Proc. Biennial Symp. Commun. (Kingston, ON), pp. C.3.4–C.3.7, June 1986

The Stability of Pitch Filters in Speech Coding

RAVI P. RAMACHANDRAN AND PETER KABAL

Abstract

The stability and performance of pitch filters in speech coding are studied. A new algorithm that estimates the pitch period is coupled with the covariance formulation of determining the predictor coefficients. Since this approach does not guarantee the stability of the pitch synthesis filter, an efficient stability test is formulated. From this, a stabilization technique that ensures a stable pitch filter is introduced. The effect of the presence of unstable pitch filters on decoded speech is investigated. The use of stable pitch filters in speech coding generates decoded speech of higher perceptual quality.

1. Introduction

In Code-Excited Linear Prediction (CELP) [1], two nonrecursive prediction filters are used to process the incoming speech signal. One is the formant predictor that removes near-sampled based redundancies. It is followed by the pitch predictor that removes distant-sampled based redundancies. At the receiver, the corresponding inverse filters, the pitch synthesis and formant synthesis filters reproduce the decoded speech. These filters are recursive and may be unstable. The autocorrelation [2] and modified covariance [3] methods calculate the coefficients of a stable formant synthesis filter. This paper addresses the stability and performance issues of the pitch filter and examines the effect of instability on decoded speech.

2. CELP Coder

In a CELP coder the residual is generated after passing the speech signal through a formant predictor ($F(z)$) and pitch predictor ($P(z)$). The predictors have transfer functions:

$$F(z) = \sum_{k=1}^{p} a_k z^{-k} \tag{1}$$

$$P(z) \quad \begin{cases} \beta_1 z^{-M} & \text{1 tap} \\ \beta_1 z^{-M} + \beta_2 z^{-(M+1)} & \text{2 tap} \\ \beta_1 z^{-(M-1)} + \beta_2 z^{-M} + \beta_3 z^{-(M+1)} & \text{3 tap} \end{cases} \tag{2}$$

The order p is typically between 8 and 16 and M is the estimated pitch period in samples. At the receiver, a pitch synthesis filter $H_P(z) = 1/(1 - P(z))$ and formant synthesis filter $H_F(z) = 1/(1 - F(z))$ are used.

The residual (after gain normalization) is compared to a set of waveforms in a codebook constructed of Gaussian random numbers with unit variance. To perform the comparison, each entry in the codebook is first filtered by $H_P(z)$ and $H_F(z)$ and subtracted from the original speech to form a difference signal. This signal is passed through a weighting filter $W(z) = (1 - F(z))/(1 - F(z/\alpha))$ where $0 < \alpha < 1$. The error is formed by squaring and averaging the filtered difference signal. The entry in the codebook that gives the smallest error represents the residual and its index is transmitted. This is equivalent to a vector quantization scheme. The codeword representing the residual is scaled by the gain factor and filtered by $H_P(z)$ and $H_F(z)$ to generate the decoded speech.

3. Covariance Formulation

In determining the coefficients of $P(z)$, the mean-square value of the residual is minimized over a frame size of N samples. This leads to a linear system of equations. In matrix form, this system

R. P. Ramachandran is with the Department of Electrical Engineering, McGill University, Montreal , Quebec, H3A 2A7.

P. Kabal is with the Department of Electrical Engineering, McGill University, Montreal, Quebec, H3A 2A7 and INRS-Télécommunications, Université du Québec, Verdun, Quebec, H3E 1H6.

of equations ($\Phi\beta = \alpha$) for a 3 tap filter is:

$$\begin{bmatrix} \phi(M-1,M-1) & \phi(M-1,M) & \phi(M-1,M+1) \\ \phi(M,M-1) & \phi(M,M) & \phi(M,M+1) \\ \phi(M+1,M-1) & \phi(M+1,M) & \phi(M+1,M+1) \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \phi(0,M-1) \\ \phi(0,M) \\ \phi(0,M+1) \end{bmatrix} \tag{3}$$

Before solving the equations, the pitch period M must be determined. Since the speech is sampled at 8 kHz, minimum and maximum values of M equal to 20 and 120 are used to cover the range of pitch periods for male and female speakers. Atal [3] estimates M by first calculating a correlation array $\tau(M)$ where:

$$\tau(M) = \frac{\phi(0,M)}{\sqrt{\phi(0,0)\phi(M,M)}} \tag{4}$$

The correlation array is searched for local maxima and parabolic interpolation is used on triplets of correlation values centered at the local maxima. Local peaks (not necessarily integer values) are located at points at which the interpolated functions are a maximum. The pitch period M is the nearest integer value of the largest local peak.

