Proc. IEEE Globecom Conf. (Orlando, FL), pp. 452-456, Dec. 1992

Rate-Distortion Function for Speech Coding based on Perceptual Distortion Measure†

Aloknath De[1] and Peter Kabal[1,2]
[1]Dept. of Elec. Eng., McGill University, 3480 University Street, Montréal, Canada—H3A 2A7
[2]INRS–Télécommunications, Université du Québec, 16 Place du Commerce, Verdun, Canada—H3E 1H6

Abstract: In [1], we have proposed a perceptual distortion measure for speech coders using an auditory (cochlear) model. This measure evaluates the neural-firing cross-entropy of the coded speech with respect to that of the original one. In this paper, the output space of the cochlear model is explored using this measure form so as to verify the existence of the pitch and the formant information. However, the prime objective of this article is to provide a rate-distortion analysis for speech coding. We evaluate a lower bound to the rate-distortion function based on this distortion measure and also compute the exact rate-distortion function using the Blahut algorithm. Four state-of-the-art speech coders with rates ranging from 4.8 kbps (CELP) to 32 kbps (ADPCM) are studied from the viewpoint of their performances with respect to the rate-distortion limits.

1 Introduction

In [1], we have used an auditory (cochlear) model to propose a distortion measure, namely the *cochlear discrimination*, for coded/distorted speech signals. This article uses this measure for two purposes. The first part describes an analysis procedure for detecting the pitch and the formant information by exploring the output space of the cochlear model and applying the discrimination measure form over some of the stages and some of the sampling times. This analysis, in turn, provides a feedback to the designer for improving any particular section of the speech coder by manifesting a new strategy or redistributing the available bits in a more efficient manner. The second part provides a rate-distortion-theoretic analysis (a lower bound as well as the direct evaluation) for the speech coder based on the introduced measure. The rate-distortion function calculates the effective rate at which the source produces information subject to the constraint that a specified average distortion is endured at the destination.

The remainder of the article is organized as follows. Section 2 briefly reviews the cochlear discrimination measure introduced in [1]. Section 3 proposes algorithms for estimating the pitch and the formant frequencies using the cochlear discrimination measure form. Section 4 defines the rate-distortion function $R(D)$ mathematically. Then it describes the evaluation procedure of $R(D)$ by characterizing the source-destination pair for speech coders, providing a lower bound to the $R(D)$ and also computing it directly using the Blahut algorithm. Finally, Section 5 compares different speech coders' performances with respect to these bounds.

† This research was supported by a grant from the Canadian Institute for Telecommunications Research under the NCE program of the Government of Canada

2 Auditory Distortion Measure

In this section, we briefly describe the formulation of the auditory distortion measure which involves mapping the *time-domain* speech signal onto a *perceptual-domain* and comparing these paramaters for the original and the coded speech.

2.1 Perceptual Representation of Speech

A cochlear model, designed by Lyon [2] and described in [1, 3], is used for mapping onto a perceptual domain where the time-place components become the fundamental bases of analysis. Fig. 1 shows the block diagram of this cochlear model.

The outer-and-middle ear effectively adds a slight high-pass response to the system. The propagation of the *basilar membrane* (BM) along the cochlea is modeled using a cascade of sixty-four biquad filters. Each *place* along the BM responds best to one frequency, termed as the *characteristic frequency* (f_c). The composite transfer function at any place is an asymmetric band-pass function for which the 3-dB bandwidth is defined as $W_{\text{ear}}(f_c) = \sqrt{f_c^2 + f_{\text{eb}}^2}/Q_{\text{ear}}$, where the ear-break frequency f_{eb} is 1,000 Hz and Q_{ear} is 8.

The movement of the BM is sensed by the inner hair cells. This is incorporated in the model with a series of half wave rectifiers (HWRs) that detect the output of each second order band-pass filter. Finally, a cascade of four automatic gain control (AGC) stages with different time constants, simulating different adaptation times in the ear, is used. These AGCs amplify the weak signals and diminish the strong signals. Coupling among the AGC stages emulate *lateral inhibition* by which the sensory neurons reduce their own gain as well as the gain of the others nearby.

The perceptual domain representation is a sequence of N-dimensional vectors at all sampling times. With each of the N cochlear stages (here, $N = 64$) and n-sampling times (n depending on the speech segment), is associated a neural converter which 'fire's (i.e., sends impulses) based on the probability-of-firing information (related to the compressed signal sensed by the hair cells).

