ELSEVIER

Speech Communication 14 (1994) 205–229

SPEECH COMMUNICATION

Auditory distortion measure for speech coder evaluation – Discrimination information approach †,‡

Aloknath De *,[1], Peter Kabal [1,2]

[1] *Department of Electrical Engineering, McGill University, 3480 University Street, Montréal, Canada H3A 2A7*
[2] *INRS-Télécommunications, Université du Québec, 16 Place du Commerce, Verdun, Canada H3E 1H6*

Received 7 January 1993; revised 9 December 1993

Abstract

In this article, we devise a fidelity criterion for quantifying the degree of distortion introduced by a speech coder. An original speech and its coded version are transformed from the time-domain to a perceptual-domain using an auditory (cochlear) model. This perceptual-domain representation provides information pertaining to the probability-of-firings in the neural channels. The introduced *cochlear discrimination information* (CDI) measure compares these firing probabilities in an information-theoretic sense. In essence, it evaluates the cross-entropy of the neural firings for the coded speech with respect to those for the original one. The performance of this objective measure is compared with subjective evaluation results. Finally, we provide a rate–distortion analysis by computing the rate–distortion function for speech coding using the Blahut algorithm. Four state-of-the-art speech coders with rates ranging from 4.8 kbit/s (CELP) to 32 kbit/s (ADPCM) are studied from the view-point of their performances (as assessed by the CDI measure) with respect to the rate–distortion limits.

Zusammenfassung

In diesem Beitrag erstellen wir ein Kriterium für die Wiedergabegüte eines Sprachkodierers das den Verzerrungsgrad quantifiziert. Originale und kodierte Sprache werden mit Hilfe eines Modells des Gehörgangs von dem Zeitbereich in eine "Wahrnehmungsbereich" transformiert. Die Darstellung in dem Wahrnehmungsbereich liefert die Zündungswahrscheinlichkeiten in den Nervenbahnen. Das vorgestellte *Gehörgangsunterscheidungsmaß* vergleicht diese Zündungswahrscheinlichkeiten in informationstheoretischer Hinsicht. Im Prinzip vergleicht es die gegenseitige Entropie der Zündungswahrscheinlichkeiten für die kodierte Sprache mit denen des Originals. Die Ergebnisse dieses objektiven Maßes werden mit subjektiven Einschätzungen verglichen. Schließlich geben wir eine Rate–Verzerrungsanalyse indem wir die Rate–Verzerrungsfunktion für Sprachkodierung mit Hilfe der Blahut Methode berechnen. Vier der modernsten Sprachkodierer mit Bitraten von 4.8 kbit/s (CELP) bis 32 kbit/s (ADPCM) werden auf ihre Leistung (gegeben durch das Gehörgangsunterscheidungsmaß) mit Blick auf Rate–Verzerrungsgrenzen untersucht.

* Corresponding author. Present address: Bell Northern Research, 16 Place du Commerce, Verdun, Quebec, Canada H3E 1H6.
† This work was supported by a grant from the Canadian Institute for Telecommuncations Research (CITR) under the NCE program of the Government of Canada.
‡ This work was presented in part at the Sixteenth Biennial Symposium on Communications, Kingston, ON, Canada, 27–29 May 1992 and also in part at the IEEE GLOBECOM'92, Orlando, FL, USA, 6–9 December 1992.

SSDI 0167-6393(94)00002-R

Résumé

Dans cet article, nous définissons un critère de fidélité pour quantifier le degré de distorsion introduite par un codeur de parole. Un signal vocal original et sa version codée sont transformés du domaine temporel dans le domaine perceptuel utilisant un modèle auditif (cochléaire). Cette représentation dans le domaine perceptuel fournit une information liée aux probabilités de décharges dans les canaux neuronaux. La mesure de *discrimination cochléaire* introduite ici compare ces probabilités de décharges au sens de la théorie de l'information. En essence, elle évalue l'entropie croisée des décharges neuronales pour la parole codée par rapport à celle de la parole originale. La performance de cette mesure objective est comparée à des résultats d'évaluation subjective. Finalement, nous fournissons une analyse débit–distorsion en calculant la fonction débit–distorsion pour le codage de la parole en utilisant l'algorithme de Blahut. Quatre codeurs de parole performants avec des débits allant de 4.8 kbit/s (CELP) à 32 kbit/s (ADPCM) sont étudiés du point de vue leurs performances (telles qu'évaluées par la mesure de discrimination cochléaire) par rapport aux limites débit–distorsion.

Key words: Auditory (cochlear) model; Distortion measure; Rényi–Shannon entropy; Discrimination information; Rate–distortion function

1. Introduction

In a typical source coding problem, a continuous-time continuous-amplitude bandlimited signal is sampled in the time domain at or above the minimum sampling rate required. This time-discretized signal with amplitude having continuous probability density function has an infinite *entropy*. To transmit the output of such a source and to recover it exactly, a communication channel of infinite *capacity* is required. In practice, every channel, due to perturbation by noise, has a finite capacity. Thus, it is not possible to transmit the output of a continuous source over any channel and recover it exactly (Blahut, 1987). Accepting the fact that there will inevitably be some distortion, a typical source coder minimizes it by removing deliberately some information which is deemed "not very important" to the destination. The extent to which the information should be removed depends on the bit-rate of the coder; the lower the bit-rate, the more information is needed to be removed.

In speech communication, the ultimate recipient of information is a human being and hence his/her perceptual abilities govern the precision with which speech data must be processed and transmitted. Thus, to reduce the amount of distortion, the speech data can be modified by an intentional removal of some information in accordance with the limitations of the auditory system. Determining "what is not very important" to the auditory system and "how the auditory system assesses" the relative importance of information is the primary task involved in devising a distortion measure for speech coders.

The sound quality of a given speech coder can best be evaluated by listening to it. However, an extensive subjective testing of speech coders is difficult to administer, time-consuming and often found to be inconsistent (due to the non-repeatability of human responses). An objective quality measure suitably defined could thus play an important role in the evaluation as well as in the design of a low bit-rate speech coder. Such a distortion measure should be computable from an original speech waveform and its coded/distorted version; and should also conform to the results of a subjective measure (Quackenbush et al., 1988). One important advantage of distortion measure is that its repeated application at different time under different environment gives the same performance. To measure the degree of correlation between the defined objective measure and a standard subjective measure, a *correlation coefficient* is often used as an indicator.

Speech coders operating at several standardized data rates are available to "match" to the capacities of communication channels. These encoders vary from the view point of coder architecture, the type of features encoded, the number of bits allocated to the features and so on. This wide variety of encoding algorithms introduces a broad range of linear and nonlinear coder distortions.

All of these distortions are not equally perceived by the auditory system. As a consequence, if we can devise a distortion measure incorporating the human perception mechanism, then that can be used to evaluate the performances of different speech coders.

The performances attained by various speech coding systems can be compared with absolute bounds derived from rate–distortion theory which provides a mathematical foundation for source coding. With a particular source and a defined distortion measure, it is possible to derive a rate–distortion function which determines the lowest achievable rate for a specified amount of the coder distortion. Defining an appropriate distortion measure can thus facilitate determining the true lower limit of the coder rate for attaining a particular speech quality.

Furthermore, a distortion measure can help the design procedure of speech coders in three ways:

(a) An analysis-by-synthesis type speech coder has two parts – an analysis stage and a synthesis stage. From the stochastic codebook, all (in the case of an "optimal" coder) or selectively chosen some (in the case of a "suboptimal" coder) entries are used along with the inverse formant and pitch filters to synthesize several coded speech signals. Finally, the index of that codebook entry is transmitted which results in the minimum distortion as measured by the defined fidelity criterion. This way, a distortion measure can be very much instrumental in selecting a "proper" excitation codebook entry.

(b) With a limited number of bits available per second, a bit allocation to different feature parameters is necessary. The bit allocation strategy adopted for an 8 kbit/s coder can neither be scaled down for a 4 kbit/s coder nor be scaled up for a 16 kbit/s coder directly. The "relative" importance of the information to be transmitted plays a significant role. In the design phase, the distortion measure can be used for improving the bit allocation policy of a particular speech coder, be it a waveform coder, an analysis-by-synthesis coder or a vocoder.

(c) While designing a speech coder, an appropriate distortion measure not only helps in making a sound bit allocation policy, but also in "populating" (also called "training") the codebook. In the training phase, determining the centroid for each class with the defined distortion measure results in the design of an "optimum" (at least in the local sense) codebook. If the distortion measure properly reflects the perceptual importance of information, then a fixed size codebook designed in this way will also be filled up with the entries which contain "perceptually more important" information.

In this article, we propose a distortion measure for speech coders using an auditory (cochlear) model. This measure is used for studying the performance of different speech coders, and also for providing a rate–distortion analysis. In Section 2, we discuss some of the existing subjective and objective (the time-domain, the frequency-domain as well as the perceptually-motivated) distortion measures. Section 3 describes the auditory system, discusses a cochlear model and defines a *perceptual-domain*. Section 4 introduces the idea of the *cochlear discrimination information*, a perceptual cross-entropy measure-based fidelity criterion for speech signals, and provides the test results with relevant remarks. Section 5 provides a rate–distortion-theoretic discussion by characterizing the source–destination pair, computing the rate–distortion function with the Blahut algorithm and studying the performances of four speech coders.

2. Existing distortion measures

Some of the existing subjective and objective distortion measures are outlined below.

2.1. Subjective measures

Subjective quality measures can be classified into two primary categories (Hecker and Williams, 1966): *utilitarian* and *analytic*. The utilitarian measures are based on a unidimensional scale whereas the analytic measures typically use more than one dimension for determining the perceived quality. With either of the classes, an extensive listener training procedure is needed to

ensure the reliability of these tests under different test environments.

The utilitarian measures often address the speech intelligibility and the articulation aspects separately. The intelligibility tests are scored by the percentage of correct understanding of the meaning conveyed by the transmitted speech, while the articulation tests are evaluated by the percentage of correct recognition of the sounds, words or sentences. The most widely used utilitarian-type subjective measure is the mean opinion score (MOS) (Quackenbush et al., 1988), in which the listeners rate the speech under test on a five point absolute scale (rate 5: imperceptible; rate 4: just perceptible, but not annoying; rate 3: perceptible and slightly annoying; rate 2: annoying, but not objectionable; rate 1: very annoying and objectionable). Since the listeners have freedom to interpret the scale-ratings in their own way, the MOS score provides an agglomerative measure value for different types of coder distortions.

One popular analytic-type measure is the diagnostic acceptability measure (DAM) (Voiers, 1977). The DAM evaluates a speech signal on sixteen separate scales (covering the signal quality, the background quality and the overall quality), all of which have a range from 0 to 100 points. Signal degradations such as fluttering (amplitude modulated speech), thin (high pass speech), rasping (peak clipped speech), interrupted (packetized speech with "glitches"), nasal; background noise such as hissing (noise masked speech), buzzing (tandemmed digital system), babbling (narrow band system with errors), rumbling (low-frequency noise masked speech); and overall qualities such as intelligibility, pleasantness, acceptability are all extensively considered in the DAM test.

