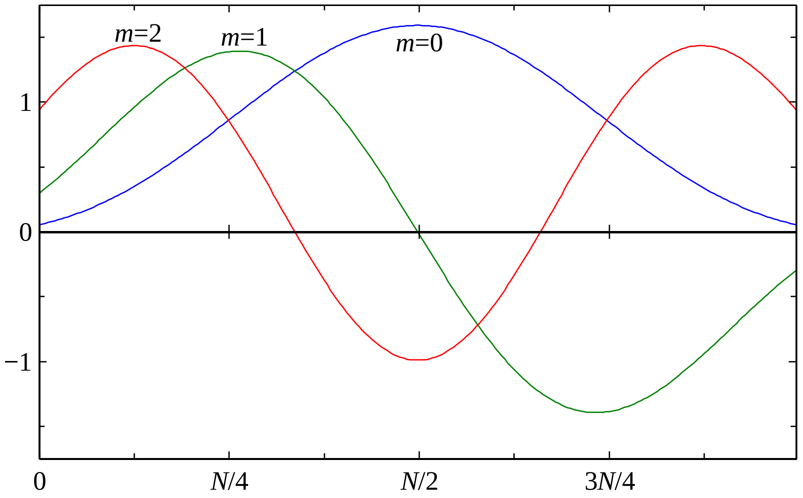

Fig. 5 Discrete prolate spheroidal sequence windows for a time-bandwidth product of $NW = 1.75$ and $N = 240$. The amplitudes have been multiplied by $\sqrt{N}$.

The final correlation estimate is

$$\hat{r}[k] = \sum_{m=0}^{M-1} \beta_m r_m[k]. \tag{12}$$

There are different choices for the weighting factors, β_m. Here we will consider only simple averaging, $\beta_m = 1/M$.

Following the development in [5], the expected value of $\hat{r}[k]$ is given as

$$E\big[\hat{r}[k]\big] = Q[k]r[k], \tag{13}$$

where $r[k]$ is the true correlation and $Q[k]$ is the average correlation of the window sequences,

$$Q[k] = \frac{1}{M} \sum_{m=0}^{M-1} \sum_{n=-\infty}^{\infty} w_m[n]w_m[n-k]. \tag{14}$$

The formulation in Eq. (13) has the form of a lag window, $Q[k]$, acting on the correlation lags. Note, however, that the true correlation values appear on the righthand side of this equation. For lag windowing to be applied, the correlation that it acts on would be an estimated correlation, calculated after applying a data window to isolate a segment of speech. For the typical frame lengths used in speech coding, the effect of the data window cannot be neglected. As such, the effect of the multiple windows acting on the data signal cannot be replaced by a lag window. A further discussion of the relationship between lag windows and multiwindow analysis can be found in [6].

Let the time-bandwidth product be $NW = 1.75$ and the window be of length $N = 240$ as before. The function $Q[k]$ is plotted in Fig. 6 for different values of M. The function $Q[k]$ can be used to measure the correlation estimation bias. The bias is the difference between the average estimated correlation value and the true correlation value. Deviations of $Q[k]$ from unity result in a larger bias. Since the first DPSS window for this time-bandwidth product is close to a Hamming window, the curve for a Hamming window would fall nearly on top of the curve for $M = 1$. The function is always positive. For $M = 2$ and $M = 3$, the function predicts a severe bias for large lags. Since the function goes negative for certain lag values, the average correlation estimate will be of the wrong sign in those regions. Note however, that for LP analysis we only need lags 0 to 10 and the penalty in bias for the difference configurations for that lag range is small.

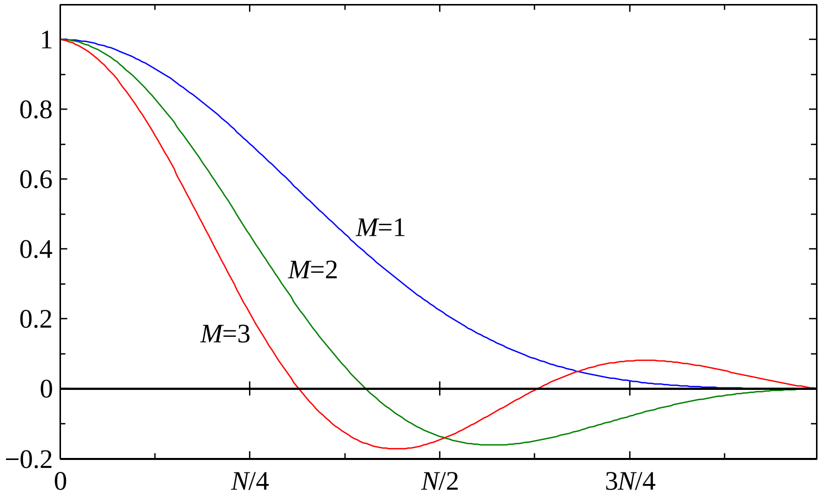

Fig. 6 The quantity $Q[k]$ for DPSS windows different values of M (time-bandwidth product $NW = 1.75$, $N = 240$).

The equivalent frequency response of the multiwindow analysis can be obtained from the Fourier transform of the sum of the correlation values ($Q[k]$). This is plotted in Fig. 7. For $M = 3$, the main lobe is noticeably flatter than for a single window ($M = 1$). For this time-bandwidth product, the first sidelobe has an attenuation of only -16.6 dB, leading to a potential for severe leakage. The sidelobes for the constituent windows add to give poor off-peak rejection. The curves for $M = 2$ and $M = 3$ show that there is a tradeoff with the number of windows. Fewer windows give lower sidelobes, but more windows reduce the variance of the spectral estimates.

Multiwindow analysis was applied to LP analysis. For the purposes of illustration, a sample frame was created as follows. White Gaussian noise was passed through a twelfth order all-pole filter

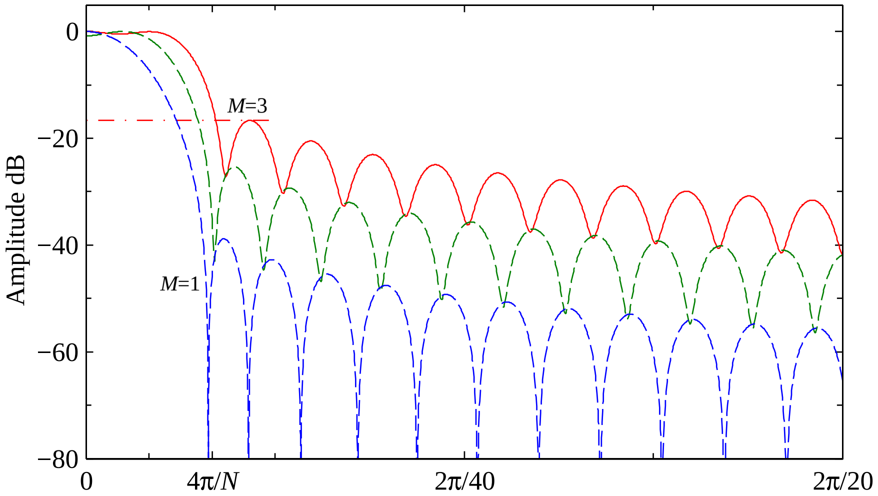

Fig. 7 Normalized frequency response for multiwindow analysis using DPSS windows ($M = 3$, time-bandwidth product $NW = 1.75$). The broken horizontal line is at -16.6 dB.