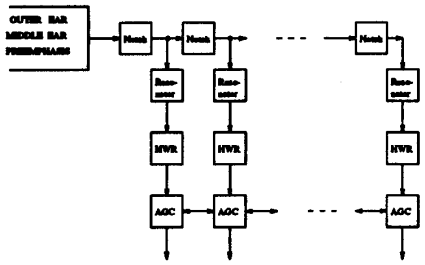

Fig. 1 Lyon's cochlear model

2.2 Cochlear Discrimination Information Measure

These neural converters may be considered as discrete information sources with an alphabet of two (i.e., firing and non-firing). Discrimination information, a powerful tool [4] for quantifying the 'closeness' of two probability distribution functions, is used to formulate the *cochlear discrimination* measure (D_α) in the Shannon entropy (with $\alpha = 1$) as well as in the Rényi entropy (with $\alpha \neq 1$) sense.

Let $p_{1|k}$ and $p_{2|k} = 1 - p_{1|k}$ be the firing and the non-firing conditional probabilities at some time t corresponding to the original speech signal conditioned on the fact that the measurement is for the k-th stage. Similarly, $q_{1|k}$ and $q_{2|k} = 1 - q_{1|k}$ are defined for the coded/distorted speech. For any stage, the neural sources corresponding to n-speech samples are assumed to form a product source whose probability distribution is the product distribution (*i.e.*, $P^n = \prod_{l=1}^{n} P_l$ and $Q^n = \prod_{l=1}^{n} Q_l$). Similarly, for any speech sample, sources corresponding to N-stages are also assumed to form a product source. Under these assumptions, $D_\alpha(P^n; Q^n|k^N) = \sum_{l=1}^{n} \sum_{k=1}^{N} D_\alpha(P_l; Q_l|k)$.

$$D(P^n; Q^n|k^N) = \sum_{l=1}^{n} \sum_{k=1}^{N} \left\{ \sum_{j_l=1}^{2} p_{j_l|k} \log \left(\frac{p_{j_l|k}}{q_{j_l|k}} \right) \right\} \quad (1)$$

$$D_\alpha(P^n; Q^n|k^N) = \sum_{l=1}^{n} \sum_{k=1}^{N} \left\{ \frac{1}{(\alpha - 1)} \log \left(\sum_{j_l=1}^{2} \frac{p_{j_l|k}^{\alpha}}{q_{j_l|k}^{\alpha-1}} \right) \right\} \quad (2)$$

One generalized form of the discrimination measure is $D_{\text{gen}}(P; Q) = \sum_{j=1}^{J} q_j f\left(\frac{p_j}{q_j}\right)$ where f is any convex function. In addition to the directed divergence form of (1) with $f(x) = x \log x$, we consider the variational distance with $f(x) = |x - 1|$. These measures, in effect, determine the amount of new information (the increase in neural source entropy) associated with the coded speech when the neural source entropy associated with the original speech is known [1].

3 Information Content Analysis

Several time-domain (e.g., zero-crossing rate, autocorrelation, average magnitude difference function) as well as frequency-domain (e.g., narrowband and wideband spectrograms) algorithms are available for estimating the pitch frequencies [5]. For this purpose, Seneff has used an auditory model and suggested a generalized synchrony detection (GSD) measure [6]. Recently, Slaney et al. [7] have introduced a perceptual pitch detector based on Licklider's 'duplex theory' of the pitch perception. Similar to the estimation of pitch, estimating the first three or four formants is also of importance especially in the speech analysis leading to a formant-based synthesis. Conventional formant estimation techniques adopt the policy of *peak-picking* in the spectral representation [5]. The GSD measure has also been used in [6] to form a 'pseudo-spectrogram' for the spectral estimation. In the following, we suggest algorithms to estimate the pitch and formant frequencies using the cochlear discrimination measure form.

3.1 Pitch Estimation

The algorithmic steps to determine pitch periods for speech frames are described below.