2.2. Objective measures

The most traditional time-domain measure is the signal-to-noise ratio (SNR) which does not correlate well (a correlation coefficient ρ of 0.24 that too measured only for the waveform coders) with subjective evaluation results (Quackenbush et al., 1988). This failure is quite obvious from the fact that a pair of antipodal tone signals has a six dB SNR difference, but the perceptual quality difference is not audible. A segmental SNR ($\mathrm{SNR}_{\mathrm{seg}}$) measure (Mermelstein, 1979) which gives an average of the SNR values [in dB units] computed over the speech segments (each segment being of the order of 128 samples) has been found to have a ρ of 0.77 across a wide range of waveform coder distortions. In segments where an original speech has almost no signal components, any amount of noise would generally give rise to a large negative SNR for that segment affecting the measure considerably. To alleviate this problem, several variations of this measure such as the frequency-weighted segmental SNR (Quackenbush et al., 1988), the granular segmental SNR (McDermott et al., 1978), etc. have been suggested.

The spectral distortion measures, in general, have been found to be more reliable than the time-domain measures as they are less sensitive to the occurrence of time misalignments between the original and the coded speech. Using the notion of one-step linear prediction error and spectral factorization, an L_p norm-based log spectral distortion (LSD) measure is defined (Gray et al., 1980) between two log spectral densities. A modified version of this measure is recently proposed in (Halka and Heute, 1992), where the measure's kernel is not the spectral distance of the input and the output signals, but the distance of the output-spectrum and the spectral representation of the nonlinear distortions. In the coherence function measure (Kubichek, 1991), the speech frames are first divided into four groups based on the four amplitude quartiles. The power spectra of the original and the coded signals and also the cross-power spectrum are computed. Finally, a measure value is given by averaging a defined objective function over all the frames and all the quartiles. The Itakura–Saito distortion (d_{IS}) measure (Itakura and Saito, 1968), which involves maximum likelihood spectral estimation, is one of the widely used distortion measures. In this measure, the power spectrum is most heavily weighted where its magnitude is the largest. Variants of this measure such as the frequency-weighted d_{IS} (Chu and Messer-

schmitt, 1982), the cosh measure (Gray et al., 1980) etc. are also available in the literature.

Among the parametric distortion measures, the log likelihood ratio distance (Crochiere et al., 1980) is one which is calculated assuming that a speech segment can be represented by an autoregressive LPC model and computing the prediction residual energies. A log-area ratio measure is defined using the reflection coefficients k_m which are relatively less spectrally-sensitive to quantization (except when their magnitudes are near unity). This measure shows a relatively better performance with a ρ of 0.62 (Quackenbush et al., 1988). In (Coetzee and Barnwell III, 1989), a multiobjective functional measure is formulated using the line spectral frequency parameters in determining the spectral peak locations, energies, bandwidths, etc. for the original and the distorted speech frames. A correlation coefficient of 0.78 is obtained with this measure. Computational efficiency and a high correlation with the L_2 norm-based LSD measure have made the cepstral distance measure very popular (Kitawaki et al., 1988). With the cepstral coefficients three times the number of LPC parameters, this measure shows a ρ of 0.80. A unifying framework for viewing different distortion measures in the cepstral domain is studied in (Lee, 1991).

A perceptually-motivated information index measure is proposed in (Lalou, 1990) which divides the spectrum into sixteen critical bands, treats them as independent channels and applies empirical frequency weights and hearing thresholds for each band. An auditory model-based algorithm which computes the probability of detection of the noise as a function of time for noise-corrupted audio and music signals is introduced in (Paillard et al., 1992). Schroeder et al. (1979) have described a method for calculating the signal degradation based on the measurable properties of the auditory perception. Motivated by this work, recently, a series of psychophysical experimental curves are invoked in (Wang et al., 1992) to define a Bark spectral distortion (BSD) measure. The original power spectral density (in Hz) is transformed to a critical band density (in Bark) and "smeared" by a prototype critical band filter. Then, the sound pressure levels (SPL) in dB is translated to the loudness levels in phons followed by a phon-to-sone conversion and a Bark spectra comparison (Wang et al., 1992). The success of this measure has exhibited the advantage of considering important perceptual events while formulating a distortion measure.

Devising a distortion measure involves conceiving a transformation operator for mapping the signals onto a "suitable" domain and formulating a comparison method in a "meaningful" sense (De and Kabal, 1992a). We argue that neither the time-domain nor the frequency-domain, in isolation, can capture all the details of the perceptual event. Accordingly, in the proposed distortion measure, several details of the auditory processing involved in speech perception are imbibed in the transformation of speech signals onto a perceptual-domain. Subsequently, these perceptual domain parameters of the original and the coded speech signals are compared in an information-theoretic sense. The fundamental difference between our approach and the BSD measure is in considering the temporal masking phenomenon and addressing the issue of the "cause" rather than that of the "effect" involved in the speech perception. In other words, instead of considering the important perceptual effects observed, we emulate the auditory system as it is and use it in the formulation of our distortion measure.

3. Auditory representation

We desire to deal with an accurate description of the human perception as far as possible. But at the same time, since the computational speed of the model is also of importance, we prefer using a functional model of the auditory system. In the following, we describe the human auditory system briefly and discuss about the perceptual-domain representation of speech signal using Lyon's cochlear model.

3.1. Auditory system

An ear consists of three sections: the outer ear, the middle ear and the inner ear. The outer

ear channelizes the sound waves into the ear canal. This 2.7 cm long ear canal acts as a quarter-wavelength open-organ pipe whereby the first resonance occurs at around 3,000 Hz. The middle ear with three dense bones (malleus, incus and stapes) acts as an acoustic transformer to match the airborne-sound impedance of the outer ear to the fluid-borne sound impedance of the inner ear. When low-frequency (< 500 Hz) sounds of more than 85–90 dB SPL reach the eardrum, an acoustic reflex occurs due to which the middle ear also provides some automatic gain control (AGC) effect (O'Shaughnessy, 1987).

The spiral-shaped cochlea (inner ear) converts the mechanical vibrations at its oval window input into the electrical excitations on its neural fiber outputs. Between the cochlear duct and the scala tympani is the basilar membrane (BM). The stiffness of the BM varies smoothly over its length. It is stiff and thin at the basal end (where the sound enters), but compliant and massive at the apical end (the ratio of stiffness between ends exceeds 100). Therefore, the cochlea near its base is most sensitive to high frequency sounds and as the wave travels down the cochlea, lower and lower frequencies are sensed. The prime feature of the cochlea is that energy in the acoustic wave is separated by frequency and each place in the cochlea responds best to one frequency, termed as its characteristic frequency (CF). There is essentially no phase delay in pressure along the BM and no significant amount of wave energy is reflected (Flanagan, 1972). The band-pass frequency responses corresponding to different places, as found by Nobel-Laureate Von Békésy, were rather broad and later Mössbaur's gamma-ray-based experiment has suggested much sharper frequency responses (Moore, 1989). These band-pass filters have an almost constant Q-factor, thereby implying a fixed ratio of the center frequency to the bandwidth for all of them.

On the top of the BM (within the organ of Corti), there are about 30,000 sensory hair cells on which the auditory neurons terminate. The fans of the cilia sticking out of the inner hair cells resist the BM motion. When the cilia are bent one way, the inner hair cells stimulate the neurons whereas no simulation is generated when the cilia are bent the other way. Thus, the inner hair cells act as half-wave rectifiers for the velocity of the BM motion (Flanagan, 1972). The frequency resolution along the BM is best at lower frequencies (apical end), whereas the time resolution is best at higher frequencies (basal end). This is primarily due to the fact that a hair-cell attached to a high-CF location on the BM "fires" (i.e., generates impulses) in response to a broader set of frequencies than does a low-CF hair cell (Lyon, 1982).

Studies on the cochlear echo and the oto-acoustic emission suggest that the BM behaves as an active system and the transfer characteristics of the BM system vary depending on the input signal level. This is attributed to the fact that the outer hair cells interact with the BM motion. Sounds with high SPLs are effectively diminished whereas sounds with low SPLs are enhanced by the "superregenerative active" mechanisms of the outer hair cells (Allen, 1985). An important aspect of hearing is the phenomenon of auditory masking in which the perception of low-energy sound is obscured by the presence of a high-energy sound (Penner, 1979). The outputs of the band-pass filters may be viewed as zero-mean "carrier" signals which are "amplitude-demodulated" by the half-wave detection nonlinearity. The phenomenon of auditory masking can thus be justified by the "threshold effect" phenomenon (Carlson, 1986) as observed in the envelope detection process of AM signals.

Effects of the outer hair cells can be emulated by automatic gain control (AGC) stages and some kind of inter-stage coupling of these AGCs can simulate the auditory masking feature. Any gain control effect (i.e., amplification or compression) is not instantaneous and the time required to adapt to any input signal is dependent on the signal level (Lyon, 1982). Depending on the resulting signal energy, the nerve endings attached to the hair cells are stimulated. This produces all-or-none electrical firings which are propagated axonally to the brain following an ascending auditory pathway (Flanagan, 1972; Allen, 1985). Unfortunately, the exact neuroelectrical representation of the sound stimuli at the higher level is not sufficiently understood.

3.2. Lyon's cochlear model

The interpretation of the auditory system as a spectrum analyzer goes back to Helmholtz (1954) in the last century. The timing or volley theory states that low sound frequencies such as those corresponding to the fundamental frequency of speech, are perceived in terms of time-synchronous neural firings from the BM apex. On the other hand, the place theory suggests that, especially for higher frequencies such as those in the formants of speech, the spectral information is decoded via the BM locations of the neurons that fire most (Geisler, 1988). Current trend in modeling the auditory system is to combine the volley theory with the place theory. Such models for representing speech in the auditory periphery falls into one of four broad classes (Greenberg, 1988): rate/place, synchrony/place, synchrony/quasi-place and synchrony/place-independent.

The rate/place representation (Sachs et al., 1988) assumes that the average rate be roughly proportional to signal amplitude over the frequency range related to the response area of an auditory nerve fiber. Although this representation, in general, functions well at low (< 50 dB) sound pressure levels (SPL) [Note: 0 dB SPL = 10^{-16} W/cm^2], it may not delineate the spectral peaks sharply in the presence of background noise even at amplitudes for which speech intelligibility is unimpaired. The synchrony/place representational form (Seneff, 1988) is based on the neural synchrony and requires the system to possess some knowledge of the tonotopic affiliation (characteristic frequency) of each fiber with which to evaluate its temporal firing pattern. The spatio-temporal responses appear as traveling waves that begin at the base of the inner ear. The traveling waves produced by different frequencies decay at CF locations in an orderly way along the spatial axis. The combined effects of the quick amplitude decay and phase shifts produce a series of discontinuities parallel to the temporal axis. The synchrony/quasi-place model (Shamma, 1985), in the form of a lateral inhibitory network, considers simultaneous activity across adjacent channels. A proposition that a spectral representation based on the synchrony need not be concerned with the tonotopic identity of the auditory nerve fibers gives rise to the synchrony/place-independent model (Ghitza, 1987). This works satisfactorily only for high (> 85 dB) SPL as the

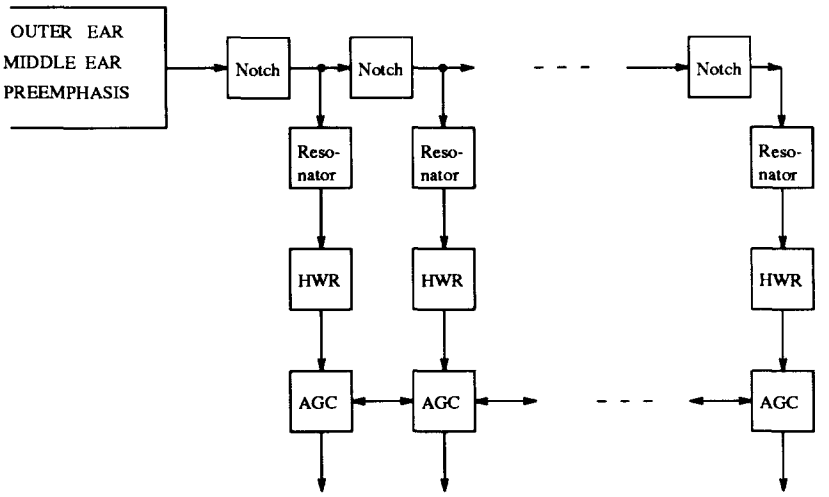

Fig. 1. Block diagram of Lyon's cochlear model ("HWR" stands for the half-wave rectifier and "AGC" stands for the automatic gain controller).

temporal information concerning formant frequencies are distributed over a broad range of frequency channels.

