Proc IEEE Int. Conf. Acoustics, Speech, Signal Processing (Hong Kong), pp. I-824-I-827, April 2003

Ill-Conditioning and Bandwidth Expansion in Linear Prediction of Speech

Peter Kabal

Electrical & Computer Engineering
McGill University
Montreal, Quebec H3A 2A7

Abstract

This paper examines schemes that modify linear prediction (LP) analysis for speech signals. First, techniques which improve the conditioning of the LP equations are examined. White noise compensation for the correlations is justified from the point of view of reducing the range of values which the predictor coefficients take on. The efficacy of the procedure is measured over a large speech database. Various techniques for bandwidth expansion of the LP spectral peaks are also examined. These include lag windowing of the correlation, windowing of the predictor coefficients, and modification of the line spectral frequencies. New formulas for the bandwidth expansion factor are given.

1 Introduction

This paper examines techniques which have been employed to modify the linear prediction (LP) analysis of speech signals. First, there are approaches which attempt to provide improved conditioning of the LP equations. Second, there are a number of approaches to bandwidth expansion of the resonances in the LP spectral models.

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.

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,

$$\varepsilon = \sum_n |e(n)|^2. \tag{3}$$

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

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

In vector-matrix notation,

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

The prediction error filter will be denoted by $A(z)$,

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

For the autocorrelation formulation, 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.

2.1 Conditioning and Predictor Coefficient Values

Of importance for implementations using fixed point arithmetic is the dynamic range of the predictor coefficients. Coefficient p_k is the sum of the products of the roots taken k at a time. Since the roots have magnitude less than one, the largest possible value for a coefficient occurs for coefficient $N_p/2$ and is $\binom{N_p}{k}$. For the case of $N_p = 10$, the predictor coefficients can be as large as 252. Many modern speech coders (G.729 [1] and SMV [2], for example) store the predictor coefficients in 16-bit fixed point, with 12 fractional bits (Q12 format). This requires that the predictor values be less than 8 in magnitude.

Problems with the predictor coefficient values will be worst for systems with singularities near the unit circle. These are the systems that are most predictable. The numerical conditioning of a system of equations can be measured by the condition number. The 2-norm condition number is the ratio of the largest eigenvalue to the smallest eigenvalue,

$$\gamma = \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{8}$$

As an example, consider a 240 sample Hamming window and 10'th order LP analysis. The input speech is filtered (modified IRS response) and sampled at 8 kHz. The condition number of the autocorrelation matrix and the maximum predictor coefficient values were measured across a data base of 25 speakers, including two children, for a total of 224,779 frames. Figure 1 below shows a histogram of the condition number expressed in power dB. The largest condition number encountered was 56.4 dB. The largest predictor coefficient generated was 11.0 for a frame with a condition number of 48.7 dB. These frames occur during normal

speech. For instance, the frame with the largest condition number occurs in male speech in the middle of the word "both". The largest predictor coefficient occurs in female speech at the end of the word "floor". In both cases, the waveforms in these regions are somewhat sinusoidal.

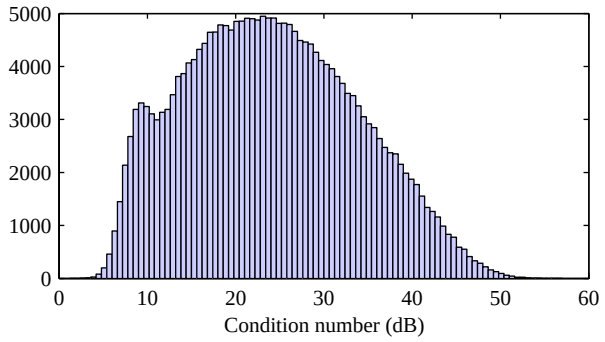

Fig. 1 Histogram of condition numbers (dB)

2.2 Power Spectrum Modification

The eigenvalues of a Toeplitz matrix formed from autocorrelation values are bounded by the minimum and maximum values of the power spectrum. The power spectrum is given by

$$S(\omega) = \sum_{k=-\infty}^{\infty} r(k) e^{-j\omega k}. \tag{9}$$

This sum involves all of the autocorrelation coefficients, not just those that appear in the correlation matrix. The condition number is related to the fluctuations in the power spectrum — a flat spectrum gives the best conditioned equations, while spectra with large dynamic ranges can give badly conditioned equations.

