Proc. Biennial Symp. Commun. (Kingston, ON), pp. 419-423, May 1992

Cochlear Discrimination: An Auditory Information-Theoretic Distortion Measure for Speech Coders†

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 this paper, our objective is to devise a fidelity criterion for quantifying the degree of distortion introduced by a speech coder. Towards this end, both original speech and its coded version are transformed from the time-domain to a *perceptual-domain* using a cochlear model. This perceptual-domain representation provides information pertaining to the probability-of-firings in the neural channels. We introduce a cochlear discrimination measure which compares these firing probabilities in an information-theoretic sense. This measure, in essence, evaluates the neural-firing cross-entropy of the coded speech with respect to that of the original one. The performance of this objective measure is compared with subjective evaluation results.

1 Introduction

Distortion measure plays a vital role in the evaluation as well as in the design of a low bit-rate speech coder. The measurement of distortion involves devising a transformation operator for mapping the signals onto an appropriate domain and formulating a suitable comparison in that domain. 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. In this article, we propose a fidelity criterion using a filter bank approach for coded/distorted speech signals. Details of cochlear (inner ear) and other auditory processing involved in the speech perception are imbibed for 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. Section 2 briefly discusses the auditory system. Section 3 describes an electrical model featuring auditory processing and defines the perceptual domain. Section 4 introduces the idea of *Cochlear Discrimination*, a perceptual *cross-entropy* measure-based fidelity criterion, for speech signals. Finally, Section 5 provides the test results with relevant remarks.

2 Auditory System

An ear consists of three sections: the outer ear, the middle ear and the inner ear [1]. Speech pressure variations are directed towards the *eardrum* by the outer ear and subsequently they are transformed into mechanical motion by the middle ear. The cochlea (inner ear) converts these mechanical vibrations into electrical excitations. The *cochlear duct* is separated from the chamber *scala tympani* by the *basilar membrane* (BM) which is stiff and thin at the basal end (where the sound enters), but compliant and massive at the apical end. Each *place* along the BM responds best to one frequency termed as the *characteristic frequency* (CF). 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. On the top of the basilar membrane (within the *organ of Corti*), there are about 30,000 sensory *hair cells* arranged in several rows along the length of the cochlea. The outer hair cells function as signal level controllers whereas the inner hair cells are the primary source of nerve pulses which are propagated axonally to the brain through neural fibers.

3 Cochlear Model

We desire to deal with an accurate description of 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. Current models of representing speech in the auditory periphery falls into one of four broad classes [2]: rate/place, synchrony/place, synchrony/quasi-place and synchrony/place-independent. In this work, we adopt a synchrony/quasi-place cochlear model suggested by Lyon [3] and described by Slaney [4]. This model, as shown in Fig. 1, essentially incorporates the best features of the *place* as well as the *volley* theory [1]. Using this model, time-domain speech signals are mapped onto a perceptual domain where time-place components become the fundamental bases of analysis.

The outer-and-middle ear effectively adds a slight high-pass response to the system. With a corner frequency of 300 Hz and a unity gain at DC, a simple first-order high-pass discrete-time filter $H_{OM}(z) = 4.7635(1 - 0.79008z)$ is designed to roughly model the effects of outer and middle ear. The cochlea is best modeled using a continuous differential equation which is very difficult to implement on a digital computer. In this work, we use discrete-place approximation and consider sixty-four stages in cascade, each of which has different frequency sensitivity representing the associated resonance and is characterized by the respective filter transfer function.

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

An implementation of the discrete-place stages involves combining a series of biquadratic notch filters that model the traveling pressure waves with a series of biquadratic resonators that model the conversion of pressure waves into basilar membrane motion. Locations of the poles and the zeros of notch and resonator filters are important. 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.

