ELSEVIER

Speech Communication 17 (1995) 39–57

SPEECH COMMUNICATION

Auditory distortion measure for speech coder evaluation – Hidden Markovian approach ☆,☆☆

Aloknath De *,[a,b], Peter Kabal [a,b]

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

Received 27 September 1994; revised 6 April 1995

Abstract

This article introduces a methodology for quantifying the distortion introduced by a low or medium bit-rate speech coder. Since the perceptual acuity of a human being determines the precision with which speech data must be processed, the speech signal is transformed onto a perceptual-domain (PD). This is done using Lyon's cochlear (auditory) model whose output provides the probability-of-firing information in the neural channels at different clock times. In our present approach, we use a hidden Markov model to describe the basic firing/non-firing process operative in the auditory pathway. We consider a two-state fully-connected model of order one for each neural channel; the two states of the model correspond to the firing and non-firing events. Assuming that the models are stationary over a fixed duration, the model parameters are determined from the PD observations corresponding to the original signal. Then, the PD representations of the coded speech are passed through the respective models and the corresponding likelihood probabilities are calculated. These probability scores are used to define a *cochlear hidden Markovian* (CHM) distortion measure. This methodology considers the temporal ordering in the neural firing patterns. The CHM measure which utilizes the contextual information present in the firing pattern shows robustness against coder delays.

Zusammenfassung

In diesem Artikel wird eine Methodologie zur Quantifizierung der Signalverzerrung vorgestellt, die durch einem mit geringer oder mittlerer Bitrate arbeitenden Sprachkoder hervorgerufen wird. Da die menschliche Wahrnehmungsschärfe die Präzision bestimmt, mit der Sprachdaten verarbeitet werden müssen, wurde das Sprachsignal unter Verwendung des Gehörschneckenmodells von Lyons in den Perzeptionsbereich (PD) übertragen. Dieses Modell liefert die Informationen zu der Abfeuerwahrscheinlichkeit in den Nervenbahnen zu verschiedenen

☆ This work was supported by a grant from the Canadian Institute for Telecommunications Research (CITR) under the NCE program of the Government of Canada.

☆☆ This work was presented in part at the Annual CITR Conference, Montréal, QC, August 1993 and also in part at the Canadian Acoustical Association Symposium, Toronto, ON, October 1993.

* Corresponding author. Contact address: Bell-Northern Research, 16 Place du Commerce, Verdun, Québec, Canada H3E 1H6.

SSDI 0167-6393(95)00016-X

Zeitpunkten. In dem hiesigen Ansatz wird ein Hidden Markov Modell benutzt, um den grundlegenden Prozess von Abfeuern/Nicht-Abfeuern zu beschreiben, der sich im Gehörgang abspielt. Wir gehen für jede Nervenbahn von einem Modell erster Ordnung mit zwei vollverbundenen Zuständen aus, die den Ereignissen von Abfeuern und Nicht-Abfeuern entsprechen. Davon ausgehend, daß die Modelle über einen bestimmten Zeitraum stationär sind, werden die Modellparameter durch die PD-Beobachtungen an dem Originalsignal determiniert. Dann werden die PD Repräsentationen der kodierten Sprache dem jeweiligen Modell zugeführt, und die entsprechenden Übereinstimmungswahrscheinlichkeiten berechnet. Diese Wahrscheinlichkeitsquoten werden dazu benutzt, eine Verzerrungsmessung mit dem Hidden Markov Modell der Gehörschnecke (CHM) zu definieren. Diese Methodologie berücksichtigt die zeitliche Anordnung der Abfeuermuster der Nerven. Die CHM-Messungen, die die kontextuelle Information aus den Abfeuermustern benutzen, zeigen sich robust gegen die Verzögerungen des Koders.

Résumé

Cet article présente une méthodologie pour quantifier la distorsion apportée par un codeur de parole à bas ou moyen débit. Puisque c'est l'acuité perceptive de l'être humain qui fixe la précision avec laquelle on doit traiter le signal de parole, celui-ci est transformé en une représentation perceptive; on utilise pour cela le modèle cochléaire (auditif) de Lyon, dont les sorties représentent la probabilité d'excitation des fibres nerveuses à un instant donné. Nous utilisons dans ce travail un modèle de Markov caché pour modéliser le processus élémentaire d'excitation/non excitation opératoire dans le système auditif. Un modèle d'ordre un, à deux états et complètement connecté est associé à chaque *canal neuronal*; les deux états du modèle représentent les événements d'excitation et de non-excitation. En supposant les modèles stationnaires sur une durée fixe, leurs paramètres sont calculés à partir des représentations perceptives du signal original. Ensuite, les représentations perceptives de la parole codée passent à travers les modèles correspondants et les probabilités associées sont calculées. Ces scores permettent de définir une mesure de distorsion à partir d'une "cochlée markovienne caché" (CHM). Cette méthode prend en compte la succession temporelle des profils de l'excitation neuronale. La mesure CHM, qui prend en compte l'information contextuelle présente dans le profil d'excitation, est robuste vis à vis du délai de codage.

Keywords: Auditory (cochlear) model; Neural firing mechanism; Hidden Markov model; Coded speech quality; Distortion measure

1. Introduction

Distortion measures play a vital role in evaluating the speech quality of coded signal synthesized by a medium or low bit-rate speech coder. Since a human being is the final information processor in speech communication, it is important to consider the major perceptual factors while devising the measure. In our work, both the original speech and its coded version are transformed from the time-domain to a *perceptual-domain* (PD). This is done using Lyon's cochlear model which considers the temporal as well as the spectral masking effects. The PD representation at the cochlear model output provides the probability-of-firing information of the neural channels at different clock times.

In (De, 1993a; De and Kabal, 1994), we have introduced and studied a *cochlear discrimination information* (CDI) measure which exploits the perceptual events at the auditory periphery. This measure compares the neural-firing information corresponding to an original speech and its coded version in a cross-entropic sense. In essence, the CDI measure computes 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. We have investigated several variants of the CDI measure applying the Rényi–Shannon entropy and symmetric/asymmetric divergence measures.

This article proposes another perceptual distortion measure, namely the *cochlear hidden Markovian* (CHM) measure (De, 1993a, 1993b). The basic firing/non-firing process operative in an auditory pathway is simulated by a hidden Markov model (HMM). Since the conversion process of the PD representation to the firing/non-firing is not exactly known, 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. All the HMMs are trained (i.e., various parameters of the HMMs are derived) with the pertinent PD observation vectors corresponding to the original speech segment. Then, for the same segment, the PD observations from the coded speech are matched (i.e., the likelihood probabilities are computed) against the derived HMMs.

The CDI measure compares the PD observations directly, whereas the CHM measure is a parametric nonlinear model-based measure. Although the CDI measure has conformed strongly to the informal subjective test results in terms of ranking coded speech signals, it has been found to be not very robust against time misalignments between the original and the coded signal. Therefore, before applying the CDI measure, it may be important to estimate and remove time-delay between the original and the coded speech signals. The CHM measure which has considered the temporal ordering in the firing pattern has shown a robustness against the coder delays. An explicit removal of the coder delays is not necessary for small delays.

This paper is organized as follows. Section 2 describes the auditory representation of speech signals and characterizes the hidden Markovian signal model. Section 3 provides some relevant background materials. Section 4 introduces a method to compute distortion for speech coders and also suggests briefly some other alternative approaches. Section 5 considers some practical aspects related to the evaluation of speech coders by the CHM measure. Section 6 provides the experimental results for speech coder evaluation.

2. Hidden Markovian neural model

This section describes how a speech signal is mapped onto a perceptual-domain; and also characterizes the neural firing process by a hidden Markov model.

2.1. Auditory representation of speech signal

In the proposed distortion measure, several details of the auditory processing are considered. We have used Lyon's cochlear model, as shown in Fig. 1, for representing the speech signal onto a PD. The outer-and-middle ear filter is modeled by a simple high-pass filter. The bandpass characteristics of the basilar membrane in the inner ear (cochlea) are simulated by sixty-four combinations of second-order notch filters and resonators. The activities of the inner hair cells are mimicked by the half-wave rectification process, while those of the outer hair cells are imitated by the automatic gain control stages. 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.