to create the signal to be analyzed. For a signal sampled at 8 kHz, frames of length $N = 240$ were formed. For a given frame, first, a single window DPSS with time-bandwidth product $NW = 1.75$ was used to estimate the correlations. Multiwindow analysis was also performed ($M = 3$) using DPSS's with the same time-bandwidth product. The LP analysis was tenth order ($N_p = 10$). The results differ for different noise sequences, but Fig. 8 shows the LP spectral fits to the data for a typical frame The ragged light line is the power spectrum computed from the data in the frame. The dotted line is the reference spectrum, i.e. the spectrum of the twelfth order all-pole filter. The smooth solid line is the LP fit using a single window. LP analysis with a Hamming window gives very similar results. The dashed line is the LP fit using multiwindow analysis. One can notice that for this data, the LP spectral fits differ considerably, with the multiwindow analysis giving a somewhat less peaky spectral fit. This was generally true, though the results change with different noise sequences. The broadening of the responses is consistent with the fact that the frequency response of the ensemble of windows gets flatter as more windows are used. This broadening smears the power spectrum.

4.2 Minimum Bias Windows

The sinusoidal windows introduced in [11] can also be used for multiwindow analysis. These windows minimize the “local bias” of the power spectral estimate. The local bias is the just the leading term of an expansion of the spectral bias, corresponding to a minimization of the second

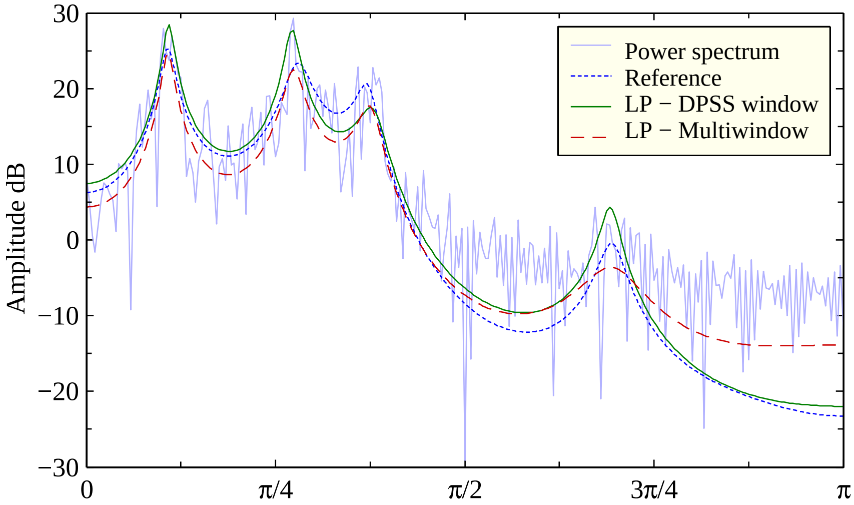

Fig. 8 LP analysis for a frame of data ($N = 240$). The power spectrum is shown by the dotted line. The solid line is the tenth order LP spectral fit using a single DPSS window (time-bandwidth product $NW = 1.75$). The dash-dot line is the tenth order LP spectral fit using multiwindow DPSS analysis ($M = 3$, time-bandwidth product $NW = 1.75$).

central moment of the power spectrum of the window response. For discrete-time windows, the window which minimizes this second moment is determined from an eigenvalue formulation. In fact, this formulation gives the orthogonal windows needed for a multiwindow analysis. As shown by in [11], the minimum bias windows can be closely approximated by the sinusoidal windows. The mth window in a multiwindow analysis is given by

$$w_m[k] = \sqrt{\frac{2}{N+1}} \sin\big(\frac{\pi(k+1)(m+1)}{N+1}\big). \qquad (15)$$

Unlike the DPSS's, the effective bandwidth increases with increasing numbers of windows in a multiwindow analysis.

5 Windows in Transform Coding

We take a digression into another application of windows. Consider a transform coding system using block-based linear transforms with overlapping windows. Let the block length be N and let the transform blocks advance by L samples. An analysis window $w_a[n]$ of length N is applied to the data before the transform. The transform coefficients are then coded. We can model the effect of coding as adding some noise to the transformed signal. An inverse transform is then applied.

The output of the inverse transform is then windowed again with a synthesis window $w_s[n]$ to give an overlap-add reconstruction.

The analysis window can be optimized for a particular application. For instance, if the transform is a DFT, the windows of the type considered heretofore are suitable. However if coding of the coefficients takes place, then a synthesis window which merges adjacent outputs can reduce block-edge effects.

At a given time point n, the input signal may appear as input to several transforms since the blocks overlap. The signal component at the output after analysis and synthesis is

$$\begin{aligned}\hat{x}[n] &= \sum_{k=-\infty}^{\infty} x[n]\, w_a[n-kL]\, w_s[n-kL] \\ &= x[n] \sum_{k=-\infty}^{\infty} w_a[n-kL]\, w_s[n-kL].\end{aligned} \tag{16}$$

5.1 Perfect Reconstruction

An additional requirement is often imposed: In the absence of modification of the transform coefficients, the output should be equal to the input. The requirement for perfect reconstruction is

$$\sum_{k=-\infty}^{\infty} w_a(n-kL)\, w_s(n-kL) = 1. \tag{17}$$

Denote the product of the analysis and synthesis windows as $w[n]$,

$$w[n] = w_a[n]\, w_x[n]. \tag{18}$$

The perfect reconstruction property can be written as a convolution,

$$\sum_{k=-\infty}^{\infty} \delta[n-kL] * w[n] = 1. \tag{19}$$

In the frequency domain, this convolution is equivalent to

$$\sum_{l=-\infty}^{\infty} \delta(\omega - \frac{2\pi l}{L})\, W(\omega) = \delta(\omega), \tag{20}$$

or equivalently

$$W(\frac{2\pi l}{L}) = \begin{cases} 1, & l = pL, \\ 0, & \text{otherwise.} \end{cases} \tag{21}$$

This is a requirement on the zeros crossings of the frequency response of the combined window.[4]

The minimum length window satisfying the perfect reconstruction property is a rectangular window of length L, $w_R[n]$. From Eq. (21), one can see that if its frequency response is multiplied by another frequency response, $P(\omega)$, the combined response, $W(\omega) = P(\omega)W_R(\omega)$ also has the perfect reconstruction property. This multiplication corresponds to a convolution in the time domain. This means we can convolve the rectangular window of length L with any other time function and still end up with a window satisfying the perfect reconstruction property,

$$w[n] = p[n] * w_R[n]. \tag{22}$$

If the sequence $p[n]$ is symmetric, the window $w[n]$ will be symmetric. As an example of this type of construction, in Appendix A.2 the modified Hann window (for N even) is shown to be expressible as the convolution of a rectangular window and another time function (a sine lobe).