In order to reduce computation, the notch and the resonator filters of each stage are integrated into a single ear-stage filter by choosing the poles of the resonator filters to be at the same locations as the poles of the succeeding notch filter [4]. The composite transfer function at any place is an asymmetric band-pass function. In conformance to various psycho-acoustical experimental data, $W_{ear}(f_c)$, the 3-dB bandwidth of a band-pass filter with center frequency f_c, is defined as $W_{ear}(f_c) = \sqrt{f_c^2 + f_{cb}^2}/Q_{ear}$, where the ear-break frequency f_{cb} is 1,000 Hz and the constant Q-factor for all the bandpass filters Q_{ear} is 8. To simulate the real situation, four successive ear-filter stages are overlapped within the 3-dB bandwidth of any one filter. The filter-stages are indexed from 1 to 64 (high to low frequencies) and the center frequency of each stage decreases by 0.25 times the bandwidth of the previous stage. $W_{ear}(f_c)$ *vs.* f_c of all the sixty-four stages has been plotted in Fig. 2.

To implement the zeros at DC for every resonator, a differentiator is required for each stage. Since all the filters used here are linear, the 'differentiator' (a term of the form $1 - z$) can be placed just once before the ear cascade. Preceding all stages of the ear-filter with a single differentiator causes the lower frequency stages to have a much lower output than the higher frequency stages. Thus, each stage is adjusted so that it has unity gain at its center frequency while within the stage, the gain is proportional to frequency. Typical frequency responses for three stage-filters have been shown in Fig. 3.

The inner hair cells act as half-wave rectifiers for the velocity of the motion of the fluid [5]. The exact shape of the half-wave nonlinearity is not obvious; there are proposals for ideal as well as soft half-wave rectification. In this work, an ideal half-wave rectifier is considered.

Experimental studies of a fully functioning cochlea indicate that the transfer characteristics of the basilar membrane are nonlinear [6]. This non-linearity is captured by the automatic gain control (AGC) stages which amplify the weak signals and diminish the strong signals. The most important adaptation mechanism in sensory systems is *lateral inhibition* by which sensory neurons reduce their own gain as well as the gain of others nearby. Lyon proposed a coupled AGC that adapts in frequency dimension and simulates lateral inhibition [3]. The gain control effect is not instantaneous and the time required to adapt to any input signal is strongly dependent on the signal level. A cascade of four AGC stages with different *time constants*, simulating different adaptation times in the ear, are used in this work. A longer time constant implies that the AGC takes longer to respond to the input. Each AGC attenuates the incoming signal so that, on an average, it remains below the *target* value corresponding to that AGC. The present model consists of four stages of AGC whose time constants and target values, respectively, are chosen as: 640 ms, 160 ms, 40 ms and 10 ms and 0.0032, 0.0016, 0.0008 and 0.0004 (relative gains).

The nerve cells (neurons) 'fire' (an all-or-none electrical spike) in response to the compressed signal as sensed by the hair cells at different places of the cochlea. These neural activity patterns for various stages, which contain information regarding the formant, the pitch and the timbre of the speech signal, can also be treated as perceptual-domain representations (values related to probability-of-firing *vs.* time). Eventually, these firing probability values obtained for an original and a coded signal are compared, in an information-theoretic sense, to obtain a simple quantitative measure related to the amount of distortion.

4 Cochlear Discrimination Information Measure

The perceptual domain representation provides 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 generates impulses based on the probability-of-firing information. These neural converters may equivalently be considered to be discrete information sources with an alphabet of two, *i.e.*, firing and non-firing. We apply the concept of discrimination information (also known as Kullback-Leibler divergence, the directed divergence, the information gain or the cross-entropy), a powerful tool [7] for quantifying the 'closeness' of two probability distribution functions, to define a distortion measure for speech coders. A cochlear discrimination measure based on the Rényi-Shannon entropy [8] is introduced below. This measure determines 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.

The sharp rise of auditory threshold for very low or high frequencies is primarily due to the fact that a smaller number of hair cells are attached with these CFs than in the mid-frequency range. This side information, which affects the given probability disributions, are taken into account by dealing with conditional probability distributions and conditional discrimination measures. 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 k-th neural channel. Similarly, $q_{1|k}$ and $q_{2|k} = 1 - q_{1|k}$ are defined for coded/distorted speech. Thus, the conditional discrimination measure becomes:

$$D_\alpha(P;Q|k) \;=\; \sum_{j=1}^{2} p_{j|k} \log\left(\frac{p_{j|k}}{q_{j|k}}\right) \quad \text{for } \alpha = 1$$

$$= \frac{1}{(\alpha-1)} \log\left(\sum_{j=1}^{2} \frac{p_{j|k}^{\alpha}}{q_{j|k}^{\alpha-1}}\right) \quad \text{for } \alpha \neq 1 \quad (1)$$