Unlike many other models, Lyon's auditory model considers the temporal as well as the spectral masking effects. The normalized cochlear model output provides the probability-of-firing information in the neural channels at different clock times. Here, the normalization is done with respect to the maximum possible output value of the four cascaded AGC blocks and the clock time is chosen to be the

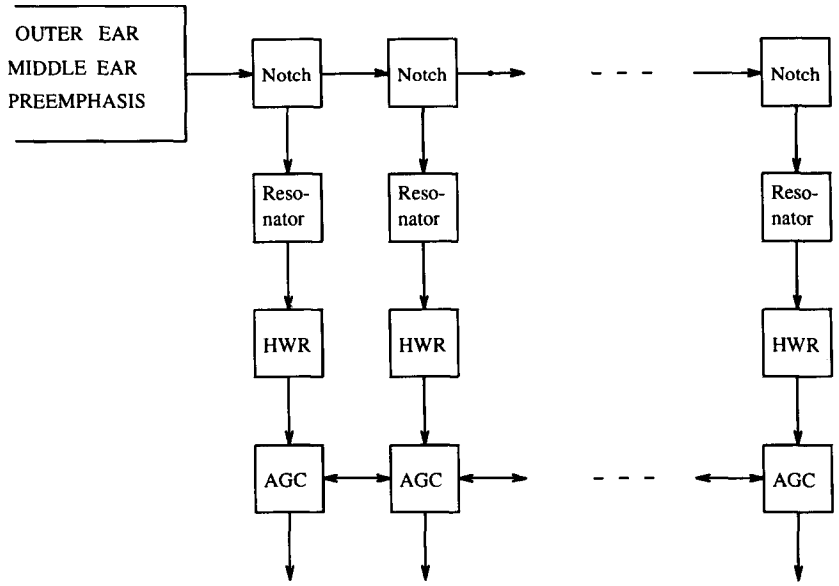

Fig. 1. Block diagram of Lyon's cochlear model ('HWR' stands for the half-wave rectifier and 'AGC' stands for the automatic gain controller).

same as the sampling time, i.e., 125 μs. A detailed description of the model is given in (De and Kabal, 1994; Slaney, 1988).

2.2. Characterization of hidden Markov model

The cochlear model output is a sequence of K-dimensional vectors (in our work, $K = 64$ corresponding to sixty-four *characteristic* neural channels) with one vector for each clock time t. The elements in each of the K-dimensional observation vectors represent the probability-of-firing information. Based on this PD representation of a speech signal, what are transmitted through neural channels to the brain are series of all-or-none electrical spikes (firings). However, the exact conversion process of the PD representation to the firing/non-firing representation is not yet known. We attempt here to capture the underlying firing/non-firing event in each channel with discrete-time series analysis.

One such analysis technique involves using a hidden Markov model for modeling the observation sequence. The time-varying observation process is considered as a concatenation of many short-time segments of a fixed duration. However, it is expected that the properties of the process change neither synchronously with every analysis duration nor abruptly from each unit to the next one. The development of an efficient optimization technique (Baum and Petrie, 1966) to estimate the model parameters so as to match the observed signal patterns has culminated in the theory of HMM-based signal representation. The success of this hidden Markov modeling technique has been proven by its application in ecology (e.g., (Baum and Egom, 1967)), text analysis (e.g., (Cave and Neuwirth, 1980)), coding theory (e.g., (Chang and Hancock, 1966)) and speech recognition (e.g., (Jelinek, 1976)).

An HMM is a doubly embedded stochastic model with an underlying process that is not directly observable (it is hidden), but can be observed through another set of stochastic processes that produce the sequence of observations. In other words, the states of an HMM are hidden and the observation is a probabilistic function of the states. The order of occurrence of observations and the correlations among

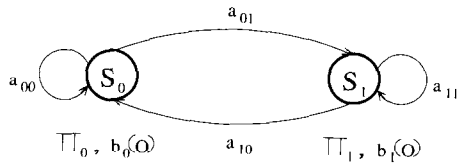

Fig. 2. A two-state fully-connected hidden Markov model (S_0 and S_1 denote the non-firing and firing states, π_0 and π_1 are the initial state probabilities, a_{ij} gives the state transition probability from a state S_i to a state S_j, $b_0(O)$ and $b_1(O)$ are the observation probability density functions for the state S_0 and S_1, respectively).

adjacent observations are suitably modeled by stochastic dependencies among the hidden states of an HMM. In the following, we characterize an HMM for our problem by selecting the model type, the number of hidden states and all the parameters associated with the model.

We consider K numbers of independent two-state ($N = 2$) fully-connected models, as shown in Fig. 2, where either state is reachable from the other one. Although in many applications, the states do not have a physical meaning; here a state S_0 corresponds to a non-firing event whereas a state S_1 corresponds to a firing event. The initial state distribution (i.e., at $t = 1$) is given as $\pi = \{\pi_i \mid i \in \mathcal{N}\}$ with

$$\pi_i = P[q_1 = S_i] \quad \text{for } i \in \mathcal{N} \text{ and } \sum_{i \in \mathcal{N}} \pi_i = 1, \tag{1}$$

where $\mathcal{N} \equiv \{0, 1\}$ and a state reached at any clock time t is denoted by q_t.

The HMM considered is of order one and hence the transition from one state to the next one occurs according to a transition probability distribution which depends only on the previous state. If we define an integer set $\mathcal{T} \equiv \{1, 2, \ldots, T-1\}$ then the state transition probability distribution $A = \{a_{ij} \mid i, j \in \mathcal{N}\}$ is given by

$$a_{ij} = P\left[q_{t+1} = S_j \mid q_t = S_i\right] \quad \text{for } i, j \in \mathcal{N} \text{ and } t \in \mathcal{T}, \tag{2}$$

where every a_{ij} coefficient (i.e., a_{00}, a_{01}, a_{10}, a_{11}) is positive, and $\Sigma_{j \in \mathcal{N}} a_{ij} = 1$ for $i \in \mathcal{N}$.

Now, we consider any one of the neural channels for which the observation is represented by $\boldsymbol{O} = O_1 O_2 \cdots O_T$. To avoid significant degradation due to any quantization process, we treat the PD representation to be continuous-valued and accordingly consider an HMM with continuous probability density functions. However, the use of a continuous pdf requires some restrictions on its form so as to facilitate reestimation of the pdf parameters (e.g., mean, variance) in a consistent manner. The pdf for each of the two states is maintained fixed regardless of when and how the state is reached. The most general representation of the pdf, for which a reestimation procedure exists (Baum and Petrie, 1966), is used here. Each state S_j is characterized by a continuous mixture pdf $b_j(x)$ of the form

$$b_j(x) = \sum_{m \in \mathcal{M}_L} c_{jm} b_{jm}(x) \quad \text{for } j \in \mathcal{N}, \tag{3}$$

where $\mathcal{M}_L \equiv \{1, 2, \ldots, L\}$ with L the number of components in the mixture and $b_{jm}(\cdot)$ is any log-concave (Baum and Petrie, 1966) or elliptically symmetric (Liporace, 1984) density. The rationale behind choosing a mixture pdf and selecting the component pdf $b_{jm}(\cdot)$ to be log-concave or elliptically symmetric is discussed later. In our present study, $b_{jm}(\cdot)$ is assumed to be a beta density function and can be written as

$$b_{jm}(x) = \frac{\Gamma(d_{jm} + f_{jm} + 2)}{\Gamma(d_{jm} + 1)\Gamma(f_{jm} + 1)} x^{d_{jm}} (1 - x)^{f_{jm}} \quad \text{for } d_{jm}, f_{jm} > 0,\ j \in \mathcal{N},\ m \in \mathcal{M}_L, \tag{4}$$

where d_{jm} and f_{jm} are the parameters associated with the density function. The beta pdf of (4) is suitable as the observations are continuous-valued between 0 and 1. Appendix A shows that the beta density function satisfies the log-concavity condition.