5.2 Optimizing the Signal-to-Noise Ratio

Let us model the coding process in the transform domain as generating noise. The noise component is different for each transform, so the noise output at time n is

$$\hat{\zeta}[n] = \sum_{k=-\infty}^{\infty} \zeta_k[n-kL]\, w_s[n-kL], \tag{23}$$

where $\zeta_k[n]$ is the noise in the nth sample after the kth transform. It is not unreasonable to assume the noise components from different transforms are uncorrelated. We will also assume that the noise terms have the same variance and the variance does not depend on n. The noise power at the output at time n is then

$$\begin{aligned} N_o[n] &= E\big[\hat{\zeta}^2[n]\big] \\ &= \sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} E\big[\zeta_k[n-kL]\,\zeta_l[n-lL]\big]\, w_s[n-kL]\, w_s[n-lL] \\ &= \sigma_\zeta^2 \sum_{k=-\infty}^{\infty} w_s^2[n-kL]. \end{aligned} \tag{24}$$

[4] The perfect reconstruction property is analogous to the Nyquist condition for no-intersymbol interference of pulses in data transmission.

The signal power is (assuming a zero mean signal)

$$\begin{aligned} S_o[n] &= E\big[\hat{x}^2[n]\big] \\ &= E\big[x^2[n]\big] \Big| \sum_{k=-\infty}^{\infty} w_a[n-kL]\, w_s[n-kL] \Big|^2 \\ &= \sigma_x^2[n] \Big| \sum_{k=-\infty}^{\infty} w_a[n-kL]\, w_s[n-kL] \Big|^2 . \end{aligned} \tag{25}$$

The signal-to-noise ratio(SNR) at time n is

$$\frac{S_o[n]}{N_o[n]} = \frac{\sigma_x^2[n]}{\sigma_\zeta^2} \frac{\Big| \displaystyle\sum_{k=-\infty}^{\infty} w_a[n-kL]\, w_s[n-kL] \Big|^2}{\displaystyle\sum_{k=-\infty}^{\infty} w_s^2[n-kL]} . \tag{26}$$

For a given choice of $w_a[n]$, the SNR is maximized (Schwartz's inequality) by choosing $w_s[n] = w_a[n]$.[5] This is just a matched-filter result. SNR maximization requires that the analysis and synthesis windows have the same shape.

In perfect reconstruction systems, the product of the two windows is often chosen to be a modified Hann window. If the analysis and synthesis windows are made equal, they become sine windows. These are the minimum bias windows discussed earlier.

An optimization of the windows for controlling the sidelobes in a transform coder, while maintaining perfect reconstruction, was carried out in [14].

6 Asymmetrical Windows

In speech processing, one can distinguish between the LP parameter extraction process and the actual processing of a frame of speech samples. Most often the analysis window used for parameter estimation is larger than the frame of speech to be processed. For symmetric windows, the centre of the analysis window is typically centred on the frame of speech, reaching both into samples before and after the speech frame. This of course means that an additional delay must be imposed for the look-ahead samples. An early paper considered the use of such asymmetrical windows [12]. Some low delay coders, for example the ITU-T G.729 coder [7], use an asymmetric analysis window. The peak of the window occurs such that the window emphasizes the more recent samples. In this way,

[5] The version of Schwartz's inequality applicable here is $|\sum_k w_1[k]\, w_2[k]|^2 \leq \sum_k |w_1[k]|^2 \sum_k |w_2[k]|^2$, with equality if and only if $w_1[k] = a w_2[k]$.

the amount of look-ahead can be reduced while keeping the peak of the window centred on the speech frame.

We will consider the window of the form used in the ITU-T G.729 speech coder,

$$w[n] = \begin{cases} \frac{1+\alpha}{2} - \frac{1-\alpha}{2}\cos(\frac{2\pi n}{2N_L - 1}), & 0 \le n \le N_L - 1, \\ \cos(\frac{2\pi(n - N_L)}{4N_R - 1}), & N_L \le n \le N_L + N_R - 1, \\ 0, & \text{elsewhere.} \end{cases} \tag{27}$$

For the G.729 window, $\alpha = 0.08$, $N_L = 200$, and $N_R = 40$. This asymmetrical window consists of the first half of a traditional Hamming window ($\alpha = 0.08$, length 400) taking up 200 samples, followed by a cosine taper of length 40. The window is plotted in Fig. 9.

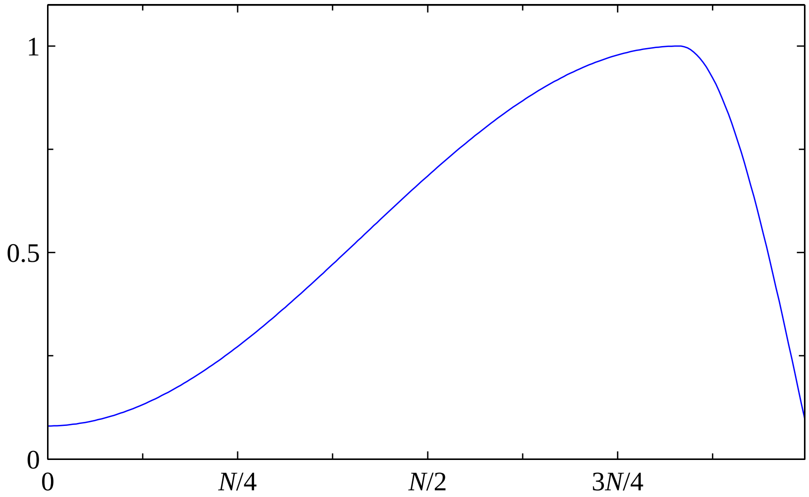

Fig. 9 Asymmetrical G.729 window ($N = 240$).

The frequency response of this window is plotted in Fig. 10. The response does not show a clearly defined main lobe — there are no zero crossings in the vicinity of the main lobe. The attenuation at $\omega = 4\pi/N$ is only 18.1 dB. For comparison, the plot also shows the frequency response of a Hamming window of the same length. It can be seen that compared to traditional symmetric windows, this asymmetric window has a very poor frequency response.