Step I: We consider a rectangular analysis window of 40 ms for each speech frame of 20 ms (160 samples) so that the successive windows overlap by 50% and atleast two pitch periods are included in the analysis window. For each sampling time, the output of each stage is compared with itself delayed by 'τ' samples (a maximum of 20 ms = 160 time lags accounting for the highest possible pitch frequency). The comparison, made to calculate the discrimination measure for the k-th stage with a measure lag τ, is given by

$$R_k(\tau) = \begin{cases} \sum_{l=1}^{160} \sum_{j_l=1}^{2} |p_{j_l|k} - p_{j_{l+\tau}|k}|, & \text{Var. Dist.,} \\ \sum_{l=1}^{160} \sum_{j_l=1}^{2} p_{j_l|k} \log \left(\frac{p_{j_l|k}}{p_{j_{l+\tau}|k}} \right), & \text{Dir. Div.}(\alpha = 1), \\ \sum_{l=1}^{160} \frac{1}{\alpha - 1} \log \left(\sum_{j_l=1}^{2} \frac{p_{j_l|k}^{\alpha}}{p_{j_{l+\tau}|k}^{\alpha-1}} \right), & \text{Dir. Div.}(\alpha \neq 1). \end{cases} \quad (3)$$

In this way, from a two-dimensional perceptual representation, we derive a two-dimensional *cross-entropogram* where the vertical direction corresponds to the place and the horizontal direction corresponds to the time lag.

Step II: To enhance the vertical structure in the cross-entropogram, a convolutional operator [-1 +2 -1] is used.

Step III: An exponential weighting $w[k]$ is applied to the k-th stage as

$$w[k] = e^{-a(k-1)/N}, \quad \text{for } k = 1, 2, \cdots, 64, \quad (4)$$

where the lowest index for k corresponds to the lowest frequency stage and so on. The number of stages N is sixty-four and the parameter a is chosen to be 7.

Step IV: The exponentially-weighted discrimination measure values for all the sixty-four stages are summed up for each of the 160 time lags. The evidences from the higher harmonics are combined this way to make the pitch estimate robust. At the same time, the contributions from the formant frequencies are minimized by providing exponentially decaying weights to the higher frequency stages. In this flattened one-dimensional cross-entropogram, the measure value becomes lowest at the time lag corresponding to the pitch period. The reciprocal of the time period at which the dip occurs gives an average F0 estimate.

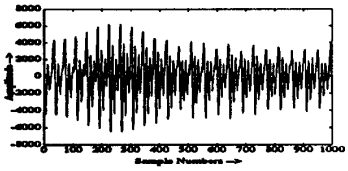

Fig. 2 Time-domain waveform of the vowel /a/ in the word 'shade' (female voice)

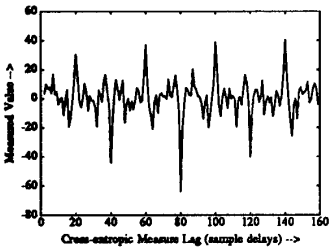

Fig. 3 One-dimensional cross-entropogram (dir. div. with α=1) for the first 160 samples of /a/

Fig. 2 shows the time-domain waveform for the vowel /a/ in the word 'shade' (female voice). We execute our pitch estimation algorithm to extract the pitch period information for the first frame (160 samples) of /a/ in that word. In the one-dimensional cross-entropogram plot (using the dir. div. with $\alpha = 1$) of Fig. 3, we observe a dip located at the measure lag of 40 samples (equivalently, 5 ms). The perceptual pitch period is thus calculated to be 200 Hz. It has been verified that the algorithm executes successfully even if the fundamental frequency or the first harmonic component is filtered out from the original signal.

3.2 Formant Estimation

For detecting the formant frequencies, steps I and II of the pitch detection algorithm are followed in the same manner. To assess the neural activity in the k_c-th stage, an exponential weighting set $\{w_c[k]\}$ is applied for the k-th stage as

$$w_c[k] = e^{-a|k-k_c|/N}, \quad \text{for } k = 1, 2, \cdots, 64. \tag{5}$$

The exponentially-weighted discrimination measure values for all the sixty-four stages are summed up for each of the 160 sample delays. The formant frequencies (F1–F3) are determined from the measured values for the dips. For our example, they are found to be 410 and 1,080 and 2,200 Hz.

4 Rate-Distortion Analysis

Now, we define the rate-distortion function $R(D)$ and characterize the *source-destination pair* for speech coding. A lower bound to the $R(D)$ is calculated with the cochlear variational distance and the Blahut algorithm [8] is applied for the direct evaluation of the $R(D)$ with the cochlear directed divergence (for α=1) and also with the variational distance.