We believe that a synchrony/quasi-place model is most appropriate for our work as it could operate satisfactorily for high, medium or even low signal levels. Consequently, we adopt one such synchrony/quasi-place model as proposed by Lyon (1982) and described by Slaney (1988). This model separates complex mixtures of sounds mainly by segregating different frequencies into different places, but also by preserving enough time resolution to separate the responses to different pitch pulses. By a detailed separation of sounds along the time and frequency dimensions, Lyon's cochlear model as shown in Fig. 1 paves way for a robust speech analysis technique. Here, we describe the model in six steps.

Step 1 (Outer-and-middle ear filter): The outer-and-middle ear effectively adds a slight high-pass response to the system. Assuming that the input speech signals are sampled at a frequency f_s of 8,000 Hz, a simple first-order high-pass discrete-time filter with a corner frequency of 300 Hz is designed to model roughly the effects of the outer and the middle ear. The frequency response of this filter $H_{OM}(z)$, plotted in Fig. 2, is given by

$$H_{OM}(z) = \frac{\left(1 - \exp\left[-2\pi\frac{300}{8000}\right]z\right)}{\left(1 - \exp\left[-2\pi\frac{300}{8000}\right]z\right)_{z=1}} = 4.76375(1 - 0.79008z). \quad (1)$$

This filter has unity gain at DC (i.e., at $z = 1$). For simplification, the AGC mechanism of the middle ear via stapedial reflex is not modeled here (Pickles, 1982).

Step 2 (Notch filters and resonators): The cochlea is best described by a continuous differential equation (Deng, 1992); however, it can be modeled by an ensemble of discrete stages in cascade. Lyon, in his proposed cochlear model, uses such discrete-place approximation. An implementation of the discrete-place stages involves combining a series of notch filters that model the traveling pressure waves with a series of resonators that

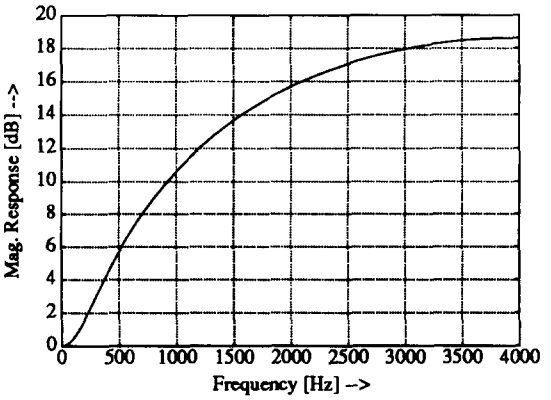

Fig. 2. First-order outer-and-middle ear filter.

model the conversion of pressure waves into BM motion (Lyon, 1982; Slaney, 1988). The notch filters operate at successively lower frequencies so that the net effect is to low-pass filter gradually the acoustic energy which are collected by the resonators corresponding to different places. We consider here sixty-four stages (covering up to 4,000 Hz) in cascade, each having a different frequency sensitivity representing the associated resonance and is characterized by the respective filter transfer function.

The notch filters and the resonators are approximated by biquadratic filter transfer functions. Each of the notch filters has a high-Q zero-pair near a low-Q pole pair whereas each of the resonators has a zero at DC with a high-Q pole pair located between the previous and the next notch filter zero-pairs. Several models of the cochlear mechanics include a micromechanical "second filter" for a resonance in the organ of Corti that contributes a zero pair slightly below the BM resonance (Hall, 1980). Presently, this not-so-well-accepted feature is left out. This can easily be incorporated in this model by putting another zero pair in the resonator section.

Step 3 (Cascade design of stage filters): The combination of the notch filters and the resonators can be implemented in cascade/parallel form as shown in Fig. 1. However, to reduce the computations, the notch and the resonator filters of each stage can be integrated into a single ear-filter

stage. The locations of the poles in the resonator filters are chosen to be at the same locations as the poles in the succeeding notch filter. This way, the zeros from each notch filter and the poles from a resonator and the next notch filter are integrated to yield a single ear-filter stage (Slaney, 1988).

The composite transfer function of each ear-filter stage is an asymmetric band-pass function. $W_{ear}(f_c)$, the 3-dB bandwidth of a band-pass filter with center frequency f_c, is defined as

$$W_{ear}(f_c) = \frac{\sqrt{f_c^2 + f_{eb}^2}}{Q_{ear}}, \tag{2}$$

where the ear-break frequency f_{eb} is 1,000 Hz and the constant Q-factor for all the band-pass filters Q_{ear} is 8. In conformance to psychoacoustical data, four successive ear-filter stages are overlapped within the 3-dB bandwidth of any one ear-filter and thus we have S_{ear}, reciprocal of the number of overlapping ear-filter stages, as 0.25. Finally, the following parameters are obtained for any ear-filter stage corresponding to a particular characteristic frequency:

$$f_{cp} = f_c, \quad Q_{cp} = \frac{f_{cp}}{W_{ear}(f_c)}; \tag{3}$$

$$f_{cz} = f_c + W_{ear}(f_c) S_{ear} Z_{off}; \quad Q_{cz} = h_{ear} \frac{f_{cz}}{W_{ear}(f_c)}, \tag{4}$$

where f_{cp} and f_{cz} are the center frequencies of the associated poles and zeros of a particular ear-filter stage having center frequency f_c. The center frequency of the associated zero is an extra stage higher than that of the pole. Thus, the Z_{off}, a factor that determines how far the zero is offset from the center frequency of the ear-filter stage, is chosen to be 1.5. Q_{cp} and Q_{cz} are the Q-factors for the corresponding poles and zeros and the parameter h_{ear}, which determines how much sharper the notch (zero) is than the resonator (pole), is selected to be 5.0.

The ear-filter stages are indexed from 1 (corresponding to the highest frequency) to 64 (corresponding to the lowest frequency) and the center frequency of each stage decreases by S_{ear} (here, 0.25) times the bandwidth of the previous stage.

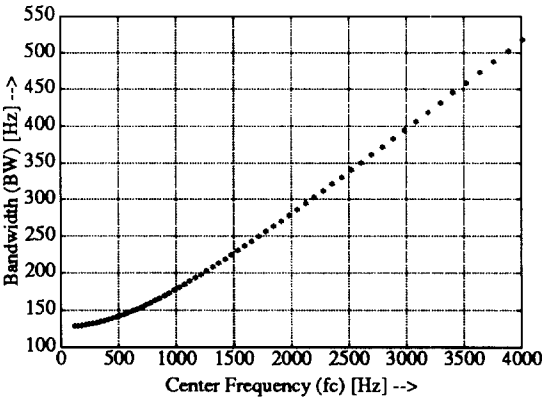

Fig. 3. Bandwidths versus center frequencies of sixty-four stages.

$W_{ear}(f_c)$ versus f_c of all the sixty-four ear-filter stages are plotted in Fig. 3 where we observe that $\lim_{f_c \to 0} W_{ear}(f_c) \to f_{eb}/Q_{ear} = 125$.

Step 4 (Other adjustments in stage filters): To implement the zeros at DC for every resonator, a differentiator is required for each stage. Since all the filtering used is linear, the differentiator (a term of the form $1 - z$) can be placed just once before the ear cascade. In addition, the differentiator is combined with a zero at the Nyquist rate $(1 + z)$ to compensate for the close spacing of the poles near $z = -1$ for high frequency. The frequency response for this combined filter is given as

$$H_{comb}(z) = 0.5(1 - z^2), \tag{5}$$

with unity gain at one-quarter of the sampling frequency.

In the cascade form, each of the ear-filter stages is implemented by a combination of two poles and two zeros. After the pole-zero integration, a pair of poles of the first stage is left aside. Thus, the ear-filter is redefined with an initial stage $H(z)$ which combines the effects of the outer-and-middle ear $H_{OM}(z)$ and the differentiator-compensator $H_{comb}(z)$ with the two poles of the first stage filter. The transfer function of this initial stage filter becomes

$$H(z) = \frac{(-0.77356 + 3.91442j)(1 - 0.79008z)(1 - z^2)}{0.67523 + 1.64342z + z^2}. \tag{6}$$

The gain of an ideal differentiator is proportional to frequency. Preceding all stages of the ear-filter with a single differentiator causes the lower frequency stages to have a much lower output than the preceding stages. While within a single stage, it is desired to add a term that is proportional to frequency, the effect of differentiator at each stage is adjusted so that it has unity gain at the center frequency of the corresponding stage. Typical frequency responses for three ear-filter stages with center frequencies as 499 Hz, 1,013 Hz and 2,509 Hz are shown in Fig. 4.

Step 5 (Half-wave rectification): The exact shape of the half-wave nonlinearity is not obvious; there are proposals for ideal as well as soft half-wave (Schroeder and Hall, 1974) rectification. In this work, an ideal half-wave rectifier is considered.

Step 6 (Coupled automatic gain controllers): The effects of the BM and the hair cell nonlinearity are taken care of adequately by lumping them into a gain control mechanism. Other nonlinear effects, such as the cubic difference tones, etc., are assumed to be relatively unimportant to normal hearing (Flanagan, 1972).

The most important adaptation mechanism in sensory systems is lateral inhibition by which the sensory neurons reduce their own gain as well as the gain of the others nearby. A logarithmic or simple non-coupled AGC mechanism does not adequately handle wide variations of energy

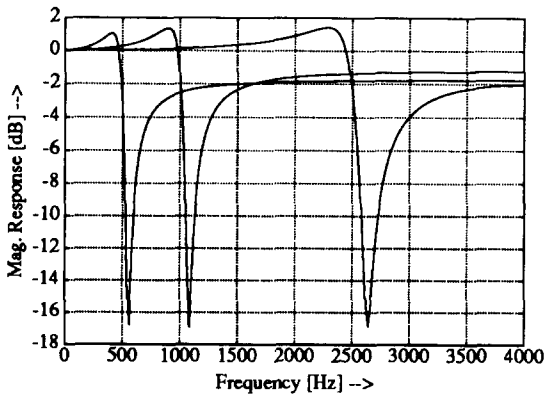

Fig. 4. Magnitude responses for three typical ear-filter stages with $f_c = 499$, 1,013 and 2,509 Hz.