It is interesting to note that the SMV coder standardized by 3GPP2 [8] cycles through three windows, each of length 240. The first is a Hamming window; the second marries the first half of a 300 sample Hamming window with the second half of a 180 sample Hamming window; and the third is very close to the asymmetrical window shown above. One can see the frequency response for two of these windows in Fig. 10. While the main lobes match to some degree, the amount of

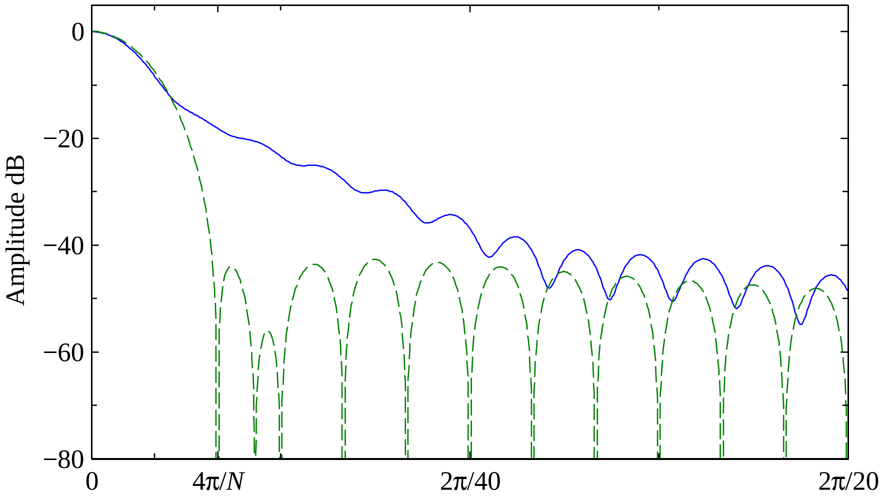

Fig. 10 Frequency response (solid line) of the asymmetrical G.729 window ($N = 240$). The broken line is the frequency response of a modified Hamming window.

leakage would vary significantly when switching windows. One can speculate that even in steady sounds, the LP parameters will have spurious changes brought about by the window switching. Whether these changes affect the overall performance is an open question.

6.1 Modified Asymmetrical Window

As an attempt to improve the response of the G.729 asymmetric window, the definition of the asymmetrical window was modified in the same way as the conventional Hann and Hamming windows were modified. First, let us look at the problems with the conventional definition of the G.729 window.

Consider the discrete variable n to be integer-spaced samples of a continuous variable t. If we look at the underlying continuous functions, we see a Hamming window starting at $t = 0$. The centre of the continuous Hamming window is at $t = N_L - 1/2$. The cosine rolloff begins at $t = N_L$ and goes to zero at $t = N_L + N_R - 1/4$. We note the following.

1. There is a half-sample "gap" between the end of the Hamming window and the start of the cosine rolloff.

2. The period of the cosine element of the Hamming window part $(2N_L - 1)$ is mismatched to the length of the pedestal (N_L).

3. The period of the cosine rolloff is $4N_R - 1$, which is incommensurate with the period of the Hamming window part for the values used in the G.729 window.

4. The Hamming window part has a pedestal, while the cosine rolloff part does not.

A modified version of the asymmetrical window can be defined.

$$w[n] = \begin{cases} \frac{1+\alpha}{2} - \frac{1-\alpha}{2}\cos(\frac{\pi(2n+1)}{2N_L}), & 0 \leq n \leq N_L - 1, \\ \beta + (1-\beta)\cos(\frac{\pi(2(n-N_L)+1)}{4N_R}), & N_L \leq n \leq N_L + N_R - 1, \\ 0, & \text{elsewhere.} \end{cases} \tag{28}$$

This modified window has a half-Hamming window, which in continuous time starts at $t = -1/2$ and ends at $t = N_L - 1/2$. The cosine rolloff starts at $t = N_L - 1/2$ and ends at $t = N_L + N_R - 1/2$. The cosine rolloff is on a pedestal of height β.

Unfortunately, these modifications do not significantly improve the frequency response.

6.2 Optimization of the Asymmetrical Window

We attempted an iterative improvement scheme. Starting from the definition of the asymmetrical window, the coefficients were modified to improve the fraction of the energy in the frequency domain within $|\omega| <= 4\pi/N$. The peak of the window was constrained to be at N_L, with a monotonic lefthand side and a monotonic righthand side. The optimization scheme took a step in one coefficient at a time within the monotonicity constraint. The order of adjustment was randomized. With optimization, the fraction of out-of-band energy reduced from 2.1% to 0.71%. The result is is a window with a flattened top (see Fig. 11). The optimized window has its centre of mass more concentrated in the middle of the window than was the case for the original G.729 window. This shift of the centre of mass works against the goal of the emphasizing samples nearer to the end of the frame. In addition, the optimized window has disconcerting discontinuities.

Recently, Chu [13] has optimized windows to maximize the average prediction error energy after LP analysis. The resulting windows take on an asymmetric form, but with a much narrower "main lobe" than the conventional asymmetric windows in, for instance, G.729.

7 Comparison of Window Properties

Table 1 shows the properties of the various windows considered above. The values given are those for a window length of $N = 240$. Values which scale with the window length are given in term of N. The table shows two DPSS windows. The first was designed to match the central portion of a Hamming window. The second was designed to have a spectral null at $|\omega| = 4\pi/N$, giving a main lobe width equal to that for a Hann or Hamming window. The properties given in the table are as follows.

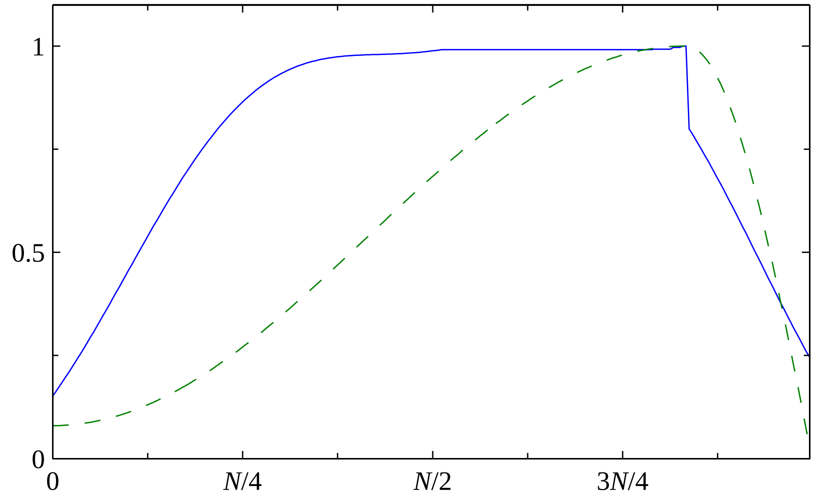

Fig. 11 Asymmetrical window ($N = 240$), optimized to concentrate energy in $|\omega| < 4\pi/N$ (solid line). The dashed line is the G.729 asymmetrical window.

1. The 6 dB bandwidth is the double-sided bandwidth measured at the half-amplitude point. For the modified Hann window, the half-amplitude point occurs exactly at $2\pi/N$, giving a 6 dB bandwidth of $4\pi/N$. The values in the table are normalized to the sampling rate. To get the value on the ω scale, multiply by 2π. To get the value in Hz, multiply by the sampling frequency in Hz. For 8 kHz sampling and $N = 240$, the 6 dB bandwidth of $2/N$ corresponds to 67 Hz.

2. The main lobe width (double-sided) is the distance between zero crossings surrounding the main lobe of the frequency response of the window. This is measured as a fraction of the sampling rate. This main lobe width measure does not apply to the asymmetric window which does not exhibit zeros in the frequency response. For 8 kHz sampling and $N = 240$, a main lobe width of $4/N$ corresponds to 133 Hz.