This measure is not a metric as it does not satisfy the conditions of (1) *symmetry* [$D_\alpha(P;Q|k)$ is not the same as $D_\alpha(Q;P|k)$ when P and Q are different] and (2) *triangle inequality* [the sum of the measures $D_\alpha(P;Q|k)$ and $D_\alpha(Q;R|k)$ may be greater, equal or smaller than $D_\alpha(P;R|k)$ for any three probability distributions P, Q and R]. Nonetheless, this measure can be used as a fidelity criterion as it is non-negative. The non-negativity of the conditional discrimination for $\alpha > 0$ can be shown in a simple manner using (1) and the inequality $\log x \geq 1 - \frac{1}{x}$.

$$\text{For } \alpha = 1: \quad D_\alpha(P;Q|k) \geq \sum_{j=1}^{2} p_{j|k}\left(1 - \frac{q_{j|k}}{p_{j|k}}\right) = 0. \quad (2)$$

$$\text{For } \alpha \neq 1: \quad D_\alpha(P;Q|k) \geq \frac{1}{(\alpha-1)}\left[1 - \frac{1}{\sum_{j=1}^{2} \frac{p_{j|k}^{\alpha}}{q_{j|k}^{\alpha-1}}}\right] \quad (3)$$

It is noted that $p_{1|k} = q_{1|k}$ (and hence also $p_{2|k} = q_{2|k}$) maximizes $X_\alpha = \sum_{j=1}^{2}(p_{j|k}^{\alpha}/q_{j|k}^{\alpha-1})$ for $0 < \alpha < 1$ and minimizes it for $\alpha > 1$. Thus, we obtain, $0 < X_\alpha \leq 1$ for $0 < \alpha < 1$ and $X_\alpha \geq 1$ for $\alpha > 1$. Hence the non-negativity of conditional discrimination (the Rényi-type) is proved. The conditional discrimination becomes equal to zero if and only if both P and Q become the same.

Conditional discrimination, as has been defined, is an asymmetric measure. Since we are interested in measuring the distortion of the coded speech with reference to the original one, we believe that this conditional directed divergence measure achieves our objective. However, conditional divergence measure $S_\alpha(P;Q|k)$, a 'symmetrized' version of conditional directed divergence, can be defined as $S_\alpha(P;Q|k) = D_\alpha(P;Q|k) + D_\alpha(Q;P|k)$.

An expected value of discrimination is calculated by assigning different weights to different channels. For any channel, 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 $Q^n = \prod_{l=1}^{n} Q_l$). Under this assumption, we can show that $D_\alpha(P^n;Q^n|k^N) = \sum_{l=1}^{n} D_\alpha(P_l;Q_l|k^N)$.

$$\begin{aligned} & D(P^n;Q^n|k^N) \\ &= \frac{1}{\sum_{k=1}^{N} w_k} \sum_{k=1}^{N} w_k \sum_{j_1=1}^{2}\sum_{j_2=1}^{2}\cdots\sum_{j_n=1}^{2}\left(\prod_{l=1}^{n} p_{j_l|k}\right)\left[\sum_{l=1}^{n}\log\frac{p_{j_l|k}}{q_{j_l|k}}\right] \\ &= \sum_{l=1}^{n}\left[\frac{1}{\sum_{k=1}^{N} w_k}\sum_{k=1}^{N}\left\{w_k \sum_{j_l=1}^{2} p_{j_l|k}\log\left(\frac{p_{j_l|k}}{q_{j_l|k}}\right)\right\}\right] \quad (4) \end{aligned}$$

$$\begin{aligned} & D_\alpha(P^n;Q^n|k^N) \\ &= \frac{1}{\sum_{k=1}^{N} w_k}\sum_{k=1}^{N}\left[\frac{w_k}{(\alpha-1)}\log\left\{\sum_{j_1=1}^{2}\cdots\sum_{j_n=1}^{2}\left(\frac{\prod_{l=1}^{n} p_{j_l|k}^{\alpha}}{\prod_{l=1}^{n} q_{j_l|k}^{\alpha-1}}\right)\right\}\right] \\ &= \sum_{l=1}^{n}\left[\frac{1}{\sum_{k=1}^{N} w_k}\sum_{k=1}^{N}\left\{\frac{w_k}{(\alpha-1)}\log\left(\sum_{j_l=1}^{2}\frac{p_{j_l|k}^{\alpha}}{q_{j_l|k}^{\alpha-1}}\right)\right\}\right] \quad (5) \end{aligned}$$