4.1 Preliminaries for Rate-Distortion Analysis

A *source-destination pair* is generally characterized by a probabilistic model of the source encoder and a fidelity criterion measuring the degradation (D) of the coded source output in reference to the original source. With any such source-destination pair, a function $R(D)$, termed as the rate-distortion function, may be associated. This function calculates the effective rate at which the source produces information under the constraint that an average distortion of D is endured at the destination. As D increases, $R(D)$ decreases monotonically and usually becomes zero at some finite value of distortion.

We consider a time-discrete source $\{X_t, P\}$ which produces *i.i.d.* outputs described by an absolutely continuous probability distribution $P(x)$ with density $p(x)$. The accuracy of reproduction of x by y is measured by a non-negative function $\rho(x, y)$, commonly termed as a distortion measure. An average distortion

$$d(q) = \int\int p(x)q(y|x)\rho(x, y)\, dx\, dy \tag{6}$$

and an average mutual information

$$I(q) = \int\int p(x)q(y|x) \log\left\{\frac{q(y|x)}{q(y)}\right\} dx\, dy, \tag{7}$$

where $q(y) = \int p(x)q(y|x)\, dx$ are assigned to every conditional probability density $q(y|x)$. Then, the $R(D)$ of $\{X_t, P\}$ with respect to the fidelity criterion is defined by $R(D) = \inf I(q)$ over $q \in Q_D$, where the set of all D-admissible conditional probability assignments is denoted by the symbol $Q_D = \{q(y|x) : d(q) = D\}$. $I(q)$ is a convex downward function of q which implies that any stationary point of $I(q)$ in Q_D must yield the absolute minimum, namely $R(D)$. For the above convex programming problem, the following parametric expressions are obtained for D and R [9]:

$$D = \int\int \lambda(x)p(x)q(y)e^{s\rho(x,y)}\rho(x, y)\, dx\, dy \tag{8}$$

and

$$R = sD + \int p(x) \log \lambda(x)\, dx, \tag{9}$$

where

$$\lambda(x) = \left[\int q(y)e^{s\rho(x,y)}\, dx\right]^{-1}. \tag{10}$$

The slope of any $R(D)$ curve at the point (D_s, R_s) is represented by the parameter s which is generated parametrically from (8), (9) and (10).

4.2 Source-Destination Pair Characterization

It may prove to be sufficiently difficult to express the nonlinear signal processing operations of the cochlear model with the help of simple mathematical operators. Thus, we take a different outlook towards the source-destination pair model shown in Fig. 4. We merge the physical speech source with the cochlear model and consider this ensemble to be the source. Moreover, a literature survey reveals that there is no uniquely accepted probability density function

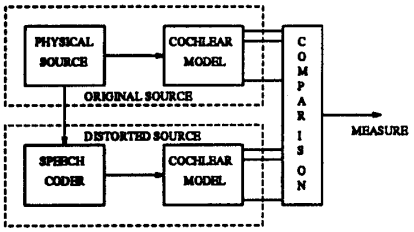

Fig. 4 Source-destination pair description

(pdf) for the physical speech source. Therefore, we are not in any further disadvantageous position by integrating the cochlear model with the speech source and determining the histogram of the firing probability at the cochlear output stages. The cochlear model output, being the probability-of-firing in various stages, assume values in the range [0,1]. We determine the histogram for the firing probability by experimenting with twelve speech utterances (six male and six female voices) of 1–2 sec. durations.

4.3 Lower Bound

We consider a widely-used form of single-letter fidelity criterion by averaging (1) or (2) over nN values. A lower bound to the function $R(D)$ is derived with a single-letter cochlear variational distance measure. Since the neural source has an alphabet of size two, the cochlear variational distance measure becomes twice the absolute magnitude error measure between the firing probabilities of the original and the coded speech signal (i.e., $\rho(x,y) = 2|x-y|$). Since this function $\rho(.)$ is of single argument such that $\rho(x,y) = \rho(x-y)$, this is also termed as a difference distortion measure.