The observations probability density function B is denoted as $B = \{b_j(x) \mid j \in \mathcal{N}\}$, where $b_j(x)\mathrm{d}x$ is the probability of observing a value O_t in state S_j at clock time t. A coefficient c_{jm} is the m-th component mixture gain in state S_j and the set $\{c_{jm} \mid j \in \mathcal{N},\ m \in \mathcal{M}_L\}$ satisfies the stochastic constraint

$$\sum_{m \in \mathcal{M}_L} c_{jm} = 1 \quad \text{for } j \in \mathcal{N} \text{ with } c_{jm} > 0 \text{ for } j \in \mathcal{N} \text{ and } m \in \mathcal{M}_L, \tag{5}$$

so that

$$\int_{-\infty}^{\infty} b_j(x)\ \mathrm{d}x = 1, \quad j \in \mathcal{N}. \tag{6}$$

3. Preliminaries

In Section 2, an HMM has been defined by describing the complete parameter set of the model. The model is represented as $\boldsymbol{\lambda} = (\pi, A, B)$, where π is the state probability vector, A is the state transition probability matrix and B is a set of two ($N = 2$) continuous mixture pdfs, each with L mixtures. In this section, we provide some preliminaries required for computing the degree of distortion (similarity) of a coded speech with reference to its original version. A forward and a backward likelihood variables and an auxiliary function are defined below.

3.1. Forward and backward likelihood variables

Let us extend the integer set $\mathcal{T}$ to $\mathcal{T}^+$ as $\mathcal{T}^+ \equiv \mathcal{T} + \{T\}$. Following Baum and Petrie (1966), a *forward likelihood variable* $\alpha_t(i)$ is then defined as

$$\alpha_t(i) = P\big(O_1 O_2 \cdots O_t, q_t = S_i \mid \boldsymbol{\lambda}\big) \quad \text{for } i \in \mathcal{N} \text{ and } t \in \mathcal{T}^+, \tag{7}$$

which gives the probability of observing the *partial* sequence $O_1 O_2 \cdots O_t$ (until time t) and reaching the state S_i at clock time t *given* an HMM $\boldsymbol{\lambda}$. Likewise, a *backward likelihood variable* $\beta_t(j)$ is defined as

$$\beta_t(j) = P\big(O_{t+1} O_{t+2} \cdots O_T \mid q_t = S_j, \boldsymbol{\lambda}\big) \quad \text{for } j \in \mathcal{N} \text{ and } t \in \mathcal{T}, \tag{8}$$

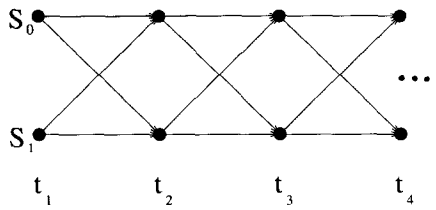

Fig. 3. A two-state trellis diagram (S_0 and S_1 denote the non-firing and firing states).

which gives the probability of observing the partial sequence $O_{t+1}O_{t+2}\cdots O_T$ (from $t+1$ to the end) *given* state S_j at time t and a model $\boldsymbol{\lambda}$.

The forward likelihood variable $\alpha_t(i)$ is initialized as the joint probability of being in state S_i at $t=1$ and an initial observation O_1, i.e.,

$$\alpha_1(i) = \pi_i b_i(O_1), \quad i \in \mathcal{N}. \tag{9}$$

With the help of the trellis diagram shown in Fig. 3, an iterative procedure is followed to compute the other forward likelihood variables from the initial one. Since $\alpha_t(i)$ is the probability of the joint event that $O_1O_2\cdots O_t$ are observed and the state S_i is reached at clock time t, the product $\alpha_t(i)a_{ij}$ becomes the probability of the joint event that $O_1O_2\cdots O_t$ are observed and the state S_j is reached at $t+1$ through the state S_i at t. Summation of this product over the possible two states S_i (for $i \in \mathcal{N}$) at time t yields the probability of reaching state S_j at $t+1$ with the corresponding partial observation sequence up to time t. Multiplication of the summed quantity by $b_j(O_{t+1})$, the probability of observing O_{t+1} at state S_j results in the forward likelihood variable $\alpha_{t+1}(j)$ for time $t+1$. This evaluation procedure can be expressed by the following recurrence equation:

$$\alpha_{t+1}(j) = \left[\sum_{i\in\mathcal{N}} \alpha_t(i)a_{ij}\right] b_j(O_{t+1}), \quad t \in \mathcal{T},\ j \in \mathcal{N}. \tag{10}$$

In a similar manner, let us now consider the backward variable $\beta_t(i)$. An initialization process arbitrary defines

$$\beta_T(j) = 1, \quad j \in \mathcal{N}. \tag{11}$$

Then, $\beta_t(i)$ is calculated recursively as follows:

$$\beta_t(i) = \sum_{j\in\mathcal{N}} a_{ij} b_j(O_{t+1})\beta_{t+1}(j), \quad t \in \mathcal{T},\ i \in \mathcal{N}. \tag{12}$$

For a given model $\boldsymbol{\lambda}$, $\beta_t(i)$ is the probability of observing the particular partial sequence from time $t+1$ to the end when it is known that the state S_i is reached at time t. To compute this, it is evident from the trellis diagram of Fig. 3 that we need to consider both the states S_0 and S_1 at time $t+1$ accounting for the possible transitions from S_i to S_j, the observation O_{t+1} in state S_j and also the partial observation sequence $O_{t+2}O_{t+3}\cdots O_T$ (being in state S_j at time $t+1$).

3.2. Auxiliary function

In order to estimate the HMM parameters, we should maximize $P(\boldsymbol{O}\,|\,\boldsymbol{\lambda})$. However, in practice, an indirect approach is adopted and an auxiliary function related to $P(\boldsymbol{O}\,|\,\boldsymbol{\lambda})$ is maximized. Following the concept of the Kullback–Leibler statistic, an auxiliary function $F(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ of two models $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$, for a given observation vector $\boldsymbol{O}$, can be defined (Juang, 1985) as

$$F(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \sum_{\boldsymbol{Q}\in\mathcal{N}^T} \sum_{\boldsymbol{M}\in\mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}\,|\,\boldsymbol{\lambda}) \log P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}\,|\,\boldsymbol{\lambda}'), \tag{13}$$

with $\boldsymbol{Q} = q_1q_2\cdots q_T$, $\boldsymbol{M} = m_1m_2\cdots m_T$, $q_k \in \mathcal{N}$ and $m_k \in \mathcal{M}_L$ for $k \in \mathcal{T}$. In the following, we show that if $F(\boldsymbol{\lambda}, \boldsymbol{\lambda}') \geqslant F(\boldsymbol{\lambda}, \boldsymbol{\lambda})$, then $P(\boldsymbol{O}\,|\,\boldsymbol{\lambda}) \geqslant P(\boldsymbol{O}\,|\,\boldsymbol{\lambda})$. The primary advantage of this technique lies in its

ability to decouple all the parameter estimation equations.

$$
\begin{aligned}
& P(\boldsymbol{O}\mid\boldsymbol{\lambda})\log\frac{P(\boldsymbol{O}\mid\boldsymbol{\lambda}')}{P(\boldsymbol{O}\mid\boldsymbol{\lambda})} \\
&= P(\boldsymbol{O}\mid\boldsymbol{\lambda})\log\sum_{Q\in\mathcal{N}^T}\sum_{M\in\mathcal{M}_L^T}\frac{P(\boldsymbol{O},Q,M\mid\boldsymbol{\lambda}')}{P(\boldsymbol{O}\mid\boldsymbol{\lambda})} \\
&= P(\boldsymbol{O}\mid\boldsymbol{\lambda})\log\sum_{Q\in\mathcal{N}^T}\sum_{M\in\mathcal{M}_L^T}\frac{P(\boldsymbol{O},Q,M\mid\boldsymbol{\lambda})}{P(\boldsymbol{O}\mid\boldsymbol{\lambda})}\frac{P(\boldsymbol{O},Q,M\mid\boldsymbol{\lambda}')}{P(\boldsymbol{O},Q,M\mid\boldsymbol{\lambda})} \\
&\geqslant P(\boldsymbol{O}\mid\boldsymbol{\lambda})\sum_{Q\in\mathcal{N}^T}\sum_{M\in\mathcal{M}_L^T}\frac{P(\boldsymbol{O},Q,M\mid\boldsymbol{\lambda})}{P(\boldsymbol{O}\mid\boldsymbol{\lambda})}\log\frac{P(\boldsymbol{O},Q,M\mid\boldsymbol{\lambda}')}{P(\boldsymbol{O},Q,M\mid\boldsymbol{\lambda})} \\
&= [F(\boldsymbol{\lambda},\boldsymbol{\lambda}')-F(\boldsymbol{\lambda},\boldsymbol{\lambda})]\geqslant 0, \qquad (14)
\end{aligned}
$$

with strict inequality except when $P(\boldsymbol{O}, Q, M\mid\boldsymbol{\lambda}) = P(\boldsymbol{O}, Q, M\mid\boldsymbol{\lambda}')$. In the above, the fact that $\log x$ is strictly concave for $x>0$ (since $\mathrm{d}^2/\mathrm{d}x^2\ (\log x) = -x^{-2} < 0$) has been used. The first inequality is the well-known Jensen's inequality whereas the second one is true by hypothesis.