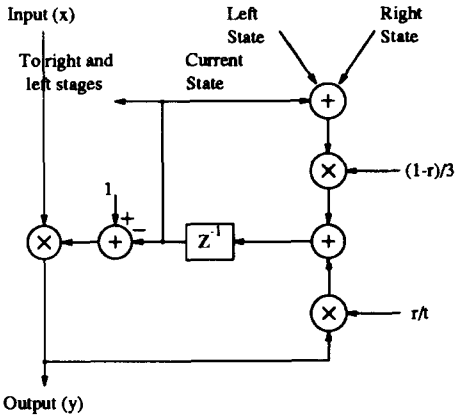

Fig. 5. A typical automatic gain control (AGC) stage.

across the frequency dimensions. Therefore, Lyon (1982) proposed a coupled AGC that adapts in the frequency domain. One such coupled AGC, as described in (Slaney, 1988) is shown in Fig. 5. Each stage is coupled directly only to its neighboring stages. However, in principle, any stage can affect all the other stages having an effect, perhaps, decaying exponentially with distance from it (Lyon and Dyer, 1986). The gain offered to an input in an AGC stage varies between 0 and 1, and this gain factor is determined based on the previous states of the current, the left and the right stages as well as the previous output value.

The time constant of the coupled AGC is made dependent on the signal level. A cascade of four AGC blocks with different time constants, simulating the different adaptation times in the ear, are used (Slaney, 1988). A longer time constant implies that the AGC takes longer to respond to the input. Each AGC attenuates the incoming signal so that, under steady-state condition, it remains below the target value corresponding to that AGC. The target parameters (t) and the time constants (τ) of the four AGC blocks, respectively, are chosen as 0.0032, 0.0016, 0.0008 and 0.0004 units (on the same scale, the amplitude of a signal with +120 dB SPL is assumed to be unity) and 640 ms, 160 ms, 40 ms and 10 ms. The r parameters as indicated in Fig. 5 are related to the τ parameters as

$$r = \exp\left[-\frac{1}{\tau f_s}\right]. \tag{7}$$

For any one of the sixty-four stages, a typical steady-state response of the four cascaded AGC blocks is depicted in Fig. 6.

3.3. Perceptual-domain representation

The outputs of the cochlear model vary over only about two orders of magnitude as the input signal varies over the entire range covering the threshold of hearing to the threshold of pain. The neurons are attached to the hair cells at different places along the cochlear partition and they "fire" (i.e., generate all-or-none electrical spikes) based on the gain-controlled signals as sensed by the corresponding hair cells. Essentially, these neural firing events are communicated from the auditory system to the brain through a large number of neural fibers. These neural pathways are termed hereafter as the "neural channels" so as to keep conformity with the other communication channels. Although these neural fibers are spread densely along the BM, since we consider sixty-four discrete-place stages, we would visualize that all the neurons could be classified into sixty-four characteristic neural channels.

The normalized cochlear model output provides the probability-of-firing information in these sixty-four neural channels at each clock time. Here, the normalization is done with respect to the maximum possible output value (i.e., 0.000213 unit as shown in Fig. 6) of the four cascaded AGC blocks and the clock time is chosen to be same as the sampling time, i.e., 125 μs. Since we do not know the exact firing process, the neural activity patterns can be presented in a cochleagram matrix form which gives the probability-of-firings in all the neural channels for all the clock times. In our work, this auditory representation is referred to hereafter as the perceptual-domain (PD). We assert that, to devise a distortion measure for speech signals, the original and the coded/distorted signal should be compared in this perceptual (time-place) domain, rather than just in the time or in the frequency domain.

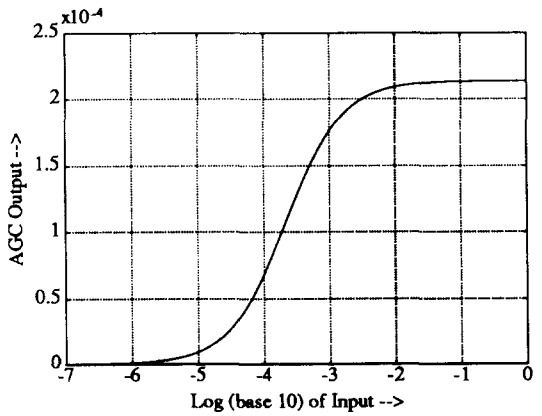

Fig. 6. A typical steady-state response of four cascaded AGC blocks.

4. Cochlear discrimination information measure

In the previous section, we have addressed the issue of representing speech signal in a perceptual-domain (PD). This PD representation is a sequence of N-dimensional (in our work, $N = 64$) vectors at the clock times within a speech signal. Each of the N neural channels may be conceived as communication channels with an input alphabet of size two, i.e., firing and non-firing. Due to the lack of our knowledge about the exact neural conversion process, we compare the probability distributions for firing and non-firing, derived from an original and a coded signal, to quantify the degree of distortion. The discrimination information which has emerged as a powerful tool (Kullback, 1959) for measuring the "closeness" of two probability density or distribution functions is applied here for defining a cochlear discrimination information (CDI) measure (De and Kabal, 1992a). The CDI measure, in effect, determines the amount of new information (the increase in neural source entropy) associated with the coded signal when the neural source entropy associated with the original speech is known or vice versa. Here, we formulate the CDI measure and study speech coder performances with it.

4.1. Distortion computation

Let P be a set of probability measures defined on a measure space $\mathscr{S}^{(J)}$ for a discrete information source with an alphabet of size J. The

Rényi–Shannon entropy $H_\alpha(P)$ for such source with $P=\{p_1, p_2, \ldots, p_J\}$ is given as (Aczél and Daróczy, 1975)

$$H_\alpha(P)=\begin{cases} -\sum_{j=1}^{J} p_j \log p_j, & \alpha=1, \\ \dfrac{1}{1-\alpha}\log\left(\sum_{j=1}^{J} p_j^\alpha\right), & \alpha\geqslant 0, \quad \alpha\neq 1. \end{cases} \tag{8}$$

It has been shown in (Aczél and Daróczy, 1975; Rényi, 1970) that

1. $H_\alpha(P)$ is a continuous positive decreasing function of α and is also continuous in P.
2. $H_\alpha(P)$ is always non-negative and $H_\alpha(P)=0$ if and only if all of the p_j's except one are equal to zero.
3. $H_\alpha(P)$ is strictly concave with respect to P for $0<\alpha\leqslant 1$; i.e., $H_\alpha(\lambda P'+\overline{1-\lambda}P'')\geqslant \lambda H_\alpha(P')+(1-\lambda)H_\alpha(P'')$, $\forall P'$, P'' and all $\lambda\in(0,1)$.
4. Convexity or concavity of $H_\alpha(P)$ with respect to P depends on J for $\alpha>1$.

Now, let us consider one neural channel for a specific clock time. Since there are only two events possible (i.e., firing and non-firing), the measure space can be written as

$$\mathscr{S}^{(2)}\triangleq\{P: P=(p_1, p_2);\ p_1, p_2\geqslant 0;\ p_1+p_2=1\}. \tag{9}$$

Appendix A shows that with $P\in\mathscr{S}^{(2)}$, $H_\alpha(P)$ is strictly concave with respect to P not only for $0<\alpha\leqslant 1$, but also for $1<\alpha\leqslant 2$. Thus, here we consider α values in the range [0, 2) which ensures a global maximum of $H_\alpha(P)$ for $p_1=p_2=1/2$.

In this work, the time-domain speech representation $\mathscr{T}$ is mapped onto the PD $\mathscr{A}$ using Lyon's cochlear model $\mathscr{C}$. Mathematically, this mapping $\mathscr{B}$ can be expressed as $\mathscr{B}$: $\mathscr{T}\xrightarrow{\mathscr{C}}\mathscr{A}$. The PD representation $\mathscr{A}$ for an original speech signal can be written in a matrix form as

$$\mathscr{A}=\begin{bmatrix} P_{11} & P_{12} & \cdots & P_{1N} \\ P_{21} & P_{22} & \cdots & P_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nN} \end{bmatrix}, \tag{10}$$

with n clock times and N neural channels. An element P_{kl} of the matrix $\mathscr{A}$ implies that p_{1kl} and $p_{2kl}=1-p_{1kl}$ are the firing and the non-firing probabilities for the k-th neural channel at the l-th clock time corresponding to the original speech signal. Similarly, let q_{1kl} and $q_{2kl}=1-q_{1kl}$ be the firing and the non-firing probabilities for the coded/distorted speech. Accordingly, the directed divergence (a form of the discrimination information measure) between P_{kl} and Q_{kl} can be written as (Rényi, 1970)

$$D_\alpha(P_{kl};Q_{kl})=\begin{cases} \sum_{j=1}^{2} p_{jkl}\log\left(\dfrac{p_{jkl}}{q_{jkl}}\right), & \alpha=1, \\ \dfrac{1}{(\alpha-1)}\log\left(\sum_{j=1}^{2}\dfrac{p_{jkl}^\alpha}{q_{jkl}^{\alpha-1}}\right), & \alpha\geqslant 0, \quad \alpha\neq 1. \end{cases} \tag{11}$$

This measure is not a metric as it does not satisfy some of the conditions required for it to be a metric: (a) the symmetry condition [$D_\alpha(P_{kl};Q_{kl})$ is not the same as $D_\alpha(Q_{kl};P_{kl})$ when P_{kl} and Q_{kl} are different]; and (b) the triangle inequality [the sum of the measures $D_\alpha(P_{kl};Q_{kl})$ and $D_\alpha(Q_{kl};R_{kl})$ may be greater than, equal to or less than $D_\alpha(P_{kl};R_{kl})$ for any three probability distributions P_{kl}, Q_{kl} and R_{kl}]. However, the satisfaction of the non-negativity condition allows it to be considered as a fidelity criterion (even though it is not a metric). We define the directed divergence measure of order α for $0<\alpha\leqslant 2$, the range in which $H_\alpha(P)$ has been shown to be concave with respect to $P\in\mathscr{S}^{(2)}$.

For simplicity, we assume that the neural firing events in different channels and at different clock times are independent. Thus, the neural sources corresponding to the N neural channels and the n clock times form a product source, i.e.,

$$\mathscr{S}=\bigtimes_{l\in\mathscr{L}}\bigtimes_{k\in\mathscr{K}}\mathscr{S}_{kl}^{(2)}, \tag{12}$$

with $\times$ as the cartesian product of the probability spaces, $\mathscr{L}\equiv\{1, 2, \ldots, n\}$ and $\mathscr{K}\equiv\{1, 2, \ldots, N=64\}$. Under this assumption, the probability distribution of the product source is the product of the probability distributions of the individual sources

(Blahut, 1987) and the directed divergence values are additive, i.e.,

$$D_\alpha(P; Q) = \sum_{l \in \mathscr{L}} \sum_{k \in \mathscr{K}} D_\alpha(P_{kl}; Q_{kl}). \quad (13)$$

The satisfaction of (13), along with the non-negativity of the directed divergence for $\alpha \geqslant 0$, are shown in Appendix B.

One generalized form for the directed divergence measure is the f-divergence (Aczél, 1978) based on which the distortion measure can be defined as

$$D_{\text{gen}}(P; Q) = \sum_{l \in \mathscr{L}} \sum_{k \in \mathscr{K}} \sum_{j=1}^{2} q_{jkl} f\left(\frac{p_{jkl}}{q_{jkl}}\right), \quad (14)$$

where $f(\cdot)$ is a convex function. This specializes to the directed divergence with $\alpha = 1$ (also known as the Kullback–Leibler divergence) if $f(x) = x \log x$; to the χ^2-divergence (Aczél, 1978) if $f(x) = (x-1)^2$; to the K-directed divergence (Lin, 1991) if $f(x) = x \log\{2x/(1+x)\}$ and to the variational distance (Rao and Nayak, 1985) if $f(x) = |x-1|$. It may be noted that there exist relationships among many of these measures (e.g., a lower bound for the Kullback–Leibler divergence in terms of the variational distance is given in (Toussaint, 1975). In this work, we also use a "symmetrized" divergence measure $S_\alpha(P; Q)$ defined as

$$S_\alpha(P; Q) = D_\alpha(P; Q) + D_\alpha(Q; P). \quad (15)$$

The divergence measures based on the entropies other than the Rényi–Shannon type can also be studied. One such common example is

$$C_{1.5}(P; Q) = \sum_{l \in \mathscr{L}} \sum_{k \in \mathscr{K}} \sum_{j=1}^{2} \left(\sqrt{p_{jkl}} - \sqrt{q_{jkl}}\right)^2, \quad (16)$$

based on the Havrda–Charvat entropy $E_{1.5}(P)$ given as (Rao and Nayak, 1985)

$$E_{1.5}(P) = 2\left(1 - \sum_{j=1}^{2} p_j^{1.5}\right). \quad (17)$$

In order to maintain the boundedness of the measure, in general, we impose a condition that the probability of firing or non-firing for the original and the coded signal cannot be a complete certainty or uncertainty; and accordingly we associate a 1^- or a 0^+ probability, as appropriate.