3. The table gives the minimum attenuation for $|\omega| > 4\pi/N$. For the asymmetrical window, since there is no null near $4\pi/N$, the minimum attenuation occurs at $4\pi/N$. For the other windows, the minimum attenuation occurs for one of the sidelobes beyond $4\pi/N$.

4. The table column labelled "sidelobe energy" gives the fraction of energy for $|\omega| > 4\pi/N$.

5. The window energy is the sum of the squared values of the window (time-domain). All of the windows have been normalized such that the interpolated window value at the middle of the window is unity. The window energy can be used to normalize the window so that it results

in an unbiased estimator for white noise inputs.

6. The pedestal height for each window is given in the table. For most of the windows, this was determined from the definition of the window. For the DPSS windows, there is no analytic expression for the window. For these windows, the pedestal height was estimated by extrapolating the window function out half a sample at either end. For the asymmetric window, the pedestal heights for the left portion of the window and for the right portion of the window are given separately.

Table 1 Properties of time windows ($N = 240$). The sidelobe attenuation corresponds to the largest sidelobe for $|\omega| > 4\pi/N$. The sidelobe energy is the fraction of the energy for $|\omega| > 4\pi/N$. Frequency values in the table are normalized by the sampling rate.

Window	6 dB Bandwidth	Main Lobe Width	Sidelobe Attenuation	Sidelobe Energy	Window Energy	Pedestal
Rectangular	$1.21/N$	$2/N$	17.8 dB	5.0 %	N	100 %
Hann	$2.01/N$	$4.01/N$	31.5 dB	0.051 %	$0.373N$	0 %
Mod. Hann	$2/N$	$4/N$	31.5 dB	0.051 %	$0.375N$	0 %
Hamming	$1.82/N$	$4.03/N$	42.7 dB	0.036 %	$0.396N$	8 %
Mod. Hamming	$1.82/N$	$4/N$	42.7 dB	0.037 %	$0.397N$	8 %
DPSS ($NW = 1.75$)	$1.84/N$	$3.89/N$	38.8 dB	0.017 %	$0.394N$	0.33 %
DPSS ($NW = 1.84$)	$1.87/N$	$4.01/N$	40.5 dB	0.012 %	$0.386N$	0.26 %
Asymmetric G.729	$1.70/N$	–	18.1 dB	2.1 %	$0.415N$	8, 0.99 %

7.1 Effect of the Window Pedestal

The usual argument for a tapered window is that as it slides across the signal, new samples are brought into play gradually. The Hamming window is a sinusoidal window sitting on a pedestal of relative height 0.08. The pedestal can cause substantial changes in the estimated LP parameters even when the window move ahead by a single sample. For comparison, we will use a Hann window, i.e. a window without a pedestal.

For this experiment, the window is advanced in time one sample at a time. Figure 12 shows the effect on a segment of male speech. The plot shows the line spectral frequencies (LSF'S) derived from a tenth order LP analysis [9]. The top plot shows the speech segment (the word "the" from the sentence "Kick the ball straight and follow through"). The second plot shows the energy under a Hann window as the window is advanced one sample at a time. The next plot shows the evolution of the line spectral frequencies using a Hann window (advanced one sample at a time). The bottom plot shows the LSF's when a Hamming window is used. The inset on the top waveform plot shows a Hamming window centred on 440 ms. This position corresponds to the roughness of the LSF tracks near 440 ms. If one superimposed the LSF tracks for the two types of windows, one observes that the "glitches" (Hamming window) in each cluster tend to occur on just one side of the smooth curve (Hann window);

In practice, the windows are advanced less often than every sample. The effect is to sub-sample the LSF tracks. It can be seen that if the sample point lies on one of the "outliers", the results for the two types of windows will be substantially different. Furthermore, the less smooth trajectories of the Hamming windowed speech parameters will be less amenable to differential coding. The effects demonstrated by this example are by no means rare — the spurious variations occur in many speech segments.

This experiment was also run with the other time windows considered earlier. For the asymmetric G.729 window, the spurious variations were as large or larger than for the Hamming window. For the DPSS window, the spurious variations were present but attenuated. This is consistent with the fact that the pedestal for the DPSS window is smaller than for the Hamming window or for the asymmetric window.

Al-Naimi *et al* [10] recognized the problems with the LSF tracks. They propose generating LP parameters at a high rate and then use an anti-aliasing filter to smooth out the variations before resampling at the frame desired rate. They show improved performance for a vector quantizer (using a moving average predictor) which is trained and operated with the smoothed LSF's. An alternative might be to use pitch-synchronous analysis with the aim of ensuring that the window location is more favourable aligned to large amplitude pitch pulses.

The results shown here indicate that the filtering proposed in [10] may be largely unnecessary. The simple expedient of using a window with no pedestal removes the spurious variations in the LSF tracks.

8 Summary

This report has examined the properties of windows used for the linear predictive analysis of speech. The suitability of a window for this purpose is a compromise between time-domain properties (length, shape, symmetry, and presence of discontinuities) and frequency-domain properties (main lobe width and shape, and sidelobe suppression). In the following, we give some suggestions. Instead of the conventional Hann and Hamming windows, use the modified versions for slightly improved properties. The DPSS window (bandwidth $3.5/N$) seems to be slightly preferable to a Hamming window. The main lobe is more compact and there is a better drop-off for the far out sidelobes, at the expense of a slightly poorer attenuation of the first sidelobe. The pedestal height is also smaller. Use the asymmetrical G.729 window only if the reduced look-ahead is of prime