If the current model is defined as $\boldsymbol{\lambda} = (\pi, A, B)$ and a reestimated model is $\boldsymbol{\lambda}' = (\pi', A', B')$, then either the initial model $\boldsymbol{\lambda}$ defines a critical point of the likelihood function (in that case $\boldsymbol{\lambda}' = \boldsymbol{\lambda}$), or the model $\boldsymbol{\lambda}'$ is better than the model $\boldsymbol{\lambda}$ in a sense that the observation sequence $\boldsymbol{O}$ is more likely to have been generated by $\boldsymbol{\lambda}'$. A positive value of the auxiliary function implies that the newly estimated model is better than the old one. From (14), we observe that when the auxiliary function reaches a critical point, $P(\boldsymbol{O}\mid\boldsymbol{\lambda})$ also reaches its local maximum. The model parameters corresponding to this point give the best possible estimate of the HMM parameters.

4. Distortion measure methodology

An original speech segment and its coded version are passed through the cochlear model to obtain the PD representations. For each of these segments, the PD observations are sequences of 64-dimensional vectors corresponding to sixty-four characteristic neural channels. A hidden Markov model is associated with each of the channels and the parameters are estimated from the PD observation sequence produced by the original speech segment. In a sense, all the sixty-four HMMs are trained with the pertinent observation vectors corresponding to the original speech segment. Then, for the same speech segment, the observations from all the coded speech signals are matched against the derived HMMs to compute the relative coder distortions. Now we describe the exact procedures for the model parameter estimation as well as the likelihood computation.

4.1. Parameter estimation

There is no optimal way of estimating the model parameters from any finite-length observation sequence. Since the closed-form maximum likelihood is not possible, the HMM parameters are (re)estimated iteratively starting from an initial estimate. To solve this problem, Baum–Welch reestimation algorithm (Baum, 1972) is used here. An application of this algorithm is equivalent to solving a mathematical optimization problem for obtaining the maximum likelihood estimates of the HMM parameters. The scheme for estimating the HMM parameters is based on the maximization of the

probability of the observation sequence given a model. This algorithm is quite powerful as it ensures a monotonic increase in the likelihood with the successive iterations of the algorithm (Baum and Petrie, 1966).

Let us now consider the calculation of $P(\boldsymbol{O}\,|\,\boldsymbol{\lambda})$, the probability of the observation sequence $\boldsymbol{O}$ given the model $\boldsymbol{\lambda}$. Assuming the statistical independence of observations, for every given state sequence $\boldsymbol{Q} = q_1 q_2 \cdots q_T$, the probability of observing $\boldsymbol{O}$ can be written as $P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \boldsymbol{\lambda})$, where

$$P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \boldsymbol{\lambda}) = b_{q_1}(O_1) b_{q_2}(O_2) \cdots b_{q_T}(O_T). \tag{15}$$

The probability of the occurrence of such a state sequence $\boldsymbol{Q}$ is given as

$$P(\boldsymbol{Q}\,|\,\boldsymbol{\lambda}) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \cdots a_{q_{T-1} q_T}. \tag{16}$$

Using (15) and (16), $P(\boldsymbol{O}\,|\,\boldsymbol{\lambda})$ can be computed as

$$P(\boldsymbol{O}\,|\,\boldsymbol{\lambda}) = \sum_{Q \in \mathcal{N}^T} P(\boldsymbol{O}\,|\,\boldsymbol{Q}, \boldsymbol{\lambda}) P(\boldsymbol{Q}\,|\,\boldsymbol{\lambda}). \tag{17}$$

The global density function of (17) with the state density defined by (3) can be rewritten as

$$\begin{aligned} P(\boldsymbol{O}\,|\,\boldsymbol{\lambda}) &= \sum_{Q \in \mathcal{N}^T} \pi_{q_1} \prod_{t \in \mathcal{T}^+} \left[a_{q_t q_{t+1}} \left\{ \sum_{m \in \mathcal{M}_L} c_{q_t m} b_{q_t m}(O_t) \right\} \right] \\ &= \sum_{Q \in \mathcal{N}^T} \sum_{m_1 \in \mathcal{M}_L} \sum_{m_2 \in \mathcal{M}_L} \cdots \sum_{m_T \in \mathcal{M}_L} \left[\pi_{q_1} \prod_{t \in \mathcal{T}^+} a_{q_t q_{t+1}} c_{q_t m_t} b_{q_t m_t}(O_t) \right], \end{aligned} \tag{18}$$

assuming the parameter $a_{q_T q_T + 1} = 1$. The direct computation of $P(\boldsymbol{O}\,|\,\boldsymbol{\lambda})$ as given by (18) involves enumerating every possible state sequence of length T. Instead, we exploit the trellis structure and use (10) and (12) for the forward and the backward likelihood parameters. In order to describe the procedure for an iterative update of the HMM parameters, we define a set of *transition likelihood variables* $\{\xi_t(i, j)\,|\,i, j \in \mathcal{N}, t \in \mathcal{T}\}$ as

$$\xi_t(i, j) = P\big(\boldsymbol{O}, q_t = S_i, q_{t+1} = S_j\,|\,\boldsymbol{\lambda}\big), \tag{19}$$

which gives the probability of observing the particular sequence $\boldsymbol{O}$, and being in the state S_i at time t and the state S_j at time $t + 1$ given the model. From the trellis diagram of Fig. 3, it can be noted that $\xi_t(i, j)$ can be written as

$$\xi_t(i, j) = \sum_{m \in \mathcal{M}_L} \alpha_t(i) a_{ij} c_{jm} b_{jm}(O_{t+1}) \beta_{t+1}(j). \tag{20}$$

We note the following relationships among the three likelihood variables as defined in (10), (12) and (19):

1. A product of the forward and the backward likelihood variables for any clock time t is shown, using (3) and (12), equal to the sum of the transition likelihood variable $\xi_t(i, j)$ over the index j.

$$\begin{aligned} \alpha_t(i)\beta_t(i) &= \alpha_t(i) \left[\sum_{m \in \mathcal{M}_L} \sum_{j \in \mathcal{N}} a_{ij} c_{jm} b_{jm}(O_{t+1}) \beta_{t+1}(j) \right] \\ &= \sum_{j \in \mathcal{N}} \xi_t(i, j). \end{aligned} \tag{21}$$

2. Using (3), (10) and (12), it is shown that a sum of the product of the forward and the backward likelihood variables, i.e., $\alpha_t(i)\,\beta_t(i)$ over i is independent of the time index t.

$$\begin{aligned}\sum_{j\in\mathcal{N}} \alpha_{t+1}(j)\beta_{t+1}(j) &= \sum_{j\in\mathcal{N}}\left[\sum_{i\in\mathcal{N}}\sum_{m\in\mathcal{M}_L} \alpha_t(i)a_{ij}c_{jm}b_{jm}(O_{t+1})\right]\beta_{t+1}(j)\\ &= \sum_{i\in\mathcal{N}} \alpha_t(i)\left[\sum_{j\in\mathcal{N}}\sum_{m\in\mathcal{M}_L} a_{ij}c_{jm}b_{jm}(O_{t+1})\beta_{t+1}(j)\right]\\ &= \sum_{i\in\mathcal{N}} \alpha_t(i)\beta_t(i) \quad \text{for } t\in\mathcal{T}. \end{aligned} \tag{22}$$

