On the Measurement of Waveform Coder Distortion Using the Log Likelihood Ratio

R. E. Crochiere

*J. M. Tribolet**

L. R. Rabiner

Acoustics Research Department
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

The log likelihood measure has been widely used in speech recognition for comparing speech signals. Recently it has been proposed as a measure for assessing the quality of coded speech. In this paper we present an interpretation of the log likelihood ratio measure within the theoretical framework of a waveform coder distortion model.

1. Introduction

The purpose of this paper is to present an interpretation of the log likelihood ratio measure [1-3] within the theoretical framework of a generalized waveform coder distortion model. As such, we first define the log likelihood ratio measure and then give an interpretation of this measure in terms of a generalized coder distortion model. Finally we discuss the implications of the results and show how they can be applied to objectively measure coder quality.

II. The Log Likelihood Ratio

The principal assumption on which the log likelihood ratio distance is based is that speech can be represented by a p^{th} order all-pole model of the form

$$x(n) = \sum_{m=1}^{p} a_m x(n-m) + G_x u(n) \tag{1}$$

where $x(n)$ is the sampled speech signal, $a_m (m=1,2,...,p)$ are the coefficients of an all-pole filter, $1/A_x(z)$, which models the resonances of the speech production mechanism, G_x is the gain of the filter, and $u(n)$ is an appropriate excitation source for the filter.

The waveform coder can be represented as shown in Fig. 1 in which $x(n)$ is the input speech, which can be modelled according to Eq. (1), and $y(n)$ is the decoded output.

The log likelihood ratio for comparing $x(n)$ and $y(n)$ can then be defined as [1]

$$l = \log\left[\frac{\mathbf{a}_x R_y \mathbf{a}_x^t}{\mathbf{a}_y R_y \mathbf{a}_y^t}\right] \tag{2}$$

where

$\mathbf{a}_x =$ LPC coefficient vector $(1, a_1, a_2, \ldots, a_p)$ for the original speech $x(n)$

$\mathbf{a}_y =$ LPC coefficient vector $(1, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p)$ for the coded speech $y(n)$

and R_y is the correlation matrix of $y(n)$ whose elements are

$$r|(i-j)| = \sum_{n=1}^{N-|i-j|} y(n) y(n+|i-j|), \quad i,j = 0,1,...,p-1 \tag{3}$$

where N is the number of samples used in the analysis (i.e., the frame size).

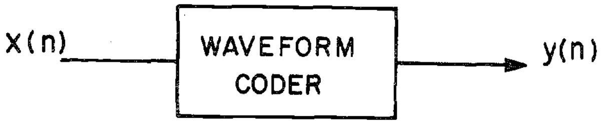

Fig. 1. Waveform Coder

An interpretation of the log likelihood ratio can be given with the aid of Fig. 2 [2]. The filter $A_y(z)$, defined as

$$A_y(z) = 1 + \sum_{i=1}^{p} \hat{a}_i z^{-i}, \tag{4}$$

is the inverse of the all-pole filter which models the spectrum of $y(n)$, and $A_x(z)$ is a similarly defined inverse filter for the signal $x(n)$. When $y(n)$ is filtered with its inverse spectral model, $A_y(z)$, the output signal corresponds to the minimum prediction error or residual error of the LPC model of $y(n)$. The energy of this residual signal (over the speech segment) is defined as α, and it can be shown that $\alpha = G_y^2 = \mathbf{a}_y R_y \mathbf{a}_y^t$. Similarly if $y(n)$ is passed through the filter $A_x(z)$, the output corresponds to another prediction residual whose energy over the same speech segment is $\beta \geqslant \alpha$, where $\beta = \mathbf{a}_x R_y \mathbf{a}_x^t$. The equality exists only when $A_y(z) = A_x(z)$. From Eq. (2) it can now be seen that the log likelihood ratio has the form

$$l = \log(\beta/\alpha) \tag{5}$$

An alternative interpretation of the likelihood measure, which is illustrated in Fig. 3, is based on the equation

$$l = \log(\beta'/\alpha') \tag{6a}$$

$$= \log\left|\frac{G_x^2}{G_y^2} \cdot \frac{\mathbf{a}_x \mathbf{R}_y \mathbf{a}_x^t}{\mathbf{a}_x \mathbf{R}_x \mathbf{a}_x^t}\right| \tag{6b}$$

where G_x and G_y are the gains of the linear predictive representations of $x(n)$ and $y(n)$, as defined in Eq. (1).