Several generalizations have been attempted for the directed divergence measure. The most promising one seems to be that of Csiszár [9] being $D_{gen}(P;Q) = \sum_{j=1}^{J} q_j f\left(\frac{p_j}{q_j}\right)$ where f is any convex function. This specializes to the directed divergence if $f(x) = x\log x$ and to the variational distance if $f(x) = |x-1|$. To avoid any unboundedness in the measure, we impose a condition that the probability of firing or non-firing for coded signal is not a complete certainty or uncertainty (a very small probability is associated).

5 Test Procedure, Results and Remarks

Since our primary goal is to 'closely' match the results of objective measure with the opinion of most of the human listeners, we conduct subjective tests and compare the objective test results with them. For descriptions and results of other distortion measures existing in the literature, see [10].

5.1 Test Procedure

Eight speech sentences, of 1-2 sec durations and spoken by male as well as female, were used for the test. An informal subjective test was administered in which the human listeners ranked five different coded signals obtained by passing a speech utterance through five different types of speech coders of rates ranging from 4.8 kbps (CELP) to 32 kbps (ADPCM). 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 four variations of the proposed fidelity criterion. These four measures were: the cochlear variational distance and the directed divergence measure with α = 1, 1.5, 2.

5.2 Test Results

A comparison of the subjective evaluation and the objective measure leads us to the following conclusions.

A. Performance of objective measures: Since there are fewer neurons attached to the low and high frequency stages and the brain perceives sounds based on an ensemble of neural information, we presume that the stages of higher and lower indices are given relatively less importance compared to the remaining stages and provide weighting factors to different stages appropriately. In Fig. 4 and Fig. 5, time-domain signals and spectrograms of an original and three coded versions of a typical speech sentence, say, '*Oak is strong and also gives shade*', are shown. It is emphasized that the lower the amount of new information (cross-entropy), better is the signal quality of the coded speech with reference to the original one. Table 1 provides average distortion values for each time-sample of the aforesaid speech utterance. All the cochlear directed divergence measures (with α =1, 1.5, 2) were found to be consistent to the subjective evaluation result in which the listeners ranked 'oakf8f' as a little (but clearly) better than 'oakf8k' whereas 'oakf8b' was evaluated as bad. However, the variational

distance measure was in contradiction with the subjective rankings when two coded signals were very close in their quality. We also note that the SNR measurement shows contradiction (incorrectly shows 'oakf8k' to be the best) whereas the cochlear directed divergence shows agreement with subjective evaluation results.

B. Effect of different entropies: The value of α in the Rényi-Shannon entropy-based directed divergence measure 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 may be useful to consider α value more than one as it increases the dynamic range of the measure values.

Measure Type	*oakf8f*	*oakf8k*	*oakf8b*
Subjective Ranking	Best	Good	Poor
Variational Distance	8.775	8.845	11.455
Dir. Div. ($\alpha = 1$)	2.720	2.755	4.270
Dir. Div. ($\alpha = 1.5$)	4.490	4.540	6.915
Dir. Div. ($\alpha = 2$)	6.750	6.810	10.165
SNR (w/o scaling) [dB]	8.724	9.178	-2.597
SNR (with scaling) [dB]	8.979	9.334	0.009

Table 1: Typical measure values for three coded signals with reference to the original sentence, '*Oak is strong and also gives shade*'.

C. Effect of gain changes: The AGC stages of the cochlear model provide signal compression and thus the output signal level varies of order two when the input signal varies over order twelve or higher. In addition to this non-linearity, the directed divergence measure also gives a non-linear effect. The directed divergence measure for one stage at a particular time (with $\alpha = 1$) has been presented in Fig. 6 where we observe that in the neighborhood of X=Y region, it is relatively small compared to those in the other regions. We speculate that a linear measurement relationship in the variational distance is responsible for its relatively poorer performance when two coded signals are very close in their perceptual quality.