3. Using (19) and applying (21), (22) and (11) subsequently, $P(\boldsymbol{O}\mid\boldsymbol{\lambda})$ can be written as the sum of the *terminal forward likelihood variables* $\alpha_T(i)$ over i, i.e.,

$$P(\boldsymbol{O}\mid\boldsymbol{\lambda}) = \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} \xi_t(i,j) = \sum_{i\in\mathcal{N}} \alpha_t(i)\beta_t(i) = \sum_{i\in\mathcal{N}} \alpha_T(i). \tag{23}$$

The logarithm of $P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}\mid\boldsymbol{\lambda}')$, the square bracketed term in (18), can be written as

$$\log P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}\mid\boldsymbol{\lambda}') = \log \pi'_{q_1} + \sum_{t\in\mathcal{T}^+} \log a'_{q_t q_{t+1}} + \sum_{t\in\mathcal{T}^+} \log c'_{q_t m_t} + \sum_{t\in\mathcal{T}^+} \log b'_{q_t m_t}(O_t). \tag{24}$$

It is seen that the HMM parameters π', A' and B' corresponding to the model $\boldsymbol{\lambda}'$ are segregated. Without any loss of generality, then the auxiliary function $F(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ of (13) can also be written in a separated form as

$$\begin{aligned}F(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \sum_{Q\in\mathcal{N}^T}\sum_{M\in\mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M}\mid\boldsymbol{\lambda})\Big\{&\log \pi'_{q_1} + \sum_{t\in\mathcal{T}^+} \log a'_{q_t q_{t+1}}\\ &+ \sum_{t\in\mathcal{T}^+} \log c'_{q_t m_t} + \sum_{t\in\mathcal{T}^+} \log b'_{q_t m_t}(O_t)\Big\}. \end{aligned} \tag{25}$$

Since $F(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ is considered as the basis for the maximum likelihood optimization procedure, separability of the individual auxiliary functions as given in Appendix B simplifies the (re)estimation procedure. Individual maximization of the first three summands subject to the constraints

$$\sum_{j\in\mathcal{N}} \pi_j = 1, \quad \pi_j \geqslant 0 \text{ for } j\in\mathcal{N}. \tag{26}$$

$$\sum_{j\in\mathcal{N}} a_{ij} = 1, \quad a_{ij} \geqslant 0 \text{ for } i, j\in\mathcal{N}. \tag{27}$$

$$\sum_{m\in\mathcal{M}_L} c_{im} = 1, \quad c_{im} \geqslant 0 \text{ for } i\in\mathcal{N}, m\in\mathcal{M}_L. \tag{28}$$

respectively, is well known. Each of the individual auxiliary functions has the same form $\Sigma_{j\in\mathcal{N}} u_j \log v_j$, which as a function of $\{v_j \mid j\in\mathcal{N}\}$ with the constraint $\Sigma_{j\in\mathcal{N}} v_j = 1$ and $v_j \geqslant 0$ for $j\in\mathcal{N}$ attains a global maximum at the single point $v_j = u_j/\Sigma_{i\in\mathcal{N}} u_i$ for $j\in\mathcal{N}$. The initial probability $\bar{\pi}$ can be reestimated as

$$\bar{\pi}_i = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i\in\mathcal{N}} \alpha_1(i)\beta_1(i)} = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i\in\mathcal{N}} \alpha_T(i)} \quad \text{for } i\in\mathcal{N}, \tag{29}$$

which is the expected frequency in state S_i at $t = 1$. Similarly, the reestimation formula for A results in a ratio of the expected number of transitions from state S_i to state S_j to the expected number of transitions out of state S_i, i.e.,

$$\bar{a}_{ij} = \frac{\sum_{t\in\mathcal{T}^+} \xi_t(i,j)}{\sum_{t\in\mathcal{T}^+}\sum_{j\in\mathcal{N}}\sum_{m\in\mathcal{M}_L} \xi_t^{(m)}(i,j)} = \frac{\sum_{t\in\mathcal{T}^+} \xi_t(i,j)}{\sum_{t\in\mathcal{T}^+} \alpha_t(i)\beta_t(i)}, \tag{30}$$

where $\xi_t^{(m)}(i,j)$ is the probability of being in state S_j at time $t+1$ and state S_i at time t with the m-th mixture component accounting for O_t, i.e.,

$$\xi_t^{(m)}(i,j) = \xi_t(i,j)\left[\frac{c_{im}b_{im}(O_t)}{\sum_{l\in\mathcal{M}_L} c_{il}b_{il}(O_t)}\right], \tag{31}$$

with $b_{im}(O_t)$ as given by (3). $\bar{c}_{im}$ is the ratio of the expected number of transitions out of state S_i using the m-th mixture component to the expected number of total transitions out of state S_i. Thus, for $i \in \mathcal{N}$ and $m \in \mathcal{M}_L$, we get

$$\bar{c}_{im} = \frac{\sum_{t\in\mathcal{T}^+}\sum_{j\in\mathcal{N}} \xi_t^{(m)}(i,j)}{\sum_{t\in\mathcal{T}^+}\sum_{j\in\mathcal{N}}\sum_{m\in\mathcal{M}_L} \xi_t^{(m)}(i,j)} = \frac{\sum_{t\in\mathcal{T}^+}\sum_{j\in\mathcal{N}} \xi_t^{(m)}(i,j)}{\sum_{t\in\mathcal{T}^+} \alpha_t(i)\beta_t(i)}. \tag{32}$$

The parameters set $\{d_{im} \mid i \in \mathcal{N}, m \in \mathcal{M}_L\}$ and $\{f_{im} \mid i \in \mathcal{N}, m \in \mathcal{M}_L\}$ can be calculated from the following two equations:

$$\sum_{r=1}^{f_{im}+1} \frac{1}{(d_{im}+r)} = -\frac{\sum_{t\in\mathcal{T}^+}\sum_{j\in\mathcal{N}} \xi_t^{(m)}(i,j)\log(O_t)}{\sum_{t\in\mathcal{T}^+}\sum_{j\in\mathcal{N}} \xi_t^{(m)}(i,j)}, \tag{33}$$

$$\sum_{r=1}^{d_{im}+1} \frac{1}{(f_{im}+r)} = -\frac{\sum_{t\in\mathcal{T}^+}\sum_{j\in\mathcal{N}} \xi_t^{(m)}(i,j)\log(1-O_t)}{\sum_{t\in\mathcal{T}^+}\sum_{j\in\mathcal{N}} \xi_t^{(m)}(i,j)}, \tag{34}$$

where the parameters d_{im} and f_{im} are assumed, for reducing computations, to take up integer values.

4.2. Distortion computation

We now discuss the CHM measure methodology. At first, we obtain the PD observation sequences from the original signal. For each of the sixty-four neural channels, we consider these PD observations for a frame of T consecutive clock times. An HMM is associated with each of such channels and the model parameters are determined starting from an initial estimate. Eqs. (29) to (34), derived based on the Baum–Welch algorithm, are used for estimating the model parameters. This technique iteratively chooses a 'better' model by maximizing $P(\boldsymbol{O}_n \mid \boldsymbol{\lambda}_n)$ where $\boldsymbol{O}_n$ is the n-th channel PD observation sequence for the original speech. After a reasonable number of iterations, the algorithm is terminated and the final model is denoted as $\boldsymbol{\lambda}_n^{(o)}$. Let the corresponding n-th channel PD observations for the coded speech be represented by $\boldsymbol{O}_n^{(c)}$. Using (23), we compute $P(\boldsymbol{O}_n^{(c)} \mid \boldsymbol{\lambda}_n^{(o)})$ for all the neural channels. This computation, in essence, evaluates the likelihood probability of the PD representation of the coded

signal against the models derived from the PD representation of the original speech. These probability scores are multiplied over all the channels. Upon taking logarithm and dividing by the number of channels (here, 64), we obtain a similarity measure for the frame. The CHM distortion measure, a negated version of the similarity measure, could be expressed as

$$\mathrm{CHM} = -\frac{1}{64}\sum_{n=1}^{64} \log P\left(\boldsymbol{O}_n^{(c)} \mid \boldsymbol{\lambda}_n^{(o)}\right). \tag{35}$$

The CHM measure for a speech utterance is computed by taking average of the CHM measures over all the speech frames.

4.3. Alternative approaches

Here, we suggest two other logical approaches for computing coder distortion although we have not carried out any test with them.