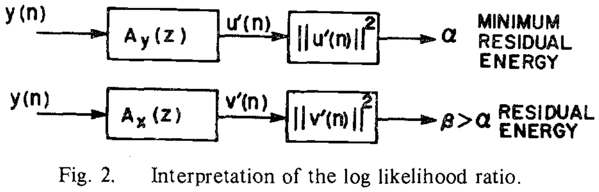

Fig. 2. Interpretation of the log likelihood ratio.

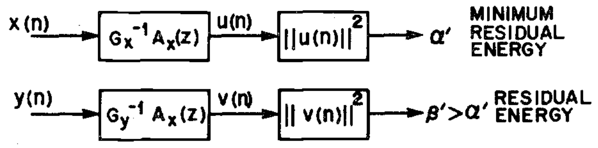

Fig. 3. An alternate interpretation of the log likelihood ratio.

*Also with the Instituto Superior Técnico, Lisbon, Portugal

CH1559-4/80/0000-0340$00.75 © 1980 IEEE

III. Waveform Coder Model

Figure 4 shows a block diagram of the waveform coder distortion model which we have investigated. This model, also used by Aaron et al [7] for a delta modulator is composed of a time-varying linear filter and an additive noise source. The rationale for this model is that the time-varying filter, $h(n)$, models the "linearly correlated" distortions in the coder (i.e. attenuation, delay, bandlimiting, reverberation) and the noise source, $e(n)$, accounts for the nonlinear distortions in the coder (i.e., additive noise, tonal noise, clicks, etc.).

Given the input $x(n)$ and the output $y(n)$ of the coder, the problem of determining the two components of the model becomes a classical system identification problem (assuming $h(n)$ to be of (finite) duration $\hat{M}$ samples) under noisy conditions [8]. Once an estimate of the filter $\hat{h}(n)$ is obtained, the "uncorrelated noise component"* $\hat{e}(n)$ can be estimated according to Fig. 5. The solution, of course, is not unique and is subject to careful interpretation.** When properly used, however, the model can provide useful insights into the dynamics of a coder.

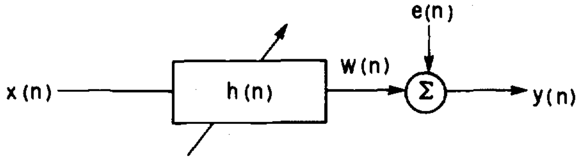

Fig. 4. A general distortion model for a waveform coder.

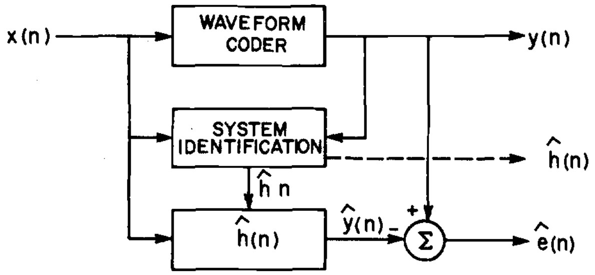

Fig. 5. Identification of the parameters of the coder distortion model.

The first issue which was considered in using this model is the rate at which $h(n)$ was allowed to vary. For the model to be perceptually meaningful, the filter should vary at a rate which is detectable to the ear as a time-varying filter, but not faster (i.e., approximately at the same rate at which formants vary in speech production). Any changes which are more rapid should show up in the noise component of the model. Based on this reasoning, 12-20 msec segments of speech were used in computing estimates of $h(n)$. Estimates were then interpolated (using overlapping segments) every 2-5 msec to ensure smooth changes of the filter.

Another important feature of the model which had to be carefully chosen was the number of samples, $\hat{M}$, in the impulse response $\hat{h}(n)$. An estimate for this value was determined by examining the ratio of the power of the "uncorrelated noise" component, $\hat{e}(n)$, to the total distortion $d(n) = y(n) - x(n)$ of the coder as a function of $\hat{M}$ the assumed length of $\hat{h}(n)$. Figure 6 shows an example of this noise power ratio, denoted as $\sigma_{\hat{e}}^2/\sigma_d^2$ (expressed in dB), as a function of $\hat{M}$ for an ADPCM coder in an overload region (where $\sigma_{\hat{e}}^2$ denotes the power of $\hat{e}(n)$ and σ_d^2 denotes the power of $d(n)$). The curves are normalized to 0 dB at $\hat{M} = 0$ since $\hat{e}(n) = d(n)$ at this point. Three speech regions were analyzed, including an unvoiced region, a semi-voiced region, and a strong voiced region. In the unvoiced region it is seen that most of the coder distortion $(d(n))$ is due to the "uncorrelated component" $(\hat{e}(n))$, whereas in the strong voiced region most of the distortion is due to the attenuation in the coder (because of overloading). Also it is seen that most of the separation of the "uncorrelated" and "correlated" components of distortion in the coder can be achieved with a two point $(\hat{M}=2)$ filter model $\hat{h}(n)$.