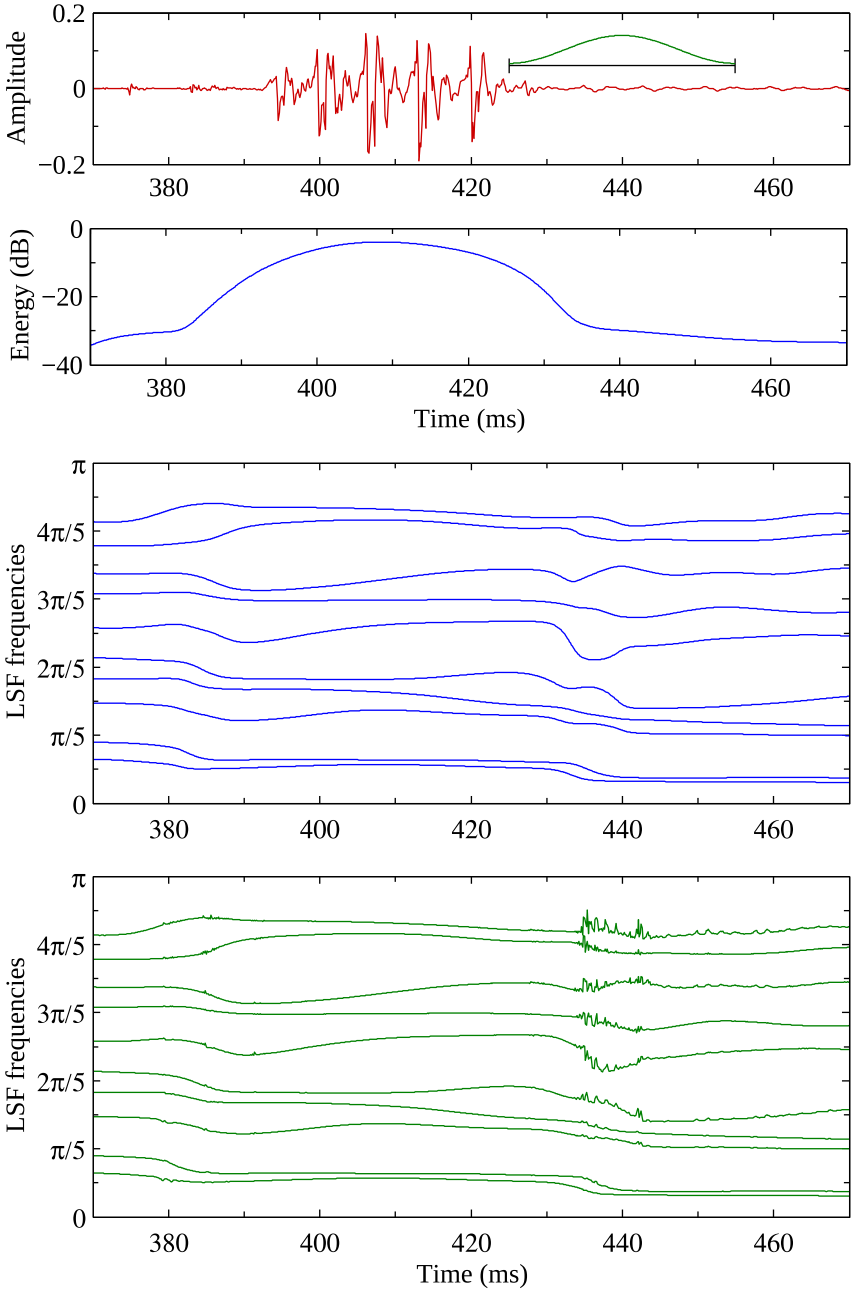

Fig. 12 Effect of the window pedestal on LP parameters. The top plot shows the speech segment being analyzed. The inset shows a Hamming window centred on 440 ms. The second plot shows the energy under a Hann window. The next plot shows the evolution of the LSF parameters when a Hann window is used (window advanced one sample at a time). The bottom plot shows the evolution of the LSF parameters when a Hamming window is used (window advanced one sample at a time).

importance. The frequency resolution and sidelobe suppression for the asymmetrical window is very poor. For most speech applications, avoid windows with significant pedestals. This points to the use of the Hann window or as a compromise the DPSS window, which has a much smaller pedestal than the Hamming window. Other choices of the bandwidth for the DPSS window may be useful for speech processing. Generally, larger bandwidths will tradeoff a larger main lobe width against better sidelobe suppression and smaller pedestal height.

Appendix A - Raised-Cosine Windows

Consider a raised-cosine window defined in continuous time,

$$w(t) = \begin{cases} \frac{1+\alpha}{2} + \frac{1-\alpha}{2} \cos(\frac{2\pi t}{W}), & |t| \leq \frac{W}{2}, \\ 0, & \text{elsewhere.} \end{cases} \tag{29}$$

This window is of length W and centred at the origin. Altering α allows the characteristics to change from a rectangular window ($\alpha = 1$), to a Hamming window ($\alpha = 0.08$), to the Hann window ($\alpha = 0$).

A discrete-time window can be created by sampling the continuous-time window. The N samples will be taken starting at $t = t_0$ and ending at $t = t_1$, where the end points are assumed to lie in the interval $[-W/2, W/2]$. The window is then

$$w[n] = \begin{cases} \frac{1+\alpha}{2} + \frac{1-\alpha}{2} \cos(\frac{2\pi}{W} \frac{nt_1 + (N-1-n)t_0}{N-1}), & 0 \leq n \leq N-1, \\ 0, & \text{elsewhere.} \end{cases} \tag{30}$$

The discrete-time Fourier transform of the raised-cosine window is

$$W(\omega) = e^{-j\omega\frac{N-1}{2}} \big(\frac{1+\alpha}{2} \operatorname{Dsinc}(\omega, N) + e^{j2\pi\frac{t_1+t_0}{2}} \frac{1-\alpha}{4} \left[\operatorname{Dsinc}(\omega-\omega_o, N) + \operatorname{Dsinc}(\omega+\omega_o, N)\right] \big), \tag{31}$$

where the function Dsinc is defined as

$$\operatorname{Dsinc}(\omega, N) = \frac{\sin(\omega N/2)}{\sin(\omega/2)} \tag{32}$$

and ω_o is $2\pi(t_1 - t_0)/(W(N-1))$. This frequency response of the window is the sum of three Dsinc functions. The function $\operatorname{Dsinc}(\omega, N)$ is periodic in ω (period 2π), takes on the value N at $\omega = 0$ and has zeros at $\omega = 2\pi k/N$ for $k = 1, \ldots, N-1$. The other Dsinc functions in the frequency response have the same zero crossing spacing, but are displaced in frequency. The zero crossings of the three Dsinc terms will coincide only if ω_o is an integer multiple of $2\pi/N$. Note also that the window is symmetric and the frequency response is linear phase if $t_0 = -t_1$.

A.1 Standard Raised-Cosine Windows

A standard formulation for a discrete-time raised-cosine window is obtained by choosing $t_1 = W/2$ and $t_0 = -t_1$, giving $\omega_o = 2\pi/(N-1)$,

$$w[n] = \begin{cases} \dfrac{1+\alpha}{2} - \dfrac{1-\alpha}{2}\cos(\dfrac{2\pi n}{N-1}), & 0 \leq n \leq N-1, \\ 0, & \text{elsewhere.} \end{cases} \tag{33}$$

For $\alpha = 0$, this is the window returned by, for instance, the `hann(N)` function call in Matlab. The end points of the discrete-time window are zero. The `hann(N-1,'periodic')` function call in Matlab returns the same window without the last zero-valued sample. The `hanning(N-2)` function call in Matlab returns the same window without either of the two zero-valued end-points.

For the standard Hann window, the zero-crossings of the 3 terms in the frequency response do not coincide. The overall frequency response has a first zero crossing which occurs between $2\pi/N$ and $2\pi/(N-1)$. The exact value must be found numerically for a given N.

The conventional Hamming window is given by Eq. (33) with $\alpha = 0.08$. This is the window returned by the `hamming(N)` function call in Matlab. The Hamming window is a Hann window of height 0.92 sitting on a pedestal of height 0.08. The frequency response of the conventional Hamming window has the problems of mismatched zero crossing terms noted above.

A.2 Modified Raised-Cosine Windows

A modified raised-cosine window can be defined with $t_1 = W/2 - W/(2N)$ and $t_0 = -t_1$, giving $\omega_o = 2\pi/N$. The modified discrete-time raised-cosine windows can be written as

$$w[n] = \begin{cases} \dfrac{1+\alpha}{2} - \dfrac{1-\alpha}{2}\cos\big(\dfrac{\pi(2n+1)}{N}\big)\big), & 0 \leq n \leq N-1, \\ 0, & \text{elsewhere.} \end{cases} \tag{34}$$

The modified Hann window can be calculated in Matlab as `h=hann(2*N+1); h=h(2:2:end)`. The modified Hamming window is returned with `h=hamming(2*N+1); h=h(2:2:end)`.