D. Speech coder evaluation: By considering the neural pathway to be a noisy channel, the subjective evaluation of speech coders has been treated as a hypothesis testing problem. 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. Let γ_n^* be the smallest probability that 'C' is determined to be samples of 'A' when it is actually samples from 'B'. This probability is smallest over all 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, γ_n^*, for all β in (0,1) and with $\alpha = 1$, can be given as [11]: $\gamma_n^* \sim exp\left[-\sum_{l=1}^{n} D(P_l;Q_l|k^N)\right]$. By carrying out a subjective evaluation test with a large number of listeners and then considering their opinions (whether 'A' or 'B') about 'C', the parameter γ_n^* was estimated. It was verified that for achieving the same probability-of-error, more samples had to be considered when $D(P_l;Q_l|k^N)$ was relatively 'typically' small.

6 Summary and Conclusions

The formulation of any distortion measure requires resolution of two important issues: (1) defining a suitable domain where the signal parameters should be compared and (2) comparing them in a meaningful sense. As far as the first part is concerned, we have argued that it is not sufficient to compare an original speech with its coded version only in the time or in the frequency domain. It is important to consider all the major perceptual events and represent the speech signals onto a joint time-place domain. For this purpose, we have used a cochlear model which incorporates the basic features of the hearing process. In the second part, these firing/non-firing probabilities were compared to determine the neural channel cross-entropy associated with the coded speech signal with reference to the original one. The cochlear discrimination measure, by conforming to subjective evaluation results, has shown promise for its use in the evaluation and the design of speech coders.

Acknowledgement: The authors would like to thank M. Slaney for providing the MacEar program.

References

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

[2] S. Greenberg, "The ear as a speech analyzer," *Jour. of Phonetics*, vol. 16, pp. 139–149, Jan. 1988.

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

[4] M. Slaney, "Lyon's cochlear model," Tech. Rep. 13, Apple Computers, 1988.

[5] J. Flanagan, *Speech Analysis, Synthesis and Perception.* Springer-Verlag, NY, 1972.

[6] J. B. Allen, "Cochlear modeling," *IEEE ASSP Mag.*, pp. 3–29, Jan. 1985.

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

[8] A. Rényi, *Probability Theory.* North-Holland, 1970.

[9] J. Aczél, "Some recent results on characterizations of measures of information related to coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 592–595, Sep. 1978.

[10] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality.* Prentice Hall, 1988.

[11] R. E. Blahut, "Hypothesis testing and information theory," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 405–417, Jul. 1974.