4.2. Experimental results

Twelve speech utterances, of 1–2 sec durations and spoken by male as well as female, were considered for the test. Digitized versions of these speech sentences (listed in Appendix C) were stored in audio-files having SNR of 50 dB approximately. Each of these original utterances were passed through six different code-excited linear prediction (CELP)-type speech coders.

No database containing various types of coded/distorted speech with accompanying MOS ratings was available to us. Also, we did not attempt to develop MOS ratings as it implies substantial cost and considerable time. Obtaining such a subjective scale involves the great difficulty of repeatability and elimination of biases and artifacts – especially without well-understood anchors. The quantization distortion unit (QDU), defined as the quantity of distortion subjectively equivalent to that of a single encoding of 64 kbit/s PCM, has often been used in practice as a distortion measure. Recent tests, however, indicate that the QDU may not be as stable and dependable as once it was thought to be (Kubichek, 1991). Considering all these aspects, we decided to administer an informal subjective test against which the objective measure results were judged.

In this subjective test, twelve listeners ranked six different coded versions (two with 8 kbit/s coders $C1$, $C2$ and four with 4.8 kbit/s coders $C3$, $C4$, $C5$, $C6$) of all the twelve speech utterances. The overall perceptual quality of the coded signals was designated as the basis for the order of their preferences. Subsequently, we carried out an objective evaluation of these coded signals with reference to the original speech signal by considering eight variations of the proposed fidelity criterion. These measures were as follows:

1. the directed divergence with $\alpha = 1$ [$D_1(P; Q)$],
2. the directed divergence with $\alpha = 1.5$ [$D_{1.5}(P; Q)$],
3. the directed divergence with $\alpha = 2$ [$D_2(P; Q)$],

4. the symmetrized divergence with $\alpha = 1$ $[S_1(P; Q)]$,
5. the variational distance $[V(P; Q)]$,
6. the χ^2-divergence $[\chi^2(P; Q)]$,
7. the K-directed divergence $[K(P; Q)]$ and
8. the Havrda–Charvat entropy-based $C_{1.5}$-divergence $[C_{1.5}(P; Q)]$.

A comparison of the informal listening test results and the objective measure values leads us to make the following remarks.

4.2.1. Performance of objective measures

In Fig. 7, the time-domain waveforms and the spectrograms of an original and three coded versions of a typical speech sentence, say, "Oak is strong and also gives shade" (with 18,800 samples), are shown. Table 1 provides average distortion measure values per clock time (with a base-10 logarithm, wherever applicable) for the aforesaid speech utterance. We also tabulate the values of corresponding SNR_{seg} as well as SNR with and without scaling ("scaling" implies multiplication of all the coded speech samples by an appropriate factor so as to maximize the SNR value).

In Table 2, we provide subjective and objective measure values per clock time for each of the sentences. The subjective rankings (6 for the best and 1 for the worst) are averaged over the rankings made by the twelve listeners. These scores are average ordinal numbers and not the absolute quality scores. For each of the twelve utterances and six coded versions, the average ranking scores are mentioned in the first column (marked "S"). As an example, if a coded signal is given a score of "6" by eight listeners, a score of "5" by three listeners and a score of "4" by one listener, the "S" value becomes $(6 \times 8 + 5 \times 3 + 4 \times 1)/12 = 5.58$.

On the other side, we have computed the eight variations of the CDI measure values. However, here we tabulate only the $D_1(P; Q)$ measure

Table 1
Different measure values for three coded signals (with three different 4.8 kbit/s speech coders) with reference to the original speech utterance F3 ("×" indicates that the objective measures for "oakf8f" and "oakf8k" do not agree with the subjective rankings)

Measure type	oakf8f	oakf8k	oakf8b	
Subjective ranking	Best	Good	Poor	
$D_1(P; Q)$	2.721	2.756	4.273	
$D_{1.5}(P; Q)$	4.492	4.540	6.916	
$D_2(P; Q)$	6.751	6.812	10.165	
$S_1(P; Q)$	2.730	2.760	4.285	
$V(P; Q)$	8.777	8.845	11.454	
$\chi^2(P; Q)$	17.326	15.486	19.111	×
$K(P; Q)$	0.795	0.806	0.909	
$C_{1.5}(P; Q)$	0.077	0.083	0.117	
SNR (without scaling [dB])	8.724	9.178	−2.597	×
SNR (with scaling [dB])	8.979	9.334	0.009	×
SNR_{seg} [dB]	6.815	7.080	−2.004	×

Table 2
Subjective and objective measure values for coded signals with reference to the corresponding original speech utterances (M1–M6 (male) and F1–F6 (female) are speech utterances [given in Appendix C], C1–C6 are speech coders, "S" denotes the average subjective ranking scores and "D_1" gives the directed divergence measure values with $\alpha = 1$)

Sent.	C1		C2		C3		C4		C5		C6	
	S	D_1	S	D_1	S	D_1	S	D_1	S	D_1	S	D_1
M1	5.75	2.569	4.92	2.662	4.17	2.703	2.58	2.741	2.58	2.744	1.00	4.931
M2	5.50	2.630	5.17	2.651	4.25	2.678	2.75	2.702	2.25	2.793	1.08	4.817
M3	5.75	2.573	5.17	2.623	4.00	2.720	2.58	2.753	2.33	2.782	1.17	4.333
M4	5.00	2.672	5.67	2.654	4.25	2.716	2.50	2.752	2.58	2.747	1.00	4.776
M5	5.75	2.578	5.17	2.627	3.83	2.692	2.67	2.725	2.50	2.759	1.00	4.833
M6	5.58	2.621	5.25	2.666	3.83	2.696	2.75	2.719	2.42	2.760	1.17	4.669
F1	5.67	2.607	5.00	2.671	4.25	2.695	2.33	2.801	2.58	2.751	1.17	4.722
F2	5.67	2.612	5.00	2.678	3.91	2.737	2.67	2.766	2.50	2.774	1.25	4.285
F3	5.50	2.619	5.17	2.648	4.25	2.721	2.50	2.756	2.25	2.771	1.33	4.273
F4	5.41	2.661	5.25	2.649	4.17	2.700	2.75	2.729	2.17	2.793	1.25	4.562
F5	5.50	2.653	5.50	2.658	3.83	2.743	2.33	2.797	2.50	2.765	1.33	4.414
F6	5.67	2.602	4.83	2.674	4.08	2.694	3.08	2.701	2.17	2.791	1.17	4.379

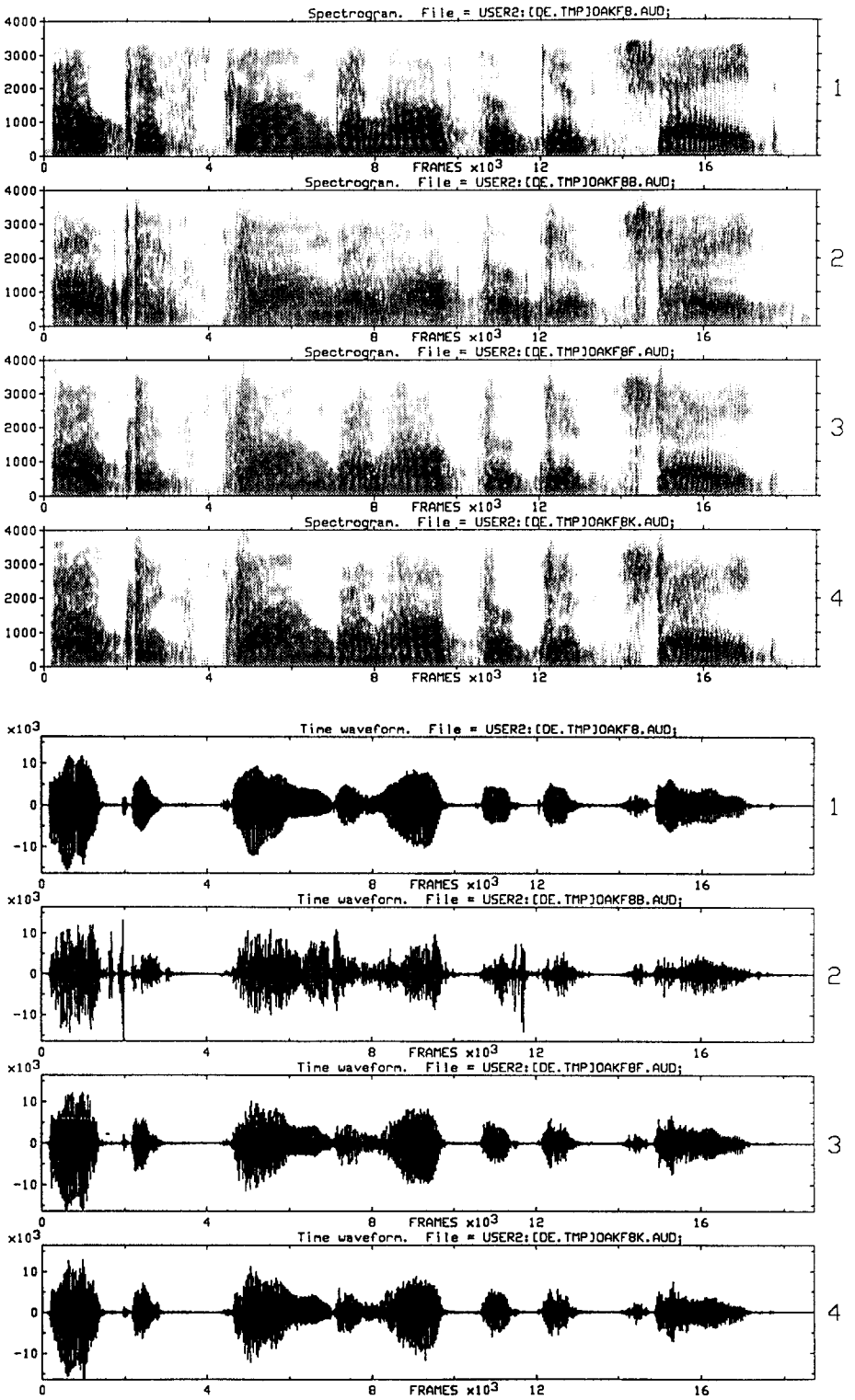

Fig. 7. Time-domain waveforms and spectrograms of an original and three coded speech signals; "*Oak is strong and also gives shade*".