4.3.1. State sequence approach

One alternative method is to determine the 'optimal' state sequences associated with the PD observation sequences of an original speech as well as its coded version. An optimality criterion chooses the state q_t that is individually most likely by maximizing the expected number of correct individual states. The individually most likely state q_t at time t is determined by computing

$$q_t = \underset{i \in \mathcal{N}}{\operatorname{argmax}} \left[P(q_t = S_i \mid \boldsymbol{O}, \boldsymbol{\lambda}) \right]. \tag{36}$$

The bracketed term, i.e., the probability of being in state S_i at time t, given the observation sequence $\boldsymbol{O}$ and the model $\boldsymbol{\lambda}$, is written for the forward-backward technique in terms of the variables $\xi_t(i, j)$ as

$$P(q_t = S_i \mid \boldsymbol{O}, \boldsymbol{\lambda}) = \frac{\sum_{j \in \mathcal{N}} \xi_t(i, j)}{\sum_{i \in \mathcal{N}} \alpha_T(i)}. \tag{37}$$

The solution simply determines the most likely state at every instant without any regard to the probability of occurrence for sequence of states. A distortion measure could be defined based on calculating the Hamming distance between the estimated state sequences for the original and the coded speech signals. There is no unique way of selecting an optimality criterion and the approach may even be modified to maximize the expected number of correct paths of pairs of states (q_t, q_{t+1}) or triples of states (q_t, q_{t+1}, q_{t+2}), etc.

4.3.2. Model distance approach

Another alternative is to estimate a model $\boldsymbol{\lambda}^{(c)}$ from the PD observations of the coded speech frame exactly the way we have estimated the model $\boldsymbol{\lambda}^{(o)}$ from the PD observation of the original speech frame. A model distance measure following the notion of discrimination information could be defined for comparing these pairs of HMMs (Juang, 1984). One such measure form is

$$D(\boldsymbol{\lambda}^{(c)}, \boldsymbol{\lambda}^{(o)}) = \sum_{n=1}^{64} \log P\left(\boldsymbol{O}_n \mid \boldsymbol{\lambda}_n^{(c)}\right) - \sum_{n=1}^{64} \log P\left(\boldsymbol{O}_n \mid \boldsymbol{\lambda}_n^{(o)}\right). \tag{38}$$

This measure is non-symmetric and a symmetrized version could be used in practice.

5. Practical considerations

A 'good' distortion measure should consider only the information relevant to perceptual events. However, the success of the measure also depends heavily on the accuracies of the implementation and the model description. Here, we discuss some practical aspects related to the evaluation of speech coders by the CHM measure.

5.1. Computational issues

The forward probability calculation is, in effect, based upon the trellis structure shown in Fig. 3. Since there are only two possible states at each time in the trellis, all the possible state sequences will remerge into one of these two nodes, regardless of the length of the observation sequence. At any time t, computation of $\alpha_t(j)$ involves only two previous values of $\alpha_{t-1}(i)$ because each of the two grid points is reached from the same two grid points at the previous time slot. For computing each $\alpha_t(i)$ and $\beta_t(j)$, it requires on the order of N^2T calculations, rather than $2TN^T$ as required by the direct calculation.

Another important issue is that computing the likelihood variables involves multiplication of many terms having values smaller than one. In a recursive procedure, each term of these variables starts to diminish towards zero exponentially and thus the number representation goes below the precision range of any machine. To circumvent this problem, the likelihood and other variables are multiplied by constants known as scaling coefficients (Levinson et al., 1983). The scaling procedure is not applied at every clock time, but once every few clock times.

5.2. Initial estimates for HMM parameters

Since a convergent reestimation procedure exists for the continuous mixture model considered here, it is theoretically possible to have arbitrary initial estimates for the HMM parameters obeying the stochastic constraints. The reestimation equations provide values for the HMM parameters corresponding to a local maximum of the likelihood function. The choice of 'good' initial estimates is thus important in making the convergence faster or ensuring the local maximum to be the global maximum of the likelihood function. In fact, some of the parameters may be very sensitive to the initial estimates (Rabiner et al., 1985a).

5.3. Training data and iterations

The PD observation sequence used for training the models has a finite length and this causes problem in determining the HMM parameters via reestimation method. An insufficient number of occurrences of different model events does not truly portray the real scenario and therefore we have to have sufficiently long training data. On the other hand, we want the model parameters to be fixed for a specific period and then vary depending on the new PD observations. Thus, the training data cannot be too long. It is emphasized that the Baum–Welch estimation algorithm needs several iterations before the convergence occurs.

5.4. Mixture processes

It is an usual practice to approximate a K-dimensional correlated random process by a mixture of few uncorrelated, K-dimensional random processes (Juang, 1985; Rabiner et al., 1985b). The number of

mixture components is heavily dependent on the degree of correlation. By assuming mixture uncorrelated processes, we effectively reduce the number of parameters to be estimated and thus make the estimates more reliable. The trade-off is clearly between the increased error in the approximation process and the increased reliability in the estimation process.

6. Experimental results

Before providing with the objective measure results, we describe the set-up procedure for some of the experimental parameters.

(i) We have trained and matched the HMMs with speech frames of 480 samples. For $N = 2$ and $T = 480$, only about 1920 computations were needed since the algorithm used was based on trellis structure.

(ii) The scaling procedure was used not at every instant, but after every ten clock times.

(iii) Although the length of the PD sequences over which the training and matching were done is 480, we overlapped each such frame with the previous frame by 50%. In other words, the observation window was shifted by 240 samples for dealing with each new model. This has allowed to have sufficiently long training data and also has facilitated the model parameters not to change drastically.

(iv) In our experiment, we have chosen models with three mixture components (i.e., $M = 3$). This has appeared to be a reasonable choice for making trade-off between the accuracy of modeling the histogram and the number of parameters to be estimated.

(v) Based on the psychoacoustic data, we have assumed the initial transition probability from a non-firing state to another non-firing state is 0.8 and that from a firing state to another firing state is 0.2. In accordance with this, the initial state probabilities were chosen to be 0.8 for non-firing state (S_0) and 0.2 for firing state (S_1).

(vi) The initial estimates for the beta pdf parameters $\{d_{im}\}$ and $\{f_{im}\}$ were chosen in such a fashion that the corresponding mean values were 0.25, 0.50 and 0.75 for $i \in \mathcal{N}$. The weighting factors $\{c_{im}\}$ were all assumed to be equal (i.e., 0.33) initially.

(vii) For any particular neural channel, the final estimate of the HMM parameters obtained for a speech frame was considered as the initial estimate of the parameters for the subsequent frame.

(viii) While solving the simultaneous equations of (33) and (34), the $\{d_{im}\}$ and $\{f_{im}\}$ parameters were allowed to take up integral values between 1 and 40. Since the exact solution could not be found, we have determined the parameter values by choosing the best pair which minimizes the sum of the square errors. One more constraint imposed on the parameters was that the mean values (given by $d_{im}/(d_{im} + f_{im})$) for three different mixture components have been kept confined to three different regions – one between 0 and 1/3, the second between 1/3 and 2/3, and the third between 2/3 and 1. This also reduced the search for best solution by making some combinations of the parameter values to be invalid.

(ix) For model parameter estimations, we have made 30 iterations for each frame of PD observations in any neural channel.

In this work, we have followed two strategies for computing coder distortions. Let us now consider determining the model parameters for the n-th neural channel. In the first strategy, while training the model, we have used only the n-th channel PD observation sequence corresponding to the original speech. We call this strategy as CHM-SC (with single channel). Table 1 shows subjective and objective measure values for six coded signals with reference to the original speech utterance. We tabulate here

Table 1
Subjective and objective measure values for six coded signals with reference to the corresponding original speech utterances ('S' gives the average subjective ranking scores and 'H' denotes the cochlear hidden Markov measure with single channel (CHM-SC))

Sentence	C1		C2		C3		C4		C5		C6	
	S	H	S	H	S	H	S	H	S	H	S	H
M1	5.75	195	4.92	225	4.17	336	2.58	358	2.58	365	1.00	420
M2	5.50	250	5.17	231	4.25	280	2.75	310	2.25	390	1.08	414
F1	5.67	209	5.00	263	4.25	300	2.33	389	2.58	371	1.17	430
F2	5.67	220	5.00	276	3.91	347	2.67	378	2.50	312	1.25	398

measure values for only four utterances. The CHM-SC measure was found to be not very satisfactory in ranking coded signals.