The frequency response of the modified raised-cosine window is

$$W(\omega) = e^{-j\omega\frac{N-1}{2}}\big(\frac{1+\alpha}{2}\,\mathrm{Dsinc}(\omega, N) + \frac{1-\alpha}{4}\big[\mathrm{Dsinc}(\omega - \frac{2\pi}{N}, N) + \mathrm{Dsinc}(\omega + \frac{2\pi}{N}, N)\big]\big), \tag{35}$$

For the modified windows, the zero crossings of all the Dsinc terms match. In fact, a Discrete Fourier Transform (DFT) of length N will have only three nonzero coefficients.

A.2.1 Frequency response of the modified Hann window

The frequency response (without the linear phase term) of the modified Hann window ($\alpha = 0$) is plotted in Fig. 13. The broken lines show the three terms that contribute to the sum (shown with a solid line). Notice that the term for the Dsinc function centred at zero (dashed line) and the terms for the other Dsinc functions (dash-dot lines) are out of phase and tend to cancel in the tails. This results in diminished sidelobes relative to a rectangular window (central Dsinc function alone).

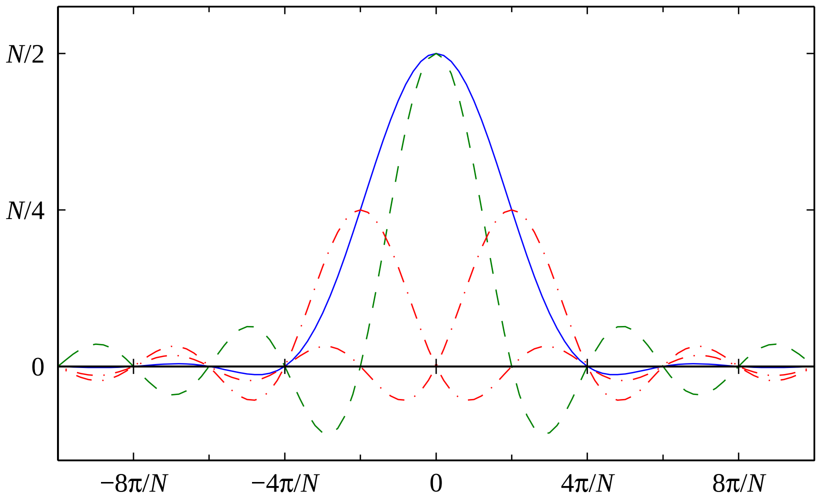

Fig. 13 Frequency response of a modified Hann window ($N = 240$). The solid line is for the overall window. The broken lines show the three additive components.

The frequency response of the modified Hann window can also be written in a product form (for N even),

$$\begin{aligned} W(\omega) &= e^{-j\omega\frac{N-1}{2}}\,\mathrm{Dsinc}(\omega, N/2) \\ &\quad\times \frac{\cos(\frac{\omega N}{4})}{\cos(\omega) - \cos(\frac{2\pi}{N})}\,\Big[\frac{1+\alpha}{2}(\cos(\omega) - \cos(\frac{2\pi}{N})) + (1-\alpha)\sin^2(\frac{\omega}{2})\cos(\frac{\pi}{N})\Big], \end{aligned} \tag{36}$$

The frequency response (without the linear phase term) of the modified Hann window is shown in Fig. 14. The two terms in the product are shown as broken lines, with the product being shown with a solid line. The first factor (dashed line) is the frequency response of a half-length rectangular window. This response has zero-crossings every $4\pi/N$. The second factor (dash-dot line) not only inserts extra zero crossings, but since it decreases in amplitude away from the main lobe, it reduces the sizes of the sidelobes in the overall frequency response.

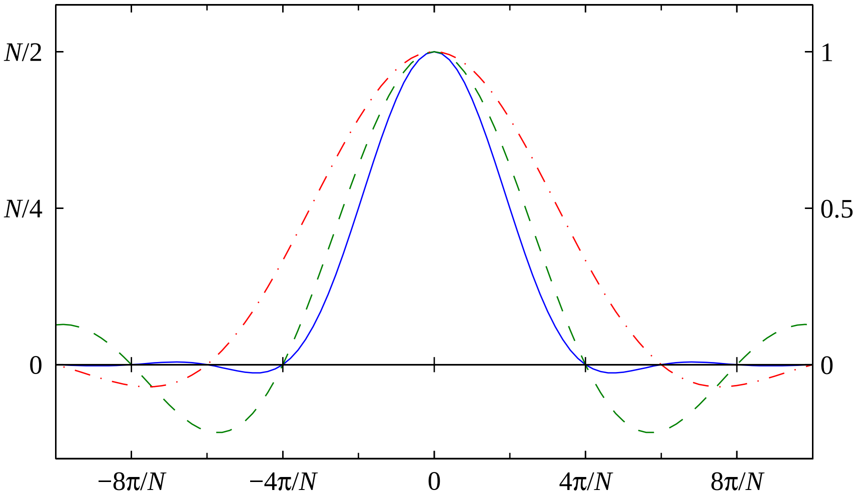

Fig. 14 Frequency response of a modified Hann window ($N = 240$). The solid line is for the overall window. The broken lines show the two product factors. The right-hand scale applies to the second product factor (dash-dot line).

The modified Hann window is the convolution of the time responses corresponding to the two factors in the frequency response. The time responses of these factors are shown in Fig. 15. The first factor (dashed line) is a half-length rectangular window (assuming N is even). The second factor (dash-dot line) is a sine-like lobe.

A.2.2 Frequency response of the modified Hamming window

The Hamming window is formed by adding a pedestal to a Hann window. This results in better cancellation of the first few sidelobes. Comparing Fig. 16 with the corresponding figure for a Hann window (Fig. 13), it an be seen that increasing the relative amplitude of the central term (dash-dot line) reduces the first sidelobe of the frequency response. However further out in frequency, the cancellation will not be as good as for the Hann window.

A.2.3 Comparison of modified and standard windows

At moderately large values of N, say around 240, the envelope of the frequency response of the standard and modified Hamming windows are largely indistinguishable. The details, of course, differ since the positions of the zero crossings differ.