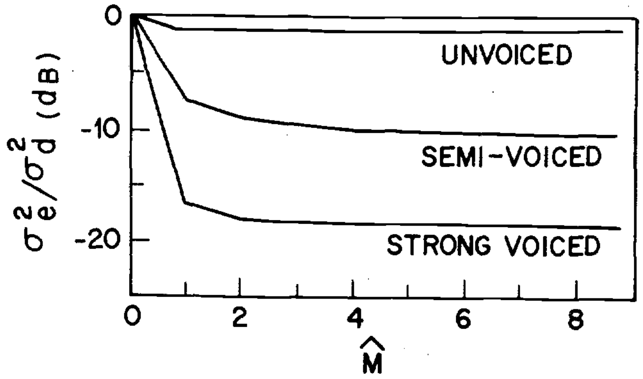

Fig. 6. "Uncorrelated noise" to total-noise ratio for the waveform coder.

This interpretation seems reasonable since it suggests that the filtering distortion for this type of coder is primarily that of a pure attenuation (due to clipping) and a spectral tilt (due to loss of high frequencies). For example, Fig. 7 shows typical frequency responses of the filter for three different input signal levels (−18 dB, 0 dB = Optimum level, and +18 dB) for the ADPCM coder in a strong voiced region.

An important consideration in choosing the filter length, $\hat{M}$, is that it should not be longer than necessary, due to the sensitivity of the system identification analysis to the coder noise [8]. For example, Fig. 8 shows the frequency response of an $\hat{M} = 6$ point filter for the same speech segment as in Fig. 7 (+18 dB condition). The large spectral variations of the filter apparently have no physical significance since the curves of Fig. 6 show clearly that little change occurs in $\hat{e}(n)$ as $\hat{M}$ goes from 2 to 6. Thus the filter coefficients are essentially those of a large class whose output is orthogonal to the error signal.

IV. An Interpretation of the Log Likelihood Ratio

By combining the likelihood ratio model in Fig. 3 with the coder model in Fig. 4 an interesting interpretation of this spectral distance can be given. Figure 9 illustrates this combination. The speech model defined in Fig. 9 is the all-pole filter $G_x/A_x(z)$ which is excited by the normalized excitation source $u(n)$. This excitation source is defined to be the normalized residual error in the LPC

* In this paper we refer to $\hat{e}(n)$ as the uncorrelated noise component of the filter model. Strictly speaking, $\hat{e}(n)$ is short-time uncorrelated with $x(n)$ for only the first $\hat{M}$ lags.

** We assume here that any fixed filtering performed in the coder, such as band limiting, is also performed on the reference $x(n)$ prior to the estimation of $\hat{h}(n)$. Therefore, $\hat{h}(n)$ as we refer to it here, only accounts for the modification of the spectral shape due to quantization and *not* to any filtering that may be performed in the coder. Later when we refer to the use of this model in the interpretation of the log likelihood ratio we will assume that the filter in the coder distortion model contains *both* the effets of fixed filtering and spectral attenuation due to quantization effects.

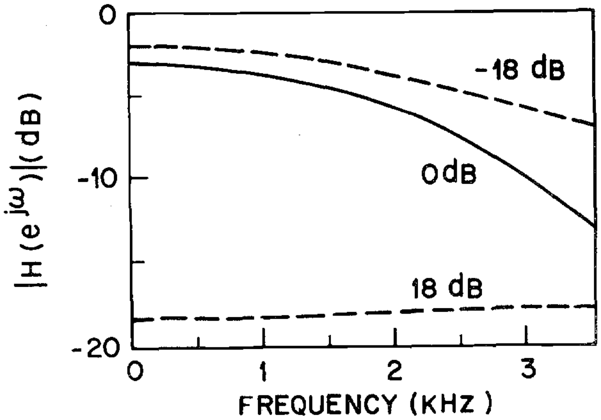

Fig. 7. Typical frequency response for a strong voiced region of an ADPCM coder.