Consider spectra with large peaks. These can be due to sinusoidal components, for instance DTMF tones used for signalling. They can, however, also occur in speech. A high pitched voice (female or child) uttering a nasalized sound can generate a surprisingly sinusoidal waveform. Problems occur because these sinusoidal components are very predictable. For pure sines, only two predictor coefficients per sinusoid are needed to achieve perfect prediction. After applying a tapered window, we no longer have a pure sinusoids, but the ill-conditioning is still present.

A simple modification to reduce the eigenvalue spread is diagonal loading of the correlation matrix (adding a positive term to the zero'th autocorrelation coefficient). In the power spectral domain, this is equivalent to adding a constant white noise term (white noise correction). This reduces the bounds on the eigenvalue spread.

The same approach is used in fractional spacing equalizers for data transmission [3]. There the problem is ill-conditioning due to over-sampling. Adaptive adjustment algorithms such as LMS are subject to having the taps wander into overload regions. The mean-square error criterion can be modified to constrain the sum of the squared coefficient values,

$$\varepsilon' = \varepsilon + \mu \boldsymbol{c}^T \boldsymbol{c}. \tag{10}$$

Taking the derivative with respect to the tap weights gives equations of the same from as earlier, but with the correlation matrix replaced by

$$\boldsymbol{R}' = \boldsymbol{R} + \mu \boldsymbol{I}. \tag{11}$$

Applying diagonal loading by multiplying the zero'th correlation value by the factor 1.0001 improves the condition numbers (see Fig. 2) by compressing the tails of the distribution. The largest condition number is now 47.6 dB and the largest predictor value is now 4.65 (within the range of ± 8 for Q12 representation).

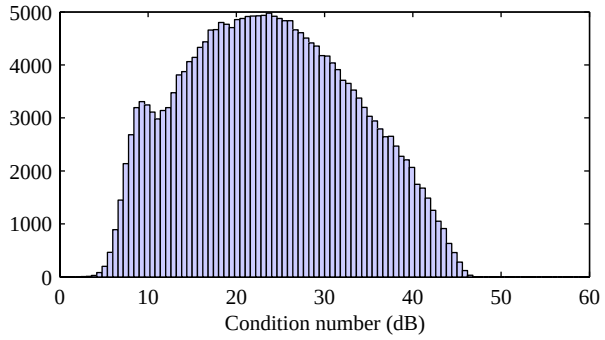

Fig. 2 Histogram of condition numbers (dB) with white noise correction

A slightly different tack to improve conditioning was taken by Atal and Schroeder [4]. They attacked the problem of ill-conditioning due to use of sharp lowpass filters. These filters leave a null in the spectrum near the half-sampling frequency. The solution proposed was to use high-frequency correction by adding the correlation for high-pass noise to the speech correlation matrix. This way, the solution is not as biased as it is when white noise correction is used.

Some wideband (16 kHz sampling rate) coders employ preemphasis of the input signal to help reduce the spectral dynamic range. They also use a higher order analysis. The preemphasis will reduce the condition number in many cases, but not for the frames which are troublesome because the waveform is nearly sinusoidal.

These approaches bias the solutions even when the equations are already well conditioned. An approach which only modifies the equations for ill-conditioned equations would be preferable. For instance, instead of adding a constant power spectrum (white noise correction), a "water-pouring" approach which fills in spectral valleys can be employed,

$$S'(\omega) = \max(S(\omega), \mu r(0)). \tag{12}$$

Alternately, correction can be applied while solving the equations. In a standard Levinson-Durbin recursion, reflection coefficients with magnitude near unity signal ill-conditioning. Such a reflection coefficient can be set to a value away from unity and remaining reflection coefficients can be set to zero (thereby setting the rest of the predictor coefficients to zero also).