It has been our understanding that the training data length was not sufficient in the CHM-SC strategy to make a reliable estimate for the model parameters. Therefore, we formulated a new strategy where three adjacent channels – the $(n-1)$-th, the n-th and the $(n+1)$-th channel PD observations – were used in alternate manners for training. This strategy has been termed the CHM-TC (with three channels). Table 2 provides subjective and CHM-TC measure values for all the twelve utterances given in Appendix C.

For the CHM distortion measure values, we have computed the logarithm (natural) of the likelihood probability scores, negated them and averaged over all the channels and all the speech frames. In Tables 1 and 2, the subjective ranking (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 all the twelve utterances and six coders, 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 $(6 \times 8 + 5 \times 3 + 4 \times 1)/12 = 5.58$. The second column (marked 'H') provides the CHM objective measure where a low value implies a better perceptual quality.

We note that with utterance M1, the C4, C5 coders and with utterance F5, the C1, C2 coders were ranked same subjectively. The CHM-TC measure has found C4 coder for M1 and C2 coder for F5 to be

Table 2
Subjective and objective measure values for six coded signals with reference to the corresponding original speech utterances ('S' gives the average subjective ranking scores and 'H' denotes the cochlear hidden Markov measure with three channels (CHM-TC))

Sentence	C1		C2		C3		C4		C5		C6	
	S	H	S	H	S	H	S	H	S	H	S	H
M1	5.75	146	4.92	188	4.17	256	2.58	314	2.58	320	1.00	408
M2	5.50	161	5.17	179	4.25	238	2.75	287	2.25	346	1.08	398
M3	5.75	157	5.17	183	4.00	261	2.58	310	2.33	304	1.17	401
M4	5.00	196	5.67	152	4.25	230	2.50	326	2.58	311	1.00	412
M5	5.75	138	5.17	170	3.83	277	2.67	301	2.50	335	1.00	421
M6	5.58	163	5.25	186	3.83	265	2.75	292	2.42	319	1.17	392
F1	5.67	154	5.00	182	4.25	244	2.33	326	2.58	307	1.17	416
F2	5.67	159	5.00	192	3.91	270	2.67	296	2.50	310	1.25	386
F3	5.50	170	5.17	177	4.25	221	2.50	319	2.25	352	1.33	381
F4	5.41	169	5.25	174	4.17	238	2.75	281	2.17	361	1.25	399
F5	5.50	162	5.50	155	3.83	272	2.33	330	2.50	304	1.33	373
F6	5.67	156	4.83	202	4.08	263	3.08	322	2.17	348	1.17	391

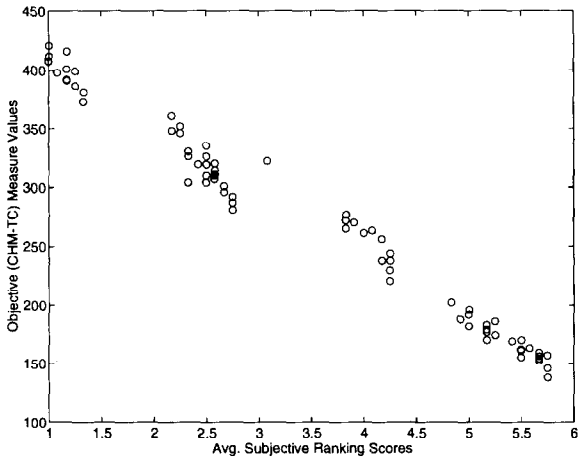

Fig. 4. Scatter plot showing subjective and objective measure values for six coded signals with reference to the corresponding original (twelve) speech utterances.

slightly better than their counterparts. Other than these tie cases, the subjective and objective measures were not in conformance only for the C4, C5 with the utterance M3. The scatter plot of Fig. 4 corresponding to the data of Table 2 shows that the subjective and objective scores, in general, agree with each other.

Unlike most of the other distortion measures, the CHM measure performs quite well without an explicit time-alignment. Table 3 provides the SNR measure as well as the CHM measure with zero, one, two and three sample delays in the coded speech. The misaligned sample places are filled in with very small (approximately zero) values. It is observed that the SNR measure which is well-known for its sensitivity to delay error varies considerably with the coder delays. On the other hand, a small sample delay does not really affect the CHM measure value.

The CHM measure involves much more computation than the CDI measure presented in (De and Kabal, 1994). However, for speech coder evaluation, the distortion measure does not need to be computed in real time. At the cost of computational complexity, the primary two advantages accrued in the CHM measure methodology are that (i) ample provisions (selecting better initial estimates, carrying out more iterations, etc.) exist for its improvement and (ii) it attempts to take time correlations into account and is fairly robust against small time shifts.

Table 3
The SNR and the cochlear hidden Markovian – three channels (CHM-TC) 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 (w/o scaling [dB])	8.724	7.391	5.619	5.117
oakf8f	CHM-TC	221	227	224	229
oakf8k	SNR (w/o scaling [dB])	9.178	7.503	6.108	7.027
oakf8k	CHM-TC	319	321	326	323

7. Summary

Determining a 'good' distortion measure for speech coding is an extremely difficult problem. At the same time, finding such a measure would surely have a significant impact on speech coding and coder evaluation procedures. We have tried to take a step towards the solution. In order to formulate a distortion measure for speech coder evaluation, we have used a physiological model for auditory processing and applied information-processing techniques from information theory.

In this article, we have introduced a cochlear hidden Markovian (CHM) measure for computing coder distortion. The basics of neural firing events have been captured by simple hidden Markov models, where the occurrence of perceptual-domain observations and correlation among adjacent observations are modeled appropriately. A two-state (one each for firing and non-firing events), fully-connected HMM has been associated with each of the neural channels.

For computing coder distortions, at first, all the HMMs are trained (i.e., the HMM parameters are estimated) with the PD observation derived from the original signal. The Baum–Welch reestimation technique has been applied to derive the HMM parameters iteratively starting from an initial estimate. The PD observations obtained from the coded speech are matched against these HMMs. A (negated) log likelihood probability score, averaged over all the speech frames and neural channels, acts as the CHM similarity (distortion) measure. This measure conforms substantially with subjective evaluation results and also exhibits robustness against time shifts.

Appendix A. Log-concavity of beta pdf

A beta density function is given as

$$b(x) = \frac{\Gamma(d+f+2)}{\Gamma(d+1)\Gamma(f+1)} x^d (1-x)^f. \tag{A.1}$$

In this appendix, we prove that this function satisfies the log-concavity condition, i.e., the logarithm of the function is concave. Taking logarithm of (A.1), we get

$$\phi(x) = \log \Gamma(d+f+2) - \log \Gamma(d+1) - \log \Gamma(f+1) + d \log x + f \log(1-x). \tag{A.2}$$

To show the log-concavity nature of (A.1), we need to show that $\phi(x)$ is concave w.r.t. x. Defining $\bar{\lambda} \equiv (1-\lambda)$, we write