For the Hann window, the situation is different. For the modified Hann window, the zeros have spacing $2\pi/N$, up to and including $\omega = \pi$. For the standard Hann window, the zeros have a slightly larger spacing, so that the last zero before π is closer to π than is the case for the modified Hann

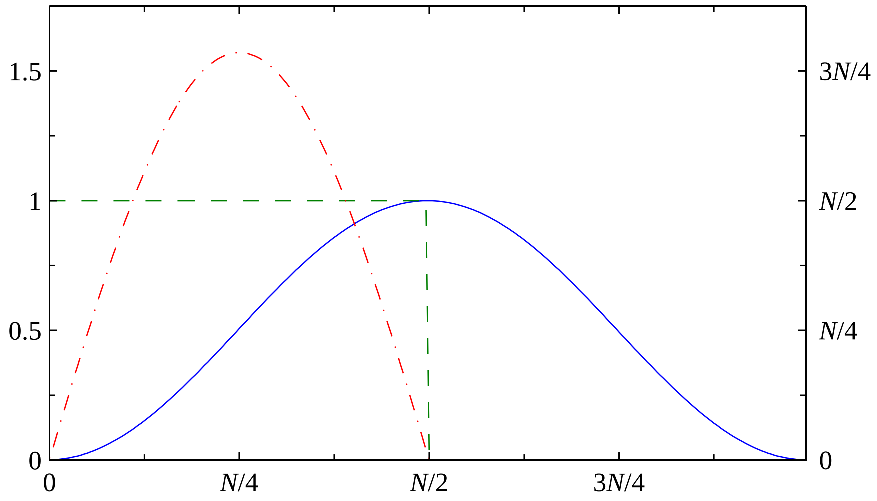

Fig. 15 Modified Hann window ($N = 240$) as the convolution of two terms. The first term is a rectangular window of length $N/2$ (dashed line). The second term is a sine lobe (dash-dot line, using the right-hand scale).

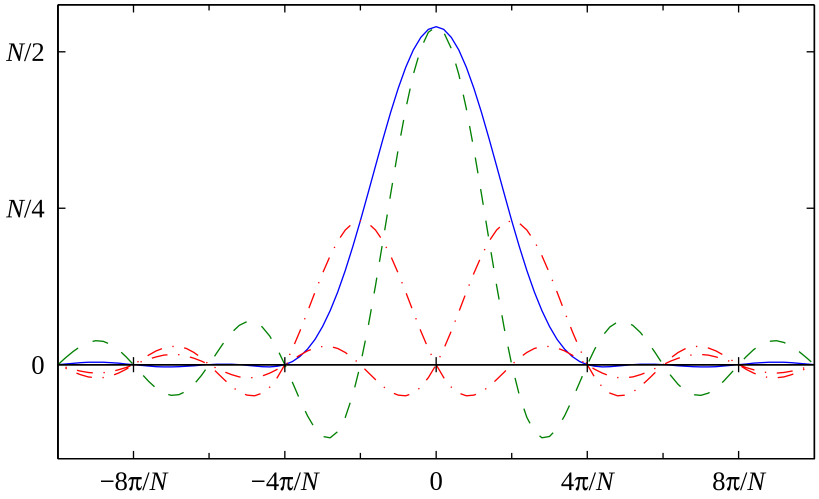

Fig. 16 Frequency response of a modified Hamming window ($N = 240$). The solid line is for the overall window. The broken lines show the three additive components.

window. This close spacing of the zeros near π results in additional attenuation near $\omega = \pi$. The difference between the windows is illustrated in Fig. 17.

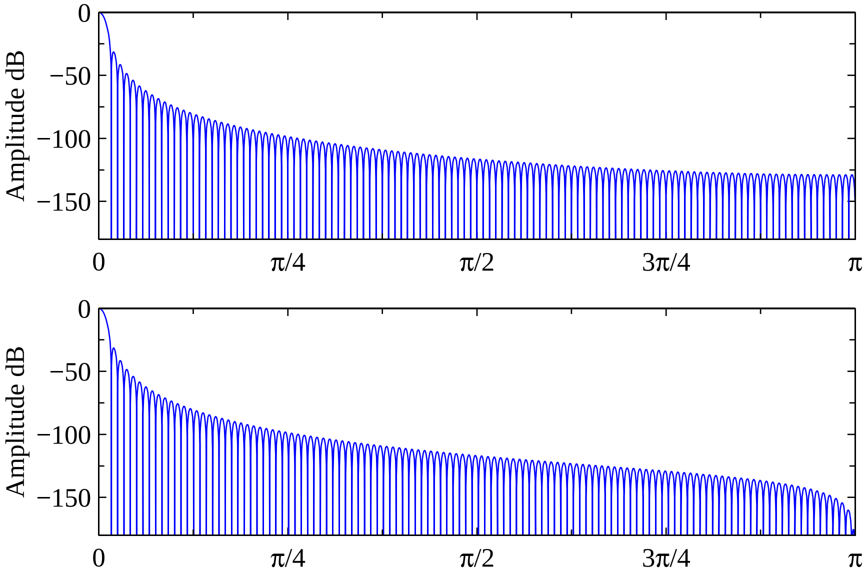

Fig. 17 Normalized frequency response of the modified Hann window (top) and standard Hann window (bottom), $N = 240$.

References

[1] P. Kabal, “Ill-conditioning and bandwidth expansion in linear prediction of speech”, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing* (Hong Kong), April 2003.

[2] S. K. Mitra, *Digital Signal Processing: A Computer-based Approach*, second ed., McGraw-Hill, 2002.

[3] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques*, Cambridge University Press, 1993.

[4] J. F. Kaiser and R. W. Schafer, “On the use of the I_0-sinh window for spectrum analysis”, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 105–107, Feb. 1980.

[5] A. Hanssen, “On multiwindow estimators for correlation”, *Proc. IEEE Workshop Statistical Signal and Array Processing* (Pocono Manor, PA), pp. 640–644, Aug. 2000.

[6] M. L. McCloud, L. L. Scharf and C. T. Mullis, “Lag-windowing and multiple-data-windowing are roughly equivalent for smooth spectrum estimation”, *IEEE Trans. Signal Processing*, vol. 47, pp. 839–843, March 1999.

[7] ITU-T, Geneva, *Recommendation G.729, Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP)*, Mar. 1996.

[8] 3GPP2 Document C.S0030-0, *Selectable Mode Vocoder Service Option of Wideband Spread Spectrum Communication Systems*, Version 2.0, Dec. 2001.

[9] P. Kabal and R. P. Ramachandran, “The computation of line spectral frequencies using Chebyshev polynomials”, *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. 34, pp. 1419–1426, Dec. 1986.

[10] K. Al-Naimi, S. Villette and A. Kondoz, “Improved LSF estimation through anti-aliasing filtering,” *Proc. IEEE Workshop on Speech Coding* (Tsukuba City, Japan), pp. 2–4, Sept. 2002.

[11] K. S. Riedel and A. Sidorenko, “Minimum bias multiple taper spectral estimation”, *IEEE Trans. Signal Processing*, vol. 43, pp. 188–195, Jan. 1995.

[12] D. A. F. Florncio, “Investigating the use of asymmetric windows in CELP coders”, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing* (Minneapolis, MN), pp. 427–430, April 1993.

[13] W. C. Chu, “Gradient-descent based window optimization for linear prediction analysis”, *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing* (Hong Kong), pp. I-460–I-463, April 2003.

[14] H. Najafzadeh-Azghandi, *Perceptual Coding of Narrowband Audio Signals*, Ph.D. Thesis, Electrical & Computer Engineering, McGill University, April 2000.