3 Bandwidth Expansion

In speech coding resonances correspond to the formant frequencies. Bandwidth expansion is the process of taking a

frequency response and broadening the bandwidths of those peaks. Such an expansion is useful in speech processing to prevent unnatural spectral peaks due to formant/pitch interactions. Modern speech coders use bandwidth expansion in one or both of two places: lag windowing of correlation values before LP analysis and modification of the LP coefficients after LP analysis.

3.1 Time Windows

The input signal is usually windowed with a tapered window (Hamming or other) prior to calculating the correlation values. The effect of the window can be described in the frequency domain as a convolution of the frequency response of the window with the frequency response of the (infinite) signal. This constitutes bandwidth expansion — the main lobe width (measured between zero crossings) for a Hamming window is $8\pi/N$ (asymptotically in N), where N is the window length. For a 240 sample Hamming window (sampling rate 8 kHz), the main lobe width of the frequency response of the Hamming window is 135 Hz between zero crossings and 44 Hz at the 3 dB points.

3.2 Correlation Windowing

Explicit bandwidth expansion prior to LP analysis is done by lag windowing the autocorrelation sequence [5, 6], often with a Gaussian or binomial shaped window. Correlation (lag) windowing corresponds to a periodic convolution of the frequency response of the window with the power spectrum.

Consider a Gaussian window which is a function of continuous time,

$$w(t) = \exp(-\frac{1}{2}(at)^2) \ . \tag{13}$$

The frequency response of this window also has a Gaussian shape,

$$W(\Omega) = \frac{\sqrt{(}2\pi)}{a} \exp(-\frac{1}{2}(\frac{\Omega}{a})^2). \tag{14}$$

The (double-sided) bandwidth measured between the 1 standard deviation points for response is $2a$. The 3 dB bandwidth is bigger by a factor $\sqrt{2\log(2)}$ (about 1.18).

The window is actually applied in discrete-time leading to frequency aliasing. However with reasonably chosen bandwidth expansion factors, the effect of aliasing can be largely ignored in the calculation of the effective bandwidth expansion. The discrete-time window is

$$w[k] = \exp(-\frac{1}{2}(\frac{\pi f_o k}{F_s})^2), \tag{15}$$

where f_o is the *two-sided* bandwidth and F_s is the sampling rate.

The G.729 and SMV coders speech coder use a bandwidth expansion of $f_o = 120$ Hz (so-called 1-σ bandwidth of 60 Hz) relative to a sampling rate $F_s = 8000$ Hz. With no white noise compensation, the largest condition number over the database is 56.0 dB with lag windowing and the largest predictor value is 10.70. This indicates that lag windowing by itself does not improve the conditioning by much. However, note that white noise compensation can itself be represented as a lag window. White noise compensation plus lag windowing can be combined into a single lag window.

3.3 Spectral Damping

Consider a digital filter $H(z)$ and a bandwidth expanded version of this filter. In many cases, it is important that the bandwidth expanded version of the filter have the same form as the original filter. For instance if $H(z)$ is an all-pole filter (as would arise from a standard LP analysis), we want the bandwidth-expanded version to be all-pole. Replacing z by z/α satisfies this requirement. Consider the all-pole filter $H(z) = 1/A(z)$. With bandwidth scaling, the filter has the same form, but with a new set of coefficients,

$$p'_k = \alpha^k p_k. \tag{16}$$

Replacing z by z/α scales the singularities of $H(z)$ inward ($\alpha < 1$) or outward ($\alpha > 1$). For a filter with resonances, choosing $\alpha < 1$ has the effect of expanding the bandwidth of the resonances.