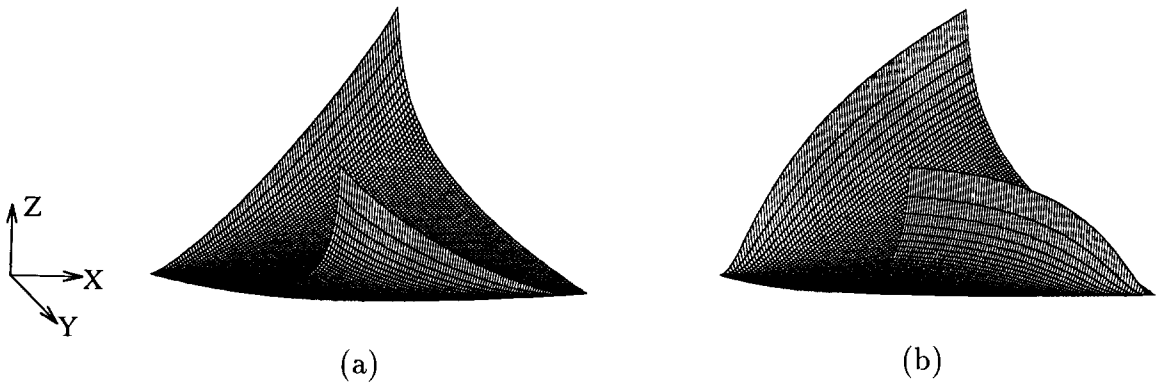

Fig. 8. The discrimination measure profiles ($J = 2$): (a) the directed divergence with $\alpha = 1$; (b) the directed divergence with $\alpha = 2$.

values (in the second column marked "D_1") as an example and make general remarks about the other measures. It is emphasized that the lower the amount of additional information (cross-entropy), the better is the signal quality of the coded speech with reference to the original one. In Table 2, we observe that with the utterance $M1$, the $C4$, $C5$ coders and with the utterance $F5$, the $C1$, $C2$ coders were ranked same subjectively. Objective measures have shown slight preference towards $C4$ coder for $M1$ and towards $C1$ coder for $F5$. Besides that, for the utterance $F4$, the subjective and objective rankings were in contradiction for the coders $C1$, $C2$.

Over the test sentences, the human rankings were found to be almost consistent with the measures $D_1(P; Q)$, $D_{1.5}(P; Q)$, $D_2(P; Q)$ and $S_1(P; Q)$; and satisfactorily consistent with the measures $K(P; Q)$ and $C_{1.5}(P; Q)$. Furthermore, the $D_\alpha(P; Q)$ class of the measures has shown conformance to subjective evaluation results where the SNR measure (with or without scaling) and also the $\mathrm{SNR}_{\mathrm{seg}}$ measure have failed. However, the $V(P; Q)$ and the $\chi^2(P; Q)$ measures often disagreed with the subjective rankings, especially when two coded signals were very close in their perceptual quality.

4.2.2. Effect of different entropies

The $D_1(P; Q)$ and the $D_2(P; Q)$ measure profiles for one neural channel at a particular clock time are presented in Fig. 8 where the X-axis is

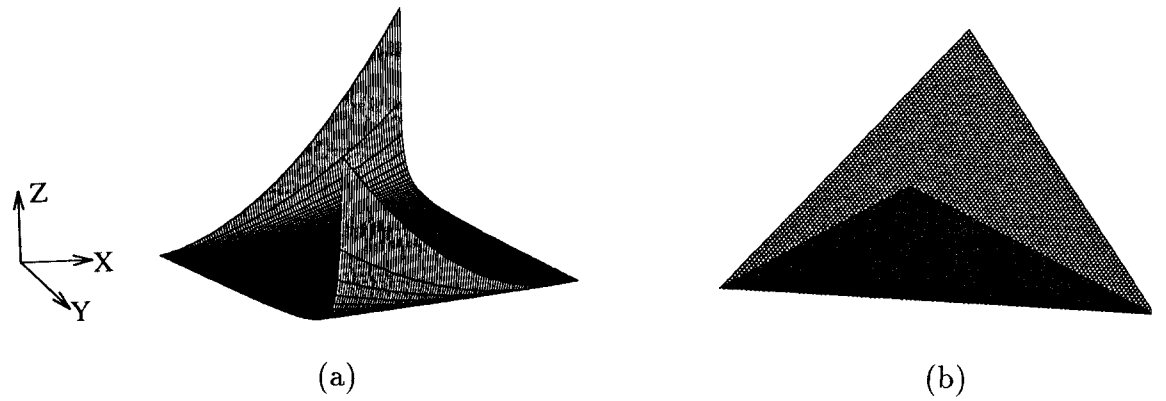

Fig. 9. The discrimination measure profiles ($J = 2$): (a) the χ^2 divergence; (b) the variational distance.

the probability-of-firing for the original signal, the Y-axis is the probability-of-firing for the coded signal in the same channel and the Z-axis is the corresponding measure. It was noticed that the value of α in the $D_\alpha(P; Q)$ measure class has a consistent but small effect on its performance. For finer classification (i.e., classifying two coded signals almost equal in their perceptual quality), it has been found to be useful to apply an α value larger than one to increase the dynamic range of the measure values. It has also been observed that the measures based on the Rényi–Shannon entropy show better performance than that based on the Havrda–Charvat entropy.

4.2.3. Effect of gain changes

The $\chi^2(P; Q)$ and $V(P; Q)$ measure profiles with the same X, Y and Z axes as in Fig. 8 are shown in Fig. 9. In addition to the AGC nonlinearity, all the measure profiles (except the $V(P; Q)$) exhibit nonlinearity and the measure values are relatively very small in the neighborhood of the $X = Y$ region. This also makes them insensitive to small gain changes. We speculate that a linear profile of the $V(P; Q)$ measure is responsible for its poor performance. Due to its broad flatness around the $X = Y$ region, the $\chi^2(P; Q)$ measure shows less sensitivity to gain changes; however, this may be the reason for its unsatisfactory performance in the coder evaluation.

4.2.4. Effect of sample delays

The CDI measures, in general, were found to be relatively less sensitive (compared to the SNR measure) to a slight time misalignment of the coded signal with respect to the original one or vice versa. For example, let us consider the coded speech signals marked "oakf8f" and "oakf8k" of Fig. 7. Table 3 provides the SNR measure (without scaling) values as well as the $D_1(P; Q)$ and the $D_2(P; Q)$ measure values with zero, one, two and three sample delays in the coded speech. These sample delays are with reference to the original signal and the misaligned sample places are filled in with zero values. In general, we observe that one sample delay does not cause much change in the CDI measure values, but two or three sample delays have considerable effect. With three sample delays, the measures show "oakf8f" to be inferior to "oakf8k" (which is aligned to the original signal) although subjectively the reverse is true.

4.2.5. Speech coder identification

By considering the neural pathway to be a noisy channel, the subjective evaluation of the speech coders can be treated as a hypothesis testing problem. Csiszár and Longo (Blahut, 1974) have shown that the probability-of-error of optimum hypothesis testers based on blocks of measurements decreases exponentially with the block length. Let us consider two coded speech of the same utterance and let γ^* be the smallest probability that "C" is identified to be the samples of "A" when it is actually the samples from "B". This probability is smallest over all the decision rules such that the probability of other type of error (i.e., "C" chosen as samples of "B" when it is actually from "A") does not exceed β. Then,

Table 3
The directed divergence (with $\alpha = 1, 2$) measure values with zero, one, two and three sample delays for the coded signal "oakf8f" and "oakf8k" with reference to the original speech sentence

Coded speech	Measure	Sample delays			
		Zero	One	Two	Three
oakf8f	SNR (without scaling [dB])	8.724	7.391	5.619	5.117
oakf8f	$D_1(P; Q)$	2.721	2.728	2.747	2.779
oakf8f	$D_2(P; Q)$	6.751	6.792	7.193	8.838
oakf8k	SNR (without scaling [dB])	9.178	7.503	6.108	7.027
oakf8k	$D_1(P; Q)$	2.756	2.762	2.791	3.128
oakf8k	$D_2(P; Q)$	6.812	6.855	7.124	8.950

γ^*, for all β in (0, 1) and with $\alpha = 1$, can be given as (Blahut, 1974)

$$\gamma^* \sim \exp\left[-\sum_{l\in\mathscr{L}}\sum_{k\in\mathscr{K}} D(P_{kl}; Q_{kl})\right]. \tag{18}$$

We conducted an experiment where the listeners were asked to listen to two coded speech sentences "A" and "B" and then a varying number of samples "C" from one of them, not known to the listeners which one, were played. In such subjective evaluation testing, there is no precise way of determining γ^*. The γ^* could be estimated by carrying out the test with a large number of listeners and then considering their opinions (whether "A" or "B") and about "C".

It would be of academic interest to investigate the validity of the relationship of (18). In our experiment, we only verified that to achieve a given probability of decision error, it required more samples (i.e., longer durations) of "C" to be played when "A" and "B" are of "near equal" quality (as indicated by our measure) compared to that required when "A" and "B" are of "substantially different" quality. Table 4 shows, for the same example sentence, the subjective identification of coders (i.e, the number of listeners out of twelve listeners correctly identified the coders) and the corresponding number of samples played. We have considered three coder pairs where $C4$–$C5$, $C3$–$C4$ and $C5$–$C6$ were ranked in the descending order from their perceptual quality "closeness" point of view. For example, let us consider the utterance $F3$. In Table 4, we observe that by playing 6,000 samples, for the $C4$–$C5$ coder pair, only one-half of the listeners could identify the coder correctly, the remaining listeners either identified wrongly or could not decide. On the other hand, with the same number of samples played, the correct coders were identified by two-third of the listeners for the $C3$–$C4$ pair and by almost all the listeners for the $C5$–$C6$ pair.

Table 4
Speech coder identification for two sentences M1 and F3 (the sample numbers played and the fraction of listeners who have correctly identified the coders are provided in the table)

Sentence	Sample nos.	C3–C4	C4–C5	C5–C6
M1	3,000	5/12	4/12	7/12
	6,000	7/12	7/12	9/12
	9,000	11/12	10/12	12/12
	12,000	12/12	12/12	12/12
F3	3,000	6/12	4/12	8/12
	6,000	8/12	6/12	11/12
	9,000	11/12	9/12	12/12
	12,000	12/12	11/12	12/12

5. Rate–distortion analysis

Rate–distortion theory essentially establishes a mathematical foundation to the source encoding problem. In general, a source–destination pair is characterized by a probabilistic model of the source encoder and a fidelity criterion measuring the degradation 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 subject to the constraint that an average distortion of D is endured at the destination. A knowledge of $R(D)$ is of considerable importance as it may prevent one from frivoling time as well as the resources to achieve an impossible task.

The function $R(D)$ with respect to a defined distortion measure is defined as (Berger, 1971)

$$R(D) = \inf_{Q\in Q_D} I(Q), \tag{19}$$

where $Q = \{Q_{v|u}\}$ with $Q_{v|u}$ as the conditional probability defined for an input alphabet u substituted by an output alphabet v; and the Q_D is the set of all D-admissible conditional probability assignments. $I(Q)$ is a convex downward function of Q which implies that any stationary point of $I(Q)$ in Q_D must yield the infimum (absolute minimum), namely the $R(D)$. As D increases, $R(D)$ decreases monotonically and usually becomes zero at some finite value of distortion.