$$\begin{aligned}
&\phi(\lambda x' + \bar{\lambda} x'') - \lambda \phi(x') - \bar{\lambda} \phi(x'') \\
&\quad = d\lambda \log(\lambda x' + \bar{\lambda} x'') + f\lambda \log(1 - \lambda x' - \bar{\lambda} x'') + d\bar{\lambda} \log(\lambda x' + \bar{\lambda} x'') + f\bar{\lambda} \log(1 - \lambda x' - \bar{\lambda} x'') \\
&\qquad - d\lambda \log x' - f\lambda \log(1 - x') - d\bar{\lambda} \log x'' - f\bar{\lambda} \log(1 - x'') \\
&\quad = d\lambda \log\left(\frac{\lambda x' + \bar{\lambda} x''}{x'}\right) + f\lambda \log\left(\frac{1 - \lambda x' - \bar{\lambda} x''}{1 - x'}\right) + d\bar{\lambda} \log\left(\frac{\lambda x' + \bar{\lambda} x''}{x''}\right) + f\bar{\lambda} \log\left(\frac{1 - \lambda x' - \bar{\lambda} x''}{1 - x''}\right) \\
&\quad \geqslant d\lambda\left(1 - \frac{x'}{\lambda x' + \bar{\lambda} x''}\right) + f\lambda\left(1 - \frac{1 - x'}{1 - \lambda x' - \bar{\lambda} x''}\right) + d\bar{\lambda} \log\left(\frac{\lambda x' + \bar{\lambda} x''}{x''}\right) + f\bar{\lambda} \log\left(\frac{1 - \lambda x' - \bar{\lambda} x''}{1 - x''}\right) \\
&\quad \geqslant d\lambda\left(1 - \frac{x'}{\lambda x' + \bar{\lambda} x''}\right) + f\lambda\left(1 - \frac{1 - x'}{1 - \lambda x' - \bar{\lambda} x''}\right) + d\bar{\lambda}\left(1 - \frac{x'}{\lambda x' + \bar{\lambda} x''}\right) + f\bar{\lambda}\left(1 - \frac{1 - x''}{1 - \lambda x' - \bar{\lambda} x''}\right) \\
&\quad = d - d\left(\frac{\lambda x' + \bar{\lambda} x''}{\lambda x' + \bar{\lambda} x''}\right) + f - f\left(\frac{\lambda(1 - x') + \bar{\lambda}(1 - x'')}{1 - \lambda x' + \bar{\lambda} x''}\right) \\
&\quad = 0
\end{aligned} \tag{A.3}$$

Since it has been shown that $\phi(\lambda x' + \bar{\lambda} x'') \geq \lambda \phi(x') - \bar{\lambda} \phi(x'')$, the beta pdf of (A.1) is proven to be log-concave.

Appendix B. Baum–Welch reestimation procedure

In our work, an auxiliary function $F(\boldsymbol{\lambda}, \boldsymbol{\lambda}')$ is considered as the basis for the maximum likelihood optimization procedure. The Baum–Welch (re)estimation procedure is used for determining different model parameters. Separability of the individual auxiliary functions has made this procedure elegant and reduced the complexity. Here, we write the expressions for individual auxiliary functions. We can rewrite (25) as

$$F(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = F_{\pi}(\boldsymbol{\lambda}, \pi') + \sum_{i \in \mathcal{N}} F_{a_i}\left(\boldsymbol{\lambda}, \{a'_{ij}\}_{j \in \mathcal{N}}\right) + \sum_{i \in \mathcal{N}} \sum_{m \in \mathcal{M}_L} F_b(\boldsymbol{\lambda}, b'_{im}) + \sum_{i \in \mathcal{N}} F_{c_i}\left(\boldsymbol{\lambda}, \{c'_{im}\}_{m \in \mathcal{M}_L}\right), \tag{B.1}$$

where

$$F_{\pi}(\boldsymbol{\lambda}, \pi') = \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M} \mid \boldsymbol{\lambda}) \log \pi'_{q_1} = \sum_{i \in \mathcal{N}} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, q_1 = S_i, \boldsymbol{M} \mid \boldsymbol{\lambda}) \log \pi'_i, \tag{B.2}$$

$$\begin{aligned} F_{a_i}\left(\boldsymbol{\lambda}, \{a'_{ij}\}_{j \in \mathcal{N}}\right) &= \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M} \mid \boldsymbol{\lambda}) \sum_{t \in \mathcal{T}^+} \log a'_{q_t q_{t+1}} \delta(q_t - S_i) \\ &= \sum_{j \in \mathcal{N}} \sum_{t \in \mathcal{T}^+} \sum_{M \in \mathcal{M}_L^T} P\left(\boldsymbol{O}, q_t = S_i, q_{t+1} = S_j, \boldsymbol{M} \mid \boldsymbol{\lambda}\right) \log a'_{ij}, \end{aligned} \tag{B.3}$$

$$\begin{aligned} F_b(\boldsymbol{\lambda}, b'_{im}) &= \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M} \mid \boldsymbol{\lambda}) \sum_{t \in \mathcal{T}^+} \log b'_{q_t m_t}(O_t) \delta(q_t - S_i) \delta(m_t - m) \\ &= \sum_{t \in \mathcal{T}^+} P\left(\boldsymbol{O}, q_t = S_i, m_t = m \mid \boldsymbol{\lambda}\right) \log b'_{im}(O_t) \end{aligned} \tag{B.4}$$

and

$$\begin{aligned} F_{c_i}\left(\boldsymbol{\lambda}, \{c'_{im}\}_{m \in \mathcal{M}_L}\right) &= \sum_{Q \in \mathcal{N}^T} \sum_{M \in \mathcal{M}_L^T} P(\boldsymbol{O}, \boldsymbol{Q}, \boldsymbol{M} \mid \boldsymbol{\lambda}) \sum_{t \in \mathcal{T}^+} \log c'_{q_t m_t} \delta(q_t - S_i) \\ &= \sum_{m \in \mathcal{M}_L} \sum_{t \in \mathcal{T}^+} P\left(\boldsymbol{O}, q_t = S_i, m_t = m \mid \boldsymbol{\lambda}\right) \log c'_{im}, \end{aligned} \tag{B.5}$$

where δ in the above expressions is the Kronecker delta function.

Appendix C. Test sentences

The reference audio files were obtained by digitally filtering the speech and sampling it at a rate of 8,000 Hz. The digital filter (255 tap FIR) applied was designed to be unity between 0 and 3,200 Hz. For the purpose of speech coder evaluation, the following test sentences (male and female voices) were used.

1. Add the sum to the product of these three.
2. Cats and dogs each hate the other.
3. Oak is strong and also gives shade.
4. Open the crate but don't break the glass.
5. The pipe began to rust while new.
6. Thieves who rob friends deserve jail.

Acknowledgements

The authors would like to thank the reviewers for their constructive comments.

References

L.E. Baum (1972), "An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes", *Inequalities*, Vol. 3, pp. 1–8.

L.E. Baum and J.A. Egon (1967) "An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology", *Bull. Amer. Meteorol. Soc.*, Vol. 73, pp. 360–363.

L.E. Baum and T. Petrie (1966), "Statistical inference for probabilistic functions of finite state Markov chains", *Ann. Math. Statist.*, Vol. 37, pp. 1554–1563.

R.L. Cave and L.P. Neuwirth (1980), "Hiden Markov models for English", in *Hidden Markov Models for Speech*, ed. by J. Ferguson, Vol. IDA-CRD, pp. 16–56.

R.W. Chang and J.C. Hancock (1966), "On receiver structures for channels having memory", *IEEE Trans. Inform. Theory*, Vol. IT-12, October, pp. 463–468.

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

A. De (1993b), "Auditory distortion measures for coded speech quality evaluation", *Canadian Acoustics*, September, pp. 105–106.

A. De and P. Kabal (1994), "Auditory distortion measure for speech coder evaluation – Discrimination information approach", *Speech Communication*, Vol. 14, No. 3, June, pp. 205–229.

F. Jelinek (1976), "Continuous speech recognition by statistical methods", *Proc. IEEE*, Vol. 64, April, pp. 532–536.

B.-H. Juang (1984), "On the hidden Markov model and dynamic time warping for speech recognition", *Bell Syst. Techn. J.*, September, pp. 1213–1243.

B.-H. Juang (1985), "Maximum-likelihood estimation for mixture multivariate stochatic observations of Markov chains", *Bell Syst. Techn. J.*, July–August, pp. 1235–1249.

S.E. Levinson, L.R. Rabiner and M.M. Sondhi (1983), "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition", *Bell Syst. Techn. J.*,, April, pp. 1035–1074.

L.A. Liporace (1984), "Maximum likelihood estimation for multivariate observations of Markov sources", *IEEE Trans. Inform. Theory*, Vol. IT-28, September, pp. 729–734.

L.R. Rabiner, B.-H. Juang, S.E. Levinson and M.M. Sondhi (1985a), "Some properties of continuous hidden Markov model representation", *Bell Syst. Techn. J.*, July–August, pp. 1251–1269.

L.R. Rabiner, B.-H. Juang, S.E. Levinson and M.M. Sondhi (1985b), "Recognition of isolated digits using hidden Markov models with continuous mixture densities", *Bell Syst. Techn. J.*., July–August, pp. 1211–1234.

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