Auditory Distortion Measures for Coded Speech Quality Evaluation[†]

Aloknath De
Dept. of Electrical Engg., McGill University, 3480 University Street, Montréal, PQ, Canada—H3A 2A7.

1 Introduction

Distortion measure plays an important role in the quality evaluation of coded speech synthesized by a medium or low bit-rate coder. The quantification of distortion involves mapping the signal onto an appropriate domain and formulating a suitable comparison in that domain. In our work, both original speech and its coded version are transformed from the time-domain to a perceptual-domain (PD) using an auditory (cochlear) model. This PD representation provides information pertaining to the probability-of-firing in the neural channels at different clock times. This article proposes two distinct approaches to process these information and measure the degree of distortion in coded speech. The remainder of the article is organized as follows. Section 2 describes the Lyon's cochlear model. Sections 3 and 4 introduce the idea of cochlear discrimination information and hidden Markovian measures; and also study their use in coded speech quality evaluation. Section 5 proposes their use in some applications of speech coder analysis.

2 Lyon's Cochlear Model

Time-domain speech is transformed onto a PD where the time-place components become the fundamental bases of analysis. This conversion is performed here using Lyon's cochlear model [1] as shown in Fig. 1. The main features of the model are outlined below (for details, please refer to [2, 3]).

A first-order high-pass filter is designed to simulate the outer and the middle ear effects. The physical structure of the inner ear (cochlea) is modeled by discrete-place ear-filter stages. The basilar membrane (BM) in cochlea is stiff and thin at the basal end (where the sound enters), but compliant and massive at the apical end. Accordingly, each place along the BM resonates most strongly with a pressure wave of a characteristic frequency associated with it. By combining notch filters and resonators, sixty-four ear-filter stages are designed where the band-pass filters have an almost constant Q-factor implying a fixed ratio of the center frequency to the bandwidth for all of them.

Observations of the BM motion indicate that the inner hair cells act as half-wave rectifiers (HWRs) whereas the outer hair cells provide a gain control effect (i.e., amplification or compression). The most important adaptation mechanism in sensory systems is the lateral inhibition by which the sensory neurons reduce their own gain as well as the gain of the others nearby. To emulate this effect, Lyon proposed coupled automatic gain control (AGC) stages. The auditory neurons attached to the hair cells 'fire' (i.e., generate all-or-none electrical spikes) depending on the strength of the gain-controlled signals. These neural firing events are communicated from the auditory system to the brain through neural fibers (termed hereafter as the 'neural channels'). In essence, the normalized (w.r.t. the maximum possible output value) cochlear model output provides the probability-of-firing information (the PD representation) in the sixty-four typical neural channels at each clock time.

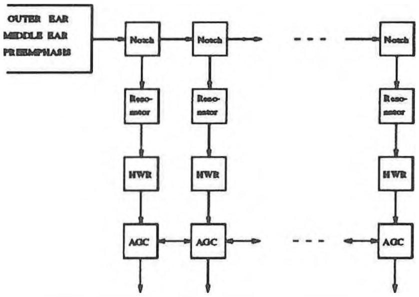

Fig. 1 Lyon's cochlear model

3 Cochlear Discrimination Information (CDI)

With each of the neural channels, is associated a neural converter which generates impulses based on the probability-of-firing information. These neural converters may equivalently be conceived as a discrete information source with an alphabet of size two, i.e., firing and non-firing. Due to the lack of our knowledge about the exact neural conversion process, the firing/non-firing probabilistic information derived from an original and a coded signal are compared to quantify the degree of distortion. Discrimination information which has emerged as a powerful tool [4] for measuring the 'closeness' of two probability distribution functions is applied here for defining the CDI measure.

This measure evaluates 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 signal is known or vice versa [5]. Different variations of this cross-entropic CDI measure, based on the Rényi-Shannon and Havrda-Charvat entropies, are investigated in [3, 5]. These measures are used for speech coder evaluation; it is found that the lower the amount of additional information, better is the signal quality of the coded speech w.r.t. the original one. The effects of different entropies, gain changes, sample delays etc. are also studied in [3]. Finally, a rate-distortion analysis is performed using the Blahut algorithm. State-of-the-art speech coders with rates ranging from 4.8 kbps to 32 kbps are studied from the viewpoint of their performances (as assessed by the CDI measure) with respect to the rate-distortion limits [6]. This analysis has indicated that there is ample scope for improving coder architectures and associated coding algorithms for a specific bit-rate transmission.