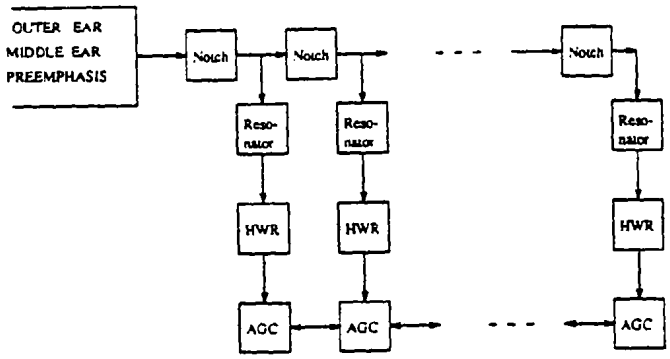

Fig. 1: Lyon's Cochlear Model

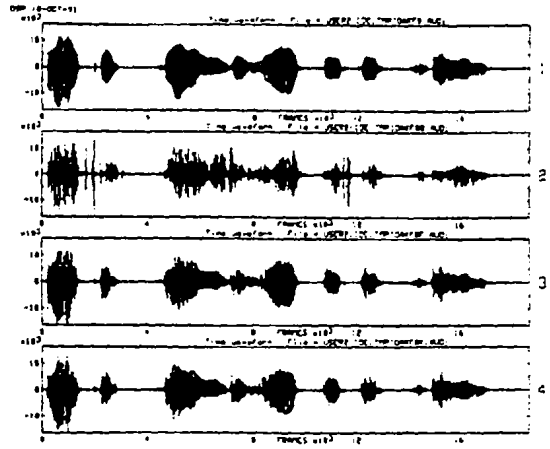

Fig. 4: Time Waveforms of an Original Speech and Its Three Coded Versions

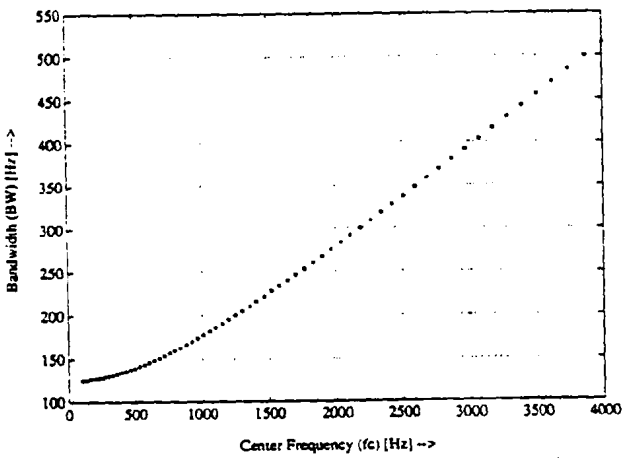

Fig. 2: Bandwidths vs. Center Frequencies for 64 Stages

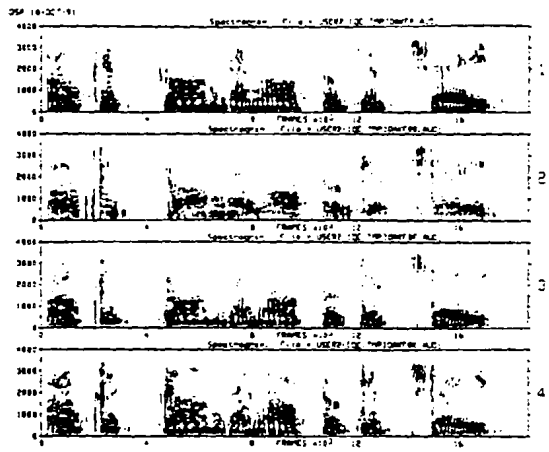

Fig. 5: Spectrograms of an Original Speech and Its Three Coded Versions

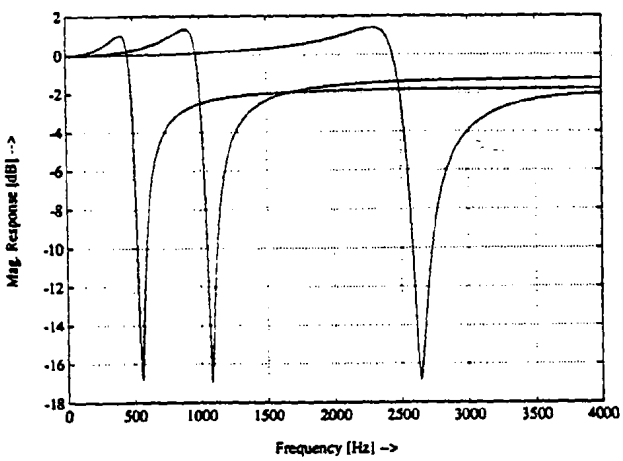

Fig. 3: Magnitude Response Plots of Stage Filters with f_c=499, 1013, 2509 Hz

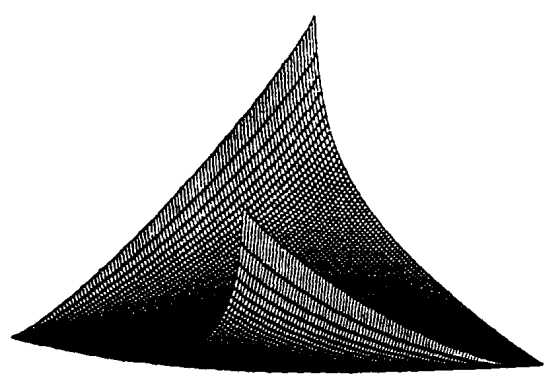

Fig. 6: Directed Divergence with α=1 (X and Y are firing probabilities for original and coded speech)