Historically, the application of the rate–distortion theory to the speech process has been hindered because of the lack of a widely accepted probabilistic model of the speech process as well as a meaningful distortion measure. The problem

is further complicated by the mathematical difficulties in evaluating the rate–distortion function even if a reasonable source–destination pair is defined. A fairly large set of probability density function models is suggested in the literature based on the first-order histograms of Nyquist samples of continuous speech waveforms. The gamma pdf based on the long-term statistics (Jayant and Noll, 1986) the Laplacian pdf based on the medium-term statistics (Noll, 1974) and the Gaussian pdf based on the short-term statistics (Richards, 1964) are among the more popular ones. An evaluation of the first-order $R(D)$ functions based on these pdfs and difference distortion measures are available in (Abut and Erdöl, 1979) and with Itakura–Saito distortion measure in (Buzo et al., 1986).

The objective here is to provide a rate–distortion-theoretic analysis for speech coders with the CDI measure. We formulate the problem by characterizing the source–destination pair precisely. Then, the $R(D)$ function is computed using the Blahut algorithm. Finally, the performances of different speech coders are studied with respect to these bounds.

5.1. Source–destination pair characterization

The cochlear model is, in essence, a highly non-linear structure with the half-wave rectifiers, the AGC stages and the coupling among them simulating the auditory spectral and temporal masking phenomena. It may prove to be sufficiently difficult to express these signal processing operations, especially the coupling of the AGC stages, with the help of simple mathematical operators. Thus, we take a different outlook towards the source–destination pair model shown in Fig. 10. We merge the physical speech source with the cochlear model and consider this ensemble to be the source. Since there is as such no uniquely accepted pdf for the physical speech source, we are not in any further disadvantageous position by integrating the cochlear model with the speech source and determining the histogram of the cochlear model outputs. These outputs, being the probability-of-firing information, assume values in the range (0, 1). The histogram for the firing-probability is determined by experimenting with twenty-four speech utterances (twelve male and twelve female voices) of 1–2 sec durations. The firing-probability histogram for each of the sixty-four neural channels could be determined separately. For simplification purpose, we have assumed all the histograms to be identical and derived only one histogram based on the probability-of-firing information obtained from all the channels.

5.2. Calculation based on Blahut's algorithm

In (De and Kabal, 1992b), we have derived analytically a lower bound to the $R(D)$ with a single-letter cochlear variational distance measure. However, with the other distortion measure forms, it becomes difficult to give an analytical solution. Moreover, these are not exact solutions; they are merely lower bounds. Here, we use the Blahut algorithm for calculating the $R(D)$ functions exactly.

We treat the probability-of-firing information to be discrete-valued with symbols from one of the 255 uniformly spaced values between 0 and 1 (i.e., $1/256, 2/256, \ldots, 255/256$). Let the input alphabet (firing-probability corresponding to the original speech) u be reproduced in terms of an output alphabet (firing-probability corresponding to the coded speech) v. Then, the algorithmic steps could be written as follows.

Step 1. An initial output probability distribution $\{Q_v\}$ is assumed, say, Q_v^0. The parameter set $\{A_{uv} = e^{s\rho_{uv}}$ is evaluated, where ρ_{uv} is the single-

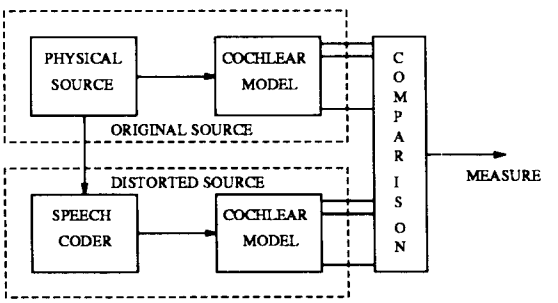

Fig. 10. Source–destination pair characterization.

letter CDI measure between the input alphabet u and the output alphabet v.

Step 2. The parameter s is chosen from the range of $-\infty$ to 0; and then Steps 3 and 4 are carried out with different values of s.

Step 3. With the values of the input probability distribution P_u (obtained from the histogram of the cochlear model ouptut) and the parameters A_{uv} the following parameters are calculated:

$$c_v = \sum_u P_u \frac{A_{uv}}{\sum_v A_{uv} Q_v}, \quad Q_v \leftarrow Q_v c_v, \tag{20}$$

$$L = \sum_v Q_v \log c_v, \quad U = \max_v \log c_v. \tag{21}$$

Step 4. If $U - L \geqslant \varepsilon$, then the previous step is repeated; otherwise, the program is terminated for this value of s by evaluating the following:

$$Q_{v|u} = \frac{A_{uv} Q_v}{\sum_v A_{uv} Q_v}, \tag{22}$$

$$D = \sum_u \sum_v P_u Q_{v|u} \rho_{uv}, \tag{23}$$

$$R(D) = sD - \sum_u P_u \log\left(\sum_v A_{uv} Q_v\right) - \sum_v Q_v \log c_v. \tag{24}$$

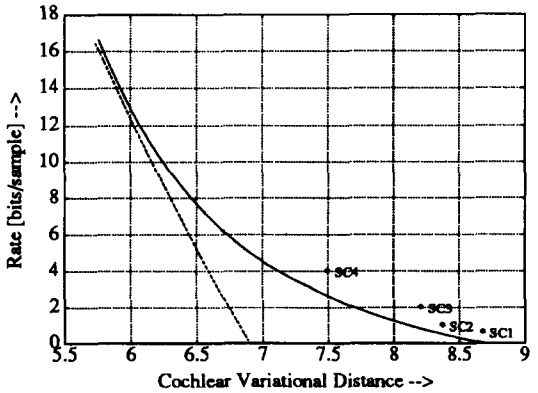

Fig. 11. Speech coder rate in bits/sample versus average cochlear variational distance measure (— — — line shows an analytically derived lower bound, ——— line shows the exact rate–distortion curve using Blahut's algorithm and four " * " points [SC1–SC4] denote the performances of four speech coders).

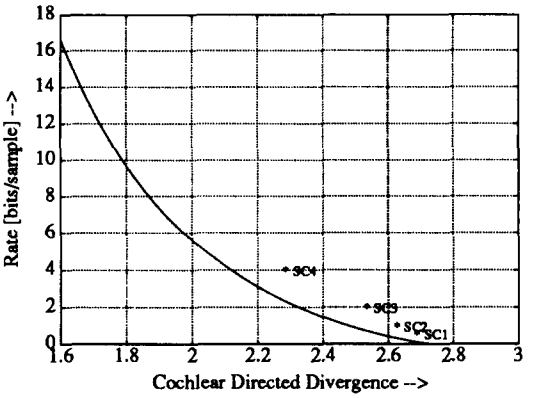

Fig. 12. Speech coder rate in bits/sample versus average cochlear directed divergence (with $\alpha = 1$) measure (——— line shows the rate-distortion curve using Blahut's algorithm and four " * " points [SC1–SC4] denote the performances of four speech coders).

Fig. 11 shows the $R(D)$ for the $V(P; Q)$ measure whereas Fig. 12 plots the $R(D)$ function for the $D_1(P; Q)$.

5.3. Measured performances of speech coders

We have considered four state-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 kbit/s) (Campbell et al., 1991), VSELP-based coder SC2 (8 kbit/s) (Atal et al., 1991), wideband CELP-based coder SC3 (16 kbit/s) (Roy and Kabal, 1991) ADPCM coder SC4 (32 kbit/s) (Jayant and Noll, 1984). For the first, second and the fourth coders with sampling rates of 8,000 Hz, sixty-four neural channels (covering up to 4,000 Hz band) were assigned. On the other hand, for the wideband coder with sampling rate of 16,000 Hz, eighty-five neural channels (covering up to 8,000 Hz band) were assigned as described in (De, 1993). Although we considered only the CELP-type speech coders for comparing the CDI measure performance with subjective assessment, we do not foresee any difficulty in applying this measure to other types of speech coders. With this understanding, we have included one ADPCM coder in this section to examine its quality with respect to the rate–distortion limit.

Twelve speech sentences of 1–2 sec durations were passed through each of the four coders to calculate the average distortion values over each sampling time. Figs. 11 and 12 plot the performances of the four speech coders (marked by "$*$") as evaluated by $V(P; Q)$ and $D_1(P; Q)$, respectively. Now, let us examine one of the figures, Fig. 12. We observe that it is possible to attain the perceptual quality obtained (measured with the $D_1(P; Q)$) by SC1 coder at a much lower rate (as low as 1.5 kbit/s). Similarly, SC2, SC3 and SC4 coder performances are achievable with almost 3.8 kbit/s, 5.4 kbit/s and 20 kbit/s, 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 kbit/s 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 kbit/s 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 and how close a specific speech coder is performing with respect to the $R(D)$ limits in terms of the perceptual quality.

6. Summary and conclusions

The formulation of any distortion measure requires resolution of two important issues: (i) defining a suitable domain in which the signal can be characterized parametrically and (ii) comparing these parameters in a meaningful sense. In this article, we have introduced and studied a distortion measure, namely the cochlear discrimination information measure, which compares the neural-firing information corresponding to an original speech and its coded version in a cross-entropic sense. An insufficient knowledge about the exact neural firing processes has prompted us to use the probabilistic information of firing/non-firing in the comparison. We have investigated several variants of the CDI measure based on different types of entropy, the associated parameters and also the cross-entropic measure form. The effects of gain changes and sample delays, etc. have also been studied. The directed divergence measure form based on the Rényi–Shannon entropy has shown very good performance by conforming strongly with informal subjective test in terms of ranking coded speech from six different coders. Subsequently, a rate–distortion analysis for speech coder has also been carried out with this measure.

While converting the time-domain speech signal into its corresponding perceptual-domain representation by an auditory model; the resonating nature of the cochlea, the perceptual nonlinearity as well as the temporal and spectral masking effects have been considered. An inclusion of the spectral masking feature has allowed the probability-of-firing information in a particular neural channel at a specific clock time to depend not only on the strength of the gain-controlled signal of that channel but also on those of the other channels. Similarly, the same probability-of-firing information depends not only on the strength of the gain-controlled signal at that clock time but also on those at the other times. Thus, the PD representation for speech signal has exploited reasonably the interdependencies at the auditory periphery level.

In the CDI measure, we have compared the cochleagram matrices (whose elements are the probability-of-firing information), element-by-element, for the original and the coded speech signals. This measure has been found to be not very robust against the coder delays. Thus, estimating and removing time-delay between the original and the coded speech are, in some sense, necessary first steps in applying the CDI measure. We have also proposed and studied a cochlear hidden Markovian (CHM) measure (De, 1993; De and Kabal, 1994), which by considering the temporal ordering in the firing pattern has shown a greater robustness against the coder delays. We emphasize that formal testing with a wide range of linear and nonlinear coder distortions and a subsequent refinement of the cochlear model would be needed to validate our distortion measure methodology. We anticipate, nonetheless, that the present framework of comparing the firing/non-firing probabilities would continue to be appropriate.