Now, a new algorithm that estimates M by minimizing an approximation to the mean-square value of the residual is formulated. In a 3 tap filter, $\beta = \Phi^{-1}\alpha$ and the resulting mean-square error is $\varepsilon^2 = \phi(0,0) - \beta^T\alpha$. The value of M is chosen so as to maximize $\beta^T\alpha$. Since a great deal of computation is required to maximize $\beta^T\alpha$, an approximation is made. After formant prediction has been accomplished, the near-sample based redundancies have been removed to a large extent. Therefore, the off-diagonal terms in the matrix Φ are neglected. Then, $\beta^T\alpha \approx \phi(0,0)(\tau^2(M-1) + \tau^2(M) + \tau^2(M+1))$. The value of M that maximizes $\tau^2(M-1) + \tau^2(M) + \tau^2(M+1)$ is chosen. For 1 and 2 tap filters, the value of M that maximizes $\tau^2(M)$ and $\tau^2(M) + \tau^2(M+1)$ respectively is chosen. No parabolic interpolation is required and a consistently higher overall prediction gain than Atal's method is achieved.

4. Stability of Pitch Filters

If $H_P(z)$ is a stable function, its denominator polynomial $D(z)$ must have all its zeros within the unit circle. This is not guaranteed by the covariance formulation. A general $D(z)$ for a pitch synthesis filter is of the form $D(z) = z^n - B(z)$ where $B(z) = \sum_{i=0}^{n-1} b_i z^i$. Although a set of necessary and sufficient conditions can be used [4], a computationally simple test based on a sufficient condition is formulated. The polynomial $z^n - B(z) = z^n(1 - z^{-n}B(z)) \neq 0$ or $z^{-n}B(z) \neq 1$ on and outside the unit circle $z = e^{j\theta}$. By the maximum modulus theorem [5], $z^{-n}B(z)$ has its maximum modulus on the contour surrounding any region in which it is analytic. Since $z^{-n}B(z)$ is a polynomial in z^{-1}, it is analytic on and outside the unit circle. Therefore, a sufficient condition for stability is that $|z^{-n}B(z)| < 1$ on and outside the unit circle. By substituting $z = e^{j\theta}$, the sufficient condition becomes $|B(e^{j\theta})| < 1$.

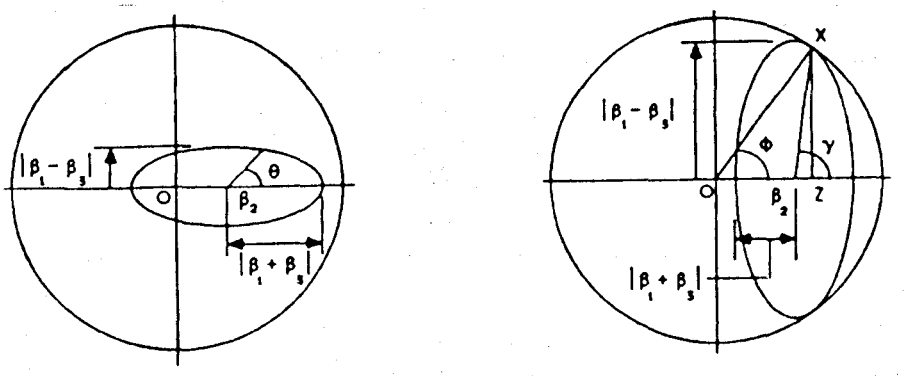
In the particular case of a 3 tap filter, $B(z) = \beta_1 z^2 + \beta_2 z + \beta_3$ and $B(e^{j\theta}) = e^{j\theta}B'(e^{j\theta})$ where:

$$B'(e^{j\theta}) = \beta_2 + (\beta_1 + \beta_3)\cos\theta + j(\beta_1 - \beta_3)\sin\theta \tag{5}$$

The expression $B'(e^{j\theta})$ defines an ellipse with center β_2 and major axis $|\beta_1 + \beta_3|$ if β_1 and β_3 have the same signs or $|\beta_1 - \beta_3|$ if β_1 and β_3 have different signs. The two cases are illustrated in Fig. 1. For now, the case $\beta_2 > 0$ is considered. Later, the analysis is extended to $\beta_2 < 0$.