If we let $s \leq 0$ and consider $\lambda(x)$ proportional to the reciprocal of $p(x)$ (i.e., $\lambda(x) = K/p(x)$), then

$$c(y) = \int \lambda(x)p(x)e^{s\rho(x-y)}\,dx = K\int e^{s\rho(z)}\,dz \leq 1. \quad (11)$$

For all $s \leq 0$ and with K satisfying the equality in (11), it has been deduced in [9] that

$$R(D) \geq h(p) - h(f_s) \equiv R_L(D,s) \quad (12)$$

with $D = \int \rho(x)f_s(x)\,dx$ and $f_s(x) = e^{s\rho(x)}/(\int e^{s\rho(z)}\,dz)$. The lower bound is dependent on the source statistics only through the differential entropy $h(p)$ of the source and on the distortion measure only through the differential entropy of f_s. The pdf of these firing probabilities can be approximated as sum of the two beta functions:

$$p(x) = A_1x^a(1-x)^b + A_2x^c(1-x)^d \quad (13)$$

The parameters of (13) are determined by matching the simulated function $p(x)$ with the histogram profile in a least-square sense subject to the constraint that the pdf satisfies the probability normalization axiom. Further, we obtain

$$\int_0^1 e^{2s|z|}\,dz = \frac{1-e^{2s}}{2|s|} \text{ and } \int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2s|z|}\,dz = \frac{(1-e^s)}{|s|} \quad (14)$$

Since $(1-e^s) > (1-e^{2s})/2$ for all $-\infty < s < 0$, we take the limits of z from $-\frac{1}{2}$ to $\frac{1}{2}$ and choose K to be $\frac{|s|}{(1-e^s)}$ so that the equality in (11) satisfies. By doing so, we weaken the lower bound to cater for the fact that z is not only dependent on $x-y$, but also on x. The maximizing value of s is the one that satisfies

$$f_s(x) = \frac{e^{2s|x|}}{\int_{-\frac{1}{2}}^{\frac{1}{2}} e^{2s|x|}dx} = \frac{|s|e^{2s|x|}}{(1-e^s)} \quad (15)$$

and the average distortion value as a parameter of s can be expressed as

$$D_s = \int_{-\frac{1}{2}}^{\frac{1}{2}} \rho(x)f_s(x)\,dx = \frac{(1-e^s+se^s)}{|s|(1-e^s)} \quad (16)$$

Thus, the lower bound of (12) becomes

$$R_L(D) = h(p) + \log\left\{\frac{|s|}{(1-e^s)}\right\} - \frac{(1-e^s+se^s)}{(1-e^s)} \quad (17)$$

with $h(p) = -0.55$ nats/symbol (as obtained from the estimated pdf). The lower bound of (17) is plotted in Fig. 5.

4.4 Blahut's Algorithm

The Blahut algorithm [8] is applied to calculate the rate-distortion functions directly for the cochlear variational distance and the directed divergence with $\alpha = 1$ (shown in Fig. 5 and Fig. 6, respectively). The algorithmic steps are as follows.

Step I: We assume that x and y take up values from one of the 255 values, uniformly spaced between 0 and 1. An initial output probability distribution Q_k (for $k = 1,\ldots,255$) is assumed, say, Q_k^0. A parameter $A_{jk} = e^{s\rho_{jk}}$ is evaluated, where ρ_{jk} is the single-letter cochlear discrimination measure between the input alphabet j and the output alphabet k.

Step II: The parameter s is chosen from the range $-\infty$ and 0, and then Steps III and IV are carried out with different values of s.

Step III: With the values of the input probability distribution P_j (obtained from the histogram of the cochlear model output) and the parameters A_{jk}, the following parameters are calculated:

$$\left.\begin{aligned} &c_k = \textstyle\sum_j P_j \frac{A_{jk}}{\sum_i A_{ji}Q_i}, \quad Q_k \leftarrow Q_k c_k, \\ &L = \textstyle\sum_k Q_k \log\, c_k, \quad U = \max_k \log\, c_k \end{aligned}\right\} \quad (18)$$

Step IV: If $U - L \geq \epsilon$, then Step III is repeated; otherwise, the program is terminated for this value of s by evaluating the following:

$$\left.\begin{aligned} &Q_{k|j} = \frac{A_{jk}Q_k}{\sum_i A_{ji}Q_i}, \qquad D = \textstyle\sum_j \sum_k P_j Q_{k|j}\rho_{jk}, \\ &R(D) = sD - \textstyle\sum_j P_j \log(\sum_k A_{jk}Q_k) - \sum_k Q_k \log c_k. \end{aligned}\right\} \quad (19)$$

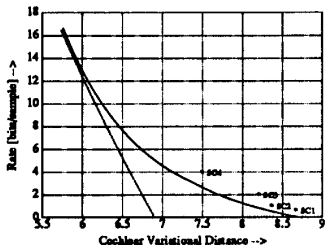

Fig. 5 Speech coder rate in bits/sample vs. average cochlear variational distance measure (- - - line shows an analytically derived lower bound, —— line shows the exact rate-distortion curve using Blahut's algorithm and four '*' points [SC1–SC4] denote the performances of four speech coders)