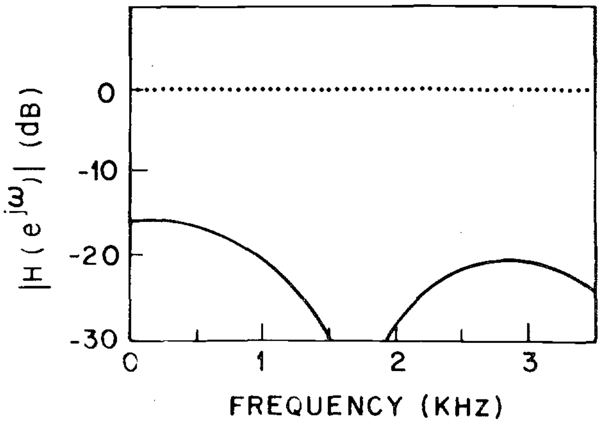

Fig. 8. Frequency response of the model when $\hat{M}$ is too large.

model of $x(n)$ and therefore the output of the model is exactly $x(n)$. The coder output $y(n)$ is filtered by the inverse filter $\frac{1}{G_y}A_x(z)$ to produce the normalized residual $v(n)$. In the absence of coder errors it is seen that $y(n) = x(n)$ and $G_y = G_x$ and therefore $v(n) = u(n)$. In this case the likelihood ratio, as defined in Eq. (6), is zero.

From Fig. 9 it can be seen that

$$v(n) = \frac{G_x}{G_y}\cdot u(n)*h(n) + \frac{1}{G_y}\cdot e(n)*a(n) \qquad (7)$$

where $a(n)$ is the impulse response of $A_x(z)$. Therefore

$$\beta' = \|v(n)\|^2 = \|\frac{G_x}{G_y}\cdot u(n)*h(n) + \frac{1}{G_y}\cdot e(n)*a(n)\|^2 \qquad (8)$$

Assuming that the noise component $e(n)$ in the coder is uncorrelated with $u(n)$, Eq. (8) can be expressed as

$$\beta' = \frac{G_x^2}{G_y^2}\|u(n)*h(n)\|^2 + \frac{1}{G_y^2}\|e(n)*a(n)\|^2 \qquad (9)$$

Also it is assumed that $u(n)$ is a spectrally flat signal, as appropriate for the LPC model of $x(n)$, and that it is normalized so that $|U(e^{j\omega})| \cong 1$. Therefore it can be shown that

$$\alpha' = \|u(n)\|^2 = 1 \qquad (10)$$

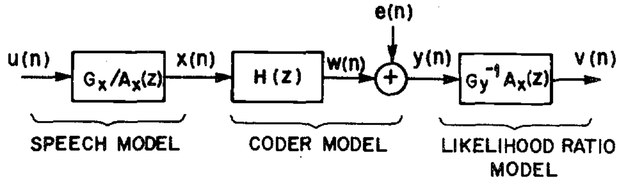

Fig. 9. Combined speech model, coder distortion model, and likelihood ratio model

and

$$\|u(n)*h(n)\|^2 = \|h(n)\|^2 \qquad (11)$$

Substituting Eqs. (9)-(11) into Eq. (6) then gives

$$l \cong \log\left[\frac{G_x^2}{G_y^2}\|h(n)\|^2 + \frac{1}{G_y^2}\|e(n)*a(n)\|^2\right] \qquad (12)$$

From the LPC model it can also be noted that the smoothed spectral estimate of $x(n)$, denoted as $|\tilde{X}(e^{j\omega})|$ is

$$|\tilde{X}(e^{j\omega})| = |G_x/A_x(e^{j\omega})| \qquad (13)$$

Finally, combining Eq. (13) with Eq. (12) gives

$$l \cong \log\left[\frac{G_x^2}{G_y^2}\right] + \log\left[\|H(e^{j\omega})\|^2 + \|\frac{E(e^{j\omega})}{\tilde{X}(e^{j\omega})}\|^2\right] \qquad (14)$$

This form clearly illustrates the properties of the log likelihood ratio, within the context of the generalized coder distortion model. The first component of l in Eq. (14) is simply associated with a dynamic gain loss and can essentially be neglected whenever the original and the coded speech are gain normalized (unless the dynamic component of the gain varies widely).

The second term in Eq. (14) has two components. The first component is due to the "correlated distortion" in the coder, denoted by the term $\|H(e^{j\omega})\|^2$ and the second term is due to the "uncorrelated" noise component $E(e^{j\omega})$ which is inversely weighted by the smoothed LPC spectrum $G_x/A(e^{j\omega})$, of the input speech signal. As seen from Eq. (14) these two components of distortion are, in effect, weighted equally in the log likelihood ratio.