3.3.1 Windowing with an exponential sequence

For a causal filter, the effect of radial scaling of the singularities is such that the impulse response coefficients are modified to become

$$h'[n] = \alpha^n h[n], \tag{17}$$

i.e., the impulse response coefficients are multiplied by an exponential (infinite length) time window. Equivalently, the frequency response of the bandwidth expanded filter is convolved with the frequency response of the window,

$$W(e^{j\omega}) = \frac{1}{1 - \alpha e^{-j\omega}} \ . \tag{18}$$

The 3 dB bandwidth of the frequency response of the window is

$$\text{BW} = 2\cos^{-1}(1 - \frac{(1-\alpha)^2}{2\alpha}) \quad \text{for } 3 - 2\sqrt{2} \le \alpha \le 1 \ . \tag{19}$$

Below the lower limit for α, the response does not decrease sufficiently to fall 3 dB below the peak. This bandwidth is the bandwidth expansion for a very narrow resonance in the LP filter.

3.3.2 Radial scaling of a bandpass filter

For an alternate derivation, consider a continuous-time bandpass filter with a single resonance

$$H(s) = \frac{s}{s^2 + (\Omega_o/Q)s + \Omega_o^2} \ . \tag{20}$$

The corresponding digital filter can be found from the continuous-time filter using a bilinear transformation,

$$z = -\frac{s+a}{s-a} \quad \text{or} \quad s = a\frac{z-1}{z+1} \ . \tag{21}$$

For complex poles, the poles get mapped to new locations $z_{1,2} = r_p e^{\pm j\omega_o}$,

$$H(z) = \frac{z^2 - 1}{z^2 - 2r_p \cos\omega_o \, z + r_p^2} \ . \tag{22}$$

Using the bilinear relationship, the 3 dB bandwidth is

$$\text{BW} = \pi/2 - 2\tan^{-1}(r_1 r_2) \ . \tag{23}$$

where r_1 and r_2 are the radii of the poles of the digital filter.

Consider scaling the z-transform

$$H'(z) = H(z/\alpha) \ , \tag{24}$$

where $0 < \alpha \leq 1$. Then the new poles have radii $r_1' = \alpha r_1$ and $r_2' = \alpha r_2$. The 3 dB bandwidth of the resonance is now

$$\omega_u' - \omega_l' = \pi/2 - 2\tan^{-1}(\alpha^2 r_1 r_2) \ . \tag{25}$$

The difference in bandwidth due to radial scaling by α is

$$\Delta_{\rm BW} = 2\tan^{-1}\Big(\frac{r_1 r_2(1-\alpha^2)}{1+\alpha^2 r_1^2 r_2^2}\Big) \ . \tag{26}$$

In the limiting case of complex poles near the unit circle, the bandwidth expansion is For $r_1 = r_2 = 1$,

$$\Delta_{\rm BW} = \pi/2 - 2\tan^{-1}(\alpha^2). \tag{27}$$

3.3.3 Comparison of Bandwidth Formulas

We now have two formulas, Eqs. (19) and (27) for the effect of radial scaling of the z-transform. This bandwidth expansion expressions derived here can be compared to approximations that have appeared in the literature. Paliwal and Kleijn [7] give the bandwidth expansion (converted to our notation)

$$\Delta_{\rm BW} = -2\log(\alpha). \tag{28}$$

This formulation can be derived from an impulse invariant transformation of a continuous-time exponential sequence. As such it ignores the aliasing effects.

For α near unity, the three expressions (Eqs. (19), (27) and (28)) give similar results. An examination of the Taylor series for the three expressions shows that they agree in value and the first derivative at $\alpha = 1$,

$$\Delta_{\rm BW} = 2(1-\alpha) + (1-\alpha)^2 + O((1-\alpha)^3). \tag{29}$$

The bandwidth expansions given by the different expressions are plotted in Fig. 3. The ordinate is the bandwidth expansion normalized to the sampling frequency, $(\Delta_{\rm BW}/(2\pi))$. From this figure we see that the simple two term Taylor series from Eq. (29) is itself a good estimate of the bandwidth expansion for useful values of α.