The ellipse must lie entirely within the circle to ensure stability. If β_1 and β_3 have the same signs, this happens if all points on the major axis are within the circle or equivalently $|\beta_1| + |\beta_2| - |\beta_3| < 1$. If β_1 and β_3 have opposite signs, the substitutions $a = |\beta_1 + \beta_3|$ and $b = |\beta_1 - \beta_3|$ are used to obtain the conditions $a + \beta_2 < 1$ and $b^2 + \beta_2^2 < 1$ that ensure that all points on the major and minor axes are within the circle. The ellipse may still touch the circle at a point of tangency X. Let b_{max} denote the value of b for which tangency is achieved. By deriving expressions for the slopes of line OX and the tangent line to the ellipse at X and setting the length of OX equal to unity, it can be shown that b_{max} can be obtained by solving $f(b_{max}^2) = 0$ where:

$$f(b^2) = b^4 + b^2(\beta_2^2 - a^2 - 1) + a^2 \tag{6}$$



(a) Horizontal Axis is the Major Axis (b) Vertical Axis is the Major Axis

Fig. 1 Illustration of Stability Test Ellipses for 3 Taps

Equation 6 is a quadratic in b^2 whose roots are real and positive. The desired solution is the larger of the roots. To check that $b < b_{max}$ (ellipse is within the circle), it suffices to check that $b^2 < a^2 + a\beta_2$ or $f(b^2) < 0$. Equation 6 need not be solved for b_{max}. If $\beta_2 < 0$, the analysis is similar and the conditions merely involve replacing β_2 by $|\beta_2|$. The implementation of the stability test for 3 tap filters is given below.

Stability Test

1. If $a \geq b$, then:
   (a) $|\beta_1| + |\beta_2| + |\beta_3| < 1$
2. If $a < b$, then:
   (a) $a + |\beta_2| < 1$
   (b) $b^2 + \beta_2^2 < 1$
   (c) (i) $b^2 < a^2 + a|\beta_2|$ or
   (ii) $f(b^2) = b^4 + b^2(\beta_2^2 \quad a^2 \quad 1) + a^2 < 0$

The test for 3 tap filters subsumes the test for 1 and 2 tap filters. In the 2 tap case, the condition $|\beta_1| + |\beta_2| < 1$ is sufficient for stability. For a 1 tap filter, the condition $|\beta_1| < 1$ is necessary and sufficient. Also, the test for 2 and 3 tap filters is necessary and sufficient in the limit of large n.

5. Stabilization Procedure

In each frame of speech (length is 80 samples), the predictor coefficients are calculated (see Eq. 3) and the stability test applied. If the filter is found to be unstable, each coefficient is scaled by a common factor c. This results in a sub-optimum predictor with coefficients $\boldsymbol{\beta}' = c\boldsymbol{\beta} = \boldsymbol{\beta} + \boldsymbol{\delta}$. The energy of the prediction residual is $\varepsilon^2 \quad \varepsilon^2_{min} + \boldsymbol{\delta}^T \boldsymbol{\Phi} \boldsymbol{\delta}$ where ε^2_{min} is the energy achieved when no scaling is applied. In order to simultaneously minimize $\boldsymbol{\delta}^T \boldsymbol{\Phi} \boldsymbol{\delta} \quad (c \quad 1)^2 \boldsymbol{\beta}^T \boldsymbol{\Phi} \boldsymbol{\beta}$ and assure a stable $H_P(z)$, a bound on c, namely, $0 < c < 1$ is imposed. The value of M is assumed to be unaltered.

In a 1 tap filter, any value of $\beta > 1$ is reset to 1 and any value of $\beta < \quad 1$ is reset to $\quad 1$. In a 2 tap filter, $c = 1/(|\beta_1| + |\beta_2|)$. In a 3 tap filter if $a > b$, then $c = 1/(|\beta_1| + |\beta_2| + |\beta_3|)$. If $a > b$ and $b^2 < a^2 + a|\beta_2|$, condition 2.(c) is already satisfied. Scaling the coefficients does not change this relationship. The value of c is chosen to force the length of the minor axis of the scaled ellipse to be equal to 1. Then, $c = 1/(a + |\beta_2|)$. Under these conditions, it can be shown that all points along the major axis lie inside the circle (condition 2.(b) is satisfied). If $b^2 > a^2 + a|\beta_2|$, the value of c is chosen to set $f(b^2) \quad 0$. Then, the scaled ellipse is tangent to the circle. The value of c is:

$$c \quad \sqrt{\frac{b^2 \quad a^2}{b^4 + b^2\beta_2^2 \quad b^2a^2}} \tag{7}$$

With this value of c, it can be shown that all points along the minor and major axes are within the circle (conditions 2.(a) and (b) are satisfied).