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

4 Cochlear Hidden Markovian (CHM) Measure

In this article, we propose another measure methodology, namely the CHM measure. Here, we attempt to capture the basics of high-level processing in the brain with simple hidden Markovian models (HMMs). We characterize the firing events by HMMs where the order of occurrence of observations and correlations among adjacent observations are modeled suitably. A two-state (one each for firing and non-firing events) fully-connected HMM is associated with each of the neural channels for a fixed-duration, short time segment (in our work, 480 samples). We consider the PD observation process for the entire original speech as a concatenation of many such small segments.

Now, let us consider the observations for any one of the neural channels for a specific time segment. An HMM for any such observation set is defined [7] by describing the complete parameter set of the model given as $\lambda = (\pi, A, B)$, where π is the state probability vector, A is the state transition probability matrix and B is a set of two continuous mixture probability density functions (pdfs), each with a few mixtures. Each component of these mixture pdfs is assumed to be a beta density function with values between 0 and 1.

For computing coder distortions, at first, all the HMMs are 'train'ed (i.e., various parameters of the HMMs are estimated) with the pertinent observation vectors corresponding to the original speech segment. There is actually no optimal way of estimating the model parameters from any finite-length observation sequence. The Baum-Welch reestimation algorithm is used to derive the HMM parameters iteratively starting from an initial estimate. The model $\lambda_n^{(o)}$ corresponding to $O_n^{(o)}$, the n-th channel observation sequence for original speech, is chosen by maximizing $P(O_n^{(o)}|\lambda_n^{(o)})$. This algorithm is quite powerful as it ensures a monotonic increase in the likelihood with the successive iterations of the algorithm.

Now, let the n-th channel PD observations of the v-th coded speech be represented by $O_n^{(v)}$. Next, we compute $P(O_n^{(v)}|\lambda_n^{o})$ for all the sixty-four channels (i.e., 'match'ed against the derived HMMs). For simplicity, we assume that the information conveyed through all the neural channels are independent and hence the likelihood probability scores are multiplied to provide a distortion measure (to be precise, a similarity measure). Experimental results have shown that the coded speech signals could be ranked (same as the subjective ordering) by this measure with fair accuracy. The effects of the iteration numbers involved in the reestimation algorithm, the initial estimates of the HMM parameters, the number of mixtures in the pdf etc. are addressed in [8].

5 Applications in Coder Analysis

For a low bit-rate speech coder, a proper bit allocation among the coder components (e.g., pitch or formant filter) is vital to achieve a good perceptual quality in the coded speech. In [6], we have described an analysis procedure for determining the pitch frequency by exploring the output space of the cochlear model. The CDI measure form is applied in comparing the outputs for each of the sixty-four neural channels with its delayed version (delayed by τ samples, τ up to 160 samples). Subsequently, a one-dimensional cross-entropogram is derived which shows the first significant dip at a τ value corresponding to the perceptual pitch period.

Next, we consider a code-excited linear predictive (CELP) speech coder which uses three-way split vector quantization for a 16-th order linear predictive coder filter parameters and allows fractional pitch lag values in the pitch predictor. In the present-day analysis-by-synthesis CELP coders, the filter parameters and the codebook entries are selected by minimizing a noise-weighted mean-square error criterion. As a second application, we have investigated the relative superiority of different noise-weighting schemes (e.g., simple noise weighting, codebook shaping filter, enhanced noise weighting).

6 Concluding Remarks

A further refinement of the cochlear model and an extensive formal testing with a wide range of linear and nonlinear coder distortions would definitely help improving the measure. Nonetheless, we emphasize that the present framework of comparing the firing/non-firing probabilities could still be maintained. Although we have not attempted to use our measure formulation in a closed-loop fashion in any speech coder, it may very well be possible to use it for 'populating' a codebook in the training phase and/or for 'selecting' an appropriate codebook entry in the transmission phase.

Acknowledgement

The author would like to express his sincere gratitude to Prof. Peter Kabal for his technical guidance and financial assistance.

References

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

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

[3] A. De and P. Kabal, "Auditory distortion measure for coded speech—discrimination information approach," *Speech Commun. (being revised for publication).*

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

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

[6] A. De and P. Kabal, "Rate distortion function for speech coding based on perceptual distortion measure," in *Proc. of IEEE Globecom'92*, pp. 452–456, Dec. 1992.

[7] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, Feb. 1989.

[8] A. De and P. Kabal, "Auditory distortion measure for coded speech—hidden Markovian approach," *Speech Commun. (being prepared for submission).*