In terms of predicting subjective quality, it is known that an equal weighting of these two components is not the most desirable [6]. However, the functional form of the log likelihood ratio seems to be a good candidate for predicting subjective quality when only one of the components of distortion is significant. For example, in waveform coders in which there is no loss of bandwidth or attenuation of certain frequency bands, this measure can be useful in predicting subjective quality [6]. Also, in the case of vocoders where the predominant form of distortion is a spectral distortion (which might be associated with the term $\|H(e^{j\omega})\|^2$) this measure has been found to be a good candidate as a predictor of subjective quality [4].

When both components of distortion are simultaneously present, however, the equal weighting of them in the log likelihood measure does not appear to be the most appropriate choice. The two components must be measured separately and then combined, with unequal weighting, to obtain a useful single measure. This has prompted an investigation of several related measures, whose definition was motivated by the model of Figure 9 and Eq. (14).

A reasonable approach to decoupling these components is to define two log likelihood measures, l_D and l_N, one associated only with the spectral distortion due to $H(e^{j\omega})$ and the other associated only with the additive noise effects $E(e^{j\omega})$. Letting $|H(e^{j\omega})| = 1$ and $G_x = G_y$, Eq. (14) reduces to

$$l_N \underline{\Delta}\, l\big|_{|H(e^{j\omega})|=1} = \log\left[1+\|\frac{E(e^{j\omega})}{\tilde{W}(e^{j\omega})}\|^2\right] \qquad (15)$$

If additive noise is not to be considered, ($E(e^{j\omega})=0$, and assuming $G_x=G_y$) Eq. (14) reduces to

$$l_D \underline{\Delta}\, l\big|_{E(e^{j\omega})=0} = \log\left[\|H(e^{j\omega})\|^2\right] \qquad (16)$$

Based on these two independent measures one may then attempt to combine them in an optimal way, so as to predict subjectively evaluated waveform coded speech quality.

A number of objective measures have been defined in the literature based on variations of the formulas of Eqs. (15) and (16). In particular, one modification that has been relatively successful [6] is a linear combination of a bandwidth measure and noise-to-signal ratio measure which has a similar functional form to that of Eq. (15).

V. Conclusions

Based on a generalized waveform coder distortion model, an interpretation of the log likelihood ratio measure was developed. The insight gained from such interpretation suggests several alternate ways of accounting for effects of coding distortions.

References

[1] F. Itakura, "Minimum prediction residual principle applied to speech recognition," IEEE Trans. Acoust. Speech and Sig. Proc., Vol. ASSP-23, pp. 67-72, February 1975.

[2] A. H. Gray, Jr. and J. D. Markel, "Distance Measures for speech processing," IEEE Trans. Acoust. Speech and Sig. Proc., Vol. ASSP-24, pp. 380-391, October 1976.

[3] R. E. Crochiere, L. R. Rabiner, N. S. Jayant, and J. M. Tribolet, "A study of objective measures for speech waveform coders," Proceedings of the 1978 Zurich Seminar on Digital Communications, pp. H1.1-H1.7.

[4] T. P. Barnwell, III, A. M. Bush, R. M. Mersereau, and R. W. Schafer, "Speech quality measurement," Georgia Institute of Technology, Report No. E21-655-77-TB-1, June, 1977.

[5] C. Scagliola, "Evaluation of adaptive speech coders under noisy channel conditions," Bell Syst. Tech. J., Vol. 58, No. 6, pp. 1369-1394, July-Aug. 1979.

[6] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A comparison of the performance of four low-bit-rate speech waveform coders," Bell Syst. Tech. J., Vol. 58, No. 3, pp. 699-712, March 1979. See also (same authors), "A study of complexity and quality of speech waveform coders," Proc. Int. Conf. on Acoust. Speech and Signal Processing, April 10-12, Tulsa, Okla., pp. 586-590.

[7] M. R. Aaron, J. S. Fleischman, R. W. McDonald, and E. N. Protonatarios, "Response of delta modulation to Gaussian signals," Bell Sys. Tech. J., pp. 1167-1195, (May-June, 1969).

[8] L. R. Rabiner, R. E. Crochiere, and J. B. Allen, "FIR system modeling and identification in the presence of noise and bandlimited inputs," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-26, No. 4, pp. 319-333, August 1978.