Theoretically, marginal stability has been assured. Scaling the value of β to 0.99 or -0.99 in a 1 tap filter assures complete stability. For 2 and 3 tap filters, calculating c as required and subtracting 0.001 from it assures complete stability by avoiding a point of tangency. Experimental results

show that the average loss in prediction gain associated with stabilization is only 0.03, 0.26 and 0.21 dB for 1, 2 and 3 tap filters respectively. These results were compiled by using different speech waveforms.

6. Effect of Instability on Decoded Speech

To examine the effect of unstable pitch filters on decoded speech, a CELP coder was simulated using a 10th order formant predictor (modified covariance method) and a 3 tap pitch predictor. Forty sample blocks of the residual were compared to a codebook of 2^{10} 1024 waveforms. The parameter α 0.8 was used in implementing $W(z)$. Decoded waveforms are shown below when the sentence "cats and dogs each hate the other" was processed. Frames having unstable pitch filters are marked by a non-zero indicator function.

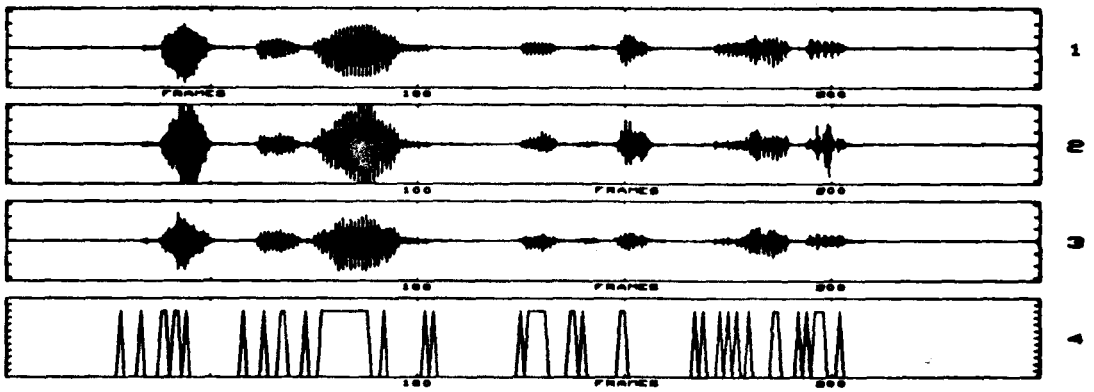


Fig. 2 Original and Decoded Signals (1) Speech Data, (2) Decoded Signal for Covariance Formulation, (3) Decoded Signal for Stabilization Technique and (4) Frames Having Unstable Filters

If a sequence of consecutive frames of high input energy have unstable filters or if the value of $\sum_i |\beta_i|$ is high, degradations in the output speech are perceptible. Frames 77 to 88 consist of high-energy voiced speech and have unstable filters. The quantization noise continues to build up, the energy of the output signal keeps rising and this noise is perceptible. Even if an unstable filter having a high value of $\sum_i |\beta_i|$ occurs in a frame of low input energy, an impulse-type distortion that is heard as a pop or click is present. This phenomenon occurs during frames 149 and 150 and frames 196 to 198. In frames 196 to 198, $\sum_i |\beta_i|$ equals 2.77, 4.02 and 2.23 respectively. When the filters are stabilized, the undesirable pops, clicks and enhanced background noise are absent. Listening tests show that waveform (3) sounds better than (2).

7. Conclusion

Decoded speech generated by a CELP coder improves in quality when stable pitch filters are used. This is accomplished by a computationally simple stabilization technique derived from an efficient stability test.

References

1. M.R. Schroeder and B.S. Atal, "Code-Excited Linear Prediction (CELP): High-quality speech at very low bit rates", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Tampa, Florida, pp. 25.1.1 25.1.4, March 1985.
2. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.
3. B.S. Atal and M.R. Schroeder, "Predictive coding of speech and subjective error criteria", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 247 254, June 1979.
4. E.I. Jury, *Theory and Application of the z-Transform Method*, John Wiley and Sons, 1964.
5. R.A. Silverman, *Introductory Complex Analysis*, Dover Publications, 1967.