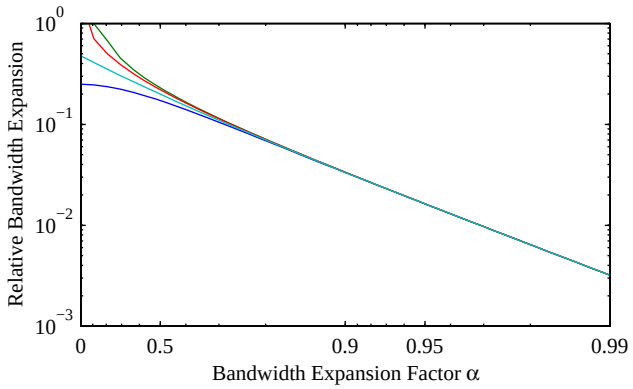

Fig. 3 Relative bandwidth expansion $\Delta_{\rm BW}/(2\pi)$ as a function of the bandwidth expansion parameter α. From bottom to top, the curves are the different estimates of $\Delta_{\rm BW}$: $\pi/2 - 2\tan^{-1}(\alpha^2)$, $2(1-\alpha)+(1-\alpha)^2$, $-2\log(\alpha)$, and $2\cos^{-1}(1-(1-\alpha)^2/(2\alpha))$

In speech coding, bandwidth expansions of 10 Hz to 25 Hz are used. With a sampling rate of 8000 Hz, these end points correspond to relative bandwidth expansion values of α of 0.996 and 0.990 (for $r = 1$).

3.4 Line Spectral Frequencies

Line spectral frequencies (LSF's) are a representation of LP coefficients. The LSF's are an ordered set of values in the range $(0, \pi)$. Closely spaced LSF's tend to indicate a spectral resonance at the corresponding frequency. Several standard coders impose minimum separation constraints on the LSF's. A bandwidth expansion scheme, albeit for use as a postfilter for speech processing, is described in [8]. In that paper, the LSF's are pushed apart by interpolating the given LSF's with a set corresponding to a flat spectrum (equally spaced LSF's).

An experiment was carried out to investigate this type of bandwidth expansion. An artificial LP filter with a single constant bandwidth resonance was created. Bandwidth expansion of the LSF's was applied. As the centre frequency of the resonance was varied, the resulting bandwidth depended on the centre frequency, varying by more than ±10%. This result shows that amount of bandwidth expansion is not consistent and depends on the frequency of the resonances.

4 Summary

This paper has demonstrated the problems of ill-conditioning of the LP equations for speech signals. The standard technique of white noise compensation ensures that the prediction coefficients values are reasonable. Bandwidth expansion to prevent abnormally narrow formant peaks can be provided by lag windowing of the correlation values and/or spectral damping applied to the prediction coefficients. A new simple formula for calculating the bandwidth expansion by spectral damping has been given.

References

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

[2] 3GPP2 Document C.S0030-0, *Selectable Mode Vocoder Service Option of Wideband Spread Spectrum Communication Systems*, Version 2.0, Dec. 2001.

[3] R. D. Gitlin, J. F. Hayes, and S. B. Weinstein, *Data Communications Principles*. Plenum, 1992.

[4] B. S. Atal and M. R. Schroeder, "Predictive coding of speech signals and subjective error criteria," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-27, pp. 247–254, June 1979.

[5] Y. Tohkura, F. Itakura, and S. Hashimoto, "Spectral smoothing technique in PARCOR speech analysis-synthesis," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-26, pp. 587–596, Dec. 1978.

[6] Y. Tohkura and F. Itakura, "Spectral sensitivity analysis of PARCOR parameters for speech data compression," *IEEE Trans. Acoustics, Speech, Signal Processing*, vol. ASSP-27, pp. 273–280, June 1979.

[7] K. K. Paliwal and W. B. Kleijn, "Quantization of LPC parameters," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), ch. 12, pp. 433–466, Elsevier, 1995.

[8] H. Tasaki, K. Shiraki, K. Tomita, and S. Takahashi, "Spectral postfilter design based on LSP transformation," *IEEE Workshop on Speech Coding*, pp. 57–58, Sept. 1987.