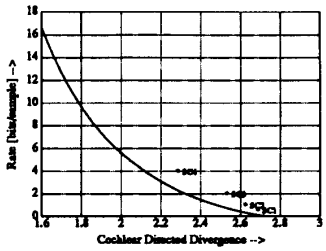

Fig. 6 Speech coder rate in bits/sample vs. average cochlear directed divergence measure (—— line shows the rate-distortion curve using Blahut's algorithm and four '*' points [SC1–SC4] denote the performances of four speech coders)

5 Measured Performances of Speech Coders

We have considered four start-of-the-art speech coders for the assessment of their average perceptual quality. These four coders (designated as SC1–SC4) were: CELP-based coder SC1 (4.8 kbps) [10], VSELP-based coder SC2 (8 kbps) [10], wideband CELP-based coder SC3 (16 kbps) [11] and ADPCM coder SC4 (32 kbps). Six speech sentences of 1–2 sec. were passed through each of the four coders to calculate the average distortion values over each sampling time (sampled at a rate of 8 ksamples/sec). Fig. 5 and Fig. 6 plot the four speech coders' performances (marked by '*') as evaluated by the cochlear variational distance and the directed divergence (with $\alpha = 1$), respectively. Now, let us examine one of the figures (say, Fig. 6). We observe that the perceptual quality obtained (measured with the cochlear directed divergence) by SC1 coder is possible to achieve with much lower rate (as low as 1.5 kbps). Similarly, SC2, SC3 and SC4 coder performances are achievable with almost 4 kbits/sec, 5.4 kbits/sec and 20 kbits/sec, respectively. From another perspective, we can say that a perceptual quality (a value of 2.575 units/sample) somewhere between those attained by SC2 and SC3 coders are attainable with a 4.8 kbps speech coder. A value of 2.485 units/sample which falls between the perceptual quality of SC3 and SC4 is theoretically achievable with an 8 kbps speech coder. Although the rate-distortion analysis does not provide with an answer to how to attain these limits, it gives an insight to what is possible in practice and how close a specific speech coder is performing with respect to the $R(D)$ limits in terms of its perceptual quality.

6 Summary

In this article, we have reviewed the cochlear discrimination measure introduced in [1]. This measure compares the neural channel cross-entropy (defined in the Rényi-Shannon entropy sense) associated with the coded speech signal with reference to the original one. The existence of the pitch and the formant information were verified using this measure. We have analytically computed a lower bound to the rate-distortion function using the cochlear variational distance. We have also evaluated the exact $R(D)$ functions for speech coding with two types of perceptual distortion measures. Four state-of-the-art speech coders were studied from the viewpoint of their performances with respect to the speech coder limits as dictated by the rate-distortion curves.

References

[1] A. De and P. Kabal, "Cochlear discrimination : An auditory information-theoretic distortion measure for speech coders," in *Proc. 16 th Biennial Symp. on Commun., Kingston, Canada*, pp. 419–423, May 1992.

[2] R. F. Lyon, "A computational model of filtering, detection, and compression in the cochlea," in *Proc. IEEE ICASSP*, pp. 1282–1285, 1982.

[3] M. Slaney, "Lyon's cochlear model," Tech. Rep. 13, Apple Computer Inc., 1988.

[4] S. Kullback, *Information Theory and Statistics.* John Wiley & Sons, 1959.

[5] D. O'Shaughnessy, *Speech Communication.* Academic Press, 1987.

[6] S. Seneff, "Pitch and spectral estimation of speech based on auditory synchrony model," in *Proc. IEEE ICASSP*, pp. 36.2.1–36.2.4, 1984.

[7] M. Slaney and R. F. Lyon, "Visualizing sound with auditory correlograms," *in submission for J. Acoust. Soc. Am.*, 1991.

[8] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, Jul. 1972.

[9] T. B. Berger, *Rate Distortion Theory.* Prentice Hall, NJ, 1971.

[10] B. S. Atal, V. Cuperman, and A. Gersho, *Advances in Speech Coding.* Kluwer Academic Pub., MA, 1991.

[11] G. Roy and P. Kabal, "Wideband CELP speech coding at 16 kbits/sec," in *Proc. IEEE ICASSP*, pp. 17–20, 1991.