Appendix A. Concavity of $H_\alpha(P)$ with $P \in \mathscr{P}^{(2)}$

It is shown in (Rényi, 1970) that the $H_\alpha(P)$ is concave for $0 \leqslant \alpha \leqslant 1$. We show here that for $J = 2$, the concavity is also satisfied for $1 < \alpha \leqslant 2$.

$$H_\alpha(P) = \frac{1}{(1-\alpha)} \log(p_1^\alpha + p_2^\alpha),$$

$$\text{where } p_2 = 1 - p_1, \quad p_1, p_2 \geq 0. \tag{A.1}$$

$$\begin{aligned}
\frac{d^2 H_\alpha(P)}{dp_1^2} &= \frac{\alpha}{(1-\alpha)} \Big\{ (\alpha - 1)(p_1^\alpha + p_2^\alpha)(p_1^{\alpha-2} + p_2^{\alpha-2}) \\
&\quad - \alpha(p_1^{\alpha-1} - p_2^{\alpha-1})^2 \Big\} / (p_1^\alpha + p_2^\alpha)^2 \\
&= \frac{\alpha}{(1-\alpha)} (p_1^{\alpha-2} p_2^{\alpha-2}) \{ \alpha - (p_1^\alpha + p_2^\alpha) \\
&\quad \times (p_1^{2-\alpha} + p_2^{2-\alpha}) \} / (p_1^\alpha + p_2^\alpha)^{-2}.
\end{aligned} \tag{A.2}$$

It is noted that for $\alpha > 1$,

$$(p_1^\alpha + p_2^\alpha) < (p_1 + p_2)^\alpha = 1. \tag{A.3}$$

Furthermore, $p_1 = p_2 = 1/2$ maximizes the expression $(p_1^{2-\alpha} + p_2^{2-\alpha})$ for $1 < \alpha \leqslant 2$. Consequently, we note that $\alpha > (\frac{1}{2})^{1-\alpha}$ for $1 < \alpha \leqslant 2$. Additionally, we observe that the denominator factor $(1 - \alpha)$ of (A.2) is negative for $\alpha > 1$. Thus, $H_\alpha(P)$ is proved to be concave in the range $1 < \alpha \leqslant 2$.

Now, we investigate the concavity for $J = 2$ and $\alpha > 2$. With sufficiently small $\delta > 0$ and $p_1 = \delta$ or $p_2 = \delta$, we obtain

$$\frac{d^2 H_\alpha(P)}{dp_1^2} > 0. \tag{A.4}$$

On the other hand, with $p_1 = p_2 = 1/2$, we have

$$\frac{d^2 H_\alpha(P)}{dp_1^2} = -4\alpha < 0. \tag{A.5}$$

From (A.4) and (A.5), we observe that for $J = 2$ and $\alpha > 2$, $H_\alpha(P)$ is neither convex nor concave.

Appendix B. The directed divergence measure based on $H_\alpha(P)$

I. Non-negativity of the measure:

$$\begin{aligned}
D_1(P_{kl}; Q_{kl}) &= \sum_{j=1}^{2} p_{jkl} \log\left(\frac{p_{jkl}}{q_{jkl}}\right) \\
&\geqslant \sum_{j=1}^{2} p_{jkl}\left(1 - \frac{q_{jkl}}{p_{jkl}}\right) \\
&= \sum_{j=1}^{2} p_{jkl} - \sum_{j=1}^{2} q_{jkl} = 0.
\end{aligned} \tag{B.1}$$

$$\begin{aligned}
D_\alpha(P_{kl}; Q_{kl}) &= \frac{1}{\alpha - 1} \log\left(\sum_{j=1}^{2} \frac{p_{jkl}^\alpha}{q_{jkl}^{\alpha-1}}\right) \\
&\geq \frac{1}{\alpha - 1}\left[1 - \frac{1}{\sum_{j=1}^{2} \frac{p_{jkl}^\alpha}{q_{jkl}^{\alpha-1}}}\right], \\
&\alpha \neq 1.
\end{aligned} \tag{B.2}$$

To show that $D_\alpha(P_{kl}; Q_{kl}) \leqslant 0$, we need to show that

$$\begin{aligned}
Y_\alpha(P_{kl}; Q_{kl}) \equiv \sum_{j=1}^{2} \frac{p_{jkl}^\alpha}{q_{jkl}^{\alpha-1}} &\leq 1 \quad \text{for } 0 \leq \alpha < 1, \\
&\geq 1 \quad \text{for } \alpha > 1.
\end{aligned} \tag{B.3}$$

We note that $p_{1kl} = q_{1kl}$ and $p_{2kl} = q_{2kl}$ maximizes $Y_\alpha(P_{kl}; Q_{kl})$ for $0 \leqslant \alpha < 1$ and minimizes it for $\alpha > 1$. Thus, the $Y_\alpha(P_{kl}; Q_{kl})$ conditions of (B.3) are met and hence the non-negativity of the divergence measure (the Rényi–Shannon type) is also satisfied. The measure becomes equal to zero if and only if the distributions P_{kl} and Q_{kl} become the same.

II. Additivity of the measure

With $w \in \mathscr{I} \times \mathscr{K}$ and $m = nN$, we obtain

$$\begin{aligned}
D_1(P; Q) &= \sum_{j_1=1}^{2} \sum_{j_2=1}^{2} \cdots \sum_{j_m=1}^{2} \left(\prod_{w=1}^{m} p_{j_w}\right)\left[\sum_{w=1}^{m} \log \frac{p_{j_w}}{q_{j_w}}\right]
\end{aligned}$$

A. De (1993), Auditory distortion measures for speech coder evaluation, PhD thesis, McGill University.

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

A. De and P. Kabal (1992b), "Rate distortion function for speech coding based on perceptual distortion measure", *Proc. IEEE Globecom '92*, pp. 452–456.

A. De and P. Kabal (1994), "Auditory distortion measure for speech coder evaluation – Hidden Markovian approach", *Speech Communication*, submitted.

L. Deng (1992), "Processing of acoustic signals in a cochlear model incorporating laterally coupled suppressive elements", *Neural Networks*, Vol. 5, pp. 19–34.

J. Flanagan (1972), *Speech Analysis, Synthesis and Perception*, (Springer, Berlin).

C.D. Geisler (1988), "Representation of speech sounds in the auditory nerve", *J. Phonetics*, Vol. 16, pp. 19–35.

O. Ghitza (1987), "Auditory nerve representation criteria for speech analysis/synthesis", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-35, pp. 736–740.

R.M. Gray, A. Buzo, A.H. Gray Jr. and Y. Matsuyama (1980), "Distortion measures for speech processing", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. ASSP-28, pp. 367–376.

S. Greenberg (1988), "The ear as a speech analyzer", *J. Phonetics*, Vol. 16, pp. 139–149.

U. Halka and U. Heute (1992), "A new approach to objective quality-measures based on attribute-matching", *Speech Communication*, Vol. 11, No. 1, pp. 15–30.

T. Hall (1980), "Cochlear models: Evidence in support of mechanical nonlinearity and second filters (a review)", *Hearing Res.*, Vol. 2, pp. 455–464.

M.H.L. Hecker and C.E. Williams (1966), "Choice of reference conditions for speech preference tests", *J. Acoust. Soc. Amer.*, Vol. 40, pp. 946–952.

H.V. Helmholtz (1954), *On the Sensations of Tone* (Dover, New York).

F. Itakura and S. Saito (1968), "Analysis synthesis telephony based on the maximum likelihood method", *Proc. 6th Internat. Cong. Acoust. Japan*, pp. C 17–C 20.

N.S. Jayant and P. Noll (1984), *Digital Coding of Waveforms: Principles and Applications to Speech and Video* (Prentice Hall, Englewood Cliffs, NJ).

N. Kitawaki, H. Nagabuchi and K. Itoh (1988), "Objective quality evaluation for low-bit-rate speech coding systems", *IEEE J. Select. Areas Commun.*, Vol. 6, pp. 242–248.

R.F. Kubichek (1991), "Standards and technology issues in objective voice quality assessment", *Dig. Signal Process.*, Vol. 1, pp. 38–44.

S. Kullback (1959), *Information Theory and Statistics* (Wiley, New York).

J. Lalou (1990), "The information index: An objective measure of speech transmission performance", *Ann. Telecommun.*, Vol. 45, pp. 47–65.

Y.-T. Lee (1991),"Information-theoretic distortion measures for speech recognition", *IEEE Trans. Signal Process.*, Vol. 39, pp. 330–335.

J. Lin (1991), "Divergence measures based on the Shannon entropy", *IEEE Trans. Inform. Theory*, Vol. 37, pp. 145–151.

R.F. Lyon (1982), "A computational model of filtering, detection, and compression in the cochlea", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 1282–1285.

R.F. Lyon and L. Dyer (1986), "Experiments with a computational model of the cochlea", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 37.6.1–37.6.4.

B.J. McDermott, C. Scagliola and D.J. Goodman (1978), "Perceptual and objective evaluation of speech processed by adaptive differential PCM", *Bell Syst. Tech. J.*, pp. 1597–1619.

P. Mermelstein (1979), "Evaluation of a segmental SNR measure as an indicator of the quality of ADPCM coded speech", *J. Acoust. Soc. Amer.*, Vol. 66, pp. 1664–1667.

B.C.J. Moore (1989), *Introduction to the Psychology of Hearing* (Academic Press, New York).

P. Noll (1974), "Adaptive quantizing in speech coding systems", *Zurich Seminar Dig. Commun., Zurich, Switzerland.*

D. O'Shaughnessy (1987), *Speech Communication* (Academic Press, New York).

B. Paillard, J. Soumagne, P. Mabilleau and S. Morissette (1992), "PERCEVAL: Perceptual evaluation of the quality of audio signals", *J. Audio Engrg. Soc.*, Vol. 40, pp. 21–31.

M.J. Penner (1979), "Forward masking with equal-energy maskers", *J. Acoust. Soc. Amer.*, Vol. 66, pp. 1719–1724.

J.O. Pickles (1982), *An Introduction to the Physiology of Hearing* (Academic Press, New York).

S.R. Quackenbush, T.P. Barnwell III and M.A. Clements (1988), *Objective Measures of Speech Quality* (Prentice Hall, Englewood Cliffs, NJ).

C.R. Rao and T.K. Nayak (1985), "Cross entropy, dissimilarity measures, and characterizations of quadratic entropy", *IEEE Trans. Inform. Theory*, Vol. IT-31, pp. 589–593.

A. Rényi (1970), *Probability Theory* (North-Holland, Amsterdam).

D.H. Richards (1964), "Statistical properties of speech signals", *Proc. IEEE*, Vol. 52, pp. 941–949.

G. Roy and P. Kabal (1991), "Wideband CELP speech coding at 16 kbit/sec", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 17–20.

M.B. Sachs, C.C. Blackburn and E.D. Young (1988), "Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus", *J. Phonetics*, Vol. 16, pp. 37–53.

M.R. Schroeder and J.L. Hall (1974), "Model for mechanical to neural transduction in the auditory receptor", *J. Acoust. Soc. Amer.*, Vol. 55, pp. 1055–1060.

M.R. Schroeder, B.S. Atal and J.L. Hall (1979), "Optimizing digital speech coders by exploiting masking properties of the human ear", *J. Acoust. Soc. Amer.*, Vol. 66, pp. 1647–1652.

S. Seneff (1988), "A joint synchrony/mean-rate model of auditory speech processing", *J. Phonetics*, Vol. 16, pp. 55–76.

S.A. Shamma (1985), "Speech processing in the auditory system II: Lateral inhibition and the central processing of

speech evoked activity in the auditory nerve", *J. Acoust. Soc. Amer.*, Vol. 78, pp. 1622–1632.

M. Slaney (1988), Lyon's cochlear model, Tech. Rep. 13, Apple Computer Inc.

G.T. Toussaint (1975), "Sharper lower bounds for discrimination information in terms of variation", *IEEE Trans. Inform. Theory*, Vol. IT-21, pp. 99–100.

W.D. Voiers (1977), "Diagnostic acceptability measure for speech communication systems", *Proc. IEEE Internat. Conf. Acoust. Speech Signal Process.*, pp. 204–207.

S. Wang, A. Sekey and A. Gersho (1992), "An objective measure for predicting subjective quality of speech coders", *IEEE J. Select. Areas Commun.*, Vol. 10, pp. 819–829.